Patrimoine 3.0
Actes du douzième colloque international sur le document électronique
Edité par **e**uropia Productions
15, avenue de Ségur
75007 Paris, France
Tel +31 1 45 51 26 07
Fax +31 1 45 51 26 32
Email: info@europia.fr
http://www.europia.fr
http://www.europiaproductions.com

ISBN : 978-2-909285-54-4

Patrimoine 3.0

Actes du douzième colloque international
sur le document électronique

21-23 octobre 2009, Université de Montréal - Canada

EUROPIA

Table des matières

Le patrimoine ethnologique et les nouvelles technologies **1**
Laurier Turgeon, Célia Forget, Louise Saint-Pierre, Martin Fournier, François Côté
Université Laval, Québec, Canada

Contexte et rapports: L'Irlande et le patrimoine irlandais **15**
Michael Buckland, Daniel Melia, et Ryan Shaw
University of California, Berkeley, CA. - E.U.

Le recours à des environnements numériques pour documenter le patrimoine bâti : une approche basée sur la complémentarité entre photogrammétrie et objet paramétrique **29**
Nathalie Charbonneau, Pierre Grussenmeyer
UMR MAP 694, Équipe PAGE, INSA de Strasbourg, France

Patrimoine photographique et dispositifs collaboratifs en ligne **43**
Nathalie Casemajor Loustau
Université du Québec a Montréal, Canada

Cartes anciennes, SIG géo-historiques et visites virtuelles : Déformations lisibles de la perception spatiale sur un corpus de cartes numérisées issues de fonds patrimoniaux et de représentations cartographiques patrimoniales hypermédia **59**
Jean-Philippe d'Erceville
Université de Lyon1 - Laboratoire ELICO (France).

Le concept de musée virtuel thématique : la collection comme visite, la visite comme lecture, la lecture comme stratégie. L'exemple du musée thématique sur l'Annonciation **77**
Kanellos, S. Daniilia
ENST-Brest, France

Manuscrits de Stendhal : Du patrimoine papier au document électronique **97**
Auriane Faure (1) , Thomas Lebarbé (1) , Cécile Meynard (2) , Aicha Touati (1)
(1) Laboratoire LIDILEM (EA 609)
(2) Equipe TRAVERSES 19-21 (EA 3748)
Université Stendhal - Grenoble 3, France

Diversité de l'Information dans les Sites de Presse **111**
Cyril Laitang, Elöd Egyed-Zsigmond, Sylvie Calabretto
INSA – Lyon, France

Connaissances prescrites ou connaissance décrites ? L'apport de la sémantique linguistique **129**
Monique Slodzian (1), Mathieu Valette (1&2)
1 CRIM (INALCO, Paris)
2 ATILF (CNRS, Nancy)

Une approche générale pour l'extraction des lignes des documents arabes anciens multi orientées **143**
Nazih Ouwayed, Abdel Belaïd
Loria, Nancy, France

Intertextual semantics generation for structured documents: a complete implementation in XSLT **159**
Yves Marcoux
Université de Montréal, Canada

Visualisations de contrôle pour la numérisation massive **171**
Rodrigo Almeida, Pierre. Henri Cubaud
Centre d'études et de recherche en informatique (CEDRIC),
Conservatoire national des arts et métiers (CNAM), France

Les nouveaux enjeux de la mise en valeur du Patrimoine scientifique et technique de la recherche dans l'espace Francophone **189**
Rachel Kamga, Khaldoun Zreik
Laboratoire Paragraphe, Université Paris 8, France

Un nouveau dictionnaire électronique structuré et évolutif pour la langue arabe **203**
Abd El Salam al Hajjar (1,2), Mohammad Hajjar (2), Khaldoun Zreik (2)
(1) Institut Universitaire de Technologie, Université Libanaise, Liban
(2) Laboratoire Paragraphe, Université Paris 8, France

La préservation par l'accès : l'approche patrimoniale de la gestion des connaissances et son instrumentation documentaire **217**
Bruno Bachimont, Stéphane Crozat,
Université de Technologie de Compiegne, CNRS UMR 6599 Heudiasyc, France

Caractériser l'information à partir des processus métiers. Méthode d'analyse et de traitement de contenus documentaires s'appuyant sur la compréhension des processus organisationnels et des métiers **229**
Noémie Musnik (1,2), Manuel Zacklad (1), Philippe Haik (2), Sylvain Mahé (2) Benoît Ricard (2)
(1) CNAM - Laboratoire DICEN, Paris, France
(2) EDF-R&D - Département STEP, Paris, France

Une médiathèque virtuelle physique **241**
Pedro Alessio, Pierre Cubaud , Boris Guillot, Alexandre Topol
Laboratoire CEDRIC, CNAM, Paris, France

Documents et Applications : CMS nouvelle génération **251**
Jean-Marc Lecarpentier, Hervé Le Crosnier, Jacques Madelaine
GREYC - CNRS UMR 6072 - Université de Caen, France

Outil de butinage du contenu des documents de collections numériques **263**
Lyne Da Sylva
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada

Extraction de termes, reconnaissance et labellisation de relations dans un thésaurus **275**
Marie-Noelle Bessagnet, Eric Kergosien, Mauro Gaio
UPPA, Laboratoire LIUPPA, Pau, France

Restructuration physique et logique de documents électroniques textuels **287**
Jean-Luc Bloechle, Rolf Ingold
Département d'Informatique, Université de Fribourg, Suisse.

Exploitation du patrimoine de documents juridiques numériques pour l'aide à la décision. Une approche de catégorisation basée sur l'étude de la structure et le contenu de documents numériques **299**
Jin Yao (1), Jacques Madelaine (1), Khaldoun Zreik (2)
(1) GREYC - CNRS UMR 6072 - Université de Caen, France
(2) Laboratoire Paragraphe, Université Paris 8, France

Design d'interfaces homme(s)-logiciel(s) et médiation des savoirs à l'ère du web 3.0 **311**
Laurence Noël, Ghislaine Azémard
Laboratoire Paragraphe, Université Paris 8, France

Analyse de la pratique de mise en document de l'information de pilotage en entreprise : le document au cœur de la structuration et de la restitution d'indicateurs **321**
Samuel Parfouru (1) Rodolphe Beck (2)
(1)EDF Recherche et Développement - Département Simulation et Traitement de l'Information pour l'Exploitation et la Production, France
(2) IBM Global Business Services, France

Intégration des Langages pour la Gestion de Documents : Une Nouvelle Etape dans l'Evolution de XML **329**
Catherine Pugin, Rolf Ingold
Département d'Informatique, Université de Fribourg, Suisse

PREFACE

Depuis 1998, CIDE propose un cycle de manifestations scientifiques sur le thème du document électronique, avec pour objectif de confronter les points de vue des différentes disciplines concernées, et de diffuser les résultats des laboratoires académiques ou industriels qui contribuent à en améliorer les usages.

Pour sa douzième édition, CiDE observe certains points de rapprochement entre les domaines du document numérique et du patrimoine. Des points qui sont d'autant plus énoncés lorsque l'on s'intéresse aux mutations que vivent ces deux domaines. Ainsi le thème retenu pour 2009 est : Patrimoines 3.0 "Patrimoines et documents à l'aire du Web 3.0"

Les stratégies, savoirs et méthodes de perception, représentation, compréhension et de traitement du document et du patrimoine connaissent des changements majeurs. Désormais l'objet document et l'objet patrimoine sont perçus (par l'utilisateur et par le chercheur) en tant qu'objets dynamiques (évolutifs), actifs (vs. Passifs) et intégrés (vs. Isolés).

Aujourd'hui le document dynamique numérique est indissociable du patrimoine et demeure en soit un objet du patrimoine.

CiDE.12 a pour objectif de présenter des travaux et d'animer des réflexions prospectives sur les patrimoines dans une problématique de Web 3.0. Ces champs couvrent tous les supports, contenus, travaux et créations liés aux patrimoines numériques, numérisés ou numérisables.

Khaldoun Zreik, Giovanni de Paoli, Ghislaine Azémard
Présidents de CiDE.12

CiDE.12

Présidents du colloque

Khaldoun Zreik, Paragraphe - Université Paris 8, France
Giovanni de Paoli, Université de Montréal, Canada
Ghislaine Azémard, Paragraphe - Université Paris 8, France

Comité du Programme

Patrick Andries, Cooptel, Canada
Bruno Bachimont , UTC, France
Abdel Belaid, LORIA UMR 7503, France
Claire Bélisle, LIRE-ISH, CNRS, France
Dinu Bumbaru, Héritage Montréal, Canada
Philippe Bootz, Université Paris 8, France
Michael Buckland, University of California, USA
Jean-Pierre Dalbéra, Ministere de la culture, France
Jacques Ducloy, DRRT-Lorraine, France
Mauro Gaio, Université de Pau, France
Patrick Gallinari, UPMC, LIP6, France
Bertrand Gervais, NT2, UQAM, Canada
Mohammad Hassoun, ENSSIB-Lyon, France
Nada El-Khouri, Université de Montréal, Canada
Jacques Labiche, Université de Rouen, France
Omar Larouk, ENSSIB-Lyon, France
Jacques Madelaine, Université de Caen, France
Michel Meimaris, Université d'Athénes GRECE
Ghassan Mourad, Université Libanaise, Liban
Florence Piron, Université de Laval, Canada
Emili Prado, Universitat Autonoma de Barcelona, Espagne
Mohammed Quafafou, Université de Marseille, France
Jean-Pierre Raysz, Jouve-R& D, France
Jean Revez, Université du Québec a Montréal, Canada
Giuseppe Richeri, Université de Lugano, Suisse
Alexandra Saemmer, Université Paris8, France
Imad Saleh, Paragraphe - Université Paris 8, France
Temy Tidafi, Université de Montréal, Canada
Eric Trupin, Université de Roeun, France
Manuel Zacklad, CNAM, France

Organisation

Laboratoire Paragraphe, Université Paris 8, France
Faculté d'Aménagement, Université de Montréal, Canada

Le patrimoine ethnologique et les nouvelles technologies web

Laurier TURGEON, Célia FORGET, Louise SAINT-PIERRE, Martin FOURNIER, François COTE

Université Laval, Québec

Internet a multiplié les possibilités de penser, pratiquer, communiquer et valoriser le patrimoine. Les nouvelles technologies Web sont particulièrement bien adaptées à la sauvegarde du patrimoine culturel immatériel qui est, par définition, difficile à saisir dans la mesure où il est immatériel. Elles permettent de conserver et surtout de communiquer très efficacement ce patrimoine, notamment par la captation et la transmission du son et de l'image. Plus qu'un simple inventaire destiné à la conservation, nous souhaitons faire une base de données multimédia virtuelle qui facilite la communication du patrimoine immatériel. Nous croyons même que la communication est le meilleur moyen de conserver le patrimoine immatériel dans la mesure où il participe à sa transmission. La transmission n'assure pas juste la conservation des traditions, elle contribue à les transformer, à les dynamiser et à les renouveler en leur trouvant de nouveaux usages sociaux. Par la même occasion, elle participe à la valorisation et à la reconnaissance de ceux qui les transmettent. Il ne s'agit pas de produire une simple archive fermée sous clé, mais de faire connaître et reconnaître les pratiques traditionnelles en tant que patrimoine dans un souci d'éducation du grand public. Les nouvelles technologies de l'information facilitent non seulement la fabrication, l'accès et la gestion des inventaires, elles suscitent de nouvelles façons de concevoir et de réaliser l'inventaire lui-même. Les communautés et les porteurs de traditions peuvent participer plus facilement au processus de collecte et de communication des données. L'accès aux données par le Web permet des appropriations et réappropriations multiples par une large gamme de personnes (communautés elles-mêmes, journalistes, muséologues, chercheurs) et favorise l'évolution des pratiques et la valorisation sociale des communautés qui en sont les détenteurs. En effet, la base de données virtuelle devient un outil dynamique de communication patrimoniale et de développement culturel et social.

Pour cette séance, trois projets de mise en valeur du patrimoine ethnologique du Québec par l'interface du Web seront présentés : l'Inventaire des ressources ethnologiques du patrimoine immatériel du Québec (IREPI), l'Inventaire du patrimoine immatériel religieux du Québec (IPIR) et l'Encyclopédie du patrimoine immatériel de l'Amérique française. Ces travaux s'appuient sur l'infrastructure de recherche de la Chaire, le LEEM (Laboratoire d'ethnologie et d'enquête multimédia), créé en 2003.

Grâce à l'usage d'équipements numériques, il nous a été possible d'enregistrer rapidement les données sonores et visuelles sur le terrain, de les transférer directement du terrain dans une base de données multimédia, de conserver et de gérer l'information efficacement et de rendre les données audiovisuelles très accessibles au grand public à des coûts peu élevés via le Web. En parallèle, nous avons aussi développé une approche participative à la cueillette et à la mise en valeur des données, qui impliquent les acteurs locaux à toutes les étapes du processus. Ce volet « recherche action » vise aussi la valorisation du patrimoine immatériel, directement sur le terrain lors des enquêtes, par la mise sur pied de projets d'exposition muséale, de présentations multimédia, de sites Web, d'encyclopédies électroniques, de routes touristiques, de festivals et de trousses pédagogiques.

À partir de l'automne 2009, fort de six années d'opération, notre groupe renouvelle ses équipements et peaufine ses méthodes, grâce à l'obtention d'une nouvelle subvention de la Fondation canadienne de l'Innovation (FCI). Celle-ci nous permet de travailler avec un équipement de dernière pointe. Nous entendons par ce biais parvenir à des captations plus fidèles des contextes étudiés, à des résultats encore plus intéressants pour le grand public, et à une meilleure validation des résultats obtenus. Comme par le passé, l'emphase sera mise sur l'emploi de technologies simples d'utilisation et relativement abordables, afin de contribuer à diffuser nos méthodes autant sur les plans local, régional, provincial, national qu'international, et pour que le grand public en tire un maximum de bénéfices.

1 Inventorier numériquement l'immatériel : le projet IREPI

Le patrimoine immatériel est par définition éphémère, fugace. Les pratiques, représentations, expressions, connaissances et savoir-faire se transmettent par le geste et la parole. S'il est possible d'archiver les données sur support numérique, il est plus difficile de les rendre

accessibles à un vaste public et surtout de les présenter de façon didactique. Le Web s'est avéré une plateforme très efficace pour faire connaître et mettre en valeur ce patrimoine.

S'inspirant de la Convention pour la sauvegarde du patrimoine culturel immatériel de l'UNESCO (2003) la Chaire de recherche en patrimoine ethnologique a développé une méthodologie d'inventaire unique afin de répondre aux grands principes édictés dans la Convention et assurer ainsi la sauvegarde et la communication du patrimoine immatériel.

La convention définit ainsi le patrimoine culturel immatériel :

> « les pratiques, représentations, expressions, connaissances et savoir-faire - ainsi que les instruments, objets, artefacts et espaces culturels qui leur sont associés - que les communautés, les groupes et, le cas échéant, les individus reconnaissent comme faisant partie de leur patrimoine culturel. Ce patrimoine culturel immatériel, transmis de génération en génération, est recréé en permanence par les communautés et groupes en fonction de leur milieu, de leur interaction avec la nature et de leur histoire, et leur procure un sentiment d'identité et de continuité, contribuant ainsi à promouvoir le respect de la diversité culturelle et la créativité humaine.»[1]

La définition de l'UNESCO propose trois principes fondamentaux qui sont autant de conditions préalables aux recherches sur le patrimoine immatériel, sur lesquelles nous avons basé notre méthodologie : soit la reconnaissance par la communauté, la notion d'un patrimoine dynamique, et le lien étroit entre le matériel et l'immatériel.

L'IREPI, notre premier projet d'inventaire numérique multimédia à avoir été créé (2004), fait la cueillette, la conservation, l'analyse, la valorisation et la diffusion du patrimoine immatériel du Québec, à l'aide de technologies audiovisuelles numériques. Des équipes d'étudiants et de professionnels de recherche sont constituées pour visiter les différentes communautés pendant plusieurs semaines et s'enquérir du patrimoine immatériel local. En réalisant des entrevues enregistrées auprès de porteurs de tradition, en filmant et en photographiant les pratiques et les savoir-faire observés, nos chercheurs recueillent les informations nécessaires à la constitution d'un dossier complet sur la ressource ethnologique inventoriée. Une fois ce dossier mis en ligne sur notre site Internet, quiconque peut le consulter et ainsi découvrir le patrimoine immatériel de chaque région du Québec.

L'innovation de cet inventaire multimédia est sans aucun doute

l'accessibilité sur le web. Plus de 675 fiches sont ainsi consultables sur le site Internet www.patrimoine-immateriel.ulaval.ca. Il est possible d'effectuer des recherches dans la base de données de diverses façons, tel que par mot clé ou par région géographique. La présentation des pratiques culturelles sur plusieurs supports (textuels, iconographiques et audiovisuels) permet à chaque région administrative du Québec de mieux connaître et de mieux exploiter ses potentiels culturels. D'autre part, notre méthodologie d'inventaire se double d'actions qui permettent de développer des partenariats régionaux et locaux et de sensibiliser les instances locales et les populations à l'importance du patrimoine immatériel comme élément contribuant au renforcement du sentiment d'appartenance et à la mise en valeur des richesses patrimoniales régionales. L'inventaire devient un outil de développement durable: les ressources identifiées viennent contribuer au développement social et économique des régions.

Le projet IREPI nous a permis de développer des collaborations dans différents pays dont la France, la Belgique, Haïti, le Congo Brazzaville et l'Égypte qui se sont inspirés de notre méthodologie pour développer leur propre inventaire national. Ce travail s'est également traduit dans le choix de la dernière thématique du congrès international d'ICOMOS (L'esprit du lieu, Québec 2008) qui mettait en lumière l'interdépendance entre patrimoine matériel et patrimoine immatériel1.

2 Le patrimoine immatériel religieux du Québec : sauvegarder l'immatériel par le virtuel

Le patrimoine religieux au Québec est menacé. La laïcisation de la société a entraîné une diminution de la pratique religieuse. Les conséquences de cette désaffection massive sont bien visibles : vieillissement des membres des communautés religieuses, fermeture des églises, fusion des paroisses faute de paroissiens et de ministres du culte et désacralisation des objets du culte qui prennent le chemin des musées. Devant l'ampleur de la crise et les enjeux culturels et mémoriaux pour la société québécoise, les pouvoirs publics et la société civile ont commencé à réagir. Reconnaissant l'importance du patrimoine religieux dans le développement et la compréhension de la société québécoise contemporaine, les communautés religieuses, les gouvernements et les citoyens ont pris en main la sauvegarde des patrimoines mobilier et immobilier.

1 Laurier Turgeon (dir.), The Spirit of Place : Between Tangible and Intangible Cultural Heritage/L'esprit du lieu : entre le patrimoine matériel et immatériel, Québec, Presses de l'Université Laval, 2009.

Mais assurer la conservation du matériel sans se préoccuper de l'immatériel ne fait plus sens aujourd'hui. Ce sont les composantes immatérielles (la mémoire, les valeurs, l'attachement) qui insufflent un sens à la culture matérielle. D'une certaine manière, le patrimoine immatériel religieux est celui qui est le plus menacé dans la mesure où il est porté par des personnes. La sauvegarde de la mémoire et des savoir-faire ne permet pas juste de conserver les éléments intangibles du patrimoine, mais aussi de mieux comprendre et préserver ses éléments tangibles. La mémoire orale, les savoir-faire, les fêtes, les rites et les coutumes sont des traditions vivantes, conservées par la simple pratique, répétée à des moments précis de la journée ou de l'année. Elles se transmettent par des personnes et lorsque les personnes disparaissent, les traditions vivantes disparaissent avec elles de manière irrévocable.

L'approche ethnographique utilisée pour recueillir les renseignements constitue la base de l'inventaire et se caractérise par une méthode qui fait appel à l'observation directe et à l'enquête orale sur le terrain. Cette approche permet de documenter les pratiques, les savoir-faire, les connaissances et les représentations d'une communauté. L'approche ethnologique s'inscrit dans une démarche à la fois réflexive et dialogique. L'enquêteur sur le terrain et les informateurs travaillent de concert. Par exemple, lors du pré-terrain, l'enquêteur rencontre la communauté et aide ses membres à dresser une liste des pratiques à inventorier. La communauté et le chercheur se questionnent sur ce qui doit être retenu. Les critères de sélection ne sont pas dictés par l'expert, mais par la communauté qui fait des choix en fonction de l'identité culturelle du groupe. Lorsque l'on aborde le domaine de la foi et des croyances, le discours officiel, celui des dogmes et des règles auxquelles doivent se conformer les croyants, peut devenir un obstacle à l'expression de différents points de vue. Cependant, la collecte des témoignages de plusieurs informateurs dans une même communauté apporte des nuances et des éclairages différents sur la vie spirituelle et quotidienne de ses membres. Les informateurs nous parlent de leur vécu et de la particularité de leurs expériences. Le discours officiel est revu à travers des individualités ancrées dans l'identité collective et la mémoire partagée.

Après avoir mené les entrevues, le chercheur traite les données recueillies. La première étape consiste à catégoriser le témoignage. Nous inspirant de l'approche bien connue et répandue des récits de vie2, nous avons défini quatre catégories de récits, dont certaines permettent de

2 Voir notamment Daniel Bertaux, Récits de vie, Paris, Nathan, 1996; Patrick Brun, Emancipation et connaissance. Les histoires de vie en collectivité, Paris, L'Harmattan, 2001 ; le Réseau québécois pour la pratique des histoires de vie (www.rqphv.org); et Carole Dornier et Renaud Dulong (dirs.), Esthétique du témoignage, Paris, Éditions de la Maison des sciences de l'homme, 2005.

combiner le matériel et l'immatériel en se préoccupant par exemple des récits de pratiques entourant un objet ou un lieu de culte. Par exemple, les rites funéraires d'une communauté pourront être consultés sous différents aspects : la symbolique du cercueil (récit d'objet), la préparation du corps du défunt (récit de pratique), les funérailles (récit de pratique) et le cimetière (récit de lieu). Nous avons défini quatre catégories de récits :

1. Les récits de lieux portent sur l'usage et le sens des espaces les plus significatifs dans chacune des communautés, les hauts lieux de l'habitat (chapelle, sacristie, jardin, grotte, réfectoire, salle d'enseignement, cimetière, presbytère, synagogue, lieu de culte, espace communautaire);

2. Les récits d'objets renvoient aux objets matériels ayant une forte valeur symbolique et identitaire, et jugés les plus significatifs pour les communautés sur le plan patrimonial (objet religieux, vêtement liturgique, habit traditionnel, mobilier traditionnel, mobilier de cuisine, etc.);

3. Les récits de vie visent à documenter des vies ou des épisodes de vie de membres de la communauté renfermant un caractère exceptionnel et donc une valeur patrimoniale (missionnaire, artiste, artisan, enseignant, etc.) ;

4. Les récits de pratiques cultuelles et culturelles regroupent les dévotions particulières, les coutumes funéraires, les pratiques liturgiques significatives, les pratiques professionnelles marquantes, les savoir-faire uniques ayant une valeur à la fois pragmatique et symbolique dans la communauté (la statuaire, la broderie, la dentellerie, la dorure, le tressage, la fabrication d'objets religieux, la fabrication de produits alimentaires, etc.).

Les propos de l'informateur sont résumés dans des fiches correspondant à l'une des catégories. Les fiches d'inventaire sont descriptives et factuelles. Chacune des fiches comprend des données nominatives (nom, adresse de l'informateur, rôle dans la communauté, etc.), et des données techniques d'inventaire (nom de l'enquêteur, indexeur, documents audio et vidéo, date des entrevues et du traitement, etc.) Chaque récit fait l'objet de descriptions textuelles: l'historique et la description de la pratique, son actualisation ainsi que ses modes de transmission.

La deuxième étape consiste à classer les données recueillies. Contrairement à la culture matérielle qui bénéficie de systèmes de classification, comme celui de Chenhall largement employé en Amérique

du Nord[3], il n'existe pas d'équivalent pour la culture immatérielle en raison du développement récent de ce champ de connaissances. Pour y remédier, nous avons adapté la grille de pratiques culturelles de Jean Du Berger.[4] Celui-ci l'avait développée comme un outil d'analyse du fonctionnement culturel. Il n'en demeure pas moins qu'elle s'avère être aussi un outil de classification très efficace car cette grille relationnelle évoque les rapports entre les différentes pratiques culturelles et démontre leur organisation et leur fonctionnement en société. En plus de contribuer à fixer le sens des mots, elle fournit une arborescence opératoire pour le patrimoine immatériel et nous permet de structurer la base de données dans un tout cohérent.

Notre approche du patrimoine immatériel religieux est culturelle et comparative. La grille de classification permet de comparer et de faire des liens entre des pratiques religieuses de même niveau, par exemple les rites de passages ou encore l'organisation religieuse, la fabrication d'objets religieux, les espaces religieux, les gestes rituels, etc. dans différentes confessions. L'internaute peut ainsi explorer la banque de données en croisant différentes données.

La cueillette et la saisie des récits par le biais de technologies audio-visuelles numériques représente un autre élément essentiel à la fois de la méthodologie de l'inventaire et de sa diffusion sur le Web.

Le patrimoine immatériel est constitué de pratiques, celles-ci sont transmises par le geste et la parole et donc rarement consignées par l'écrit. Même lorsqu'elles sont écrites, il est souvent très difficile, voire impossible, de les reproduire en raison de l'absence des nombreux détails nécessaires à leur reconstitution. Les documents multimédias associés à la fiche descriptive (supports photo, audio et vidéonumériques) permettent de contextualiser les récits. Les gestes, l'intonation et l'émotion vécue par l'informateur en relatant ses expériences, donnent au récit une autre dimension. Puisque le patrimoine immatériel est transmis oralement, l'ajout de documents multimédia donne un nouveau sens, personnifie les valeurs de la communauté et présente le récit de façon didactique et accessible à un vaste public.

La Chaire collabore avec la Direction du patrimoine et de la muséologie du ministère de la Culture, des Communications et de la Condition

3 Robert G. Chenhall, Nomenclature for Museum Cataloguing: A System for Classifiing Man-made Objects, Nashville, TN, AASLH Press, 1978 ; James R. Blackaby, Patricia Greeno, and The Nomenclature Committee (ed.) Robert G. Chenhall, Revised Nomenclature for Museum Cataloging: Revised and ExpandedVersion of Robert G. Chenhall's System for Classifying Man-Made Works, Nashville, TN, AASLH Press, 1988.

4 Jean Du Berger, Grille de pratiques culturelles, Québec, Septentrion, 1997.

féminine du Québec (MCCCF) afin d'intégrer sa base de données sur le patrimoine religieux immatériel à la banque de données ministérielle qui recèle déjà une grande quantité d'informations sur le patrimoine immobilier (bâtiments et sites) et mobilier (meubles, œuvres d'art, vêtements, artefacts) religieux. L'internaute aura alors la possibilité de cliquer sur une fiche d'inventaire d'une église classée et d'y trouver des informations sur l'architecture et également sur tous les biens patrimoniaux mobiliers et immatériels associés, contenant des fiches descriptives des principales œuvres artistiques et artisanales accompagnées de photos, d'images en 3D, et d'enregistrements audio-visuels. Par un simple clic, l'internaute accédera aux récits de lieux, d'objets, de pratiques et de vie. Cette banque de données offrira une vision complète et intégrée du patrimoine. Connue sous l'acronyme PIMIQ (Patrimoine immobilier, mobilier et immatériel du Québec), elle représentera, à notre connaissance, la première banque de données informatisées du genre au monde. À terme, l'intégration de données du patrimoine immatériel religieux sur le site Web du MCCCF, le Répertoire du patrimoine culturel du Québec, donnera un accès direct aux documents textuels, iconographiques, sonores et vidéo des différentes traditions religieuses sur leurs lieux de culte, leurs objets et leurs pratiques cultuelles et culturelles.

Plus qu'un simple inventaire destiné à la conservation, la Chaire a développé une base de données multimédia virtuelle qui facilite la communication du patrimoine immatériel religieux. La communication est le meilleur moyen de conserver le patrimoine immatériel religieux dans la mesure où il participe à sa transmission. La transmission n'assure pas seulement la conservation des traditions, elle contribue à les transformer, à les dynamiser et à les renouveler en leur trouvant de nouveaux usages sociaux. Par la même occasion, elle participe à la valorisation et à la reconnaissance de ceux qui les transmettent. Il ne s'agit pas de promouvoir le culte, mais de faire connaître et reconnaître les pratiques traditionnelles--cultuelles et culturelles-- et leurs artisans en tant que patrimoine, dans un souci d'éducation du grand public. L'usage d'équipements d'enregistrement électroniques, de bases de données numériques et des applications Web pour exploiter ces bases a contribué à révolutionner les pratiques de l'inventaire du patrimoine immatériel. Les nouvelles technologies de l'information facilitent non seulement la fabrication, l'accès et la gestion des inventaires, elles suscitent de nouvelles façons de concevoir et de réaliser l'inventaire lui-même. Les communautés et les porteurs de traditions peuvent participer plus facilement au processus de collecte et de communication des données. L'accès aux données par le Web permet des appropriations et réappropriations multiples par une large gamme de personnes

(communautés elles-mêmes, journalistes, muséologues, chercheurs) et favorise l'évolution des pratiques et la valorisation sociale des communautés qui en sont les détentrices. En effet, l'inventaire virtuel devient un outil dynamique de communication patrimoniale et de développement culturel et social.

En plus d'aider directement les communautés dans l'identification de leur riche patrimoine immatériel, la mise en ligne des récits des communautés contribuera à une meilleure connaissance des traditions religieuses qui ont façonné le Québec.

3 L'Encyclopédie du patrimoine culturel de l'Amérique française et les nouvelles tendances web

L'Encyclopédie du patrimoine culturel de l'Amérique française est avant tout un projet de diffusion des connaissances contemporaines sur le patrimoine, incluant les nouvelles manières de le concevoir, de l'étudier et de le communiquer. Dans cette encyclopédie diffusée exclusivement sur Internet depuis avril 2008, le web est non seulement un moyen novateur de mieux communiquer toutes les dimensions du patrimoine, qu'elles soient matérielles ou immatérielles, textuelles, visuelles ou sonores, intellectuelles ou émotives, mais il est également une source de réflexion stimulante sur la relation entre le patrimoine et les gens qui le vivent.

Depuis plusieurs années, les campagnes de mise en ligne de millions de documents textuels, visuels, sonores et audiovisuels ont rendu accessibles plusieurs collections d'archives et de musées qui constituent le fondement d'un important patrimoine collectif. Les bâtiments, les lieux, la faune et la flore, les festivals et autres événements culturels, qu'ils soient des attractions locales ou des éléments du patrimoine mondial, sont également présents sur le web et, de ce fait, accessibles comme jamais par le passé. Cette accessibilité accrue de multiples éléments du patrimoine offre des possibilités nouvelles et change les perspectives dans ce domaine en effervescence.

L'Encyclopédie puise dans ces banques de données en ligne afin de sélectionner l'information la plus pertinente pour compléter ses articles. Elle participe également à ce processus de diffusion en numérisant elle-même nombre de documents multimédia inédits qu'elle met à la disposition des internautes. Elle vise à accroître la connaissance et la compréhension du patrimoine. En effet, grâce aux textes de nos articles, rédigés par des spécialistes, qui décrivent et expliquent des éléments

majeurs et parfois méconnus du patrimoine des francophones d'Amérique, ainsi que l'histoire de leur formation et de leurs transformations, nous rendons disponibles ces connaissances à travers le monde. De plus, grâce aux nombreux documents multimédia qui permettent aux internautes de prendre contact plus intimement et plus directement avec un lieu, un bâtiment, une œuvre d'art, un savoir-faire, un rituel, un accent, une personne, le patrimoine prend vie, bien que de façon virtuelle. Rappelons cependant que le web est un puisant incitatif à visiter et à participer en chair et en os au patrimoine parfois découvert par l'entremise d'Internet.

3.1 Les initiatives en cours

À la jonction des perspectives du web 3.0 et des approches multidisciplinaires en sciences humaines, l'Encyclopédie s'efforce d'offrir une documentation intégrée sur le patrimoine. Non seulement des auteurs de disciplines diverses rédigent nos articles : ethnologues, historiens, littéraires, biologistes, gestionnaires, et autres, afin de couvrir les trois grandes catégories du patrimoine reconnues par l'UNESCO (immatériel, matériel et naturel), mais l'approche que nous privilégions pour décrire et analyser le patrimoine – la patrimonialisation – amène les auteurs à réfléchir à la convergence de plusieurs facteurs. Le patrimoine se forme en effet sous l'influence de valeurs culturelles dominantes, qui se transforment dans le temps, d'acteurs sociaux divers, organisés ou non, et répond à besoins économiques, sociaux et culturels de la collectivité qui varient eux aussi au fil du temps. Ainsi, l'Encyclopédie présente une information « convergente » sur l'évolution dynamique du patrimoine. Cette approche s'inscrit dans les réflexions actuelles les plus pointues sur les phénomènes humains, et dans les perspectives de développement du web.

Une autre pratique d'intérêt de l'Encyclopédie, en lien avec le web 3.0, consiste à joindre à chaque article de l'Encyclopédie une documentation multimédia qui en facilite la compréhension fine et détaillée. Bien sûr, des illustrations permettent de voir les sujets dont il est question dans les articles. Cet usage est fort répandu. Mais des documents audiovisuels, des chansons, des articles de journaux, des œuvres d'art et des témoignages sonores s'ajoutent aux illustrations. Cet ensemble de documents multimédias sélectionnés pour leur pertinence donnent accès à la profondeur culturelle du patrimoine décrit dans les articles. À l'inverse, le texte des articles facilite la compréhension et la contextualisation des documents multimédias présentés en lien avec les articles. Ce travail de recherche, de sélection et de présentation de documents complémentaire aux articles demande patience et réflexion. Car la convergence des informations sur un sujet donné, en vue d'en faciliter la compréhension et

d'en approfondir la connaissance, n'est pas évidente à établir. Cette pratique développée dans l'Encyclopédie, et la réflexion qui la sous-tend, sont propices au développement d'un web plus « intelligent », qui serait davantage en mesure de rassembler une information variée et pertinente sur un sujet donné, alors que cette information se trouve aujourd'hui le plus souvent disséminée au travers d'innombrables sites, très peu connectés les uns aux autres. Notre expérience à ce niveau suggère que les progrès souhaités dans le développement du web 3.0 représentent un défi de taille.

3.2 Les développements à venir

L'Encyclopédie tente de tirer profit des développements rapides du web qui offrent constamment de nouvelles possibilités. La numérisation 3D, par exemple, dans laquelle s'engage l'Encyclopédie grâce à des appareils de numérisation maintenant portatifs, permettra un contact inégalé avec les objets du patrimoine conservés dans les musées, même si ce contact n'est que virtuel. En effet, les visiteurs des institutions muséales n'ont que très rarement la possibilité de manipuler les objets qui sont exposés, alors que la technologie 3D leur permettra d'observer ces objets sous tous les angles, à leur guise, en « manipulant » les images 3D diffusées sur notre site. Non seulement cette technique donnera-t-elle accès à tous les détails des objets, mais elle s'adaptera de surcroît aux intérêts et aux impulsions de chacun. Elle accroît donc la qualité et l'étendue de nos rapports aux objets du patrimoine, sans risque de détérioration de ceux-ci, grâce à la médiation du web.

L'Encyclopédie explore également le potentiel web dans le domaine du patrimoine immatériel, principalement dans la section de l'Encyclopédie destinés spécifiquement aux jeunes de 14-16 ans. Nous réalisons à leur intention divers modules interactifs dont le plus ambitieux porte sur la démocratie, en tant que valeur et savoir-faire clé de notre société. Ce module s'articule autour d'une simulation de haut niveau (serious gaming) des pratiques démocratiques actuelles et émergentes. Sur la base d'informations résumant l'évolution de la démocratie au Québec, son fonctionnement, ses institutions et son impact sur la société, depuis l'instauration du gouvernement responsable (XIX^{e} siècle) jusqu'aux tendances les plus récentes (notamment l'utilisation du web lors de la récente campagne du président américain Barack Obama), ce module proposera aux participants de relever le défi suivant. Il s'agira de résoudre un problème de nature complexe par le biais des processus démocratiques : soit la conciliation du développement économique et de la protection de l'environnement dans une perspective durable. Les données de base de cette simulation reflèteront la diversité des enjeux, des acteurs et des opinions présents dans la société.

Les données statistiques sur les choix privilégiés par les participants à ce « jeu sérieux » deviendront progressivement le principal élément de la prise de décision démocratique qui permettra de résoudre le problème posé. Celle-ci évoluera donc au fur et à mesure que les participants s'additionneront. Elle sera également influencée par divers modes de scrutins qui seront proposés aux participants (majorité simple, système proportionnel, choix multiples énumérés en ordre de priorité sur les bulletins de vote, et autres). Enfin, elle illustrera clairement comment une opinion personnelle peut se transformer en une position influente au niveau collectif à travers l'engagement politique, la communication média et d'autres formes d'action publique tel le réseautage web (Facebook, etc.) Des données réelles reposant sur l'histoire, ainsi que des exemples récents de conciliation économie/environnement bien documentés, orienteront la simulation et canaliseront le parcours ludique. Ce module interactif s'avèrera donc à la fois un lieu d'information sur la problématique proposée et sur le processus démocratique. Elle servira à consolider la connaissance et la valeur de notre patrimoine démocratique en rappelant l'impact de la pratique démocratique sur les transformations sociales, culturelles et institutionnelles. Enfin, elle permettra de simuler des voies démocratiques en émergence qui pourraient se matérialiser bientôt dans notre société.

En considérant que ce « jeu sérieux » offrira de plus une excellente base à des animations de groupe sur les processus démocratiques, par exemple en classe, on constate que le web permet dans ce cas-ci une intégration très poussée de plusieurs facettes complémentaires de ce phénomène social complexe et important qu'est la démocratie, et ce à un coût raisonnable. Seul le web permet aujourd'hui un processus d'apprentissage interactif aussi global.

En explorant diverses possibilités nouvelles du web, l'Encyclopédie remplit pleinement son mandat d'éclairer le dynamisme du patrimoine, ce phénomène en constante transformation qui accompagne l'évolution de la société.

4 Passage à un nouveau paradigme technologique : le LEEM 2

La sauvegarde d'informations ethnologiques, réalisée encore récemment sur supports analogiques (bandes magnétiques et films), exigeait des équipements lourds, de longs séjours sur le terrain, des conditions de conservation particulières (salles à température et à humidité contrôlées) et des coûts élevés. Nous avons transformé ces modes de recherche. Nos

équipements numériques nous ont permis d'innover en renouvelant les méthodes d'enregistrement, de conservation, d'étude et de valorisation du patrimoine immatériel. Plus encore, c'est notre approche intégrée de ces différentes technologies qui nous a permis d'innover dans ce domaine.

Nous avons placé l'utilisation de technologies numériques au coeur de nos pratiques, depuis la collecte jusqu'à la sauvegarde en passant par la diffusion. Nos travaux sont fondés sur la mise à profit d'appareils numériques portables, sélectionnés sur la base de leur simplicité d'utilisation. Ces outils simples et efficaces, capables de produire des contenus de grande qualité avec un minimum de ressources humaines, nous permettent d'alléger de manière considérable la gestion de nos opérations. Notre succès doit beaucoup à cette approche fondée sur la simplicité. Elle nous permet à la fois d'alléger la formation des chercheurs, le transport des équipements sur le terrain et leur entretien. Elle facilite de plus la propagation de notre méthode, autant en Occident que dans les pays en voie de développement. Enfin, ce parti pris nous met à l'abri d'une trop grande dépendance envers des techniciens et des ingénieurs, ce qui s'avère souvent coûteux pour des équipes en sciences humaines.

Nos premières années d'expérimentation nous ont permis de développer, valider et peaufiner nos méthodes. Nous sommes désormais prêts à passer à une seconde étape, où nos projets de recherche gagneront en appui technologique. Après une première phase qui revisitait par le numérique des pratiques ethnologiques classiques (photographie, film, enregistrement sonore monocanal ou stéréo), nous mettrons maintenant à profit des outils directement issus de l'ère de l'informatique multimédia: numérisation 3D et captations audiovisuelles immersives.

L'ajout de la numérisation 3D couleur aura un impact majeur sur nos activités d'enquête et d'inventorisation. Cette technologie consiste à enregistrer la forme et la couleur d'un objet à l'aide d'un appareil à balayage laser. Le patrimoine immatériel est le plus souvent inextricablement lié au patrimoine matériel. Nous avons donc besoin d'archiver les traces artefactuelles des pratiques et rites que nous étudions, de nous référer aux objets pour comprendre les idées et contextes qui les ont fait naître. Pour l'instant, ces besoins sont partiellement pris en charge par la photographie. Mais un artefact tridimensionnel offre une représentation infiniment plus juste, et comporte une importante valeur ajoutée sur le plan de la diffusion, car l'objet virtuel peut être observé à l'écran sous tous ses angles et transmis par voie électronique comme tout autre fichier numérique.

Même si la numérisation 3D couleur existe depuis des années, ce n'est qu'avec le lancement du balayeur laser VIUscan en 2008, par la compagnie canadienne Creaform, que son potentiel peut à notre avis s'actualiser dans le domaine du patrimoine culturel. Aucun appareil avant celui-ci ne permettait de se déplacer sur le terrain ou dans une réserve de musée avec un numériseur 3D couleur portable. Les artefacts devaient être transportés dans l'un des rares laboratoires équipés à cette fin, des opérations coûteuses qui prenaient plusieurs jours par objet, et impliquaient l'emballage méticuleux et le transport de chaque objet. Or, le nouvel appareil permet la numérisation en couleur de dizaines d'artefacts par jour, voir davantage, et cela in situ, donc sans emballage et transport préalable des objets. Ce nouveau paradigme facilite de manière radicale les opérations et s'accorde à la prédilection du LEEM pour des technologies simples d'utilisation.

Nous nous investirons donc au cours des prochains mois au développement de standards et meilleures pratiques pour le travail sur le terrain, l'archivage, l'analyse et la diffusion des données 3D. Un projet pilote, dans le cadre de l'Encyclopédie du patrimoine culturel de l'Amérique française, sera d'ailleurs mené au cours des prochains mois. Ce projet pilote devrait permettre de valider notre méthodologie.

Une autre technologie qui s'ajoute à nos équipements et qui bonifiera notre programme de recherche est la captation audiovisuelle panoramique. Jusqu'à maintenant, nos enquêteurs cadraient les scènes à filmer, pointaient le micro dans des directions précises. Désormais, à l'aide d'un micro ambiophonique et d'une caméra vidéo LadyBug, nous pourrons au besoin enregistrer des paysages sonores multicanaux (5.1) et des vidéos immersives. Ces médias nous permettront de capter sur 360 degrés le son et les images de lieux porteurs de patrimoine immatériel.

La numérisation 3D et la captation audiovisuelle panoramique ouvrent tout un univers de défis sur le plan de la diffusion sur le Web. Profitant des bases solides que nous offrent les projets IREPI, IPIR et l'Encyclopédie du patrimoine culturel de l'Amérique française, notre équipe procédera progressivement à différentes expériences, de manière à pouvoir proposer sous peu des solutions pertinentes dans le domaine. Mais nous pouvons déjà annoncer que ces solutions privilégieront des formules simples et accessibles à tous, une optique qui s'est jusqu'ici révélée efficace à la fois pour les chercheurs de notre Chaire et pour les publics qui consultent ses productions sur le Web.

Contexte et connexions: L'Irlande et le patrimoine irlandais

Michael BUCKLAND (1), Ryan SHAW (1), Daniel F. MELIA (2)

(1) School of Information, University of California, Berkeley (2) Celtic Studies Program, University of California, Berkeley

Mots-clés : Documents, Irlande, meta-données, patrimoine, recherche, vocabulaire.

Keywords: Documents, Ireland, metadata, cultural heritage, search, vocabulary.

Résumé : Le savoir nécessite qu'on s'informe du contexte et de ses connexions. Comment faciliter la compréhension du contexte historique et culturel d'un patrimoine ? A l'Université de Californie, Berkeley, l'Electronic Cultural Atlas Initiative et la School of Information développent un environnement en ligne qui reproduit le milieu éducatif d'une collection de référence dans une bibliothèque. Une interface « Context finder » facilite les recherches dans les sources recommandées. Le système « Context builder » ajoute des liens à une source d'explication du texte en XML. La technique « Context provider » ajoute en sens inverse au texte ces liens vers la source d'explication. Ainsi un réseau explicatif s'établit.

Abstract: Understanding depends on knowing context and relationships. How can learning the historical and cultural context of cultural heritage be facilitated ? At the University of California, Berkeley, researchers in the Electronic Cultural Atlas Initiative and in the School of Information are developing online services to resemble the educational environment of a library reference collection for use by anyone reading any text online. A « Context finder » interface supports search in recommended resources. A « Context builder » system adds links to an explanatory resource into the text in XML. A « Context provider » technique adds to a resource reversed links back to the text.

1 Introduction

Francis Bacon écrivit « Savoir, c'est pouvoir », *quia ignoratio causae destituit effectum* : pour obtenir des résultats, on doit comprendre les rapports entre les choses, comment elles sont liées, les unes avec les autres. Dans le domaine des patrimoines, il faut prendre garde au contexte

historique et culturel. Sinon, les objets culturels, matérials ou numériques, ne signifient pas grande chose.
Depuis plusieurs années deux groupes à l'Université de Californie, Berkeley, travaillent les techniques numériques permettant de donner accès et compréhension de ressources informationnelles sur l'héritage culturel. L'Electronic Cultural Atlas Initiative (ECAI, Initiative pour Atlas Culturel Electronique, <ecai.org>)) [1] , est une collaboration internationale qui promeut des techniques pour l'analyse de données culturelles géospatiales et géotemporelles. A la School of Information Ray Larson, Fredric Gey et Michael Buckland dirigent un groupe appelé Metadata Research Program. Ce groupe poursuit des travaux sur les problèmes de récupération de l'information, notamment comment exploiter les métadonnées pour la recherche multilingue, géographique, et historiques, à travers les genres numériques diverses.
L'ECAI et le Metadata Research Group collaborent pour rendre plus efficace la consultation d'ouvrages de référence dans l'environnement Web [1][2][3]. Nous vous présentons l'organisation des ressources explicatives relatives à l'histoire et au patrimoine irlandais pour lecteurs des périodiques numérisés au Centre for Data Digitisation and Analysis, The Queen's University, Belfast, dirigé par Paul Ell[2].

2 Patrimoine et document

2.1 Le passé, l'histoire, et patrimoine

Il convient de distinguer le passé, ce qui est arrivé autrefois, l'histoire, les récits, toujours imparfaits, multiples, et partiaux du passé, et le patrimoine, ces éléments du passé et de l'histoire que nous retenons. Le passé est parti et n'est plus connaissable que par les histoires, les documents, et les traditions. Le patrimoine est ce que nous avons aujourd'hui du passé. Le patrimoine est ce que nous retenons, soit hérité soit choisi, entre mémoires, traditions, et objets. Notre patrimoine est la culture et les objets culturels que nous avons absorbés, retenus ou construits. Nous façonnons notre patrimoine et notre patrimoine nous forme.

2.2 Document et savoir

Quand nous utilisons n'importe quelle technologie numérique nous ne nous occupons pas directement de conceptions abstraites, mais de données, de textes, et d'autres objets concrets. La technologie est nécessairement matérielle. Donc nous nous occupons indirectement du

1 ECAI : < http://ecai.org >

2 CDDA : < http://www.qub.ac.uk/research-centres/CentreforDataDigitisationandAnalysis/ >

savoir. Nous nous occupons directement de signes, de représentations de la connaissance, d'objets que nous considérons comme significatifs. On pourrait dire que nous nous occupons de documents, mais de documents dans n'importe quelle forme. Les documents ne sont pas seulement faits de texte. Parler de « document » de cette façon n'est pas original. En 1937 l'Institut International de Coopération Intellectuelle, une organisation crée par la Société des Nations, a collaboré avec l'Union Français des Organismes de Documentation, à fin de définir des termes techniques, y compris « document » :

> Document : Toute base de connaissance, fixée matériellement, susceptible d'être utilisée pour consultation, étude ou preuve. Exemples: manuscrits, imprimés, représentations graphiques ou figurées, objets de collections, etc. [4, page 234].

Suzanne Briet (1894-1989), bibliothécaire et documentaliste [5][6], a avancé ce concept de « document » en 1951 dans son manifeste *Qu'est-ce que la documentation?* [7][8] Elle déclare, tout d'abord, que « Un document est une preuve à l'appui d'un fait ». Ensuite, elle explique qu'un document est « . . . tout indice concret ou symbolique, conservé ou enregistré, aux fins de représenter, de reconstituer ou de prouver un phénomène ou physique ou intellectuel. » [7, 7]. Par conséquent on ne peut pas considérer que le métier de documentaliste s'occupe de textes, mais, plutôt, de toute espèce de preuve, de témoignage, d'évidence et que cette preuve (« le document ») est de forme concrète et non pas abstraite. Remarquons que Briet a employé le mot « indice ». Le mot « indice » veut dire qu'un objet ne devient une preuve (un document) que si cet objet est situé en rapport avec d'autres preuves, des autres documents. C'est à dire que les documents doivent être arrangés les uns par rapport aux autres. Aujourd'hui nous voulons déléguer autant que possible les tâches documentaires aux logiciels.

Une approche plus contemporaine serait de dire que le sens est construit par l'observateur et que tout objet pourrait, dans certaines situations, être preuve, être un « document ». Donc tout objet peut devenir signifiant. Tout objet concret peut être un document. Nous retenons deux suppositions de Briet: Que tout objet peut être un document; et que la documentation concerne relations entre ces objets [9].

2.3 Document et patrimoine 3.0

Il convient de considérer un document numérique à trois niveaux :

Direct : Un document numérique existe et on peut l'examiner sur le Web.

Interne : L'analyse d'un document (ou collection de documents) par l'exploration et l'analyse de données (data-mining, cluster analysis, natural language processing, etc.).

Externe : Les rapports entre un document (ou collection de documents) et son contexte.

Si « Patrimoine 1.0 » signifie l'acquisition d'objets et de données numériques dans le cadre des bibliothèques et musées, et si l'esprit Web 2.0 permet la participation de tout le monde, on peut proposer que pour « Patrimoine 3.0 » il faut créer la cyberinfrastructure des liens et rapports entre l'individu, les objets culturels, et leurs contextes.

3 Savoir et contexte

Selon Bacon, savoir n'est pouvoir que si on comprend comment une chose est rapportée à son contexte. Pour Briet ce qu'un document signifie dépend de ses rapports avec son contexte. Alors, comment faciliter la compréhension du contexte et des rapports de n'importe quel sujet ?

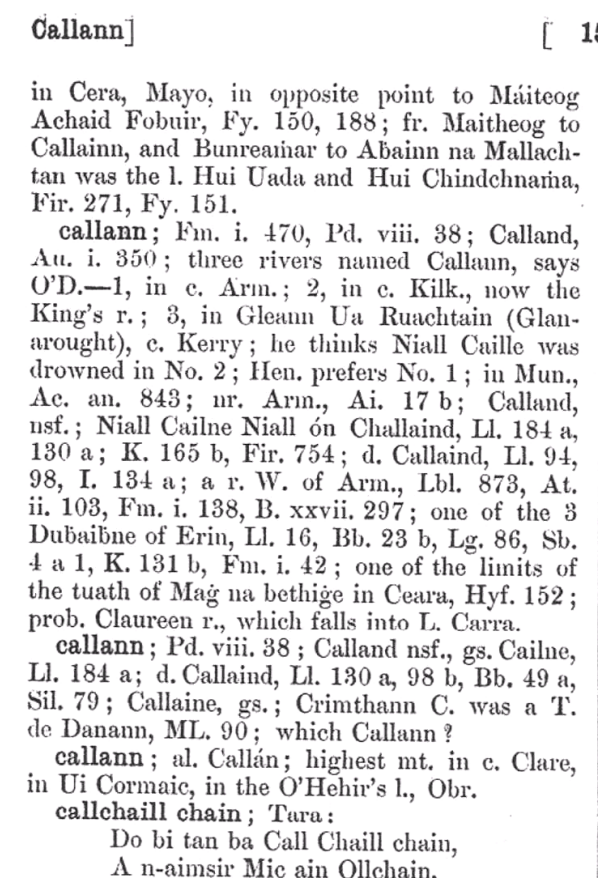
3.1 Ouvrages de référence

Pendant des siècles, on a développé plusieurs genres d'ouvrages de référence, par exemples, des bibliographies et des catalogues de documents, les dictionnaires biographiques, les cartes et dictionnaires géographiques, les histoires et chronologies d'événements, les encyclopédies et les dictionnaires. Par conséquent, la salle de bibliographies d'une bibliothèque constitue un milieu admirable, plein de ressources dignes de confiance pour s'instruire sur n'import quel domaine, y compris le patrimoine.

C'est la même Suzanne Briet qui, en 1934, a fait naître à la Bibliothèque Nationale la Salle de Catalogues et Bibliographies. Elle a sélectionnés les ouvrages de références les plus utiles, les a retirés des rayons sans libre accès, et les a rassemblés dans une salle avec libre accès bien adaptée pour les lecteurs [10]. Malheureusement, le service si éducatif d'une collection d'ouvrages de référence d'une bibliothèque n'existe guère sur le Web. Nous avons besoin d'un tel service [11].

Notons qu'une version numérique d'un ouvrage imprimé ne suffit pas. L'adoption de technologies numériques implique deux étapes. D'abord on utilise des techniques numériques pour obtenir les mêmes résultats de manière plus efficace ; ensuite les capacités du logiciel sont exploitées à fin de développer de nouveaux services plus utiles. Considérons un dictionnaire géographique de l'Irlande, le *Onomasticon Goedelicum* par Edmund Hogan [12] qui a indexé les noms géographiques mentionnés dans plusieurs documents anciens (Fig 1). Cet œuvre essentielle, parue en 1910, symbolise les ressources imprimées : le texte est très abrégé et peu commode. Les exemplaires sont assez rares mêmes dans les bibliothèques. Une version numérisée du texte imprimé se trouve sure le Web[3].

3 The Locus project : < http://www.ucc.ie/locus/>



Callann] [1

in Cera, Mayo, in opposite point to Máiteog Achaid Fobuir, Fy. 150, 188; fr. Maitheog to Callainn, and Bunreamar to Abainn na Mallachtan was the l. Hui Uada and Hui Chindchnama, Fir. 271, Fy. 151.

callann; Fm. i. 470, Pd. viii. 38; Calland, Au. i. 350; three rivers named Callann, says O'D.—1, in c. Arm.; 2, in c. Kilk., now the King's r.; 3, in Gleann Ua Ruachtain (Glanarought), c. Kerry; he thinks Niall Caille was drowned in No. 2; Hen. prefers No. 1; in Mun., Ac. an. 843; nr. Arm., Ai. 17 b; Calland, nsf.; Niall Cailne Niall ón Challaind, Ll. 184 a, 130 a; K. 165 b, Fir. 754; d. Callaind, Ll. 94, 98, I. 134 a; a r. W. of Arm., Lbl. 873, At. ii. 103, Fm. i. 138, B. xxvii. 297; one of the 3 Dubaibne of Erin, Ll. 16, Bb. 23 b, Lg. 86, Sb. 4 a 1, K. 131 b, Fm. i. 42; one of the limits of the tuath of Mag na bethige in Ceara, Hyf. 152; prob. Claureen r., which falls into L. Carra.

callann; Pd. viii. 38; Calland nsf., gs. Cailne, Ll. 184 a; d. Callaind, Ll. 130 a, 98 b, Bb. 49 a, Sil. 79; Callaine, gs.; Crimthann C. was a T. de Danann, ML. 90; which Callann?

callann; al. Callán; highest mt. in c. Clare, in Ui Cormaic, in the O'Hehir's l., Obr.

callchaill chain; Tara:

Do bi tan ba Call Chaill chain,
A n-aimsir Mic ain Ollchain.

Figure 1 : Page typique de l'Onomasticon Goedelicum par Edmund Hogan.

Un dictionnaire géographique comme celui d'Hogan peut être transformé dans l'environnement Web. Les détails abrégés seraient développés. On préciserait la latitude et la longitude de chaque endroit à fin de permettre des visualisations cartographiques et des analyses géographiques [13][14]. Les références signalétiques aux textes cités par Hogan pourraient devenir des liens au paragraphe cité (si le texte visé est disponible sur le Web) ou aux cotes de placement dans les bibliothèques voisines. On pourrait chercher chaque localité sur les cartes numérisées anciennes ou modernes. Chaque personnage mentionné pourrait être relié aux dictionnaires biographiques. Dans un contexte de « Patrimoine 3.0 » un livre comme le « Hogan » n'est plus le même plus ouvrage.

Chaque lien utile de « Hogan » à un autre ouvrage peut être noté et ajouté en XML. Ainsi « Hogan » est enrichi constamment. Egalement, ces liens peuvent être établis en sens inverse et de cette manière l'ouvrage cité citera lui même « Hogan ». Comme cela un réseau se bâtit, un réseau de rapports et de contextes.

3.2 Contexte et vocabulaire

Un vocabulaire évolue au sein d'une communauté et d'un domaine discursif. Chaque contexte culturel ou scientifique développe sa terminologie spécialisée, son propre dialecte. Donc pour s'instruire sur n'import quel contexte il faut employer la terminologie spécialisée du domaine [15].
De plus, dans les systèmes d'indexation les termes sont très souvent des adaptions plus ou moins artificielles de la langue courante (par exemple: « God -- Knowableness -- History of Doctrines -- Early Church, ca. 30-600 ») ou alors une notation artificielle est employée (par exemple « 330 » signifie « Sciences Economiques » dans la Classification Décimale de Dewey). Ce sont des systèmes pour coder le savoir (KOS). On a reconnu depuis longtemps que les systèmes d'indexation sont des langues. On parle aujourd'hui de « métadonnées », mais avant « métadonnées » on parlait de « langages documentaires », « langages d'indexation », ou bien « métalangages » (c.f. Maurice Coyaud [16]).

4 Conditions requises

4.1 Ressources recommandées

Les moteurs de recherche (comme Google) découvrent des ressources touchant n'importe quel sujet mais ils manquent de sélectivité. On devrait employer de préférence les ressources les plus dignes de confiance. Donc il faut établir un ensemble d'ouvrages soigneusement choisi. Evidemment le choix dépend du sujet et, également, de l'utilisateur. C'est la même tâche que le développement d'une collection d'ouvrages de référence dans une bibliothèque.

4.2 Partage de ressources de pair à pair (peer to peer)

Un service éducatif de type « Patrimoine 3.0 » doit déléguer la recherche aux logiciels autant que possible. Le partage de ressources de pair à pair est effectué par des protocoles comme Z39.50 (ISO 23950 Information retrieval (Z39.50) : Application service definition and protocol specification), SRU (Search/Retrieve via URL), et CQL (Common Query Language).
De plus, les moteurs de recherche n'indexent que le Web superficiel (the open web) et n'atteignent pas le Web profond (the deep web), pour lequel on a besoin de recherches de pair à pair (federated search). Mais, jusqu'à maintenant la plupart des ressources de patrimoine électroniques ne supportent pas encore les protocoles de pair à pair. Les guides bibliographiques ne disent pas encore quels protocoles sont utilisables avec quelles ressources.

4.3 Service recommandeur de termes de recherche

Tout système de recherche parcourt de multiples vocabulaires [17]. Même quand un texte non-édité est parcouru avec une requête simple, au moins deux vocabulaires son présents :
1. Le vocabulaire de l'auteur du document, ou bien les vocabulaires de plusieurs auteurs; et
2. Le vocabulaire du chercheur.
Dans les systèmes opérationnels actuels, on trouve simultanément plusieurs vocabulaires. Un catalogue de bibliothèque, par exemple, contient au moins trois vocabulaires en plus des précédents :
3. Le vocabulaire d'indexation du documentaliste, qui modifie ou complémente le vocabulaire de l'auteur.
4. Les renvois -- EM (Employer); EP (Employé pour); etc. -- pour harmoniser ou corriger le vocabulaire des documentalistes;
5. Le vocabulaire de la personne effectuant la recherche formulé selon par les exigences de la requête.
Il y a toujours des vocabulaires multiples en jeu et cette multiplicité est une cause fréquente d'erreurs. Un chercheur peut employer le terme A et un auteur a employé le terme B, même s'ils voulaient indiquer le même signifiant -- des synonymes. Cependant, il est possible que tous les deux employaient le terme A pour indiquer deux sens différents -- des homographes.
Les vocabulaires intermédiaires (que ce soit celui du documentaliste, d'une requête formulée, ou la structure de renvois) normalisent l'usage de termes afin de rectifier des discordances. Les termes d'indexation du documentaliste rectifient le titre donné par l'auteur en représentant le sujet du document à travers un vocabulaire standardisé. Les chercheurs expérimentés savent comment modifier leurs requêtes ou celles des autres d'une façon que le système y réponde utilement.
Il y a autant de re-représentations que de transitions d'un vocabulaire à un autre. Chacune de ces re-représentations présente une opportunité de rectifier les dissonances entre chercheur et document, mais aussi la possibilité de nouvelles dissonances. Un bon intermédiaire de recherche (humain ou informatisé) sait adapter sa terminologie au vocabulaire du système. L'accès efficace aux sources spécialisées exige des services recommandeurs de termes de recherche [18].
Les bases de données ont un seul index créé pour la base entière même si celle-ci couvre plusieurs domaines discursifs. A cause de cette multiplicité, on voudrait trouver autant d'indexes que de communautés d'utilisateurs [19].

4.4 Extrêmement convenable

Les éducateurs et les bibliothécaires se plaignent que les services les plus faciles, surtout Google et le Wikipédia, sont les services les plus

fréquemment choisi au lieu des ressources les plus dignes de confiance. La solution est évidente: Il nous faut construire de services aussi convenables que Google et le Wikipédia et aussi fiables que les ouvrages de référence imprimés. C'est une question d'art, de génie documentaire.

5 Un projet de documentation patrimoniale

Depuis quelques années une équipe à l'Université de Californie, Berkeley, travaille sur les techniques nécessaires pour l'exploitation avec la moindre difficulté les meilleures ressources disponibles à travers l'internet [20]. Nous voulons construire trois outils :

1. Le « Context Finder » trouve le contexte de toute chose. Si on rencontre quelque chose d'intéressant, on reçoit un éclaircissement d'une ressource de confiance avec deux « clics ». Le lecteur clique sur un nom dans le texte en ligne; l'ordinateur pose la question si c'est le nom d'une personne ou d'un endroit et répond par une liste brève de ressources recommandées adaptée au texte et, en principe, à la compréhension du lecteur. Un deuxième clic sélectionne la ressource et un lien dynamique lance un recherche automatique et présente le résultat.
2. Le « Context Builder » construit le contexte trouvé. Il insère les détails de la recherche, de la ressource exploitée, et de l'explication dans le texte même en format XML. Ainsi ces données sont prêtes pour la lecture prochaine et le lecteur prochain.
3. Le « Context Provider » récupère les liens en tout sens. Le « Context builder » permet l'accumulation de liens dynamiques à sens unique entre un texte et les ouvrages explicatifs. Ces liens peuvent être insérés aussi en sens inverse à fin d'enrichir les ouvrages explicatifs avec les liens aux textes qui mentionnent ces mots comme indiqué dans l'exemple du « Hogan ».

6 Un prototype d'interface

Notre premier prototype d'interface est montrée dans les figures 2 bis 4.

Figure 2 : Un étudiant lit en ligne le page 341 d'un livre numérisé de l'Open-Access Text Archive de l'Internet Archive[4]: The Story of the Irish Nation par Francis Hackett (New York: The Century Co., 1922). Le logiciel de l'interface discerne automatiquement les noms propres qui apparaissent dans cette page : Edward Martyn, George Moore, William Butler Yeats, etc., et génère le liste à droite. Le lecteur qui veut s'informer sur Yeats met le curseur sur ce nom soit dans la liste soit dans la page et « Yeats » s'allume.

4 Internet Archive Open-Access Text Archive : <http://www.archive.org/details/texts>

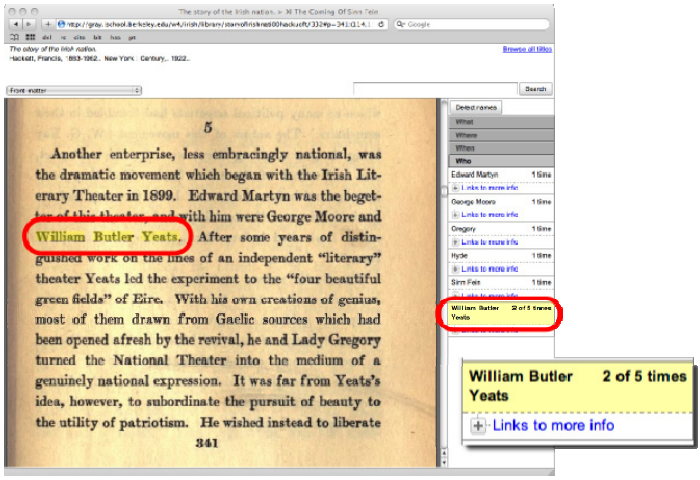


Figure 2 : L'interface « Context Finder » : Le lecteur met le curseur sur le nom William Butler Yeats.

Figure 3 : Le lecteur clique sur le choix « Links to more info » au-dessous de « William Butler Yeats » et l'interface offre une liste de liens aux ressources recommandées.

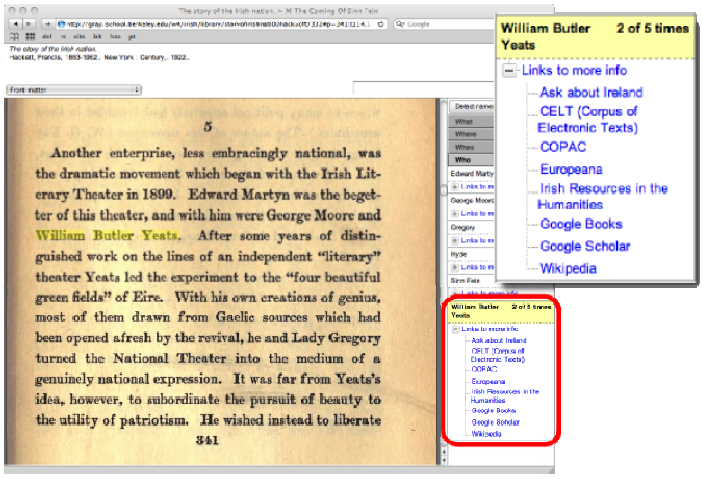


Figure 3 : Le « Context Finder » : L'interface liste des ressources à consulter sur William Butler Yeats.

Figure 4 : Le lecteur sélectionne « Europeana », clique ce lien et reçoit les résultats d'une recherche dans « Europeana » sur Yeats en temps réel. Notons que c'est l'interface elle-même qui interroge « Europeana ». Le

lecteur ne fait rien que cliquer deux fois. C'est une recherche très facilité !

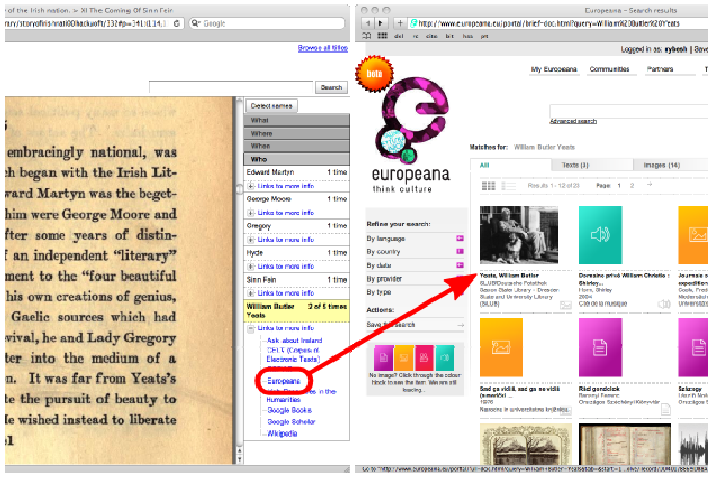


Figure 4 : Le « Context Finder » recheche William Butler Yeats dans Europeana.

Cet exemple simple est assez limité et nous travaillons des raffinements :

- Le logiciel ne discerne pas parfaitement les noms de personnes et les noms de lieu ;
- Une autre interface expérimentale permet au lecteur de corriger les identifications fausses ou ambigües du logiciel ;
- Nous n'avons pas encore mis en œuvre les topiques (« What ») ou le temps (« When ») ; et
- La liste de ressources à consulter est limitée.

Ce prototype d'interface est cependant prometteur quant à la possibilité de construire un "Context finder ».
L'image de la page est accompagnée par le texte en XML. Ce texte XML peut être rédigé et chaque nom propre est noté en XML de même pour les noms de lieu et tout mot ou phrase intéressant. De plus, à côté de chaque mot ou phrase on peut noter aussi en XML l'identification d'une ressource explicative convenable. Ainsi la page est de plus en plus préparée pour la lecture prochaine et le lecteur prochain. De telles pages annotées offrent au lecteur l'explication du contexte et des rapports de tout ce qu'il lit. Ce système d'annotation pour clarifier les mots et indiquer les ressources explicatives constitue notre « Context builder ».
Les liens enregistrés par le « Context builder » sont tous à sens unique : du texte à une ressource à consulter. Si ces liens étaient récoltés et inversés on pourrait ajouter ces liens à sens inverse : de la ressource à

consulter aux textes où ce nom ou mot est mentionné à la manière de « Hogan ». Notre « Context provider » est notre rêve de « Patrimoine 3.0 ».
Veuillez visiter http://metadata.berkeley.edu/demos pour voir ce prototype et d'autres exemples de prototypes d'interface.

7 La documentation du patrimoine irlandais

Les études irlandaises n'existent guère comme domaine scientifique. Il y a des études nombreuses touchant l'histoire de l'Irlande, l'archéologie, le géographie humaine, la langue gaélique irlandaise, la littérature (en anglais et en gaélique), les beaux arts, la musique, la politique, et beaucoup d'autres sujets, mais l'édition de ces études est fragmentaire et dispersée. Surtout, les périodiques des sociétés régionales des 18e, 19e et 20e siècles contiennent des articles toujours d'intérêt. La plupart de ces périodiques manquent d'indexation et la bibliographie de ces matières est inachevée. Elles sont rarement disponibles dans les fonds de bibliothèques, et on ne les trouve guère en dehors d'Irlande. Maintes ressources numérisées existent, mais les normes et les protocoles qui assurent l'accès convenable sont peu adoptés. De même ceux qui font des recherches sur des sujets relatifs à l'Irelande sont dispersés dans de multiples établissements universitaires. Ces conditions sont caractéristiques de la documentation des patrimoines.
En 2007-2009 le Centre for Data Digitisation and Analysis a numérisé quatre-vingts périodiques irlandais (environ 600,000 pages) et cette collection numérisée constitue la « Ireland Collection » de JSTOR[5] [21]. Nous examinons comment lier les détails de ces articles (et autres textes du patrimoine irlandais) aux ouvrages de référence. Quoique notre projet traite de l'Irlande, les techniques qui sont efficaces dans la documentation du patrimoine irlandais seront utiles pour gérer les autres ressources patrimoniales.

8 Conclusion

Comprendre quelque chose exige qu'on s'informe du contexte et des rapports. Alors comment faciliter la compréhension du contexte historique et culturel d'un patrimoine?
Dans l'ère des imprimés on rassemble les ouvrages de référence les plus digne de confiance, on navigue parmi eux et on adapte la recherche selon la terminologie particulière de chaque ressource. Si un lecteur explique à un bibliothécaire ce qu'il recherche, un bibliothécaire qualifié sait naviguer dans les ouvrages pour retrouver les faits ou documents

5 < http://www.jstor.org/templates/jsp/_jstor/templates/info/about/archives/aboutCollections/aboutIreland.pdf>

pertinents. Actuellement, un pareil environnement n'existe pas encore en ligne.
L'Université de Californie, Berkeley, a le projet de développer un tel service en trois phases. Une interface « Context finder » permet les recherches convenables dans les ressources recommandées. Un système « Context builder » note les liens à chaque ressource explicative dans le texte du lecteur en XML. La technique « Context provider » ajoute ces liens en sens inverse de la ressource vers les textes. Ainsi un réseau explicatif s'accumule.
Comme toujours, les tâches sont quelque peu déléguées au logiciel. L'interface assume le rôle du bibliothécaire et doit être sensible aux demandes du lecteur, mais aussi prendre des initiatives pour arriver au but. C'est un rôle selon Briet « . . . comme le chien du chasseur – tout à fait en avant, guidé, guidant." [22, 43]

Remerciements : Le project « Context and Relationships : Ireland and Irish Studies » est subventionné par l'Advancing Knowledge programme (Award PK-50027-07) du National Endowment for the Humanities et l'Institute of Museum and Library Studes, Washington, DC. Nos collègues Paul Ell, Fredric Gey, Matthew Holmberg, Ray Larson, Barry Pateman, et Jeanette Zerneke participent aussi dans ce projet et nous remercions Dara Hellman et Michel Menou de leur aide.

9 Références bibliographiques

[1] Support for the Learner: What, Where, When, and Who. http://ecai.org/imls2004/

[2] Bringing Lives to Light: Biography in Context. http://ecai.org/imls2006/

[3] Context and Relationships: Ireland and Irish Studies. http://ecai.org/neh2007/

[4] Anon. La terminologie de la documentation. Coopération Intellectuelle 77 (1937): 228-240.

[5] M. Buckland Le centenaire de "Madame Documentation": Suzanne Briet, 1894-1989. Documentaliste: Sciences de l'information 32, no. 3 (Mai/Juin 1995): 179-181.

[6] S. Fayet-Scribe. Women professionals in France during the 1930s. Libraries and the Cultural Record 44, no 2 (2009): 201-219.

[7] S. Briet. Qu'est-ce que la documentation? EDIT, Paris, 1951. http://martinetl.free.fr/suzannebriet/questcequeladocumentation/

[8] S. Briet. What is Documentation? Scarecrow Pr., Lanham, MD, 2006. http://ella.slis.indiana.edu/~roday/what%20is%20documentation.pdf

[9] M. Buckland. What is a « digital document » ? Document numérique 2 (juin 1998) : 221-230. http://people.ischool.berkeley.edu/~buckland/digdoc.html

[10] S. Briet. La nouvelle Salle des catalogues à la Bibliothèque nationale. Bulletin du bibliophile et du bibliothécaire, NS, 17. année (20 Oct., 1938): 437-442.

[11] M. Buckland. Library reference service in a digital environment, Library and Information Science Research 30, no 2 (2008): 81-85. http://people.ischool.berkeley.edu/~buckland/libref.pdf

[12] E. Hogan. Onomasticon goedelicum locorum et tribuum Hiberniae et Scotiae; an index, with identifications, to the Gaelic names of places and tribes. Hodges, Figgis, Dublin. 1910. http://publish.ucc.ie/cocoon/doi/locus

[13] M. Buckland, A. Chen, F. C. Gey, R. R. Larson, R. Mostern & V. Petras. Geographic Search: Catalogs, Gazetteers, and Maps. College & Research Libraries 68, no. 5 (Sept 2007): 376-387. http://www.ala.org/ala/mgrps/divs/acrl/publications/crljournal/2007/sep/Buckland07.pdf

[14] J. L. Zerneke, M. Buckland & K. Carl. Temporally Dynamic Maps: The Electronic Cultural Atlas Initiative Experience. Human IT 8.3 (2006): 83–94. http://www.hb.se/bhs/ith/3-8/jzmbkc.pdf

[15] M. Buckland. Naming in the Library: Marks, Meaning and Machines. In: Nominalization, Nomination and Naming in Texts. C. Todenhagen et W. Thiele, eds. Stauffenburg, Tübingen, 2007, pp. 249-260.

[16] M. Coyaud. Introduction a l'étude des langages documentaires. Klincksieck, Paris, 1966.

[17] M. Buckland. Forme, Signification, et Structure des Systèmes de Sélection du Savoir. Deuxiême colloque du chapître de l'International Society for Knowledge Organization, ISKO99, Lyon, France, Oct 21-22, 1999. http://people.ischool.berkeley.edu/~buckland/lyon-fr.html

[18] M. Buckland, A. Chen, F. C. Gey & R. R. Larson. Search Across Different Media: Numeric Data Sets and Text Files. Information Technology and Libraries 25, no 4 (Dec 2006): 181-189. http://www.ala.org/ala/mgrps/divs/lita/ital/252006/number4decembera/buckland.pdf

[19] M. Buckland, H. Jiang, Y. Kim et V. Petras. Domain-Based Indexes: Indexing for Communities of Users. In: 3e Congrès du Chapitre français de L'ISKO, 5-6 juillet 2001. Filtrage et résumé informatique

de l'Information sur les réseaux. Paris: Université Nanterre Paris X.. 181-185 http://metadata.sims.berkeley.edu/papers/ISKObuck.pdf

[20] M. Buckland et R. Shaw. 4W vocabulary mapping across diverse reference genres. In: Culture and Identity: Proceedings of the Tenth International ISKO Conference 5-8 August 2008 Montréal, Canada. Ed. by C. Arsenault and J. T. Tennis. Würzburg, Germany: Ergon Verlag, 151-156. http://people.ischool.berkeley.edu/~buckland/ISKO08.pdf

[21] Digital Library of Core E-Resources on Ireland http://www.jisc-collections.ac.uk/catalogue/ireland_eresources

[22] S. Briet. Bibliothécaires et documentalistes. Revue de la documentation 21 (1954): 41-45.

Le recours à des environnements numériques pour documenter le patrimoine bâti : une approche basée sur la complémentarité entre photogrammétrie et objet paramétrique

Nathalie CHARBONNEAU, Pierre GRUSSENMEYER

UMR MAP 694, Équipe PAGE, INSA de Strasbourg, France

Mots-clés : patrimoine bâti, photogrammétrie, librairie de composants, objets paramétriques, programmation 3D, ouvertures

Keywords: built heritage, photogrammetry, component library, parametrical object, 3D programming, openings

Résumé : Cet article expose une méthode qui a été développée afin de tirer profit de la complémentarité pouvant exister entre objet paramétriques et photogrammétrie, dans le cadre de projets de restitution architecturale. Nous explorons la possibilité d'élaborer la maquette numérique d'un artéfact par le biais d'un processus en trois étapes; i) d'abord sélectionner de façon interactive le type des composants, ii) ensuite, effectuer une mise à l'échelle automatique à partir d'un nuage de points de base densité et finalement iii) effectuer les retouches nécessaires pour adapter la topologie de l'objet virtuel au cas de figure observé. Pour ce faire, nous avons développé un environnement numérique faisant appel aux applications Maya[1] et Photomodeler[2]. Notre étude de cas porte sur les détails architecturaux composant les ouvertures dans les façades de bâtiments résidentiels. Nous retraçons la démarche de restitution à partir d'un exemple de fenêtre à guillotine dont la configuration est typique du patrimoine bâti montréalais.

Abstract: This paper describes the methodology of an ongoing research project concerned with the possibility of making use of parametrical objects and photogrammetric techniques within the framework of an architectural restitution

1 Maya: http://usa.autodesk.com/adsk/servlet/index?id=7635018&siteID=123112

2 Photomodeler : http://www.photomodeler.com

project. We explore the possibilities of elaborating the 3D model of an artefact by means of a three steps procedure: i) select the components' type in an iterative way, ii) put to scale the resulting model, automatically retrieving the relevant data from a points cloud and iii) retouching the model in order to adapt the object topology to the scenario under study. With this in view, we implement a digital environment making use of the following commercial software: Maya1 and Photomodeler2. Our case study deals with architectural details adorning openings in residential buildings. By way of example, we describe the restitution process of a sash window, typical of the Montreal built heritage.

1 Introduction

Le but d'un projet de restitution architecturale est habituellement de documenter un élément patrimonial et de présenter l'information tridimensionnelle de façon conviviale, afin de permettre l'extraction de données pertinentes par différentes catégories d'intervenants. Dans cette optique, le bâtiment (ou l'ensemble architectural) est modélisé afin de générer une maquette numérique. Dans le cadre du présent exposé, nous portons notre intérêt sur l'une des facettes du travail du modélisateur, soit la restitution des détails architecturaux.

Pour documenter les éléments appartenant au patrimoine bâti, les technologies numériques offrent une vaste panoplie de solutions. Dans nombre de projets de restitution, on a recours aux outils de dessin de logiciels CAD pour élaborer la maquette numérique de bâtiments. Il est cependant notable que, lorsqu'il s'agit de modéliser certains éléments, tels que les détails architecturaux, développer un modèle géométrique distinct correspondant à chacun des cas de figure existant peut constituer un véritable 'travail de moine', dépendamment de l'ampleur du projet.

Au cours des dernières années, plusieurs stratégies alternatives ont été développées par les chercheurs, pour pallier à ce problème. L'objet paramétrique est celle qui retient ici notre intérêt ; il s'agit d'une stratégie basée sur le concept de modules réutilisables. En effet, ce type d'approche permet de transformer la configuration d'un objet en modifiant de manière interactive les valeurs numériques associées aux paramètres. Dans le cadre de projets de restitution, les objets paramétriques, auxquels on a attribué les valeurs qui convenaient, sont ensuite insérés dans la maquette du bâtiment que l'on cherche à restituer. Une librairie d'objets interactifs peut contribuer à optimiser le travail du modélisateur, ce qui se traduit par un considérable gain de temps. Le lecteur trouvera un exemple de ce type d'approche sur le site GOP [3] qui présente les travaux menés par le CRAI [1].

Pour attribuer aux différents paramètres les valeurs qui conviennent, il faut nécessairement connaître les dimensions de l'objet réel. Or, arriver à

3 GOP: http://anabar.crai.archi.fr/~chevrier/Gop/gop.html

connaître les dimensions d'un élément architectural peut poser problème selon l'accessibilité de cet objet. Dans le cas des ouvertures, il arrive qu'on doive effectuer toute une série de mesures fines sur des détails situés en hauteur. Par ailleurs, il arrive qu'on doive mesurer à maintes reprises des éléments semblables (linteaux, appui, petits bois, etc.) sur des ouvertures de même type ou de types similaires. Ces problèmes peuvent ralentir, voire entraver, le déroulement des travaux de restitution numérique. C'est pourquoi nous cherchons à proposer une méthode visant d'une part à augmenter ce que l'on pourrait appeler 'l'accessibilité virtuelle' de l'objet et, d'autre part, visant à systématiser la façon de consigner les caractéristiques de l'objet et d'effectuer les relevés.

Un paramètre important dans la définition d'une méthodologie de relevé, et dans le choix d'une technologie, est le niveau de complexité de l'objet [2]. Notre étude de cas sur les ouvertures dans les bâtiments résidentiels nous amène à analyser des objets d'une géométrie relativement peu complexes, mais dont les différentes 'instances' sont déclinées avec de multiples variations. Le peu de complexité de ces objets découle du fait qu'un ensemble relativement restreint de primitives géométriques et de dimensions suffisent habituellement à en décrire la géométrie. Pour un type donné d'ouverture, ces dimensions forment un ensemble que l'on peut définir a priori (largeur et hauteur de l'ouverture, largeur et hauteur du linteau, etc.).

Dans ce contexte, il est permis de présumer que le recours à l'objet paramétrique pour consigner les caractéristiques de l'objet, ainsi que l'utilisation de nuages de points pour extraire les données numériques nécessaires, pourrait optimiser la démarche du modélisateur. Nos travaux reposent sur la possibilité de combiner les avantages de ces deux méthodes, soit la flexibilité de l'objet paramétrique et la précision, en termes de dimensionnement, propre à l'application de techniques de photogrammétrie. Dans cette optique, nous avons développé un environnement numérique faisant tour à tour appel à deux applications, Maya [1] et Photomodeler [2], la première pour le développement d'algorithmes rendant possible l'élaboration d'objets paramétriques, la seconde pour permettre le recours aux techniques de photogrammétrie[4].

Au cours de l'exposé qui suit, nous retraçons la démarche qui nous a permis de combiner le potentiel de deux approches distinctes mais complémentaires. À partir d'un exemple concret (la génération de la maquette numérique d'une fenêtre), nous expliquons de quelles façons les données sont saisies, traitées et exploitées. Cette démarche, qui est présentée au cours des prochaines sections, se déroule en trois phases, soit la sélection des composants, la mise à l'échelle automatique et les modifications finales.

4 Notons que pour être en mesure de travailler dans un tel environnement numérique, il est essentiel que l'utilisateur ait une connaissance pratique de ces deux logiciels.

1.1 Première étape : la sélection des composants

Nous travaillons depuis plusieurs années à formaliser, par le biais d'algorithmes, la logique de composition d'artéfacts appartenant au patrimoine bâti [3]. En nous référant à une typologie donnée, l'expertise développée nous permet maintenant de décrire des ensembles de cas de figure sous forme de modèles fédérateurs. Ceux-ci ont pour mandat de permettre la génération (semi-automatique) de maquettes numériques représentant un maximum de cas de figure divers, mais inter-reliés au niveau formel. Ce type d'approche offre la possibilité de développer des objets paramétriques avec lesquels l'utilisateur interagira par le biais d'interfaces graphiques. Le mandat de cette interface est de permettre à l'utilisateur de sélectionner les différents composants (et type d'organisation) de l'artéfact et de transmettre ces informations au système qui, lui, génèrera le modèle 3D.

Figure 1. Fenêtre à guillotine

Nous appliquons maintenant cette approche au processus de restitution architecturale. Dans le cadre de la démarche que nous retraçons, les ouvertures du bâtiment sont restituées en tant qu'entités autonomes. Dans cette logique, les maquettes partielles, représentant les diverses ouvertures, seront ultérieurement insérées dans la maquette globale du bâtiment. L'artéfact que nous utilisons pour exposer ce processus est une fenêtre à guillotine (voir figure 1). Il s'agit là d'un type d'ouverture

couramment employé dans les habitations montréalaises de type duplex ou triplex construites au début du XXème siècle.
Pour reconstituer la géométrie de cet artéfact, la première phase en est une d'observation. L'utilisateur doit sélectionner les différents composants et transmettre ces informations au système par le biais de l'interface graphique. Pour le développement de cette dernière, nous avons eu recours au langage MEL (Maya Embedded Language). Nous avons privilégié le paramètre de type non numérique et opté pour un mode de contrôle offrant la possibilité de choisir de façon itérative entre diverses options mutuellement exclusive. Les différentes alternatives proposées sont illustrées de façon schématique dans l'interface à proximité de *radio buttons.* Ceux-ci permettent à l'utilisateur de sélectionner de façon conviviale les différents composants de l'objet à restituer (voir figure 2).

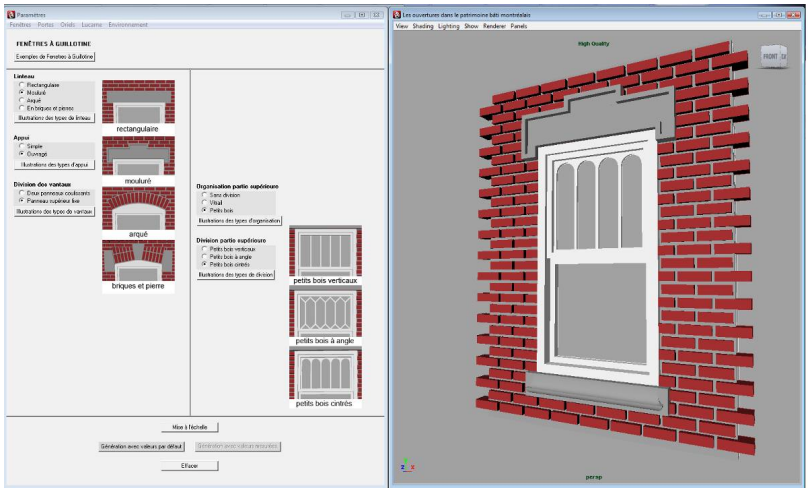


Figure 2. Interface basée sur des radio-buttons

Par le biais de ces ensembles de *radio-buttons*, l'utilisateur sélectionne les valeurs (non-numériques) qu'il convient d'assigner aux différents paramètres, à savoir le type de linteau, le type d'appui, le type d'organisation des petits bois, etc. L'utilisateur est en mesure d'amalgamer différents types de composants pour arriver à une configuration qui soit, le plus possible, en conformité avec la réalité observée. Au terme de cette étape, la maquette 3D de l'artéfact est générée par l'environnement Maya en utilisant les dimensions par défaut. L'utilisateur peut alors évaluer la maquette résultante, afin de vérifier la conformité de la configuration avec l'artéfact à l'étude (au niveau des types de composant). Cet examen visuel pourra révéler quelques

disparités entre le modèle et l'artéfact. Nous reviendrons sur cette éventualité dans la troisième partie de l'exposé.

1.2 Deuxième étape : la mise à l'échelle

Suite à la phase d'observation et de sélection des composants, on passe maintenant à l'étape de mise à l'échelle, qui mettra à contribution les techniques de photogrammétrie. À partir du moment où tous les paramètres ont été déterminés, l'utilisateur a accès à l'option 'Générer un nuage de points'. Cette dernière, lorsqu'activée, entraîne l'apparition d'une nouvelle fenêtre présentant de façon graphique la stratégie à adopter pour élaborer le nuage de points en question. La stratégie recommandée différera selon le cas de figure sur lequel travaille l'utilisateur. En d'autres termes, pour chacun des composants sélectionnés précédemment, un croquis démontre quels points devront être référenciés ainsi que la nomenclature appropriée.

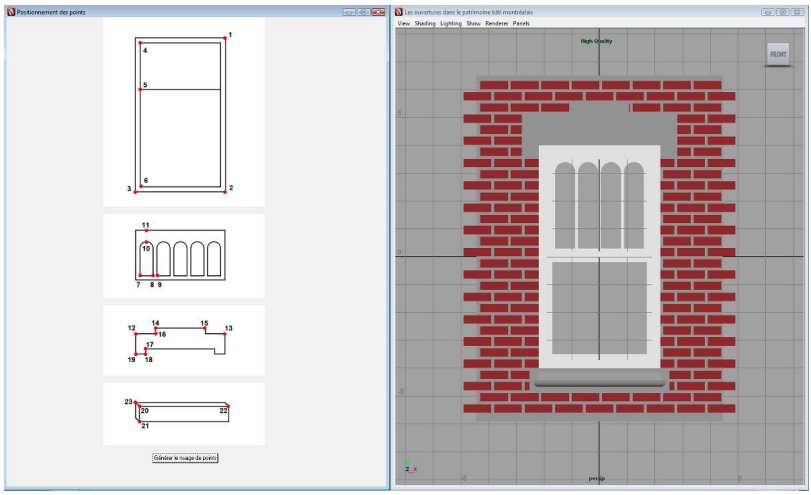


Figure 3. Illustrations démontrant la stratégie à adopter pour élaborer le nuage de points

L'utilisateur pourra se référer à ces croquis tout au long du processus d'élaboration du nuage de points. Au moment jugé opportun par l'utilisateur, l'application Photomodeler est activée à partir de l'interface principale. Comme nous le savons, ce logiciel permet de générer des nuages de points à partir de plusieurs photographies d'un même objet, prises sous des angles de vue qui diffèrent. L'utilisateur sera appelé à sélectionner les photographies de l'artéfact sur lesquelles il souhaite

travailler. Il s'agira pour lui d'identifier chacun des points apparaissant sur les croquis, et ce, sur les différentes photographies[5].

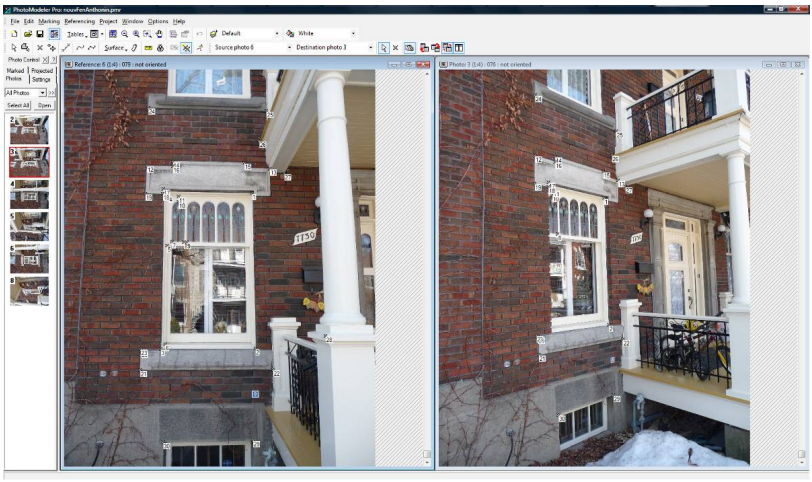


Figure 4. Le nuage de points référencés dans l'application Photomodeler

L'aspect sémantique est ici crucial. En effet, chacun des points doit être identifié et nommé conformément aux croquis et consignes fournies dans l'interface principale. Par exemple, dans le cas de la fenêtre à guillotine, les points #1 à #3 sont utilisés pour consigner la position des sommets délimitant l'ouverture; le point #1, le sommet supérieur droit du rectangle formant l'ouverture, le point #2 le sommet inférieur droit, le point #3 le sommet inférieur gauche et ainsi de suite… Le respect de la nomenclature est de toute première importance puisque, par la suite, les points #1 et #2 seront automatiquement utilisés pour calculer la hauteur de l'ouverture, les points #2 et #3, la largeur, etc.

Dans notre exemple, les points #4 à #6 sont utilisés pour consigner les dimensions des panneaux, les points #7 à #11, les dimensions des petits bois, les points #12 à #19, les dimensions du linteau et les points #20 à #23, les dimensions de l'appui. Notons que, quel que soit l'élément architectural à l'étude, il est toujours important de référencer également quelques points (ici #24 à #30) répartis sur le reste de la façade afin d'optimiser les calculs effectués par le système pour le positionnement des points dans un espace tridimensionnel. Il est par ailleurs possible d'effectuer des vérifications durant l'élaboration du nuage de points ; on pourra alors s'assurer que certains points sont colinéaires ou que les axes

5 Notons ici que les points à référencer devront apparaître sur plusieurs photographies. Cet aspect constitue l'une des limites de la méthode puisqu'il arrive que la végétation interdise l'accès visuel à certaines zones de la façade.

définis par quatre points sont perpendiculaires, conformément à la logique du modèle.
Le nuage de points ainsi constitué est un nuage de base densité puisqu'il compte tout au plus une trentaine de points. Lorsque l'utilisateur a terminé de référencer tous les points utiles, le nuage de points est exporté sous forme de fichier script en langage MEL, cette fonctionnalité étant disponible dans le sous-menu 'Export model' du logiciel Photomodeler. Chaque point identifié est alors considéré comme un *locator,* c'est-à-dire un repère dans l'espace tridimensionnel. Les *locators* se voient automatiquement attribuer les mêmes noms que ceux qui ont été donnés dans Photomodeler; le point #1 devient le "locator1" et ainsi de suite...
Sous Maya, chaque entité existant dans l'espace tridimensionnel d'une scène est liée à un nœud *transform* par la relation enfant/parent; il s'agit de graphes directionnels acycliques (ou *DAG nodes*). Ce nœud *transform* reçoit un nom et des attributs; il gère les transformations de l'entité, soit sa position, son orientation et sa taille. (Pour plus d'informations sur les *DAG nodes*, voir Gould [4]). Le script qui est exporté vers Maya établit les liens entre *transforms* et *locators*. Pour les points #1 et #2 de notre exemple, la relation est formalisée de la façon suivante :

```
createNode transform -n "locator1";
  setAttr ".t" -type "double3" -0.811823 0.399546 -4.26539;
createNode locator -n "locatorShape1" -p "locator1";
  setAttr -k off ".v";
createNode transform -n "locator2";
  setAttr ".t" -type "double3" 1.03848 0.410492 -3.95695;
createNode locator -n "locatorShape2" -p "locator2";
  setAttr -k off ".v";
```

Lorsque le fichier texte est importé dans l'environnement numérique principal, chacun de ces *transforms* est interrogé, de façon à récupérer la position en x, y et z des différents *locators*. Les diverses coordonnées sont ensuite transformées en vecteurs, afin de rendre ces données exploitables sous MEL.

```
float    $p1x    = `getAttr  locator1.translateX`;
float    $p1y    = `getAttr  locator1.translateY`;
float    $p1z    = `getAttr  locator1.translateZ`;
vector $v1 = <<$p1x, $p1y, $p1z >>; // le sommet supérieur
droit du rectangle formant l'ouverture
float    $p2x    = `getAttr  locator2.translateX`;
float    $p2y    = `getAttr  locator2.translateY`;
float    $p2z     = `getAttr  locator2.translateZ`;
vector $v2  = <<$p2x, $p2y, $p2z >>;// le sommet inférieur
droit du rectangle formant l'ouverture
```

Une fois les données transformées, il s'agit de soustraire un vecteur de l'autre afin d'obtenir la longueur du vecteur résultant. Cette valeur correspond à la distance entre les deux points traités. Pour chacun des

couples de points pertinents, ces opérations se font automatiquement, sans que l'utilisateur ne soit impliqué dans le processus.

```
vector $vResultant = $v1 - $v2;
float $hauteur = mag ( $vResultant ); //la hauteur de
l'ouverture
```

Les distances entre chacun des couples de points constituent un ensemble de proportions : par exemple la proportion entre la hauteur et la largeur du linteau, la proportion entre la largeur des petits bois et leur espacement, etc. Pour effectuer la mise à l'échelle à proprement parler, il faut nécessairement avoir mesuré au moins une distance. Dans l'exemple qui nous occupe, la base de l'ouverture a été mesurée (1.085 m). Cette valeur a été assignée à l'espacement existant entre les points #2 et #3, dans l'application Photomodeler, suite au positionnement des différents points.
Si, en raison de l'inaccessibilité de l'objet, il est impossible d'effectuer cette mesure, il est parfois envisageable de l'évaluer, notamment dans le cas d'une ouverture perçant un mur de briques. En effet, la largeur de l'ouverture varie habituellement entre 3 et 5 briques (3, 3½, 4, 4½, 5). La dimension des briques étant standard (20.5 x 6 cm) il est possible de déduire la valeur de la base de l'ouverture (l'épaisseur du mortier étant évaluée à 1cm). Cette dimension (mesurée ou déduite) étant connue, le système est en mesure de déduire toutes les autres.
Voici une liste de toutes les distances extraites à partir du nuage de points de notre exemple :

Pour l'ouverture :	• hauteur et largeur, • hauteur du panneau supérieur, • hauteur du panneau inférieur, • largeur, espacement et décalage vertical des petits bois,
Pour le linteau :	• hauteur et largeur de l'élément, • hauteur et largeur du redent supérieur, • hauteur et largeur du redent inférieur,
Pour l'appui	• hauteur, largeur et profondeur de l'élément.

Il est important d'établir, dès le début de l'élaboration du système, le niveau de détail recherché. Dans notre exemple, nous prenons le parti de nous restreindre à une quinzaine de dimensions extraites à partir du nuage de points. Les valeurs attribuées au détail fin, telles que dimension des moulures, largeur du châssis et du cadre (etc.) sont attribuées par défaut. Cependant, dans le cas d'une ouverture très ouvragée, le détail fin (tel que les moulures et bas-reliefs) pourra être pris en considération dans la mesure où il aura été possible de s'approcher de l'objet pour le

photographier. Ceci étant dit, il est évident que le niveau de détail requis est intimement lié à la finalité du modèle.
Suite aux calculs, un algorithme effectue des vérifications pour évaluer la cohérence des données extraites à partir du nuage de points; il pourra déceler certaines incongruités dans la logique constructive. Dans notre exemple, le système déduit que la hauteur de l'ouverture fait 1.762 m. Or, dans un cas comme celui d'une ouverture perçant un mur de brique, il faut prendre en considération la question du calepinage. Lorsque l'ouverture est générée avec les dimensions par défaut, la hauteur de l'ouverture est calculé par l'expression suivante (int * ($hauteurBrique + $hauteurMortier)) dans laquelle "int" est un entier représentant le nombre de rangs de briques. Lorsque la valeur déterminé par le système pour la hauteur de la fenêtre ne correspond pas à un nombre entier de rangs de briques, on peut en déduire qu'il y a imprécision, soit dans l'estimé de la dimension de la brique, soit dans le nuage de points. Dans notre exemple, la fenêtre ferait 25.18 briques de hauteur, ce qui est improbable. Le système arrondira cette valeur à l'entier le plus près (25) et, le cas échéant, ajustera la hauteur des panneaux en conséquence, tout en respectant les ratios. (Notons qu'ici la dimension standard de la brique ne sera jamais réajustée afin de préserver la cohérence au niveau de la façade dans son entièreté.)
Un autre type d'erreur peut surgir si la nomenclature n'a pas été respectée. Le système détecterait par exemple que l'espacement entre les petits bois est inférieur à la largeur de ceux-ci, ou encore que la largeur du redent est supérieur à la largeur totale du linteau. L'algorithme effectue donc un ensemble de vérifications. En cas d'anomalie, soit il procède automatiquement à des rectifications afin que l'output soit cohérent, soit il affiche un message indiquant que des erreurs ou des imprécisions peuvent remettre en cause la validité du modèle généré.
En somme, les valeurs non numériques assignées par l'utilisateur ont un impact sur le type de nuage de points qui sera requis. L'analyse de ce nuage de points, quant à elle, permet d'assigner automatiquement les valeurs numériques aux dimensions. Par la suite, ces données sont traitées par l'algorithme en vue de déceler des anomalies. C'est le flux de l'information à travers les procédures conditionnelles qui fait la richesse d'une telle approche, basée sur le recours conjoint à la photogrammétrie et à la programmation fonctionnelle. Les mécanismes de vérification inclus dans l'algorithme permettent de valider (ou invalider partiellement selon le cas) les valeurs numériques générées automatiquement à partir du nuage de points. Bref, il s'agit de coller le plus possible à la réalité physique dont rend compte le nuage de points, et ce, tout en respectant la logique des principes de composition de l'artéfact.
Au terme du traitement des données extraites du nuage de points, l'utilisateur a accès à l'option 'Mettre à l'échelle'. L'activation de cette

option provoque l'affichage, dans l'environnement Maya, de la maquette numérique résultante.

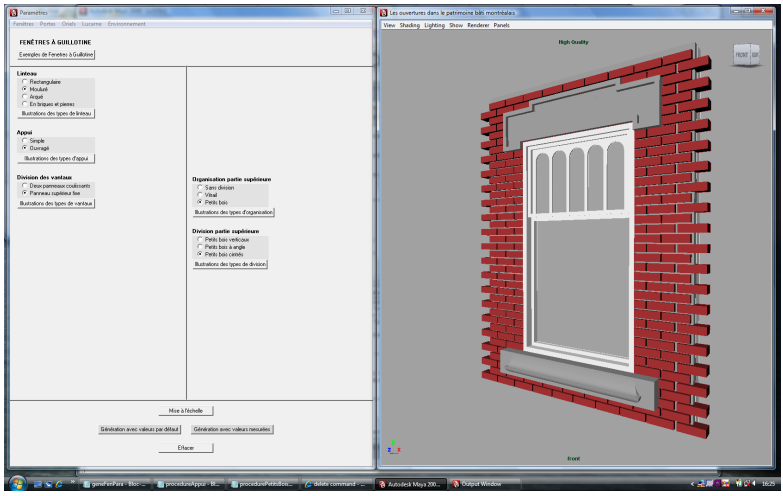


Figure 5. Saisie d'écran du modèle 3D à l'échelle

1.3 Troisième étape : les modifications et l'insertion

On peut supposer que, dans une majorité de cas, il sera nécessaire d'effectuer certains ajustements et de modifier légèrement la topologie de l'élément, afin de rendre le modèle plus conforme à la réalité. Lorsque les types de composants disponibles dans l'environnement numérique ne correspondent pas exactement ce dont l'utilisateur a besoin, deux pistes sont envisageables :

(i) À la fin du processus de restitution, l'utilisateur pourra procéder à des manipulations manuelles, dans l'environnement Maya, pour modifier la topologie de l'objet ; il y aura ajout, retrait ou modification de certains éléments et détails.

(ii) Si le projet de restitution se fait en parallèle avec la démarche de recherche, le prototype d'environnement numérique pourra évoluer à mesure que le programmeur remaniera cet environnement pour y inclure un plus grand nombre de types et de cas de figures.

Au terme du processus de restitution de l'artéfact, et après avoir fait les ajustements nécessaires, il sera possible d'insérer cet élément, ici l'ouverture, dans la maquette du bâtiment complet (et éventuellement d'insérer cette dernière dans la maquette d'un ensemble architectural).

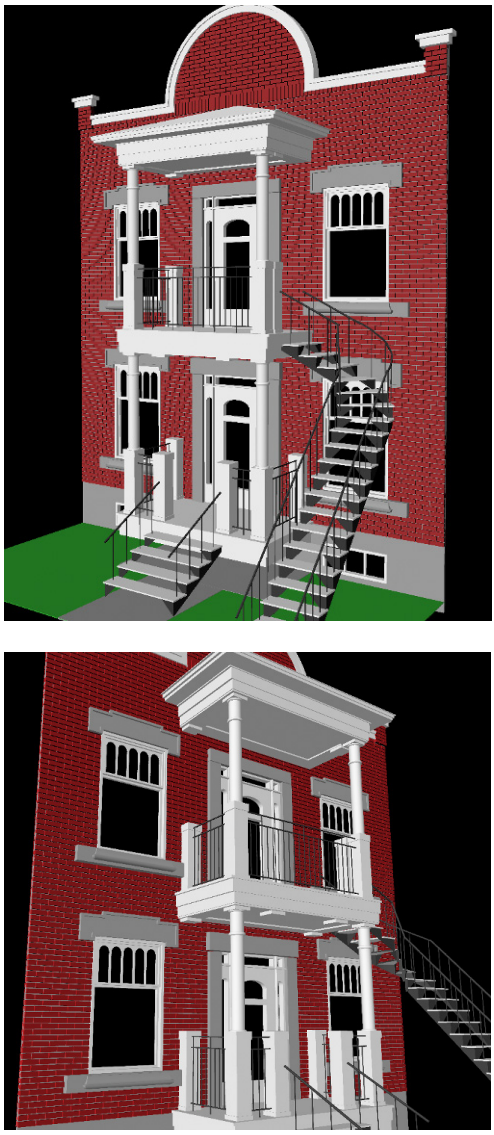


Figure 6. Maquette numérique de la façade d'une habitation de type duplex dans laquelle a été insérée l'ouverture

2 Conclusion

Dans le cadre d'un projet de restitution architecturale, une approche de ce type peut contribuer à optimiser la démarche du modélisateur dans la mesure où l'on arrive à identifier un leitmotiv dans la trame architecturale. Le cadre bâti à restituer doit être caractérisé par la

récurrence de certains éléments dont les différentes 'instances' sont déclinées avec de multiples variations. Le modélisateur devra établir une typologie de ces éléments architecturaux en fonction des travaux à effectuer et de l'ensemble architectural à l'étude. De plus, la librairie élaborée à partir de cette typologie devra contenir un bon nombre d'objets paramétriques et chacun de ceux-ci couvrir une gamme relativement vaste de cas de figure. Cette diversité est essentielle. En effet, lorsque le modélisateur aura recours à ce type de système, plus il se verra contraint de retoucher la géométrie de l'objet, plus la pratique nous éloignera du principe du module réutilisable à la base de ce type d'approche.
Notons en outre que, pour optimiser la performance d'un tel système, il semblerait préférable de travailler dans un premier temps uniquement sur des éléments architecturaux clairement définis (tels qu'ouvertures, corniches, pilastres, etc.) et non sur une façade ou un bâtiment dans son entièreté. D'une part, il est plus difficile d'élaborer une typologie de façade en raison de la multiplicité des cas de figure. D'autre part, étant donné que l'aspect sémantique est crucial, il est important que le nuage de points demeure de base densité. Si la densité venait à augmenter au-delà d'une certaine limite (une centaine de points par exemple), le nuage pourrait devenir difficile à gérer par l'utilisateur.
Il est clair que l'élaboration d'un tel système – typologie, objets paramétriques, interface – demande un investissement de temps non négligeable. Cependant, s'il s'agit de restituer un ensemble de bâtiments dans lequel on décèle une homogénéité relative de la forme architecturale, ce type d'approche pourra porter fruit. Il faut cependant demeurer conscient du fait que le gain de temps réel ne pourra être évalué qu'en comparant la méthode 'traditionnelle' (impliquant le recours aux fonctionnalités des logiciels CAD) et la méthode proposée. Ce gain pourrait croître proportionnellement au développement du système; plus ce dernier est étoffé (en termes de types d'ouvertures et de types de composants), plus il y aurait théoriquement optimisation du processus de restitution numérique. Seule la mise à l'essai d'un tel système, sur divers projets de restitution portant sur un cadre bâti relativement homogène, pourra nous renseigner sur l'efficacité de l'approche proposée.

Remerciements : *Nous remercions le Fonds Québécois de Recherche sur la Société et la Culture (FQRSC), qui a subventionné ce projet de recherche dans le cadre du programme de Bourse de Recherche Postdoctorale.*

3 Références bibliographiques

Chevrier, C. et Perrin, J.P. (2009). Generation of architectural parametric components, Proceedings of CAAD future 2009, Montréal, June 17-19, pp. 1005-118.

Grussenmeyer, P. (2008). Photogrammétrie et lasergrammétrie pour la documentation du patrimoine culturel. In Patrimoine et Enjeux, p. 15-34. Paris : Europia Productions.

Charbonneau, N., Boulerice D. et Booth, D. (2006). Understanding Gothic Rose Windows with Computer-Aided Technologies. Actes de la conférence eCAADe24, Volos, p.770-777.

Gould, A. D. (2002). Complete maya programming. San Francisco: Morgan Kaufmann.

Patrimoine photographique et dispositifs collaboratifs en ligne

Nathalie Casemajor Loustau

Université du Québec à Montréal - Université Lille 3

Mots-clés : photographie, Internet, indexation sociale, user generated content, Web 2.0, Bibliothèque et Archives Canada, Flickr

Keywords: photography, Internet, social indexing, user generated content, Web 2.0, Library and Archives Canada, Flickr

Résumé : Cette contribution porte sur la diffusion en ligne des collections photographiques de Bibliothèque et Archives Canada, et plus particulièrement sur les projets invitant les internautes à collaborer à la description des images. Quelles sont les spécificités du médium photographique du point de vue de la description collaborative sur Internet ? Quelles sont les modalités de collaboration proposées par BAC sur son site internet ? Quelles figures de l'usager et quels résultats ces dispositifs font-ils apparaître ? L'analyse montre que si des opportunités existent pour mettre les connaissances des utilisateurs au service de la documentarisation des fonds photographiques, le taux de participation et la qualité des contributions dépendent surtout des modalités de gestion du projet et de l'animation de l'activité des internautes.

Abstract: This paper deals with the Library and Archives Canada photographic online collection; focusing more precisely on its collaborative projects which enable users to participate in describing the pictures. What are the specificities of photographic collections regarding social indexing? How did LAC set the conditions for users to collaborate on its website? What user profiles and results have emerged from these projects? The analysis shows that opportunities for integrating users' knowledge into the collections exist, but the participation rate and its quality clearly depend on project management and facilitation frames.

1 Introduction

De plus en plus de bibliothèques, de musées et de centres d'archives utilisent aujourd'hui le réseau Internet pour expérimenter des projets invitant les utilisateurs à partager, à s'approprier et à commenter des collections photographiques historiques. Sous la forme d'albums, d'expositions virtuelles ou de bases de données en ligne, ces projets surfent sur la vague du Web 2.0 avec l'espoir de faire circuler plus largement ce patrimoine visuel et d'utiliser le savoir des internautes pour redocumentariser [1] les fonds photographiques.

Cette contribution est issue d'un travail de thèse en sciences de l'information et de la communication portant sur les médiations du patrimoine photographique sur Internet. Nous avons traité ce problème selon deux axes d'analyse : (1) les modalités de construction d'une représentation des fonds (procédures de sélection des objets à mettre en ligne, opérations de numérisation, édition et construction de parcours d'interprétation sur le site internet) et (2) l'émergence de nouvelles formes de relation entre institutions patrimoniales, publics et fonds photographiques en ligne. C'est sur ce deuxième point que va porter l'article.

Dans le cadre de notre recherche doctorale, nous avons étudié deux institutions conservant d'importantes collections photographiques (à la fois en terme de volume et de valeur historique des fonds) : la Médiathèque de l'architecture et du patrimoine (France) et Bibliothèque et Archives Canada (BAC). Pour chacune d'entre elles, nous avons procédé à une analyse de contenu du site internet et à une série d'entretiens sur place auprès de conservateurs, d'archivistes et de responsables des projets multimédias.

Nous limiterons ici le propos au cas de BAC en nous intéressant aux dispositifs collaboratifs de son site internet[1]. Quelles sont les spécificités du médium photographique du point de vue de la description collaborative sur Internet ? Quelles sont les modalités de collaboration proposées par BAC sur son site ? Quelles figures de l'usager et quels résultats ces dispositifs font-ils apparaître ?

2 La contribution des internautes à la documentarisation des fonds photographiques

Les opérations de documentarisation des fonds photographiques peuvent être particulièrement longues et délicates (difficultés à identifier le support photographique, multiplicité des objets, lieux et personnages

1 Bibliothèque et Archives Canada. <http://www.collectionscanada.gc.ca/index-f.html>.

représentés, nombre d'artefacts se chiffrant en milliers voire en millions dans certaines collections). Les utilisateurs peuvent-ils contribuer à mieux connaître ces fonds ?

2.1 La collection photographique de Bibliothèque et Archives Canada

Encadrée par le Ministère du Patrimoine Canadien, Bibliothèque et Archives Canada est une institution née en 2004 de la fusion de la Bibliothèque Nationale du Canada et des Archives Nationales du Canada. La section « Art et photographie » de BAC conserve autour de 25 millions de photographies. Sur ces 25 millions, seulement 500 000 environ ont été décrites individuellement (pour les autres, il n'existe qu'une description générale au niveau du fonds). Le catalogage est en effet une opération très longue et coûteuse. Selon E. Klijn et Y. de Lusenet, « cataloguer individuellement les éléments d'une collection de 536 000 photographies nécessiterait environ 30 000 jours de travail, ce qui représente à peu près 136 années de catalogage continu ! » [2].
La collection photographique de BAC rassemble aussi bien des archives issues du gouvernement fédéral (ministère de la Défense nationale, ministère des Affaires indiennes et du Nord canadien, Office national du film du Canada, etc.) que des albums compilés par des amateurs de photographie (officiers, gens d'affaires, hommes et femmes ordinaires), des donations ou encore des acquisitions (les fonds du studio Topley et du célèbre portraitiste canadien Yousuf Karsh en sont deux exemples parmi les plus connus).

2.2 Les difficultés de catalogage des fonds photographiques

Devant la diversité des sujets représentés en image et le volume massif des fonds conservés à BAC (la photographie étant, ne l'oublions pas, un « art du multiple »), l'inventaire et le catalogage restent lacunaires. Or, si de manière générale les notices descriptives sont essentielles pour la gestion de l'accès aux fonds documentaires, la photographie entretient un rapport particulier au texte du point de vue de son interprétation. Selon les mots G. Freud [3], les légendes qui commentent la photographie « peuvent en changer la signification du tout au tout ».
Les fonds de collectionneurs privés présentent davantage de difficultés à cataloguer que ceux des établissements à but commercial, dans la mesure où ces derniers dressaient un inventaire de leurs fonds à des fins de gestion. Dans ce cas, un ensemble d'informations déjà disponibles sur les images est transféré à BAC en même temps que la collection. Mais il en va autrement dans le cas des fonds de collectionneurs privés. En effet, avec la disparition du collectionneur s'évanouissent souvent de

nombreuses informations permettant d'identifier les images et leur contenu.
Par exemple, l'historien des chemins de fer Andrew Merrilees a constitué une importante collection représentant la majorité des chemins de fer nord-américains. Faute d'une connaissance spécialisée des archivistes dans ce domaine très pointu, des milliers de photographies sont simplement décrites par le mot « train », sans pouvoir distinguer le type de locomotive représenté ou la compagnie qui l'exploitait.

2.3 Tirer parti des connaissances des utilisateurs

Selon un archiviste de BAC, il arrive très fréquemment que des usagers en sachent plus que lui sur certains fonds. Quelqu'un lui a par exemple fait remarquer au sujet d'un lot de photographies de guerre que les descriptions fournies par l'auteur lui-même à l'époque avaient été modifiées pour ne pas divulguer des informations stratégiques à l'ennemi. Selon lui, il est difficile d'affirmer qu'une description fournie par un archiviste soit plus valable qu'une description fournie par un usager spécialiste du domaine. L'expertise étant par définition une compétence ciblée, c'est la multiplicité des contributions des utilisateurs et leur complémentarité qui fait la force de ce type de collaboration.
La description et la mise en ligne d'une partie de la collection photographique de *BAC* sur son site internet a été l'occasion de susciter la collaboration des usagers en mettant en place des applications de commentaire des images. Certains utilisateurs se manifestent d'eux-mêmes, comme cet internaute passionné par les chemins de fer qui a envoyé au directeur de la section « Art et photographie » un courriel avec une liste de 3 000 corrections concernant les notices descriptives des photographies en ligne.
Par ailleurs, la diffusion et le partage de photographies sur Internet a donné lieu à de nombreuses expériences innovantes sur des plates-formes intégrant des outils issus de la mouvance Web 2.0 : possibilité d'étiqueter (tagger) les images, de sélectionner des portions de photographies pour y poster des commentaires, etc. Plusieurs raisons peuvent être convoquées pour expliquer l'attractivité des fonds photographiques en termes de projets collaboratifs en ligne : la dimension visuelle de la photographie et la « lisibilité » dans laquelle elle se donne à voir, sa forte présence dans la culture médiatique contemporaine, la grande popularité de la pratique photographique et son rôle traditionnel en tant que support du souvenir personnel, familial et collectif.

3 Dispositifs collaboratifs et figures du public

Le site internet de BAC présente deux formats d'accès à ses fonds photographiques : des bases de données thématiques (« Photographies », « Photographies : les infirmières canadiennes », « L'Office national du film du Canada ») et une dizaine d'expositions virtuelles (citons par exemple « Visages de guerre », « Hommages à Karsh : Maître du portrait », « William James Topley : Réflexions sur un photographe de la Capitale »).

3.1 Quelles modalités de participation ?

Pour évaluer les modalités de participation offertes aux utilisateurs sur ce site, nous avons défini six indicateurs à partir de l'observation d'un ensemble d'autres sites patrimoniaux[2] : (1) la possibilité de publier en ligne des commentaires sur les photographies, (2) d'envoyer un commentaire sur un produit en ligne, (3) de bloguer sur le site de l'institution, (4) de créer des albums personnalisés visibles sur le site par d'autres utilisateurs, (5) de participer à l'indexation sociale des photographies (étiquetage social, folksonomy) et enfin (6) la collaboration avec le site Flickr et l'utilisation des outils collaboratifs de ce site.

L'observation a fait apparaître que le site de BAC propose trois de ces modalités, soit la moitié des indicateurs que nous avons identifiés. Par ailleurs, un autre projet est en cours autour de l'exposition virtuelle « William James Topley ». Il s'agit de permettre aux utilisateurs d'ajouter des informations aux notices descriptives des photographies en ligne. Mais il semble que le développement de ce projet ait été placé en attente suite à un changement à la direction de l'établissement.

Quel sont les dispositifs collaboratifs actifs sur le site de BAC ?

- Premièrement, la plupart des expositions virtuelles de BAC proposent une rubrique « Commentaire », invitant les utilisateurs à envoyer un commentaire à propos de l'exposition. Cette rubrique contient un formulaire électronique intitulé « Dites-nous ce que vous pensez du site X ». Dans le cas de l'exposition « Vision photographique du Canada », la coordinatrice des projets multimédias nous a signalé qu'elle tend à recevoir de nombreux commentaires lorsqu'un produit vient d'être lancé, puis ce flux se tarit progressivement.

- Deuxièmement, dans le cadre d'une exposition virtuelle intitulée « Le trèfle et la feuille d'érable », BAC offre à la consultation un

2 Médiathèque de l'architecture et du patrimoine, Musée Nicéphore Nièpce, Paris en images, Musée McCord. Bibliothèque nationale de France.

album d'images sur le site de partage de photographies en ligne Flickr. Ce site propose de nombreuses modalités d'interaction pour les usagers : poster des commentaires à propos du profil de l'institution, d'un album ou d'une photographie en particulier, ajouter une note directement sur l'image, ajouter des étiquettes descriptives à propos d'une photographie. Nous évoquerons en détail ce projet dans la dernière partie de l'article.

- Troisièmement, les expositions virtuelles « Visages de guerre » et Un visage, un nom » offrent la possibilité de publier des commentaires sur les photographies. Examinons plus en détail cette dernière modalité de collaboration.

3.2 Deux cas d'expositions virtuelles

L'exposition virtuelle « Visages de guerre » présente des photographies d'hommes et de femmes ayant servi dans les forces armées canadiennes durant la Deuxième Guerre Mondiale. Ce projet comprend un volet invitant les internautes ayant participé à ces événements historiques à partager leurs souvenirs : « nous espérons que ces images vous rappelleront des souvenirs et nous vous invitons à nous en faire part » peut-on lire sur le site.
Les usagers sont ainsi sollicités pour ajouter des commentaires dans la base de données de photographies liée à l'exposition virtuelle. Lors d'un entretien mené à BAC, la coordinatrice des projets de numérisation a justifié la pertinence de ce projet en mentionnant un attrait du public ciblé pour le partage de leurs souvenirs : « les militaires aiment bien pouvoir regarder les photos et dire ça c'est telle chose, identifier les personnes, le contexte, etc. ». Toutefois, un sondage de la base de données[3] laisse penser que la participation des utilisateurs est faible : sur 210 images visionnées dans la base, seules 3 ont été commentées.
L'autre projet appelant la participation des internautes s'intitule « Un visage, un nom ». Il est destiné à trouver le nom des Inuits représentés dans certaines collections photographiques de BAC. Ce projet vise spécifiquement la participation des utilisateurs appartenant à la communauté inuite (en particulier du Nunavut). Ils sont invités à collaborer en identifiant les personnes représentées sur les photographies :

> Reconnaissez-vous quelqu'un parmi ces gens ? Si vous avez des renseignements concernant une photo, veuillez remplir le formulaire en ligne. Nous vous remercions de votre aide et de l'intérêt que vous avez manifesté à l'égard de ce projet.

3 Consultation des trois premières photographies affichées pour chaque photographe recensé dans la base, soit 210 images sur 2500.

Cette initiative affiche un objectif d'intégration de la mémoire de communautés culturelles minoritaires dans une représentation collective de l'identité nationale. Dans ce cas de figure, l'enjeu est de remettre en contexte des fonds photographiques produits dans un contexte de domination politique et culturelle et de revaloriser le mode de représentation des autochtones dans les fonds photographiques d'État.
Ces photographies ont été prises par des fonctionnaires fédéraux ou des photographes professionnels embauchés par le gouvernement. Un grand nombre des portraits ne mentionne pas l'identité de la personne photographiée, ou présentent des erreurs dans l'orthographe des noms. N. Aglukkaq, une aînée ayant participé au projet, témoigne dans le texte de présentation de l'exposition virtuelle : « j'ai cherché si longtemps une photographie de mon père et je me suis aperçue que le nom indiqué sur la légende de la photo n'était pas le sien. Son véritable nom n'était pas mentionné correctement ».
Toutefois, l'exposition virtuelle n'offre pas de dispositif de commentaire des images en ligne, et c'est par courriel que les internautes transmettent les informations dont ils disposent. En termes de résultats, le site indique que plus d'une quarantaine de photographies ont été identifiées grâce à Internet mais la personne responsable du projet à BAC nous a signalé que la grande majorité des identifications n'a pas été faite par l'intermédiaire de la diffusion sur Internet, mais en se rendant auprès des communautés ou via un partenariat avec un journal local (le journal *Nunavut News North*, qui publiait chaque semaine des portraits issus de la collection de *BAC*).

3.3 Figures de l'usager acteur

Les dispositifs de médiation du patrimoine comme ceux proposés par BAC sur son site internet façonnent un programme d'usage des contenus en ligne [4] et modèlent en creux des figures de l'usager internaute.

3.3.1 L'usager témoin

La première figure qui semble émerger de « Visages de guerre » et de « Un visage, un nom » est celle de l'*usager témoin*. Ces deux expositions invitent en effet l'internaute à partager sa mémoire et son vécu personnel à propos des photographies publiées en ligne.
Dans le cas du projet « Visages de guerre », il s'agit d'amener les anciens militaires à témoigner d'un événement historique majeur auquel ils ont pris part. Par exemple, le sous-lieutenant L. Brooks, un artiste engagé dans la marine et représenté dans une photographie de 1945 a lui-même commenté cette image : « After the war, the war artists moved to Ottawa to finish their paintings of the war. Leonard Brooks - November 19th, 2008 P.S. I was certainly dashing back then! ».

Dans l'autre exposition virtuelle, il s'agit de puiser dans la mémoire individuelle et familiale des utilisateurs pour identifier des membres de leur communauté. Ce projet vise spécifiquement à inscrire une mémoire vivante (portée par les individus) dans l'archive pour la conserver et la transmettre aux générations futures, lorsque les derniers témoins auront disparu.
Dans ces deux expositions, la mémoire individuelle des usagers est sollicitée afin de nourrir les cadres d'une mémoire collective [5] conservée dans l'archive photographique et mise en circulation sur le Réseau. Cette démarche rejoint la réflexion de Maurice Halbwachs sur la mémoire collective:

> C'est qu'en général l'histoire ne commence qu'au point où finit la tradition, au moment où s'éteint ou se décompose la mémoire sociale. […] alors le seul moyen de sauver de tels souvenirs, c'est de les fixer par écrit en une narration suivie puisque, tandis que les paroles et les pensées meurent, les écrits restent [6].

Ainsi au témoignage de l'objet patrimonial (comme vestige du passé) s'arrime le témoignage de la mémoire vécue, inscrite sous forme de texte accompagnant et donnant sens à l'image.

3.3.2 L'usager expert

La deuxième figure qui est ressortie de l'observation est celle de l'*usager expert*, c'est-à-dire un internaute qui met son expertise au service de la connaissance des photographies publiées en ligne. C'est particulièrement le cas dans « Visages de guerre » où par exemple les régiments d'appartenance des soldats peuvent être difficiles à identifier pour les archivistes. Une photographie présentant la légende « an unidentified Canadian Scottish regiment » a ainsi été commentée par un internaute qui a ajouté l'information suivante : « These are Calgary Highlanders; they are easily identified by the red triangle overtop of the blue rectangle on the sleeve - the 1942 era insignia of The Calgary Highlanders ».
Plus largement, les connaissances apportées par les usagers peuvent toucher une grande variété de domaines : des événements historiques (comme dans le cas de « Visages de guerre »), l'identification de catégories d'objets (tel un type de locomotive dans la collection Merrilees), de personnages (recherche des noms de personnes photographiées dans « Un visage, un nom »), de lieux représentés (rue, ville, etc.) ou tout autre aspect se rapportant au contexte de production de la photographie (son auteur ou sa date par exemple).

4 Le cas de BAC sur Flickr

Nous allons à présent évoquer un troisième projet mis en place par BAC pour permettre aux utilisateurs de commenter les photographies et

d'apporter leur expertise au sujet de la description des images. Il s'agit de la publication de certaines photographies de sa collection sur Flickr.

4.1 Flickr et « Les Organismes publics »

Flickr est un site web de partage de photographies (et de vidéos) en ligne qui hébergeait en mars 2009 plus de trois milliards de photographies. Grâce à sa simplicité d'utilisation et à ses multiples fonctionnalités de partage et de commentaire des images, *Flickr* est souvent considéré comme « l'un des sites exemplaires du Web social » [7]. Il semble remarquablement illustrer les principes généraux du Web 2.0 : une activité reposant sur le contenu produit par les utilisateurs (*user generated content*), une pratique d'étiquetage des images, mais surtout la constitution d'une « communauté virtuelle » d'utilisateurs [8] instaurant les ferments d'un réseau social.

C'est la Bibliothèque du Congrès (États-Unis) qui a initié la collaboration avec Flickr dans le cadre d'un projet plus large intitulé « Les Organismes publics » (The Commons). Le 16 janvier 2008, elle y a publié plus de 3 000 photographies anciennes issues d'un fonds consacré à l'histoire des États-Unis avec l'objectif d'améliorer l'accès aux collections, mais aussi d'enrichir la qualité des métadonnées associées aux photographies en permettant « au grand public d'apporter des informations et des connaissances » [4]. Aujourd'hui, plus de vingt-cinq établissements ont rejoint cette initiative.

4.2 « Le trèfle et la feuille d'érable »

Le projet de *BAC* sur cette plate-forme est bien plus modeste, et ne s'inscrit pas dans le cadre formalisé des « Organismes publics ». Il s'agit simplement de l'ouverture d'un compte au nom de l'institution, comme le font de plus en plus de bibliothèques, de centre d'archives et de musées. Un des pré-requis de *BAC* pour mettre en ligne des photographies sur un site externe était de pouvoir garantir une plate-forme bilingue français-anglais, ce que permet *Flickr*.

Les objectifs du projet sont de faciliter l'accès aux collections (« explorer de nouvelles façons d'améliorer l'accès »), de développer les modalités de communication avec les usagers (« parrainer le dialogue et augmenter l'interaction ») et enfin de solliciter la contribution des utilisateurs (« explorer comment les usagers interagissent avec des collections numérisées dans des environnements qui favorisent la contribution de commentaires et de renvois »). Par ailleurs, sur son site internet, BAC se déclare « enthousiasmé[e] par les occasions que les communautés sociales de partage de contenus multimédias offrent aux Canadiens de

4 Extrait du texte de présentation des « Organismes publics » sur le site Flickr.
<http://www.flickr.com/commons?phpsessid=ea7b4da468f5935f24b65f41dbfc356f>.

discuter et de contextualiser une importante sélection de notre histoire collective ».
BAC a mis en ligne un échantillon de 84 images en lien avec l'exposition virtuelle « Le trèfle et la feuille d'érable » hébergée sur son propre site. Issu principalement des collections J. Topley, J. Ballantyne et G. Heriot, cet ensemble de documents (des photographies d'Ottawa, de Montréal, de Toronto, de Québec et d'autres endroits au Canada et en Irlande, mais aussi des affiches, des cartes, des imprimés, des dessins, des peintures et des gravures datant des années 1860 aux années 1920) a été constitué à l'occasion du Symposium d'études irlandaises 2008, afin d'illustrer la présence du patrimoine documentaire canado-irlandais dans les collections de BAC.
Ces images ne sont pas protégées par copyright (le site de BAC indique que : « tout le monde peut voir les images, les partager avec d'autres et les redistribuer gratuitement, à la condition que l'utilisateur de l'image mentionne la source ») et à chacune est associé un ensemble d'informations (en anglais et en français) comprenant le titre de l'image, son auteur, sa date de création, son numéro de référence à *BAC*, le lieu représenté et la mention de la source.
Quels sont les modes d'interaction proposés aux utilisateurs sur *Flickr* ? Ils peuvent tout d'abord ajouter une image à leurs favoris sur leur compte personnel ou envoyer la photographie à un ami par courriel. Ils peuvent également « bloguer » la photo (c'est-à-dire l'insérer dans un blogue qu'ils auront paramétré à l'avance sur leur compte personnel) ou l'ajouter dans un groupe dont ils sont membre (par exemple, un groupe dédié au photojournalisme de rue). Ils peuvent ajouter un commentaire directement sur l'image, en dessous de l'image, et aussi au niveau plus général de l'album d'image ou du profil de l'institution. Enfin, les internautes peuvent attribuer des étiquettes descriptives (*tags*) permettant de classer les images en fonction de thématiques, ajouter des contacts à leur compte (comme d'autres sites de réseautage social tels que *Facebook* et *Myspace*) et créer et participer à des groupes d'utilisateurs.

4.3 Des résultats modestes

Examinons à présent les résultats de la participation des usagers au 1er avril 2009, soit cinq mois après le lancement du projet. Au moment où nous avons effectué l'observation du site, trois fonctionnalités n'avaient pas été utilisées par les internautes : l'ajout d'images dans un groupe, l'ajout de témoignages sur le profil de *BAC* et l'ajout de commentaires sur l'album. Pour les fonctionnalités utilisées, le taux de participation s'avère faible :

- 8 images ajoutées en favoris ;
- 1 note ajoutée sur une image ;

- 3 images taggées par 3 utilisateurs différents pour un total de 20 étiquettes ;
- 5 images commentées par les utilisateurs, pour un total de 9 commentaires.

Par ailleurs, l'activité des utilisateurs (favoris, tags, notes et commentaires confondus) s'est concentrée sur 13 images et 16 utilisateurs au total ont participé à ces différentes activités. Parmi ces utilisateurs, nous avons pu déterminer que 8 sont canadiens, 2 irlandais, un britannique et un australien.
Intéresserons-nous de plus près à la teneur de ces commentaires ajoutés en dessous des images. On peut en distinguer plusieurs types :

- des commentaires d'appréciation des images (« This is my adopted hometown. Lovely old photo! », « It would be amazing to travel back in time for a day just to visit ») ;
- des commentaires d'appréciation du projet dans son ensemble, qui sont tous positifs et encourageants (« Well done », « love it! », « progressive », « great », « excellent! », « great site ») ;
- des requêtes : demande d'information (« Who is the gentleman in the bottom corner? ») et demande d'amélioration du service (« let's have higher resolutions please! ») ;
- une correction d'information : suite à la remarque d'un utilisateur (« Is this photograph from near Grant`s Causeway, or the Giant`s causeway ») le titre d'une des photographies a été corrigé (une erreur dans la légende originale avait transformé « Giant » en « Grant »).

Si les commentaires d'appréciation du projet sont très positifs, en définitive peu de nouvelles informations sur les photographies ont été récoltées (un seul commentaire a apporté un correctif). En reprenant les objectifs initiaux du projet, on peut dresser le bilan suivant.

- Un premier objectif était de solliciter la contribution des utilisateurs, or le taux de participation des internautes est faible.
- Un second objectif consistait à « parrainer le dialogue et [à] augmenter l'interaction ». Deux utilisateurs ont posé des questions à propos d'une des images de l'album, auxquelles *BAC* a répondu sur le ton informel de la conversation. Par ailleurs, plusieurs fils de discussion autour d'un document présentent une interaction entre l'institution et les utilisateurs d'une part, et entre les utilisateurs eux-mêmes d'autre part. Malgré le peu de commentaires postés par les utilisateurs, le dispositif en place semble tout de même offrir des conditions favorables à une communication dialogale.

– Un troisième objectif était d'« explorer de nouvelles façons d'améliorer l'accès ». Pour évaluer cet objectif, on peut se demander par exemple si la mise en ligne de documents sur *Flickr* a permis de toucher d'autres internautes que ceux qui consultent le site officiel de l'institution. Nous avons contacté l'ensemble des utilisateurs ayant commenté une image de *BAC* ou ajouté l'une d'entre elles dans leurs favoris (sauf un utilisateur qui avait masqué son identifiant). Sur les 15 personnes contactées, 10 nous ont répondu. Il est apparu que 6 n'avaient jamais navigué sur le site internet de *BAC* avant de consulter son profil sur *Flickr*. Quatre d'entre elles expliquent avoir découvert l'album de *BAC* par hasard en naviguant sur *Flickr*. Par exemple, l'une d'elles écrit : « I was doing a search of local material from my hometown of Cobh, Ireland when the search brought up a link to LAC collection ». Si l'échantillon et le nombre de réponses obtenus sont restreints, il semble néanmoins que la mise en ligne de photographies sur *Flickr* permette de les faire circuler parmi d'autres publics que le public habituel de l'institution patrimoniale.

En fin de compte, il ressort que le projet présente un bilan mitigé : une faible participation et peu de nouvelles informations collectées sur les photographies, mais des opportunités de dialogue, d'interaction et d'ouverture à de nouveaux publics qui pourraient être développées.

4.4 Opportunités et modalités de gestion du projet

On peut chercher à comprendre les raisons de ces résultats modestes. Le public n'est-il pas intéressé par ce type d'initiative ? Il semble pourtant que les commentaires postés expriment un réel intérêt de la part des internautes. Un utilisateur que nous avons contacté explique par exemple que commenter des images de collections publiques sur Flick est pour lui une sorte de hobby : « Every now and then I go into the 'commons' to tag images ». De plus, la manifestation spontanée de certains internautes pour corriger des notices sur le site de BAC fait apparaître un potentiel intéressant en termes d'apport d'information sur les photographies. Pour prendre un peu de recul vis-à-vis de ces résultats, mettons-les en perspective avec deux autres projets de même nature : celui de la Bibliothèque du Congrès et le projet PhotosNormandie, qui vise à améliorer l'indexation d'un fonds de photographies historiques sur la Bataille de Normandie (2 763 photographies ont été mises en ligne en janvier 2007).

Deux jours après son lancement, le compte de la Bibliothèque du Congrès totalisait déjà plus d'un million de pages vues, 500 photos commentées, et 1 200 photographies ajoutées en favoris. Environ trois mois plus tard, le profil de la bibliothèque affichait 11 000 « contacts » d'utilisateurs,

plus de 3 500 commentaires avaient été postés sur les images par plus de 1 400 utilisateurs[5], et la bibliothèque avait mis à jour une centaine de descriptions de photographies grâce aux commentaires des utilisateurs. Environ 55 000 étiquettes avaient été ajoutées (10 000 étiquettes différentes, avec pour certaines images l'atteinte du seuil limite de 75 étiquettes). Selon une interview de deux responsables du projet[6], l'équipe de la bibliothèque a été impressionnée par la teneur des échanges entre les utilisateurs, au point qu'un lien vers les discussions les plus pertinentes sur Flickr a été créé dans certaines notices documentaires du site de la bibliothèque.

Dans le cas de PhotosNormandie, en septembre 2008, 3 806 descriptions avaient été complétées, corrigées et mises à jour[7]. Un groupe nommé « Discussions sur PhotosNormandie » a été créé par les responsables du projet et sept mois après son lancement, il comptait 39 membres dont une dizaine de participants réguliers. Les résultats très positifs de ce projet mené avec des moyens bien moindres que ceux de la *Bibliothèque du Congrès* témoignent de l'importance de bâtir une communauté d'utilisateurs autour du compte Flickr. Un premier moyen pour favoriser la participation au projet semble être de développer le nombre de « contacts » affiliés au profil de l'institution, de manière à tisser un réseau d'utilisateurs et de les tenir informés de l'actualité du projet (signaler les ajouts de nouvelles images ou de nouveaux outils de participation) et de les inviter à s'impliquer chacune de ces occasions.

Mais selon des responsables du projet de la Bibliothèque du Congrès, le versement en une seule fois d'un trop grand nombre de nouvelles images sur Flickr peut décourager les contributeurs. Ils ont plutôt opté pour un versement hebdomadaire de 50 nouvelles images[8]. Par ailleurs, les résultats du projet PhotosNormandie semblent montrer que la création d'outils facilitant le travail d'élaboration collective (groupe de discussion dédié au projet de l'établissement public, participation à des groupes de discussion tiers) permet d'offrir un autre espace d'échange que celui du profil de l'institution, d'élargir les perspectives du projet et de favoriser la dissémination sociale des images au sein du réseau.

Il existe toutefois des difficultés et des limites dans ce type d'initiatives, notamment du point de vue de la participation des usagers à l'enrichissement des métadonnées. Selon des responsables du projet de la *Bibliothèque du Congrès*, le travail d'exploitation et d'évaluation des commentaires laissés par les utilisateurs est long et fastidieux, et la

5 E. Bermes, « LC+Flickr : Bilan d'une expérience 2.0 », 2008, sur le blog Figoblog. <http://www.figoblog.org/node/1921>. Consulté le 2 avril 2009.

6 « LC and Flickr - 3 months later », posté le 27 mars 2008 sur le blog HangingTogether (consacré à l'actualité des archives, bibliothèques et musées). <http://hangingtogether.org/?p=401>. Consulté le 2 avril 2009.

7 Le nombre plus grand que celui des photos s'explique par le fait que certaines légendes ont été corrigées plusieurs fois.

8 Ibid.

bibliothèque a affecté pas moins de 12 personnes au développement de cette initiative. Des membres de l'équipe expliquent que :

> Dans cette perspective, l'étiquetage social ne consiste pas à laisser les autres faire le travail de catalogage. Il s'agit plutôt de susciter des conversations qu'il faut ensuite suivre de près pour en extraire les informations les plus pertinentes[9].

Finalement, si ce type de projet exploite les outils collaboratifs du Web 2.0, il les intègre dans une démarche institutionnelle « documentaire et rédactionnelle », apparentée à « un travail collectif avec un objectif de production, et non à une folksonomie caractéristique du web social »[].

5 Conclusion

Les différentes observations que nous avons effectuées montrent que la figure du public des fonds photographiques patrimoniaux tend à se déplacer, s'éloignant de la représentation d'un récepteur passif d'un savoir « savant » pour intégrer la représentation d'un collaborateur potentiel à la production des savoirs sur ce patrimoine. En miroir, la figure de l'institution patrimoniale est elle-même amenée à évoluer : reconnaissant les limites de ses compétences d'indexation et la compétence des usagers, BAC cherche à mettre en place un rapport aux publics moins vertical et moins hiérarchisé.

Dans le cas des initiatives de BAC que nous avons présentées, il s'agit toutefois de projets pilotes dont l'ampleur et les résultats restent limités. Il semble que le développement du Web 2.0 et son application au secteur patrimonial aient suscité un fort engouement de la part des institutions patrimoniales, tout comme une attente de la part des publics. Avec son projet sur Flickr, BAC a cherché autant à tester ces nouvelles potentialités qu'à se positionner du point de vue de sa stratégie de communication institutionnelle comme un acteur innovant à la pointe des tendances du Web. Les résultats modestes de cette initiative tendent à montrer que les moyens humains à investir pour que le dispositif fonctionne et soit efficace en termes de redocumentarisation des fonds se révèlent importants et ont pu être sous-estimés.

Il convient de replacer cette nouvelle conception de la relation entre BAC et ses usagers dans une dynamique plus large d'émergence de nouvelles figures du citoyen et de redéfinition des relations entre les institutions culturelles et leurs usagers. Le type de relation aux publics construit sur les sites webs patrimoniaux s'inscrit au croisement d'un paradigme muséal et d'un paradigme des technologies multimédia [10] : il est d'une part influencé par l'évolution de la représentation du public des institutions culturelles (dans un mouvement vers la « démocratie

9 Traduction de l'auteur. Ibid.

culturelle », l'habilitation des usagers à construire un lien personnalisé et créatif avec les biens culturels et patrimoniaux) et d'autre part, il est influencé par les représentations du public des médias interactifs, autrement dit une conception des « interactants » [11] comme participants actifs à la production de contenus en ligne.
Malgré tout, ces déplacements ne constituent pas des transformations subites et radicales, mais plutôt des tendances qui s'incarnent dans des projets de médiation collaborative tels que ceux décrits plus hauts, et qui ne se substituent pas à des formes de médiation plus traditionnelles, même sur Internet.

6 Références bibliographiques

R. T. Pédauque (collectif). La redocumentarisation du monde, Cépaduès Éditions, Toulouse. 2007.

E. Klijn, et Y. de Lusenet. SEPIADES. Cataloguing photographic collections, European Commission on Preservation and Access, Amsterdam. 2004, p. 9.

G. Freund. Photographie et société, Éditions du Seuil, Paris. 1974.

C. Tardy, J. Davallon et Y. Jeanneret. Les médias informatisés comme organisation des pratiques de savoir, Organisation des connaissances et société des savoirs : concepts, usages, acteurs, Toulouse. 2007, pp. 169-184.

M. Halbwachs. Les Cadres sociaux de la mémoire, Albin Michel, Paris. 1994 [1925].

M. Halbwachs. *La mémoire collective*, Albin Michel, Paris. 1997 [1950], p. 130.

P. Peccatte. Une plate-forme sociale pour la redocumentarisation d'un fonds iconographique, Actes de la deuxième conférence Document numérique et Société, Éditions ADBS, Paris, 2008.

H. Rheingold. Les communautés virtuelles, Addison-Wesley, Paris, 1995.

P. Peccatte. *Ibid*, p. 6-7.

G. Vidal. Internautes citoyens et consommateurs, *2001 bogues : globalisme et pluralisme. Usages des TIC*, Presses de l'Université de Laval, Sainte-Foy. 2003, pp. 325-339.

T. Bardini et S. Proulx. Entre publics et usagers : la construction sociale d'un nouveau sujet communiquant, *Médiations sociales, systèmes d'information et réseaux de communication,* SFSIC, Metz. 1998, pp. 267-274.

Cartes anciennes, SIG géo-historiques et visites virtuelles :

Déformations lisibles de la perception spatiale sur un corpus de cartes numérisées issues de fonds patrimoniaux et de représentations cartographiques patrimoniales hypermédia.

Jean-Philippe d'Erceville

Université de Lyon (France) - Université Claude Bernard Lyon1 - Laboratoire ELICO EA 4147

Mots-clés : carte, cartographie, patrimoine, représentation spatiale, espace-temps, modèle, déformation, perception, SIG, modèle numérique de terrain

Keywords: ancient map, cartography, heritage, spatial representation, time space, model, deformation, perception, GIS, digital model of ground

Résumé : Cet article tend à définir les usages des technologies hypermédia, appliquées à un type particulier de collections: les fonds anciens de cartes et plans. Il s'agit pour nous de déterminer en quoi la lisibilité d'une collection de cartes, issues d'un fonds ancien, peut subir des déformations de la perception et de la représentation spatiale. L'objectif est de mener une réflexion sur les enjeux conceptuels d'une articulation entre cartes issues d'un fonds patrimonial et reconstruction numérique d'un patrimoine, déterminé par son ancrage dans l'espace et le temps, soit une évolution permettant de passer des cartes anciennes, en tant que documents primaires intégrés à des collections, aux strates d'informations spatialisées au cœur d'un système d'information géohistorique, pour enfin s'interroger sur les principes d'un modèle de visite virtuelle en 3D, dans lequel s'implique l'usager.

Abstract : This article will tend to define the manners of hypermedia technologies, applied to a particular type of collections: the ancient collection of maps and plans. It is a question for us to determine in what the legibility of a collection of maps, stemming from a ancient fund, can undergo deformations of perception and spatial representation. The objective is to lead a reflection on the conceptual stakes in a joint between maps, stemming from a patrimonial fund, and digital reconstruction of an heritage, determined by its anchoring in the space and time, to wit an evolution allowing to make the link between ancient maps, as primary documents integrated into collections, and spatial information's strata in the heart of a geohistoric GIS, to wonder finally about the principles of a 3D virtual tour, in which involves the user.

1 Introduction

En abordant la question de l'évolution des technologies Web, sous l'angle de la médiation d'un patrimoine, il nous est apparu nécessaire de nous interroger sur l'objet de patrimonialisation [1] que nous allons traiter : la carte. Si les cartes sont des représentations de l'espace dans le temps, elles sont avant tout un outil et un support de médiation, une interprétation humaine du réel, construction ou re-construction intellectuelle autant que médium, porteur d'une intention, témoignant d'une évolution des représentations, mais aussi des théories et discours sous-jacents [2]. Se pose alors la question de la construction lisible et intelligible de l'information patrimoniale spatialisée, élaborée sur le mode graphique, par la numérisation des fonds patrimoniaux, en termes de perception, de production et de réception.
En quoi la numérisation et donc l'acte de réédition des cartes, procèdent-ils d'une articulation entre le traitement d'information patrimoniale spatialisée et la communication de cette information dans une situation d'énonciation donnée ? En quoi pouvons-nous considérer ce processus comme une réécriture ?

Notre article se construit en deux partie :

- La carte, en tant que construit, comme objet de patrimonialisation et de réédition numérique critique,
- La carte numérique en trois dimensions, en tant que représentation ou re-construction d'un patrimoine selon le temps et le lieu dans lesquels elle s'inscrit.

Distinguer ces deux espaces de réflexion, c'est mener une analyse sur deux processus bien distincts dans le temps, correspondant à deux révolutions de l'écriture cartographique : la première a eu lieu à la fin du XVIème siècle [3], la deuxième, plus récente, dans les années 1990 avec l'émergence du numérique. Nous verrons que ces deux révolutions fonctionnent à l'inverse l'une de l'autre dans la modification de la perception de l'espace, documentaire et réel, qu'elles produisent. C'est néanmoins la dernière révolution qui nous intéresse. En effet, la carte ancienne numérisée devient représentation d'un type de document, intégré au patrimoine écrit, autant que représentation d'un contenu patrimonial spatialisé. Ces deux types de représentation ne peuvent être analysés selon la même méthode.

Notre réflexion s'appuie sur des études de cas précis dont voici une brève présentation, support de notre analyse exploratoire et conceptuelle :

- Le catalogue numérique commenté des cartes de la ville de Lyon (France) [4] : *Forma Urbis*©[1]
- Le projet de SIG (Sytème d'Information Géographique) géo-historique de la ville de Lyon©

en tant que processus de patrimonalisation des collections. Il s'agit ici de deux projets conçus selon deux intentions différentes et utilisant deux types de techniques différentes, que nous étudierons dans la première partie de cet article .

Le premier projet, aboutit en 1999, soit il y a dix ans, à l'édition d'un ouvrage de type traditionnel avec éditorial, préface, sommaire et index. Il est ensuite mis en ligne et présente, sous une forme paginée, les plans généraux de la ville de Lyon entre le XVIème et le XIXème siècle. Il s'agit là d'un catalogue d'exposition conçu par un service d'archives et non une bibliothèque mais la constitution d'un catalogue critique issu d'une institution dont l'enjeu est la conservation et la diffusion de fonds patrimoniaux porte un questionnement proche de celui qui occupe les conservateurs de fonds anciens en bibliothèque tout autant que les conservateurs des musées. L'intention est clairement de développer la science de l'iconographie dans la cité et la connaissance même de l'histoire scientifique, artistique, politique et sociale de la cité, par l'étude de ses représentations. Son fonctionnement demeure celui de la linéarité du codex et sauf à fournir des images en grands formats par hyperliens et popups, son hypertextualité reste limitée.

Le projet de *SIG géohistorique de la ville de Lyon*© est actuellement en développement par l' ISIG (Imagerie et Système d'Information Géographique), sous l'égide de l'UMR 5600, Environnement-ville-société (Université Lyon 2 - Lyon 3, ENS LSH Lyon). Il s'agit d'un projet de création d'un SIG issu d'une approche géo-historique de l'espace urbain lyonnais et donc du patrimoine architectural urbain, à partir de cartes du XVIIème au XIXème siècle, mises en concordance avec des cartes contemporaines. Il porte l'ambition d' "une compréhension renouvelée des processus à l'œuvre dans [l'évolution de l'espace construit lyonnais] : acteurs, financements, logiques sociales, foncières, économiques, diachroniques."2
Ce projet, utilisant les ressources du numérique, comporte les caractéristiques des systèmes d'information géographique classiques. La

1 Réf. : Archives municipales, 1999, Forma Urbis: les plans généraux de Lyon XVIème XIXème siècle, (coll. dossiers des archives municipales, n°10), ISBN 2-908949-18-0 [Disponible en ligne] URL: http://www.archives- lyon.fr/old/fonds/plan-g/plan2.htm, (consulté le 15/12/2008)

2 Bernard Gauthiez, Un système d'information géographique sur l'espace urbain à Lyon aux XVIIe-XVIIIe siècles, UMR 5600 Environnement ville société, url: http://umr5600.univ-lyon3.fr/siglyon.html

possibilité de superposer différentes couches d'information et donc de construire une information en temps réel en fonction de la requête de l'utilisateur. Encore faut-il que ce dernier soit familiarisé avec cet outil. L'intention, clairement énoncée par ses producteurs est de fournir un outil de lecture des bases de données géographiques constituées préalablement avec des urbanistes, des historiens, des géographes ainsi que des professionnels de la modélisation spatiale et de la géovisualisation en trois dimensions. Le groupe de travail n'écarte nullement ce type de représentation.

Nous analyserons, dans la deuxième partie de cet article, les applications ancrées dans les pratiques de médiation culturelle devenues relativement courantes: les visites virtuelles, en tant que reconstruction virtuelle d'un espace intégré au patrimoine.

– La visite virtuelle *Marseillenet©*[3] et le projet *Rome Reborn©*[4]

Ces deux projets semblent totalement différents dans le mode de représentation et dans l'interface qu'ils proposent. La visite virtuelle *Marseillenet©* est issue d'une application commerciale destinée à être utilisée de la même manière, quelque soit le lieu que l'on visite. L'interface est stable et identique selon les lieux. L'application est destinée à promouvoir un espace touristique tout autant qu'un objet ou un patrimoine immobilier par le biais de la photographie panoramique. On peut ainsi l'utiliser pour faire une visite d'appartement à louer. Nous verrons que le mode de perception qu'il engendre, peut, malgé des différences très nettes avec le deuxième projet, lui être comparé dans les limites qu'il impose à l'usager.

Le projet *Rome Reborn©*[5], quant à lui, tend à « reproduire » la Rome antique (de 1000 av. JC à 550 ap. JC) sous la forme d'un modèle numérique de terrain en trois dimensions. Le projet, commencé en 1997, a beaucoup évolué, en plus de 10 ans, et en est actuellement, selon les informations disponibles, à sa version 2.0. La démarche mise en oeuvre n'est pas du tout la même que celle du projet marseillais. Il s'agit en effet d'une réelle démarche de production de connaissances historiques, dans un cadre universitaire, pour la construction d'une interface destinée à la consultation par le grand public. Il semble que cette interface sera, dans l'avenir, intégrée aux applications de GoogleEarth® . Nous disposons de

3 Visite virtuelle http://www.marseillenet.com/ de Marseille.

4 Ce projet, commencé en 1997, est développé par l'UCLA et l' IATH, Cultural Virtual Reality Laboratory (CVRLab), l'expérience UCLA Technology Center (ETC), le Reverse Engineering (Indaco) Lab au Politecnico di Milano, l'Institut Ausonius du CNRS, l'Université de Bordeaux-3, et l'Université de Caen, en partenariat avec GoogleEarth® et IBM® . La création d'un modèle numérique en trois dimensions de la ville de Rome illustre son développement urbain au cours de l'histoire sous la forme d'une visite guidée virtuelle.

5 Détails du projet sur : http://www.romereborn.virginia.edu

peu d'informations sur la compositon du modèle, hors de ce qui en a été dit lors de la conférence Siggraph 2008 de Los Angeles. En revanche, de nombreuses vidéos de démonstrations permettent de mieux cerner l'outil et ses capacités ainsi que l'interface.
Pour fixer le cadre méthodologique, il nous a semblé pertinent de nous intéresser aux pratiques cartographiques dans un futur proche et donc de produire une analyse exploratoire de type constructiviste et conceptuelle sur des objets ou applications numériques en cours de développement.

2 Collection de cartes numérisées et patrimonialisation du document cartographique:

Analyser des cartes issues d'un fonds patrimonial, c'est faire le choix de construire ou de reconstruire une collection de "traces" historiques d'une perception du monde et donc un discours qui se développe en tant qu'il a pour unique objet la lisibilité d'un espace défini dans le temps. La collection ne peut s'envisager que comme un projet, dont les modes de conception dépendent en partie de la matérialité des objets et de leur capacité à devenir des supports d'analyse critique. La numérisation des fonds patrimoniaux, dans le cadre d'une gestion électronique de documents n'en est plus à ses balbutiements et pourtant, les fonds de cartes anciennes restent encore relativement peu traités en tant que traces ou plus précisement en tant que signes d'un projet de représentation du monde. C'est par la numérisation et un traitement qui dépasse le cadre d'une simple numérisation de conservation qu'il est possible de reconstruire ce discours.
Les applications hypermédia, en tant que modes de reconstruction peuvent permettre, dans l'avenir, de concevoir le développement d'une réelle analyse critique, de corpus topocentrés hors du cadre de la culture locale, développée autour des "hauts-lieux" d'un espace défini socialement, de sorte à restituer bien plus que les images d'Epinal liées au développement touristique d'une région. La culture collective locale peut alors être dépassée, pour faire du patrimoine un réel objet de recherche, un construit. En cela, la possibilité de construire des strates d'informations cartographiques, à partir de fonds patrimoniaux, peut permettre de faire d'un document d'archives, un projet de recherche utilisant les ressorts du numérique. La superposition d'ensembles cohérents d'informations géohistoriques permet la perception des représentations construites sous la forme de palimpsestes, mais aussi une réécriture du territoire par la mise en relation de documents dont le processus même d'écriture, à son origine, est le palimpseste. Il est fondamental de pouvoir retrouver la trame de construction de ce type de document, en effectuant la même navigation intellectuelle que celle qui a

engendré cette représentation du monde. Recontruire un discours cohérent avec la méthodologie appliquée durant la phase d'écriture de la carte du XVIème ou du XVIIème siècle, c'est mener un processus réflexif sur la production et le mode de publication des cartes imprimées. C'est retrouver les plaques de cuivre gravées et les superposer en couches d'informations et de culture, c'est également re-produire le geste de l'imprimeur et concevoir la carte, non sur le mode d'un résultat analysable en tant qu'objet fini, mais en tant que processus d'écriture.
La stratification des informations, par la superposition de couches cartographiques diachroniques sur le modèle du calque, permet cette reconstruction mais ne pourra suffir à concevoir un espace de réflexion sur la carte, disponible et intégrable par des usagers non rompus aux systèmes d'informations géographiques.
La compréhension d'un objet dans la reconstruction d'un espace-temps autre que celui de la réception peut être facilitée par l'apport de couches d'informations portées *a posteriori* de la production des cartes. C'est toute l'importance d'une structure de métadonnées, implantées lors de la reconstruction du discours, qui est alors mise en jeu. La compréhension des enjeux d'une surimpression d'informations critiques, de repères balisant le produit semble le moyen le plus intuitif de reconstruire non seulement un objet "intelligent", mais un réel projet de développement des connaissances, tout autant que de transmission de ces connaissances. C'est l'acte même de médiation du savoir, sur un patrimoine encore peu connu du grand public, en deçà des fantasmagories et fétichismes que peut engendrer la réception d'une carte ancienne, comme celle de tout ouvrage ayant traversé le temps. La construction des modèles de métadonnées devient alors un second projet qui vient se greffer sur celui de la constitution de la collection. L'enjeu d'une hypermédiatisation de ces nouveaux construits se complète d'un travail prospectif à réaliser sur le modèle même de l'interface, de l'écran et de sa lecture. Il est, semble-t-il, nécessaire de ne pas concevoir une application uniquement sur le mode de la diffusion mais bien de la reconstruction intelligible d'une collection.
La diffusion seule ne peut qu'engendrer une incompréhension de l'objet même et de sa finalité, de l'intention qu'il porte. En effet, si l'on peut concevoir la visée politique d'une diffusion plus large des collections patrimoniales, c'est avant tout par une compréhension des projets qu'elle porte que se fait une réelle médiation du patrimoine. En dehors de toute analyse marxiste telle que W. Benjamin l'a développée dans son article "L'oeuvre d'art à l'ère de sa reproductibilité mécanisée", il ne s'agit pas de déposséder une élite de sa mainmise sur l'authenticité de réels chefs d'oeuvres graphiques, mais bien de construire un objet complexe tout autre. C'est ce que nous étudions dans notre réflexion sur la numérisation et la patrimonialisation des objets de sciences tels que les cartes.

3 Question de perception de l'espace : déformation et représentation d'un patrimoine urbain par la cartographie dynamique en trois dimensions.

3.1 Analyse conceptuelle

Nous commençons cette analyse par une approche permettant de déterminer ce qui, dans le discours que porte la carte, nous "parle" de la perception de l'espace orienté vers la découverte d'un patrimoine archictural et urbain. Plusieurs questions se posent alors:
-Quelles sont les déformations lisibles de la perception de l'espace engendrées par « l'intrusion mouvante » du signifiant incarné dans le patrimoine cartographique ? Quel mode de perception transparait par cette implication du regardant ou plus précisément de son "double virtuel" dans la lisibilité du document ?
-Quelle différence de perceptions de l'espace s'opère-t-il entre une représentation d'un espace ou d'un territoire dans le temps de cet espace et sa re-construction *a posteriori?*
Analyser les cartes numériques c'est, avant tout, comprendre le mode de fonctionnement de quatre grandes familles de représentations : la carte ancienne numérisée, le SIG, le système GPS et le modèle numérique de terrain parfois appelé visite virtuelle, dans le cadre d'applications destinées à construire des connaissances à destination du grand public.

3.1.1 Quatre modes de représentation spatiale:

Il s'agit ici de montrer que la carte numérisée, et lisible sur écran, ne fonctionne pas sur le même mode de perception de l'espace historique, mais également du temps historique. Pour cela, nous nous appuyons sur l'idée d'une désynchronisation de la navigation dans l'espace et le temps reconstruits et sur la multiplicité des modes de représentation: topocentrique (aréolaire et géolocalisé), anthropocentrique (perception virtuellement localisée), odologique (itinérant) et géométrique (planaire).
Le premier mode représentation de l'espace, représentant le patrimoine spatialisé, est tout autant objet de patrimonialisation que représentation du patrimoine et outil de reconstruction. Les cartes anciennes issues des fonds patrimoniaux que nous avons évoqueés dans la première partie de cet article, sont avant tout le résultat d'un transfert de support permettant un enrichissement *a posteriori*. L'avantage majeur de cette médiation de document est de constituer une base d'informations contemporaines du patrimoine sur lequel elles porte. L'inconvénient majeur, quant à lui, est que ces documents ne portent pas l'intention d'une pérennité des connaissances portant sur le patrimoine architectural et urbain. La notion de patrimoine n'étant, à l'époque de l'écriture de ces cartes, pas

totalement inexistante mais encore très peu définie. C'est par le traitement et l'enrichissement, issu de l'archéologie et du travail des historiens, que se construit la connaissance des éléments présentés sur ces cartes, souvent erronés quant à la position strictement topographique des lieux. L'intéret majeur consiste en la représentation des lieux par les contemporains de la construction de ces espaces devenus patrimoine. Ils en indiquent l'importance sociale dans le temps de l'écriture des cartes qui les représentent. La numérisation de ces cartes, outre l'enrichissement, a pour vocation la conservation et une difusion plus large des ces objets. Les institutions qui conservent les objets originaux diffusent ainsi une vision de l'espace dans un temps donné. Celui qui visionne cet espace au travers de son écran pense le percevoir tel que le cartographe le percevait. On se rapproche, là encore, des fanstasmagories liées au concept d'authenticité alors même que cette dernière a disparu, suite au processus de numérisation. Parler de désynchronisation de la navigation dans l'espace est ici inapproprié puisque c'est le processus même de consultation de la carte qui importe. L'objectif n'est pas, pour le grand public de "percevoir" un lieu mais de "voir" un document (ce qui ne signifie pas le lire).

Tout comme dans le cadre de la réédition des cartes anciennes, par la reproduction et l'ajout d'un appareil critique sous la forme de livre en version électronique, le SIG géohistorique intègre la notion de strates d'informations et se comporte en fonction des interrogations de base de données (comparables à des index ?) à la manière du palimpseste. Les cartes anciennes se voient attribuer un maillage de données. Ce qui intéresse le cartographe n'est plus alors de multiplier les "couches" d'information mais d'en comprendre les entrelacements, les imbrications de sorte à pouvoir déterminer les unités de mesure de la ville qui ne seront pas les mêmes selon l'utilisateur du système: l'urbaniste ou l'archéologue ne peuvent segmenter le territoire de la même manière et cependant le travail sur les cartes anciennes leur fournit une base de connaissances empilables. En considèrant les bases de données, à l'origine de la création graphique, superposables aux différentes cartes anciennes, il devient nécessaire de s'interroger sur la définition de ce qu'est un appareil critique. Pouvons-nous considérer les bases de données géoréférencées comme un appareil critique? La réponse, qui pourra paraître péremptoire, semble être négative. L'appareil critique nécessite l'intervention d'une interprétaion humaine. Mais, si nous nous reportons à la définition même de ce qu'est une carte, telle que nous l'avons donnée en introduction, à savoir, une interprétation humaine du réel, alors, une nouvelle question se pose. Il s'agit de ne pas confondre interprétation d'une perception de l'espace (construction de la carte et extraction du matériau sémiotique à partir du réel) et interprétation d'une représentation de l'espace perçu (travail d'analyse critique de la carte). La

fonction première du SIG réside dans la première de ces interprétations mais la capacité du SIG géohistorique réside justement dans la capacité de ses applications à intégrer, non seulement la construction d'un espace perçu, mais également une représentation spatiale déjà construite à savoir une carte ancienne. C'est dans cette capacité que réside la fonction d'appareil critique du SIG géohistorique. Quant au mode de perception qu'il impose, on pourra le définir, en fonction des quatre catégories déjà nommées, comme majoritairement planaire puisque son fonctionnement, reposant en partie sur le vectoriel, se construit selon le mode géométrique. Il est à l'opposé même de l'antrhropocentrique dans la mesure où la quantité d'information n'est pas assimilable par l'esprit humain dans le cadre d'une vue synoptique. C'est par la lecture et non par la "vue" en tant qu'acte que se produit le processus d'interprétation de la carte, qui ne trouve son centre que dans le regard de l'utilisateur et non dans son mode de représentation. Il s'agit là d'une différence majeure avec les deux outils suivants: le GPS et le modèle numérique de terrain en 3D (dans le cadre de son utilisation en tant que visite virtuelle, car les modèles numériques de terrain peuvent accéder à un niveau de complexité graphique et d'abstraction aussi important que le SIG) .
La technologie GPS se distingue des outils précédants en tant que son interprétation porte, en soi, la finalité de l'outil, orienté vers l'action et le déplacement. Si la finalité impose une modalité de perception de l'espace clairement odologique, la représentation, quant à elle, et nous le redéfinirons plus avant dans la suite de cet article, se rapporte clairement à une modalité anthropocentrique et aréolaire. Elle pourrait se définir ainsi:

– "Je suis un point fixe et central. L'espace est mouvant autour de moi", autrement dit : "Je ne traverse pas l'espace . C'est lui qui me traverse"

et:

– "La portée de ma perception s'arrête là où se fixent les limites de la vision humaine dans le réel".

La perception de l'espace est alors centrée sur l'individu utilisateur, ou sur le point qui le représente, et anticipe sur le mouvement de celui-ci dans le réel, il représente plus d'éléments que l'utilsateur ne peut en percevoir dans la réalité de son déplacement. L'utilisateur devient capteur. Mais le temps du déplacement du signe graphique est synchrone (dans la limite d'intervale de temps de transfert des ondes) avec le temps du déplacement de l'utilistaeur dans le réel. En revanche cet outil devient totalement prescriptif et quitte le domaine de la représentation en cela qu'avec l'usage de voix de synthèse, la représentation graphique disparait pour devenir une indication de direction .

Le modèle numérique de terrain, orienté vers la découverte d'un patrimoine, et plus largement les visites virtuelles, marque un changement majeur dans la perception de l'espace patrimonial et du temps historique. Les deux modalités de perception de l'espace décrites précédemment, intègrent un paramètre supplementaire: le temps. Non seulement l'espace est mouvant autour d'un centre qui est le signe de l'utilisateur, mais le temps est lui aussi mouvant. Nous le traiterons de manière plus précise par la suite. Il est également utile de remarquer que, contrairement à la technologie GPS, les limites de la perception de l'espace sont celles de la représentation.
Cette brève présentation et redéfinition des quatre familles d'outils nous permet d'établir des correspondances entre les différentes modalités de représentation graphique du patrimoine et les déformations qu'elles impliquent dans la perception de l'espace patrimonial, par l'usage des visites virtuelles et donc par la réception de ces représentations.

3.2 La pensée de l'espace : vers une dé-localisation et dé-synchronisation de la navigation dans l'espace historique reconstruit.

Nous avons indiqué que la cartographie avait connu deux grandes révolutions, marquant un déplacement dans le mode de perception de l'espace. Il est remarquable d'observer que si la première révolution a marqué une nette coupure avec le mode anthropocentrique aristotélicien de perception et de représentation de l'espace, l'avènement du numérique marque un très net retour vers cette représentation centrée sur l'œil de l'usager et sur les deux modalités premières du mode de perception anthropocentrique, définies auparavant dans la description des applications de visites virtuelles. Ces modalités, dues à un type de représentations, viennent se surajouter au support de visualisation: l'écran dont Jeanneret nous indique la fonction d' "idéal narcissique"[10, p.135] en tant que mode d'écriture et en tant que développement d'un imaginaire.
Il ne s'agit cependant pas, ici, de traiter cet aspect sous l'angle d'une progression ou d'une régression des représentations mais sous l'angle d'une évolution cyclique du mode de perception.

3.2.1 Dé-localisation

Pour démontrer les mécanismes de cette évolution cyclique, il convient de revenir à la première des deux révolutions de l'écriture cartographique: Nous sommes à la moitié du XVIème siècle. Les cartes sont réalisées sous forme de vues cavalières construites sur un mode pictural (Fig.1) proche de la photographie. Ce qui est représenté est en adéquation avec ce qui est perçu, dans la limite des techniques utilisées. Puis une évolution commence à se développer et le cartographe se

représente lui-même dans le cadre du tableau qu'il peint. Son point de vue change. La perception se fait alors événement distinct du sujet qui perçoit, séparant l'œil du sujet, et donc ce qui est perçu, de ce qui est représenté.
L'analyse de ce phénomène a été considérablement traitée par Merleau-Ponty dans sa *Phénoménologie de la perception* [7, p. 240]. Ainsi, il écrit:

> *« La pensée objective ignore le sujet de la perception [...] Elle ne se donne pas d'abord comme événement dans le monde [...] mais comme une re-création ou une re-constitution du monde à chaque moment »*

Le cartographe continue néanmoins à travailler selon une approche déterminant une "étendue-qualité" (Fig.1) et non une "étendue-quantité". Dans son *Essai sur la connaissance approchée* [8, p. 50-51] Gaston Bachelard traite de la mesure qualitative comme « premier contact avec la notion d'espace », il en fait une véritable connaissance qualitative de l'étendue, une « étendue-qualité antécédente à l'étendue-quantité ». Reprenant les propos du psychologue William James, il traite de la « voluminosité comme d'une qualité commune à toutes les sensations [dont il extrait ce qu'il appelle] la sensation primitive d'espace».
En s'incluant dans la représentation de l'espace perçu, il se situe dans un rapport de contenant à contenu. Or, Merleau-Ponty, dans l'ouvrage déjà cité, exclut la perception de l'espace de ce rapport qui ne peut, selon lui, exister qu'entre des objets. Se rapportant aux propos de Kant, il détermine une ligne de partage entre « l'espace comme forme de l'expérience externe et les choses données par cette expérience » [7, p. 281]

Le cartographe se situe dans ce rapport de contenant à contenu et cette approche détache la représentation de son centre de perception : l'œil du cartographe. Le XVIIème siècle deviendra l'ère de la représentation quantitative et quittera le mode de perception anthropocentrique qui persistait depuis Aristote. Il est remarquable que l'avènement des technologies numériques, issues des jeux vidéos et des technologies Web, représente un retour vers ce mode de perception anthropocentrique, primitivement qualitatif (Fig.2), jouant sur l'identification de l'usager à son avatar numérique et donc recentrant la perception sur l'homme. Si le cartographe du patrimoine en trois dimensions se situe dans ce rapport de contenant à contenu, tel qu'on peut l'observer dans l'application Romereborn©, c'est, semble-t-il en partie du à la pratique, déjà intégrée dans ses shémas intellectuels, de l'utilsation des technologies numériques de positionnement. Il met en actes, tout comme son lecteur, la figure de son "hexis numérique", tel que peut le définir F. Georges dans sa *Sémiotique de la représentation de soi dans les dispositifs interactifs* [11,

p. 5-7], en tant que “territoire intérieur qui s’informe en l’écran ” , résultat d' “observation abstractive informée ” de niveau zéro, en tant que sa matérialisation à l'écran peut n'être qu'un simple point. Il est alors défini par l'action, par le mouvement. Sans mouvement, il devient impossible de le délimiter en tant que point parmi les autres éléments de la représentation. Cet Hexis, objet virtuel défini par l'action, peut donc être inclu dans un contenu. En l'occurence, une représentation d'un espace. Ses mouvements ne sont, dans cet “environnement” pas forcément soumis aux lois de la physique élémentaire et l'hexis numérique peut donc devenir cet ”œil”, cet objet de perception tout autant que d'action, sans obéir aux lois de la gravité . Peu importe alors que l'avatar puisse “voir”, en plongée, un espace virtuel dont il peut intégrer chaque “point”, voire chaque pixel, de la représentation. L'utilisateur devient un “point de vue” et donc un capteur virtuel d'espace virtuel. Tout autant que le “Citizen[...] as sensor[...]” de M. Goodchild dans sa géographie 2.0, dite “néo géographie”. La visite du patrimoine peut alors être envisagée sous l'angle d'une reconstruction évoluant au fil des actes et donc des déplacements de l'avatar au sein de l'application.

Figure 1. Vue cavalière de la ville de Lyon en 1548, par Androuet du Cerceau (Archives de la ville de Lyon) - Gravure en taille douce (AML. 2PH 250/194)6

6 URL in [4 (VN)]: http://www.archives- lyon.fr/old/fonds/plan-g/41.htm

*Figure 2 . Une **"vue"** (trad. de view) du centre de Rome le long de la rivière Tibre, du Théâtre Marcellus et de la colline du Capitole (à gauche) jusqu'au cirque Maximus (à droite).*[7]

Se pose alors une question qui ne s'applique pas uniquement à l'espace de la visite mais également, nous le verrons plus tard, au temps. Il s'agit de déteminer la clôture de l'espace perçu. Le mode de représentation semble bien proche de celui de la narration, à la différence que l'écriture non figurative, comme la parole, ne peutt exprimer la clôture de l'espace[8].

Lors d'une opération de numérisation de cartes issues de fonds anciens, comme c'est le cas du *Forma Urbis*, La clôture semble déterminée par le cadre même de la carte, par son support. Qu'en est-il de la visite virtuelle? Si nous observons les différentes démonstrations du projet Romereborn© et que nous l'appliquons en fonction de l'application finale: GoogleEarth®, il semble, *a priori,* que les limites de la perception de l'espace du patrimoine se fixent sur l'objet de représentation de départ, à savoir : le globe terrestre. Peut-être est-il nécessaire cependant d'aller plus loin dans l'étude de cette clôture de la perception. Explorer l'espace de la Rome antique, c'est déjà déterminer une barrière qui n'est pas celle de la Rome actuelle. La première clôture se fixe donc non seulement sur la ville en tant que lieu mais également en tant que symbole. Délimiter les frontières de la ville antique, observable à l'écran, peut sembler moins évident qu'il n'y parait. Les limites fixées par l'archéologie correspondent à une vérité scientifique qui peut être la plus pertinente. Mais est-ce vraiement de celà dont il s'agit lorsque l'on aborde la clôture de la

7 Image courtesy Barry Minor, IBM. Model © 2008 The Regents of the University of California - Image © 2008 The Board of Visitors of the University of Virginia tiré de:
http://www.romereborn.virginia.edu/rome_reborn_2_images/gallery/thumbs/RR1.1/RR_1.1_Tiber_view.jpg

8 Dixit C. Herrenschmidt, Possibilités de la cartographie et écritures [in] Journée d'étude: « Le métier de cartograpge hier et aujourd'hui », 28 mai 2009, Enssib

perception du lieu ? Les limites de la perception ne sont-t-elles plutôt celles de la technique qui permet l'observation de l'espace intégré aux patrimoine ?

En effet, si la clôture de l'espace, sur une carte ancienne, en est le cadre, qu'en est-il sur une représentation numérique dans laquelle se déplace le regard ? Il semble possible de déterminer, comme cadre possible de la perception, les limites du déplacement de l'usager à l'intérieur de l'application, mais, en s'attachant à décrire, dans le même temps, le projet *Romereborn*© et le projet *marseillenet*®*,* nous pouvons déterminer que la perception obéit à la fusion des quatre modalités de perception de l'espace reconstruit déjà mentionnées: aréolaire, anthropocentrique, odologique et géométrique. Cependant, chacune de ces deux applications intègre une différence de degré dans les modes de perception de l'espace qu'elle impose au regardant. La projet *marseillenet*® fixe des points sur une carte dont la fonction première est d'indiquer les haut-lieux de la ville et ne se départit pas de son impact touristique. Ces "hauts lieux", socialement déterminés, et conçus comme des "monuments" appartenant au patrimoine architectural ou naturel de la cité, sont alors des points d'ancrage de la visite. Si la carte conserve la dimension planaire de la visite virtuelle, le déplacement entre ses points, c'est à dire l'itinéraire, n'entre pas dans la logique de perception de l'espace marseillais et obéit aux fantasmagories liées à l'histoire du lieu ou à sa fréquentation. La perception qui en découle est totalement aréolaire. La clôture de la perception de l'espace est celle de la position du regardant, fixée au centre du haut lieu. Aucun déplacement n'est alors possible. Ce mode de perception, lié à la technique de la photographie panoramique n'est pas totalement absent du projet portant sur la Rome antique car, même si dans ce deuxième cas, la modalité majeure est celle de l'itinéraire et du déplacement, c'est également dans sa capacité à représenter l'intérieur des "haut-lieux" que se construit la perception aréolaire centrée, tout autant sur l'avatar, que sur le lieu (fig. 3 et 4). Les limites de la perception restent celles de l'avatar qui en est le centre. C'est la présence d'un centre qui en fait une représentation aréolaire, quelque soit la capacité de ce centre à se déplacer dans le monde virtuel visité. Toute représentation aréolaire et figurative devient, par là même, anthropocentrique.

3.2.2 Dé-synchronisation

La visite virtuelle de l'espace marseillais constitue une représentation, dont la technique même, impose un mode de perception temporelle, si ce n'est synchrone avec celui du monument, du moins, contemporain de l'époque de l'utilisateur.

Figure. 3 Une vue aérienne de l'intérieur de la Basilique de Maxentius et Constantine.[9]

Figure. 4. Vue panoramique par procédé photographique de l' intérieur de la cathédrale de la Major à Marseille

La photographie, représente une capture du lieu dans son état actuel. Il n'y a donc pas, pour le concepteur de l'application, à se poser la question de la reconstruction du lieu. La question de l'image et donc de ses attributs symboliques, semble bien plus prépondérante. Il s'agit, bien plus que de "montrer" le lieu, de le symboliser. Ce type de visite semble se détacher de la représentation pour devenir métonymie d'un lieu dont

9 http://www.romereborn.virginia.edu/rome_reborn_2_images/gallery/thumbs/RR1.0/BasMax01.jpg

l'intéret reste l'attractivité touristique. La temporalité de la représentation tient alors bien plus dans l'acte futur de la visite réelle du monument, que dans sa visite virtuelle dans le temps de la consultation sur l'application. Ce type de représentation porte une fonction quasi prescriptive intimant à l'usager, si ce n'est l'ordre, du moins le désir de la visite réelle d'un monument dont la forme est la même dans le temps présent. Ce temps devient une marque sociale d'un besoin de la cité d'attirer des visiteurs. L'objet en est le réel et non sa représentation. L'objectif est ainsi de rapprocher le visiteur du lieu et non, comme cela peut être le cas dans certaines structures, de l'éloigner de l'objet de patrimonialisation. La temporalité de la perception se fait donc dans la capacité de l'utilisateur à se projeter dans le temps d'une visite réelle du site.

La représentation spatiale en trois dimensons, quant à elle, plonge l'usager dans un "monde"[10] qui se veut fidèle à celui du temps évoqué. Les usagers, comme vu précédemment, sont utilisateurs de plus en plus nombreux de systèmes GPS et peuvent donc naviguer sans difficultés particulières sur un système orienté vers la perception d'un espace qui se construit en fonction de leur propre position synchrone dans le temps et l'espace de leur déplacement. Mais il bien plus anxiogène (et fascinant) de pouvoir effectuer un voyage dans une "collection", dont la compréhension passe obligatoirement par l'intégration, dans ses propres schémas intellectuels, de la désynchronisation entre le temps de la navigation et celui de l'espace dans lequel on navigue. La démarche est proche de celle de la science fiction ou de l'astronomie, dite grand public. L'espace dans lequel on reconstruit le discours n'est pas celui du temps dans lequel on évolue. Le processus se construit alors selon le mode de la *Time machine* de H.G. Wells, même si la désynchronisation se fait dans le sens inverse. Plus encore, on se rapproche de l'astronomie, science de la cartographie céleste, qui a amené ses praticiens à visualiser, non seulement l'espace, mais les temps révolus, de sorte à remonter jusqu'au Big Bang. En s'éloignant de la surface terrestre, ils ont pu remonter les époques. De même, en éloignant la représentation du réel, les concepteurs de visites virtuelles "reconstructives", ont pu effacer les outrages du temps sur les artefacts traités par l'archéologie, façonnant un espace reconstruit sans parfois posséder d'autres référents que des vestiges ou ruines, voire, une simple hypothèse de leurs existences dans le temps. Il ne s'agit pas de montrer ce qui est, mais ce qui a été, en en faisant un évènement du présent. Là encore, ne se rapproche-t-on pas du mode de la narration, voire de la fiction ? Il est peut- être un peu hasardeux d'établir ce lien, dans la mesure où la représentation produite par les concepteurs du projet Romereborn© montrent un souci de vérité scientifique, mais la

10 Il est possible de se référer à la Théorie des mondes possibles (cf: D. Lewis car la reconstruction implique une formulation des hypothèses historiques dues aux travaux des archéologues et historiens de l'architecture antique.

discipline même de la reconstruction historique porte en son sein la capacité à développer un imaginaire de l'histoire qui ne peut être totalement effacé de la représentation.Nous ne pouvons traiter, sur ces quelques pages, le rapport étroit entre représentation virtuelle et narration du temps historique, par la graphique, mais cette question, vaste, en engendre d'autres, qu'il serait interressant de voir figurer dans les interrogations portées sur le patrimoine: celle du rapport entre la clôture de la perception du temps de la navigation et la clôture du temps historique imposée par la représentation virtuelle de ce patrimoine.

4 Conclusion

Ce bref article ne peut que poser de nombreuses questions, sans prétention d'apporter des réponses définitives à l'étude d'un ensemble d'éléments du patrimoine qui, comme nous pouvons le constater, passe en grande partie, par la représentation graphique spatialisée. Nous n'apportons pas de solutions techniques permettant une amélioration des processus de patrimonialisation des fonds de cartes anciennes mais une réflexion sur l'évolution cyclique des représentations et de la perception de ce qu'est le patrimoine, avec une hypothèse bien plus qu'un constat : le retour à une perception anthropocentrique de l'espace, liée à ce que certains auront appelé la postmodernité et le retour à une forme d'image destinée à “impressionner” (au sens de “s'inscrire dans l'affect de” et non de “remplir d'admiration”) l'usager. Dans une société de l'image, la multiplication des représentations en 3D n'est pas dénuée d'un certain retour à la compréhension d'une esthétique des objets de sciences tels que les cartes. Si l'on veut établir un lien entre les époques, tel que nous l'avons fait au cours de ces quelques réflexions, il semblerait que la 3D, dans sa représentation du patrimoine soit ,en quelque sorte, un équivalent de la perspective du quattrocento, ce qui, dans sa réalisation même, n'est pas erroné, puisque la représentation en 3D sur un écran en deux dimensions est en réalité une adaptation de la perpective, dans la perception de l'espace ainsi que dans les déformations qu'elle génère. Il nous a semblé important de définir le cadre d'une réflexion future plus avancée, les modalités de la perception spatiale et temporelle du patrimoine, par la réception d'un mode de réécriture ou de réédition de ce patrimoine spatialisé, présenté graphiquement à l'écran, et dont l'avenir peut passer par la représentation holographique dans le réel sur les lieux mêmes où reposent les vestiges des temps passés.

Remerciements : Nous remercions Mme Lallich-Boidin, professeure en SIC (Université de Lyon1, directrice de thèse), et M. Guichard (MCF-ENSSIB, co-dir. de thèse), qui nous ont aidé à faire avancer notre réflexion et nous ont encouragé dans le processus de recherche sur les cartes en SIC.

5 Références bibliographiques

[1] J. Davallon, *Le don du patrimoine : une approche communicationnelle de la patrimonialisation*, Hermès Lavoisier [coll. Communication, Médiation et construits sociaux (dir. Yves Jeanneret)], Paris, 2006, 222 p., ISBN : 2-7462-1436-9

[2] Ch. Jacob, *L'Empire des cartes : Approche théorique de la cartographie à travers l'histoire*. Belin, Paris, 1993

[3] A. Koyré., H. Michel, V. Ronchi [et al.], *La science au seizième siècle, colloque de Royaumont (1957)*, Hermann [coll. Ecole pratique des hautes études: Histoire de la pensée], Paris, 1960 344 p.

[4] Archives municipales, *Forma Urbis: les plans généraux de Lyon XVIème XXème siècle*, (coll. Les dossiers des archives municipales, n°10), Lyon, 1999 ISBN 2-908949-18-0 Version Numérique (VN) : Archives de Lyon : *www.**archives-lyon**.fr* [Disponible en ligne] URL: http://www.archives-lyon.fr/old/fonds/plan-g/plan2.htm, (consulté le 15/12/2008)

[5] W. Benjamin , 1939, « L'œuvre d'art à l'ère de sa reproductibilité mécanisée », in W. Benjamin *Œuvre complète* vol.III, Gallimard [coll. Folio/essai], Paris, 2000, p. 269-316.

[6] H. Fondin. *Le traitement numérique des documents*, Hermès, Paris, 1998, 382 p.

[7] M. Merleau-Ponty , 1976, *Phénoménologie de la perception*, Gallimard [coll. Tel], Paris, réed. 2001, p.240-344

[8] G. Bachelard, 1987, *Essai sur la connaissance approchée*, Vrin (6ème éd..: 2006) [coll. Textes Philosophiques], Paris, 2006, p.47-68

[9] R. T. Pédauque, *La redocumentarisation du monde*, Cépadues ed. , Toulouse, 2007, 213 p.

[10] Y. Jeanneret, « Économies de l'écran : discours, pratiques et imaginaires entre visible et invisible », in : [sous la dir. de.] Roelens, N. et Jeanneret Y., *L'imaginaire de l'écran / Screen Imagery*, Ed. Rodopi, Amsterdam-New-York, 2004 p.141-162

[11] F. Georges, *Sémiotique de la représentation de soi dans les dispositifs interactifs: l'hexis numérique ,* thèse de doctorat, Sciences de l'art, Université Paris I Panthéon-Sorbonne, décembre 2007, 467p.

Le concept de musée virtuel thématique : la collection comme visite, la visite comme lecture, la lecture comme stratégie.

L'exemple du musée thématique sur l'Annonciation.

Ioannis Kanellos (1), Sister Daniilia (2)

(1) Telecom Bretagne, Département Informatique, CS 83818 29238 Brest cedex 3, France
(2) Ormylia Foundation, Art Diagnosis Centre, 63071 Ormylia, Greece

Mots-clés : Musée virtuel thématique, ontologies locales, points de vue, représentation des connaissances à profondeur variable, stratégies de lecture, scénarios de visite, détail et interpicturalité.

Keywords: Thematic virtual museum, local ontologies, points of view, variable depth knowledge representation, reading strategies, visiting scenarios, detail and interpicturality.

Résumé : L'article discute les idées directives d'un musée virtuel thématique, en matière de représentation des connaissances (RC) et d'implémentation. Le cas d'étude est le musée sur l'Annonciation (www.annunciation.gr). Nous abordons le problème de la RC suivant plusieurs points de vue et à profondeur variable ainsi que le besoin d'une modélisation des ressources faisant la part tant au détail qu'à l'interpicturalité. Nous expliquons l'importance de reprendre la notion de visite d'un musée virtuel dans celle de lecture. Nous exposons enfin une structure des données susceptible de servir les stratégies qui sous-tendent trois genres de visite (découverte, étude et approfondissement). Nous concluons par une discussion sur quelques enjeux concernant le développement de musées virtuels aujourd'hui.

Abstract: The paper discusses the leading ideas of knowledge representation (KR) and implementation of a thematic virtual museum. The study case is the Annunciation museum (www.annunciation.gr). We analyse the problem of multi-point of view and of variable depth KR as well as the need to model the resources in a way that may encapsulate both details and interpicturality relationships. We explain the significance to treat with the notion of a visit of a virtual museum as a special case of a reading procedure. We finally outline the data structure likely to give evidence to strategies supporting three different genres of visit (discovery, study and deepening). We conclude with some topics underlying the development of virtual museums nowadays.

1 Introduction : les hésitations du musée virtuel

Comment se fait-il qu'un récit aussi court et aussi fragmentaire que celui sur l'Annonciation[1] a reçu tant de considération et a connu une telle postérité dans la production artistique occidentale depuis tant de siècles ? Que signifie-t-il, que transmet-il de si important pour inspirer décidément tant de générations et susciter tant d'intérêt créatif, même chez des artistes non chrétiens, voire non croyants ?

La question apparaît sans doute insolite. Elle l'est, d'ailleurs. Elle coordonne une finalité. En fait, elle reformule une demande ancienne pour une représentation des connaissances respectueuse de nos habitudes de lecture. En la posant, nous entendons, plus particulièrement, attirer l'attention sur un aspect remarquable qui sous-tend le traitement de l'information dans le cadre des musées virtuels. On notera, par exemple, qu'on ne saurait trouver de réponse à cette question à aucun musée virtuel ; ni même au-delà des musées virtuels, peut-être, nonobstant la diligence de maître Google.

Malgré tant d'investissements, de fracas et de publicité, il s'avère que nous nous ennuyons fréquemment en visitant des musées virtuels. Ou, peut-être, sentons-nous quelque frustration, souvent inavouée mais pas moins réelle, de sinistre mémoire. On se demande, en réalité, si l'investissement qui vise à réaliser un musée virtuel apporte vraiment une valeur ajoutée à la problématique de l'héritage culturel ; notamment, s'il contribue véritablement sur le plan de l'accessibilité aux œuvres. L'ennui que nous ressentons est probablement l'indice de quelque désir de connaissance invalidé. On observe, d'ailleurs, une certaine parenté des critiques avec celles qui se sont adressées naguère à diverses réalisations de e-Learning.

Les problématiques du « e-Museum » ne sont pas encore stabilisées ; vraisemblablement, ne le seront-elles jamais, tant elles dépendent des technologies d'où elles tirent leur raisons d'être. Cependant, elles s'affichent d'emblée des volontés éducatives [6, 7] qui les rendent sensibles aux mêmes réfutations. Tout comme le « cours virtuel », le terme « musée virtuel » désigne prioritairement un domaine de mutation de plus de nos pratiques culturelles par l'avènement des technologies de l'information et de la communication. Il ne doit cependant pas être compris en termes exclusivement techniques. L'idée de musée virtuel va plus loin que la simple opportunité et les effets de mode, bien plus loin

1 Le texte entier est contenu dans 13 petites lignes dans la version du manuscrit grec et se trouve au seul évangile de Luc (ch. 1, lignes 26 à 38). On trouve aussi, dans la tradition islamique, un récit apparenté, dans la Sourate 19, intitulée « Marie (Maryam) » ; là, la description est encore plus courte : elle tient en 6 lignes au plus (16 à 21) !

que les préoccupations d'une ingénierie asservie à son temps. De façon minimaliste, le musée virtuel se veut, tout d'abord, un patrimoine numérique sur une collection d'œuvres ; il est souvent vitrine ou reproduction d'expositions réelles, transitoires ou permanentes ; il se transforme, parfois, en site éducatif ; il peut être un univers entièrement virtuel, même avec des œuvres virtuels ; il intègre des jeux pour petits et grands (les fameux « serious games »), etc. Ses multiples transfigurations recoupent nombre d'initiatives qui semblent centrales dans une société comme la nôtre, mieux comprise, désormais, comme une société de la Communication, de l'Information et de la Connaissance. Globalement, il désigne des enjeux complexes mais porteurs, relevant de l'héritage culturel, dans une économie numérique orientée par les services.

Curieusement, alors que le crédit technologique semble augurer des possibilités illimitées, une étude que nous avons menée en amont, portant sur plus de cinquante musées, a montré qu'environ 9 fois sur 10 on a affaire à des musées virtuels qui sont des calques de musées réels, i.e. de musées qui se trouvent physiquement implantés en un lieu et opèrent sur un mode plutôt classique. Étrange constatation alors que la notion de musée virtuel était censée nous délivrer des contraintes d'unité du lieu qui pèsent sur les collections. L'intérêt d'un musée virtuel, son essence même, est précisément cette ouverture qu'il permet en matière d'accessibilité diversifiée à des collections d'objets choisis, prioritairement pour leur caractère culturel[2]. Or, aujourd'hui encore, le concept de collection reste majoritairement enfermé dans la logique de l'institution, probablement parce que le marché était déjà formaté et occupé par des musées classiques qui, naturellement, ont eu plus des moyens que d'autres pour développer des vitrines en ligne. Ces musées réels et plus ou moins fortunés, ont compris les TIC comme des moyens complémentaires pour promouvoir leurs actions culturelles ; incidemment, aussi, pour soutenir leur business et leur image à un moment où le monde muséal semblait subir des mutations profondes.

Cela ne serait peut-être pas fâcheux, ni même nuisible en matière de développement du concept si, même avec une telle limitation, l'opportunité virtuelle ne s'avérait décidément castrée. Certes, on voit ci et là des innovations technologiques séduisantes investir peu à peu un métier nouveau, celui de muséologue (et de muséographe) numériques : des bases de données multimédia, des jeux éducatifs [19, 22, 25, 26], des

2 Rappelons la définition de l'UNESCO d'un musée : « une institution permanente, sans but lucratif, au service de la société et de son développement, ouverte au public et qui fait des recherches concernant les témoins matériels de l'homme et de son environnement, acquiert ceux-là, les conserve, les communique et notamment les expose à des fins d'études, d'éducation et de délectation ». Elle reste aussi valide dans le cas des musées virtuels. Bien entendu, le musée virtuel peut aussi concerner des objets de n'importe quelle nature, fonction et vocation.

plateformes de type 2.0 pour la création de nouvelles sociétés muséales [19, 20, 22, 26], de la réalité virtuelle [18, 21, 23, 24], des interfaces plus attractives, esthétiques [27] et plus ergonomiques, des techniques de traitement d'image attrayantes, de la synthèse de parole recevable [21, 28], des protocoles de recherche évolués, etc. Cependant, l'ensemble reste globalement contenu dans des interrogations communes qui relèvent de la conception d'un site commercial plus ou moins standard. Tristement, le thème de la visite, prioritaire pour un musée, qui deviendrait visite virtuelle dans le cas d'un musée virtuel, se voit dégradé, entièrement fondu qu'il apparaît dans une navigation usuelle. Même les standards émergents, plus ouverts à des réseaux sémantiques, ne lui réservent pas une place importante [3]. Par ailleurs, la conception de la structure des données ne permet pas une flexibilité suffisante pour opérer des changements significatifs. En fait, dans la quasi-totalité des musées virtuels calques, on rencontre, dirait-on, une pensée unique qui impose divers régimes de restriction :

1. Le visiteur du musée virtuel de ce genre n'a pas la possibilité de voir toutes les collections qui existent dans le musée réel mais seulement une sélection, plus ou moins étendue, plus ou moins en correspondance avec son désir de visite, une sélection décidée par avance par le concepteur du site.
2. Il ne peut pas non plus réaliser sa visite comme il aurait souhaité, i.e. en naviguant dans les collections proposées de manière qui reflèterait encore quelque chose de ses propres pratiques ; en effet, dans la mesure où il doit suivre un scénario de visite unique, également fixé, il ne peut jouir que de ces formes de plasticité que permet la structure de l'hyperdocument (navigation relationnelle et associative ; et souvent à vue).
3. Le visiteur ne peut même pas formuler une demande d'information complémentaire sur les œuvres proposées, qui sont informées de la même manière pour tout le monde, i.e. sans égard au profil du visiteur, à son niveau, à son histoire ou à son objectif de visite.
4. Enfin, il ne peut pas toujours s'approcher des œuvres, pour examiner certains de leurs détails ; les œuvres, dans leur présentation picturale, sont données, le plus souvent, avec des résolutions qui correspondent à une vue plutôt lointaine, même après quelques agrandissements possibles permis par les outils de visualisation. De l'autre côté, il ne dispose pas non plus de moyens de (re)contextualisation des œuvres.

L'intérêt semble, donc, dès la mise en place d'un tel musée virtuel, compromis ; la notion de visiteur reste au niveau de la métaphore pour parler d'un utilisateur d'un système de gestion de bases de données. On en saisit les raisons : de tels musées sont bâtis sur un modèle économique commun qui leur offre, certes, les fondations mais aussi les limitations. À part, donc, quelques cas d'approfondissement emblématiques, que l'on trouve dans certains musées calques disposant des moyens [3], et qui proposent des études à part qui se surajoutent à la structure globale du site, le reste est souvent réduit à une base de données, plus ou moins étendue et soignée, plus ou moins conviviale et appropriable, plus ou moins renseignée et affinée, avec les fonctionnalités classiques en matière d'interrogation et de navigation. Le musée virtuel, pourrait-on dire, s'est également abandonné aux charmes de la pensée « BD » revenant du coup à une posture que l'on croyait à jamais révolue, somme toute patrimoniale et conservatoire. Tout le reste doit s'inscrire dans le cadre de cette pensée. Qui fait office de norme.

La pensée « BD » est simple ; elle repose. Elle n'offre pas seulement une liberté hors normes mais aussi une liberté voulue hors les normes. O tempora ! O mores !

Ainsi, le concept de musée virtuel renvoie aujourd'hui, majoritairement et typiquement, à l'idée d'imitation. Il assure, assurément, son rôle de lieu d'émergence de réseaux sociaux, par la constitution de diverses communautés d'intérêts notamment ; mais il reste faillible devant l'exigence de formuler une proposition respectueuse des volontés et des pratiques culturelles. Il rejoint les prérogatives d'un service facilitateur de commerce électronique ; mais il trahit les espoirs qui voulaient de lui un outil d'accès à la connaissance.

Le concept de visite, nous revenons, semble intéresser peu les concepteurs des musées virtuels, qui restent dans une optique de communication et de démonstration. Le concept de lecture encore moins. En effet, l'organisation des documents se réalise généralement sans tenir compte des pratiques de lecture[4]. Décidément, notre question inaugurale restera à jamais sans réponse dans la mesure où elle convoque un niveau

3 Comme la rubrique des « Œuvres à la loupe » du musée du Louvre [24], le sous-menu « Masterpieces » du Rijksmuseum [21], la section « Explore » du British Museum [18], etc.

4 Remarquons que le paradigme d'un musée virtuel reçoit toujours de la plupart des muséographes et des muséologues une méfiance non dissimulée. Il n'est peut-être pas sans rapport avec notre analyse. La logique des bases de données impose un cadre qui fossilise les pratiques muséales en matière d'organisation d'une exposition, bâties traditionnellement sur la notion de visite. C'est comme si, dirions-nous, dans un musée traditionnel, on abattait les cloisons pour en faire un immense espace où tous les œuvres seraient ensemble et où tout parcours deviendrait possible.

de lecture qui n'est pas envisageable dans un musée virtuel typique de notre époque.

2 La visite comme (stratégie de) lecture

Cette brève promenade dans ce que nous trouvons de commun et de récurrent dans les musées virtuels de nos jours visait à poser rapidement les jalons d'une critique pouvant nous orienter vers quelque développement alternatif. Déjà, on relève que si la première mission de la notion d'exposition virtuelle est de s'affranchir de l'emprise de la matérialité d'un musée réel, la notion de collection doit également se chercher des principes de constitution sur une base différente ; disons, par exemple, thématique. Autrement dit, le mouvement de résistance culturelle dans ce régime quelque peu totalitaire des musées virtuels vitrines d'aujourd'hui, produits sur la logique du couple « institution de support et pensée BD », s'appellerait probablement musée virtuel thématique. Ailleurs, on l'appelle imaginaire [2]. Son rapport au musée virtuel typique serait, en un sens, le rapport de l'intension à l'extension.
Le musée virtuel sur le thème de l'Annonciation que nous présentons ici (www.annunciation.gr) en sera notre support en tant que cas d'étude et d'application[5]. Il s'agit d'un musée qui présente une collection d'œuvres représentant le thème de l'Annonciation dans la tradition iconographique byzantine.

Pour sa conception, notre départ était la notion, certes vague mais fondatrice, de pratique de visite [5]. Plus précisément, il s'agissait pour nous d'interroger, d'entrée en scène, le thème de la visite et de chercher à le revitaliser au sein d'une théorie de l'interprétation [16, 14]. Cet engagement vient du fait que nous avons toujours reconnu en la notion de visite d'un musée (réel certes, mais, à plus forte raison, virtuel) une déclinaison de plus de la notion de lecture [15]. Nous défendons une approche pleinement interprétative des affaires sémiotiques qui finit par rehausser le rôle de la réception dans toute écologie communicative. Par conséquent, nous avons essayé d'implémenter dans cette première version du musée sur l'Annonciation des exigences de représentation et de présentation des connaissances commandées par le concept de stratégie de lecture.

La représentation numérique assure, on le sait, la transmutation de l'objet à exposer en une constellation de fragments d'information. L'objet

5 Il a été développé en commun par la Fondation Ormylia et TELECOM Bretagne. Le développement a été assuré par la société Trinity Systems (www.trinitysystems.gr). Il est le résultat de nombreux travaux menés depuis plusieurs années dans le Art Diagnosis Centre de cette Fondation et à TELECOM Bretagne.

matériel présenté, qui était hier encore contraint, et de façon non négociable, dans l'unité du lieu, du temps et de la visite, s'efface, dans un musée virtuel, devant ce qui le remplace et le « dit » dans un discours nouveau, généralement multimédia et, heureusement, partiellement à la portée de la machine. On a certes perdu l'objet mais son spectre, sa structure informationnelle, permet de définir plusieurs régimes de transition et d'intégration supplétifs. La dissociation entre substance et information, on le sait aussi, libère la structure des connaissances qui décrit un objet et ouvre, plus avant, à des possibilités d'évolution inédites, dans des espaces qualitatifs, indépendants et combinables à désir.

Précisément, dans le cas du musée sur l'Annonciation, l'importance que nous accordons aux interprétations normées tend à n'en retenir que des espaces qui restent en accord avec des pratiques attestées en Histoire de l'Art. Plus techniquement, les parcours de lecture sont bâtis sur cinq points de vue qui constituent autant d'espaces d'analyse d'une œuvre d'art[6]. Le terme « point de vue » doit ici être entendu de manière technique : il désigne une localité homogène dans la représentation des connaissances. Pour une icône représentant le thème de l'Annonciation, par exemple, nous avons fini par en distinguer cinq : (i) Description, (ii) Esthétique, (iii) Contexte de production et d'exposition, (iv) Exploration technique (i.e. physico-chimique) et, ce qui synthétise les précédents, (v) Interprétation.

L'art concerné dans cette tradition picturale est un art plutôt stylisé, maniéré souvent, généralement narratif, essentiellement figuratif et, bien sûr, fortement normé. On pourrait sans doute généraliser ces points de vue à d'autres genres artistiques qui ne se reconnaissent pas nécessairement sous ces caractéristiques génériques. Contentons nous ici de remarquer simplement que ces points de vue ne sont rien d'autre que des ontologies locales, thématiques en fait, qui organisent de façon cohérente de grands domaines de connaissance et sont issues des pratiques d'analyse de ce genre pictural. Elles se ramifient immédiatement, et donnent naissance à diverses sous-ontologies de spécialité. Par exemple, le point de vue Esthétique décline quatre sous-ontologies au total : Composition, Expressions, Postures & Gestes, Couleurs & Lumière ; l'Exploration technique cinq : Support & Préparation, Dessin, Technique & Pigments, Stratigraphie et, enfin, Palette (cf. un aperçu sur la Figure 5 ci-dessous). Peu importe leur nombre et leur nature, ce sont ces ultimes qui constituent le squelette de l'ontologie du domaine tout en conditionnant la notion de parcours.

6 Certes, d'autres auraient pu aussi bien convenir. La question n'est pas le nombre, ni même la qualité, mais la reconnaissance de régimes d'analyse cohérents, attestés par des discours d'analyse dans le domaine de l'Histoire de l'Art. Et, évidemment, leur potentiel de représentation dans une machine !

Globalement, la structure des connaissances se déploie suivant un schéma de représentation multi-niveau (Figure 1).

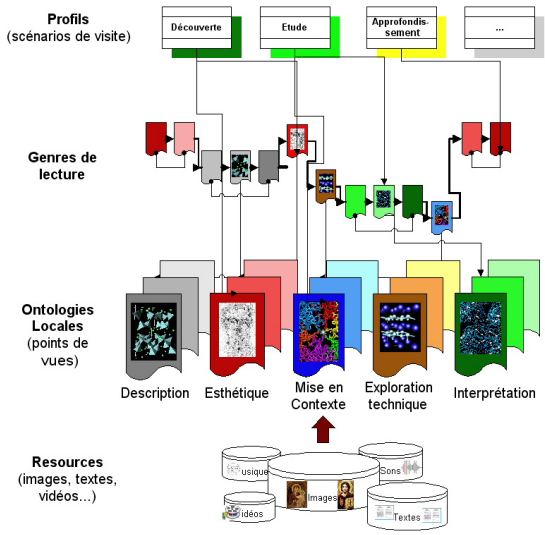


Figure 1 : Les niveaux d'organisation des connaissances dans le musée virtuel sur l'Annonciation.

3 Ontologies locales et genres de lecture

Les ontologies locales (points de vue et sous-ontologies des points de vue) deviennent dans la suite la matière pour définir, précisément sur elles, les genres (i.e. les formes) de lecture. La notion de genre est capitale dans la mesure où elle assure le pont entre la structure des données et la scénarisation de la visite [8, 9, 10]. Elle renvoie à un niveau de codification des normes. Ici, il s'agit de normes de lecture.

Pourquoi lit-on ? Comment lit-on ? Peu de travaux se sont attachés à ces questions qui passent ordinairement pour évidentes ou triviales. En tout cas, elles ne font pas l'objet d'études systématiques. Il est pourtant facile de constater que l'on ne fait pas la même chose lorsqu'on lit pour découvrir, pour s'inspirer, pour s'informer, pour vérifier quelque chose, pour approfondir ou affiner, pour se rafraîchir la mémoire, pour étudier, pour copier ou apprendre par cœur un morceau, pour acquérir des connaissances, pour comparer… Chaque genre de lecture mobilise des facultés cognitives différentes dans la mesure où il fait appel à des procédures de sélection, d'organisation et d'évaluation différentes. Mais il n'y a pas que la finalité de la lecture qui importe : il y a aussi ses

temporalités, i.e. ses contraintes de réalisation dans un référentiel temporel qui s'offre pour cadre à une pratique. On n'applique pas la même stratégie de lecture lorsqu'on « scanne » un document, lorsqu'on l'examine rapidement en le feuilletant, lorsqu'on a le temps pour le lire avec intention et attention ou, enfin, lorsqu'on lui consacre son temps, sa vie même, pour l'étudier et l'approfondir [11, 13]. La lecture n'est pas une donnée unique : elle dépend d'un grand nombre de paramètres, et elle se configure différemment suivant la temporalité qui lui est assignée pour se réaliser ; c'est cette dernière qui spécifie la stratégie en œuvre.

Quoi qu'il en soit, il apparaît important de réserver, dans la conception même de l'architecture des connaissances, une place importante aux genres de lecture en tant que traces de pratiques attestées. Autrement, la lecture est asservie à une raison qui avance sur un mode associatif et qui édifie, sur la logique des bases de données, un paradigme de lecture presque indéplaçable. Dans notre cas, ces genres de lecture sont des constructions qui se réalisent au-dessus du niveau des points de vue. Nous nous expliquons.

Un genre de lecture est une codification des normes de réception en vigueur dans une communauté. Il permet le déclenchement des stratégies de lecture adaptées à une écriture donnée. Il permet, en un sens, de réunir dans le même projet écriture et lecture. Cette normativité a une incidence particulière sur l'organisation des informations ; elle prend la forme d'une surdétermination de structures glanées dans divers points de vue. Ces idées, qui ne concernent certainement pas que le musée virtuel, ont été maintenues au centre de nos préoccupations lors du design du musée que nous avons développé. Plus précisément, nous avons choisi trois formes de visite, qui nous ont semblé fondamentales, pour illustrer le concept de genres de lecture :

- La visite « Plaisir ou Découverte ». Elle est bâtie sur le profil de l'amateur de l'art. La temporalité s'y trouve sans contraintes. L'interruption peut s'opérer à tout moment sans nuire l'intention globale. La collection entière est organisée en sous-collections autour d'un prototype. Le visiteur entre par le prototype et procède à la visite d'une partie du musée en évoluant d'une œuvre à une autre suivant son goût et son désir, essentiellement par association. La notion structurante de cette approche de la collection du musée est la similitude ; cette dernière reste modulable au sens où les critères de sa définition et leur combinatoire sont laissés au goût du visiteur. L'information sur les œuvres est ici volontairement réduite ; elle concerne le lieu et l'époque de l'œuvre, le cycle iconographique dans lequel elle

appartient, le support de réalisation, le courant artistique, le peintre, le type thématique et les prototypes lui étant associés. La reconfiguration de la visite suivant un autre prototype ou suivant d'autres critères de visite est à tout moment possible.

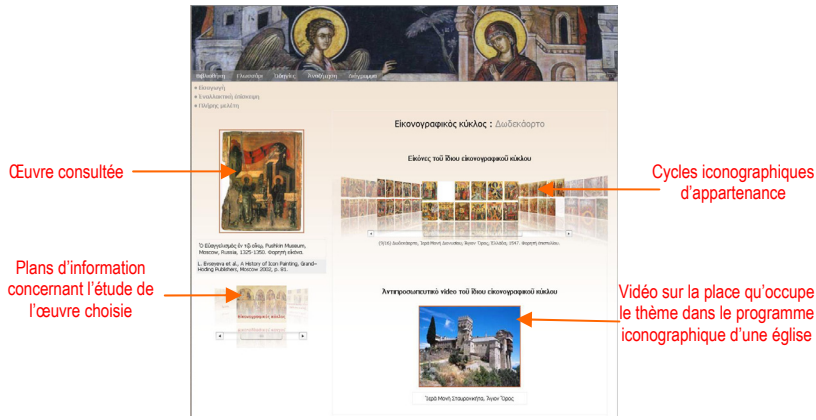


Figure 3 : Une étape de la visite « Découverte ». Le visiteur aborde ici l'œuvre sous l'aspect de son rapport à la narration. Des exemples des cycles iconographiques de son appartenance ainsi qu'un exemple de la place qu'elle possède dans la réalisation des fresques d'une église sont données à titre d'exemple.

- La visite « Étude ». Elle est conçue sur le profil de l'étudiant. Contrairement à la précédente, qui se met en place de manière incrémentale, ici le visiteur aborde les œuvres par classes qui sont produites par des topiques d'étude prédéterminées. Par exemple, on pose des questions concernant le thème de l'Annonciation diachroniquement, diatopiquement, historiquement, contextuellement (suivant le peintre, le support, l'école et le style, le type thématique, le cycle iconographique et narratif d'appartenance etc.), des éléments descriptifs, esthétiques, etc. Autrement dit, alors que la précédente était une visite personnalisable pendant laquelle l'accent était mis sur l'individualité de l'œuvre et l'idiosyncrasie du parcours, ici c'est l'interpicturalité (le rapport des tableaux entre eux) qui organise la lecture suivant plusieurs catégories. La visite « Étude » vise, en quelque sorte, la reconquête d'une identité de l'œuvre à travers les formes de sa « socialité » textuelle et picturale qui lui assignent une place et un rôle dans l'histoire de la production artistique.

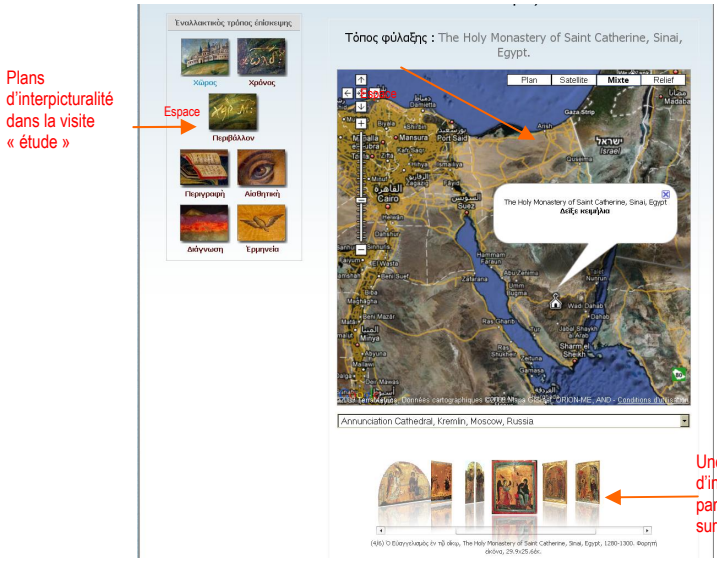


Figure 4 : Une étape dans la visite « Étude ». Le visiteur aborde généralement le contenu de l'œuvre comme la résultante d'un « topos inter-iconique ». Ici, il a demandé les œuvres avec le même thème conservées au monastère de Sainte Catherine au Sinaï (Égypte).

- La visite « Approfondissement ». Elle est construite sur le profil du spécialiste (le peintre, le restaurateur, le conservateur, l'expert en Histoire de l'Art, etc.). Elle a surtout une visée paradigmatique. On propose, en effet, des cas typiques d'étude qui ont été poussés bien plus loin que les précédents. L'information, sur tous les points de vue, se voulait ainsi aussi riche que possible et toujours susceptible de s'améliorer par des apports ultérieurs. On arrive ici même à des études physico-chimiques, des analyses plus étendues en matière d'esthétique, des développements historiques ou philosophiques soutenus, etc. Trois œuvres ont été choisies selon une volonté de diversification ; entre autres, parce qu'ils relèvent de traditions, de styles, d'époques ou des techniques différentes. Cette forme de visite est, en un sens, l'aboutissement des deux précédentes. Elle constitue un véritable cours d'initiation à l'art de l'iconographie byzantine. Et permet, rétrospectivement, d'étudier les œuvres abordées par une visite de type « Découverte » ou « Étude » différemment.

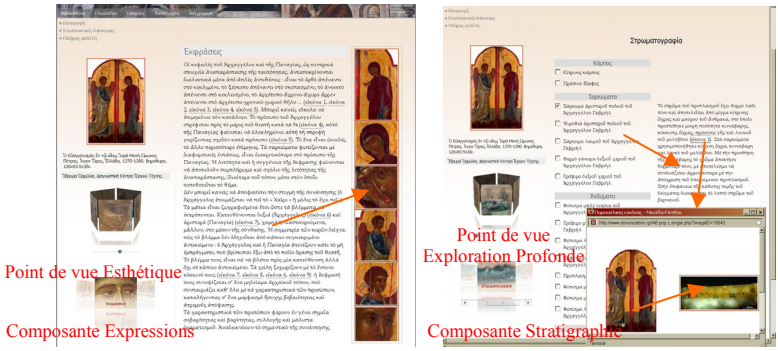


Figure 5 : Deux étapes dans la visite « Approfondissement ».

À gauche, le visiteur a choisi le point de vue Esthétique et, plus particulièrement, sa composante concernant les Expressions. Le texte propose une analyse systématique des expressions des visages des figures (Ange et Marie). Des images intégrées dans le texte illustrent les propos ; on peut les consulter avec des résolutions très hautes pouvant ainsi aborder les œuvres à travers un grand nombre de détails.

À droite, le visiteur a choisi, pour la même œuvre, le point de vue d'une Exploration Profonde et, plus particulièrement, sa composante concernant la Stratigraphie, souhaitant, par exemple, étudier le pinceau du peintre ; il a sélectionné un point du tableau (carnation du pied de l'Ange) ; un court texte explique la préparation de la palette et explicite les couches de couleurs sur ce point précis. Une image de stéréo-microscope en propose une vision détaillée ; on peut également l'étudier en très haute résolution faisant ainsi dévoiler au grand jour des éléments importants de la technique du peintre.

Les genres de lecture et les scénarios de visite sont des notions jumelées. Certes, un genre de lecture convoque des finesses qu'un scénario de visite, qui discrétise, fixe et opérationnalise, ne saurait capter. D'une certaine manière, les scénarios de visite ne sont que des particularisations de genres de lecture, réalisables dans une machine. Il ne reste pas moins que leur coordination est significative dans un musée virtuel. Dans le cas du musée sur l'Annonciation, les trois genres de lecture choisis se projettent à autant de scénarios de visite (Figure 6). Pour résumer :

- le genre de lecture traduit, d'un côté, les normes de réception culturelle, telles que nous aurions souhaité retrouver en nous promenant dans un musée virtuel ; il est le cadre d'une naturalité qui vient du respect des pratiques ;

- de l'autre, il sert de fondement pour la scénarisation de la collection, i.e. sa mise en valeur à travers des parcours de visite ; il s'agit d'une étape indispensable, aux sources d'un design ergonomique sans conflits avec nos habitudes sémiotiques.

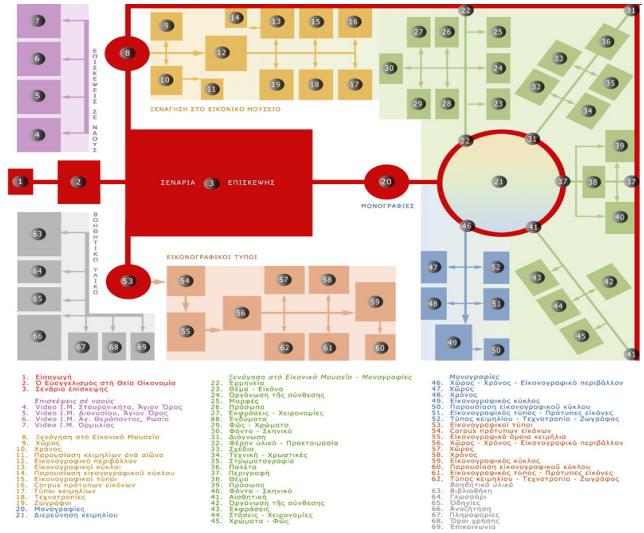


Figure 6 : Scénarios de visite. Les différentes parties (couleurs différentes) schématisent les différents parcours de visite ; elles correspondent à un genre de lecture particulier.

4 Interpicturalités et détails : Le proche, le lointain et les actifs de la lecture

On parle peu de lecture ; mais on ne parle pas non plus davantage des pratiques d'analyse des œuvres. En Histoire de l'Art, il semble se dégager un vague invariant que l'on pourrait qualifier d'invariant du « regard à distance ». C'est un regard durable qui a depuis longtemps nourri nos manières de voir. En effet, nous avons pris l'habitude de regarder une œuvre depuis une certaine distance – disons, pour un tableau de taille moyenne, depuis une distance de deux à cinq mètres habituellement. L'occupation de l'espace dans un musée réel et les consignes de présentation y sont sans doute pour quelque chose.

On n'imagine généralement pas combien cette habitude, qui n'a rien de naturel, influe dans nos façons de voir, de lire et de comprendre. C'est en réalité un trait qui contraint fortement nos catégories de réception. Il

convoque une forme d'énaction qui fixe d'emblée l'horizon des potentialités d'analyse. Cependant, comme le remarque Daniel Arasse dans son remarquable traité sur le détail [1], il y aurait, à côté de cette approche des œuvres à distance, qui fonde une Histoire de l'Art typique, une autre Histoire de l'Art, toute autre peut-être, faite de regards rapprochés, qui choisissent l'abord de l'œuvre sur la base d'un détail. Ce serait, précisément, une Histoire de l'Art du détail. Elle complèterait la première en proposant des lectures originales parfois en réinventant le contenu de l'œuvre ou, au moins, en l'éclairant différemment. L'idée d'Arasse semble être celle d'une possibilité d'organiser la lecture d'un tableau à travers un élément singulier, un élément distinctif en quelque sorte, contrairement à une pratique qui serait fondée sur ce qui rassemble et classe les œuvres. Par la porte du détail on engage « une lecture non pas devant mais aussi dans le tableau ». Le singulier guide autant une lecture que le général ; seulement, il la guide différemment. Il rectifie aussi ou il complète le rapport entre savoir et voir : il ne s'agit pas de « soumettre le voir au savoir » mais de constituer un savoir aux sources même du voir. En d'autres termes, le détail ouvrirait à des cheminements qui sont susceptibles d'aboutir à des interrogations autrement productives sur l'œuvre, mettant en lumière des contenus et des valeurs imperceptibles à un regard arrêté à une distance fixe et contraint de voir à travers la grille de similitudes qui unifient le vu avec le déjà vu.

On pourrait même dire que le rapprochement, qui vise à construire des stratégies de lecture sur les détails, pondère la vision statique de l'œuvre en la confrontant à une vision dynamique qui correspondrait à un mouvement vers l'avant ; ce faisant, il établit un régime de proximité généreux pour repenser l'interprétation sous forme d'opposition ou d'affinement.

Serait-ce suffisant ? On le comprend, il y aurait aussi une troisième opportunité pour la lecture. Celle, également dynamique, qui correspondrait à un mouvement vers l'arrière et dont l'objectif serait de réintroduire l'œuvre dans ses sociétés (d'appartenance et de référence). En s'éloignant de l'objet exposé, en réalité, on le recontextualise. La troisième Histoire de l'Art tirerait, donc, sa légitimité de la transmutation de l'intertextualité (qui régit les textes) en « interpicturalité ». Elle serait, concrètement, fondée sur le rapport d'une œuvre à d'autres œuvres. Ces diverses sociétés fonctionnent, dans une certaine manière, comme la notion de voisinage en Topologie. Elles apportent des déterminations venant de l'extérieur de l'œuvre. Elles constituent les écologies nécessaires pour établir une identité située de l'œuvre.

En récapitulant :

- le rapport au détail serait un rapport intrinsèque, un rapport, somme toute situant ; il procèderait d'une logique méréologique ; sorte de synecdoque picturale, il chercherait à éclairer le tout (voire un autre tout) à travers la partie et même l'élément singulier ;
- le rapport à un espace interpictural, au contraire, témoignerait de l'inscription de l'œuvre à des globalités sémiotiques supérieures ; il procèderait d'une logique de l'identité sociale ; mais il resterait, lui aussi, cohérent avec le principe herméneutique de la détermination du local par le global.

Les trois visions semblent cependant nécessaires pour la lecture d'une œuvre. Elles se complètent et indexent le voir et le comprendre sur le projet de regard dynamique et mouvant. Il y a, certes, beaucoup de détails dans une œuvre ; virtuellement, une infinité. Il y a, aussi, beaucoup des sociétés interpicturales pour une œuvre ; virtuellement, une infinité, également. Parfois, même, détail et interpicturalité se renvoient mutuellement. En effet, on comprend souvent un détail à l'aide d'une mise en perspective de l'œuvre dans une société particulière d'œuvres (par exemple, la présence de l'enfant déjà formé dans le ventre de Marie au moment même de l'Annonciation) ; on légitime souvent une société d'œuvres parfois par un détail (la représentation du thème de l'Annonciation à l'intérieur d'une scène de théâtre à l'italienne).

Les trois regards qui s'en suivent (le regard qui fixe à distance, le regard qui explore en s'approchant et le regard qui s'éloigne pour recontextualise) semblent tous nécessaires pour penser la modélisation d'un musée virtuel. Ils sont à la source des factures interprétatives dont ils constituent la cause efficiente. Ils ne désignent pas seulement des possibilités de réorientation ou d'affinement d'une visite mais permettent, par ailleurs, d'établir une grille de comparaison entre lectures.

Techniquement, ils exigent des qualités spécifiques du corpus des images de la collection : ils doivent concerner tous les points des vues. Dans le musée virtuel sur l'Annonciation, ils prennent la forme des choix techniques sur les reproductions des œuvres : les images sont de haute et de très haute résolution, précisément pour supporter autant que possible le mouvement vers l'avant ; ils sont complétés par des reproductions de bon nombre de détails choisis qui prennent des formes différentes suivant les visites. Les relations d'interpicturalité, de l'autre côté, débordent sur le genre pictural de la collection, et embrassent des formes inédites, comme les cycles narratifs d'appartenance (l'art byzantin est essentiellement narratif, ce qui donne aussi l'opportunité de superposer interpicturalité et intertextualité), le support de réalisation, la fonction liturgique, la place

dans un programme iconographique d'une église, etc. Les deux premiers genres de visite (Découverte et Étude) en font massivement usage.

Mais la dynamique interprétative exige aussi un régime de transition entre les regards ; autrement dit, le visiteur doit pouvoir passer d'un regard à un autre de manière naturelle. Certes, tout ne doit pas être possible ; mais le nécessaire, en matière d'interprétation, est souvent ce qui procède de la norme ; et ce nécessaire doit être toujours possible. C'est la raison pour laquelle les visites dans le musée sur l'Annonciation se recoupent souvent permettant des glissements d'une lecture à une autre.

5 Éléments d'implémentation et d'interaction Homme/Machine

Arrivés à ce niveau, il ne nous reste qu'à décrire la mise en scène. Ou la mise en œuvre, qui n'en est pas moins une seconde scénarisation, superposée, dérivant des préoccupations muséologiques.

Le problème de l'ergonomie est largement prioritaire. Les deux risques à évaluer et à minimiser à chaque phase sont l'extrême fragmentation (beaucoup d'actions demandées au visiteur pour recueillir l'information qu'il souhaite sur une œuvre, en matière de forme, de profondeur et d'extension) ou l'extrême intégration (tout gît devant lui, comme une classe d'objets surdéterminée). Sans doute, tout est ergonomique et l'absolument non ergonomique n'a jamais existé ; ou alors sous forme de farce. On sait, par ailleurs, la difficulté d'évaluer l'ergonomie ainsi que la réussite d'inventions notoirement non ergonomiques. Dans notre cas, il s'agissait d'un travail amont sur la facilité de l'appropriation du site. Le « clic » est, nous en convenons, tout puissant, mais pas toujours nécessaire et parfois doté de pouvoirs qui, par leurs conséquences, dépassent son objectif en brouillant les pistes de la visite projetée ; plus même, il construit des chemins de lecture qui ne correspondent pas à l'intention première ; ni à des normes de progression, d'ailleurs. De l'autre côté, une grande complexité de l'interface, qui demanderait une compétence informatique quelque peu élevée, serait un facteur d'exclusion ; alors que l'argument majeur des musées virtuels est précisément leur engagement en faveur de l'inclusion culturelle.

Ce sont là deux des raisons, principales, qui nous ont amenés à bâtir l'ensemble du musée avec un nombre restreint de composants, nécessitant peu de compétences informatiques et obéissant, globalement, à la même logique (trois au total : un de visualisation, et deux pour la navigation horizontale/verticale (réalisés comme « flowlist » et

« carrousel »)). L'objectif était au fond de disposer des outils intuitifs pour réaliser des parcours de graphe de manière transparente (le visiteur, en manipulant des objets 3D navigue de manière plutôt ludique dans une structure de données pourtant complexe, mais sans en avoir l'impression de se perdre).

Quant au reste, l'architecture du site est une architecture relativement classique « client/serveur ».

6 Conclusion

Notre vie chemine souvent à travers des collections. Autour de nous, des « choses », sensibles ou non, s'organisent en sociétés. Mais pas toutes seules ; ni par une règle imposée ou par quelque principe transcendant : seulement par nos propres pratiques. La pratique la plus profonde est celle qui s'efface parce qu'entièrement dissoute dans nos manières de vivre. Notre regard porte ainsi, secrètement, l'inscription d'une pensée de nos expériences de collections. Comment sont-elles ? Sûrement pas des expériences de logique relationnelle à la manière des bases des données. Plus encore que la collection, c'est notre visite qui raconte notre histoire. La lecture que nous faisons des pièces qui composent la collection, et mieux : notre façon de la lire comme un tout.

Le musée virtuel renouvelle cette vision avec enthousiasme. Mais quel musée virtuel ? Tous les musées virtuels ne se valent pas, on l'a vu, la notion de transmission culturelle ayant été spoliée par les catégories conceptuelles d'une logique relationnelle issue d'un paradigme informatique unique. La mise à disposition des collections sans égard aux pratiques de lecture, qui fait d'elles des objets identiques pour tout le monde, découpés des procédures d'interaction sémiotique normées, constitue sans doute une avancée dans l'immense productivité des bases de données ; mais sa contrepartie culturelle reste douteuse [4]. Apprend-on plus avec les musées virtuels ? Y a-t-il des utilisateurs qui consacrent du temps pour parcourir les musées virtuels typiques (i.e. les musées virtuels qui reproduisent des musées réels) dans un projet de découvrir des œuvres ? De les étudier ?

Reste la chance des musées thématiques. Du moins, s'ils ne dégradent pas, eux aussi, le concept de visite. Ils constituent, à notre sens, une incontestable offre culturelle. Et un authentique lieu de convergence des intelligences. On dirait, en s'appropriant les termes de Malraux [12], que le musée virtuel thématique se trouve dans la transcendance, alors que les outils actuels, qui visent diverses formes d'inventaire des fonds des musées, se situent dans l'immanence. Derrière cette différence, il y a

toute une différence dans le rôle qu'on accorde à l'interaction et à l'anthropocentrisme. La logique des bases de données en est seulement un pis-aller.

On remarquera, précisément, qu'il n'y a pas d'outils spécifiques pour la mise sur pied d'un musée virtuel. Les seuls produits qu'on trouve au marché proposent des versions améliorées de gestion et de présentation de bases de données multimédias ; on n'y trouvera pas les notions de scénario de visite, de profil de visiteur et de niveau d'exploration des collections, que l'on doit, éventuellement, construire a posteriori[7]. Ou alors, ils sont concernés par diverses formes de management des fonds[8] ; ils ne se soucient point de la création d'authentiques musées virtuels ; autrement dit, ils ne sont pas des outils pour configurer, maquetter et présenter en ligne, pour un public visé, une collection dans une optique muséographique et muséologique. D'ailleurs, pour le moment, il ne semble pas exister d'évaluation fiable de tels systèmes.

Peut-être une idée pour une direction en matière de développement. L'artisanat réclamera pour longtemps des outils qui préservent et pérennisent l'expertise acquise.

Remerciements : Nous remercions la Fondation Ormylia pour nous avoir permis la mise en place de ce musée, qui s'appuie sur un important travail mené en amont depuis plusieurs années. Nous remercions aussi Trinity Systems pour avoir consacré beaucoup de temps de développement hors contrat pour améliorer le site.

7 Références bibliographiques

[1] Arasse, D. : *Le Détail. Pour une histoire rapprochée de la peinture*. Flammarion, 2008.

[2] Bernier, R : Les musées sur Internet en quatre tableaux. *Archée Cybermensuel*, 2001 (4 articles, section Cyberculture) http://www.archee.qc.ca/index.htm

[3] *CIDOC* – ISO standard 21127:2006

[4] Deloche, B. : *Le musée virtuel.* PUF, 2001.

7 Comme, entre bien d'autres, le GallerySystem : www.gallerysystems.com/products/emuseum_FR.html. Voici comment la société néozelandaise Infospects (www.infospecs.co.nz/prdct-e-museum.html) présente son outil : "The E-Museum system is a TextWorks database designed for organisations whose main requirement is to catalogue and provide easy online search access to their collection." La préoccupation "Data Base", est prioritaire ; le musée virtuel sombre dans ses déclinaisons.

8 Comme le très médiatisé MuseumPlus (www.zetcom.com/fr/museumplus/introduction).

[5] Doering, Z. : Strangers, Guests, or Clients ? Visitor Experiences in Museums. *Curator*, 42(2), 1999, pp. 74-87.

[6] Hooper-Greenhill, E. : Museums and Education: Purpose, Pedagogy, Performance. Kindle Edition, 2009.

[7] Hooper-Greenhill, E. : Museums and the shaping of knowledge. Kindle Edition, 2007.

[8] Kanellos I., Le Bras Th., Kervella Ph., Guiziou E., Hatala A., Clochet S., Branelec O. : The Knossos project: Constructing Multi-point-of-view and Practice-adapted Interfaces for Indexing and Retrieving Information through Iconic Corpora, Proceedings of the ICTTA'04, Damas, Syrie, avril 2004, pp. 372-3 (article complet dans le CD-ROM de la conference).

[9] Kanellos I., Le Bras Th., Miras F., Suciu I. : Le concept de genre comme point de départ pour une modélisation sémantique du document électronique. Actes du Colloque International sur le Document Électronique (CIDE'05), Beyrouth, Liban, avril 2005, pp. 201-216.

[10] Le Bras, Th., Kanellos, I., Suciu I., s. Daniilia: The course as hermeneia: when interpretations leads the modeling of e-learning systems. *iPED, Researching academic futures*, Coventry (UK), 2007, pp. 45-55.

[11] Le Bras, Th : *Étude et mise en place d'un modèle générique de cours en EIAH. Typologie formelle des genres et des styles de cours. Moyens informatiques de réalisation.* Thèse de doctorat, Université Européenne de Bretagne, décembre 2008.

[12] Malraux, A. : *Le musée imaginaire*. Gallimard (Folio, Essais), 1996.

[13] Miras, F. : *Ergonomie de lecture et feuilletage électronique*. Thèse de doctorat, TELECOM Bretagne et Université de Bretagne Sud, janvier 2008.

[14] Pearce, S. : Interpreting Objects and Collections. Routledge (Leicester Readers in Museum Studies), 1994.

[15] Laneyrei-Dagen, N. : *Lire la Peinture. Dans l'intimité des Œuvres*. Larousse, 2002.

[16] Rastier, F. : *Sémantique Interprétative*. PUF, 2009 (3e édition).

[17] *Annunciation virtual museum :* www.annunciation.gr

[18] *British Museum* : www.britishmuseum.org

[19] *Brooklyn Museum* : www.brooklynmuseum.org

[20] *Powerhouse Museum :* www.powerhousemuseum.com

[21] *Rijksmuseum :* www.rijksmuseum.nl

[22] *Tate Museum :* www.tate.org.ok

[23] *Museo del Prado :* www.museodelprado.es

[24] *Musée du Louvre : www.louvre.fr*

[25] *Musée de Pompei :* www.pompeiisites.org

[26] *Métropolitan Museum :* www.metmuseum.org

[27] *Musée de l'Hermitage :* www.hermitagemuseum.org

[28] *Eternal Egypt* (www.eternalegypt.com)

Manuscrits de Stendhal : Du patrimoine papier au document électronique

Auriane FAURE(1), Thomas LEBARBÉ(1), Cécile MEYNARD(2), Aïcha TOUATI(1)
(*1) Laboratoire LIDILEM (EA 609)*
(2) Equipe « TRAVERSES 19-21 » (EA 3748)
Université Stendhal – Grenoble 3

1 Du manuscrit à la plateforme en ligne

La Ville de Grenoble possède la quasi totalité des manuscrits laissés par Stendhal à sa mort, soit environ 20 000 feuillets. Cet ensemble constitue l'un des plus importants fonds de manuscrits littéraires modernes en France, et à ce titre représente un élément précieux du patrimoine culturel et scientifique.
L'Etat français et les collectivités régionales et locales ont investi des sommes importantes depuis le début du XXème siècle pour acquérir les différents documents du fonds : Pour ne prendre qu'un exemple récent, 6 cahiers des journaux de Stendhal dit « cahiers Bérès » ont été achetés pour 900 000 euros en décembre 2006. Il a d'ailleurs aussi fallu faire appel à des mécènes privés pour réunir une telle somme : se pose alors le problème légitime de la mise à disposition du public de ces manuscrits. Il n'est évidemment pas imaginable de laisser tout un chacun consulter les manuscrits, précieux et fragiles, qui ne doivent faire l'objet que de consultations ponctuelles et justifiées afin d'être préservés pour les générations futures.
La 1ère solution consiste à se contenter d'une simple numérisation des pages, les images étant ensuite mises en ligne. La Bibliothèque municipale de Grenoble, qui a numérisé récemment le fonds des manuscrits de Stendhal (entre 2007 et 2009) a ainsi donné au public la possibilité de feuilleter les pages des cahiers Bérès[1]. Mais cette mise à disposition s'avère d'un intérêt limité, car le manuscrit se réduit à n'être qu'un bel objet pour l'utilisateur curieux qui doit se contenter de tourner

1 http://www.bm-grenoble.fr/patrimoine/acces-aux-collections-numerisees.htm. La bibliothèque a également mis en ligne une reproduction de la belle édition diplomatique de Gérald et Yvonne Rannaud chez Klincksieck, qui offre à l'utilisateur la possibilité de consulter les images des pages et leurs transcriptions, et de faire des recherches simples dans le texte. Mais aucun enrichissement scientifique n'a été ajouté et la liberté de parcours de l'utilisateur reste limitée.

des pages sans pouvoir toujours lire les pattes de mouche de l'écriture stendhalienne (comme le note Almuth Gresillon, l'utilisateur se trouve ici dans une situation d'esthète, à « regarder le manuscrit comme on regarde un tableau »[2]) ni forcément comprendre la logique de l'organisation interne des documents ou de leur rattachement à des ensembles plus vastes. Aucune recherche n'est par ailleurs possible dans les textes. L'intérêt scientifique, et même culturel, de la consultation reste donc limité.

C'est dans ce souci de valorisation du fonds (le rendre lisible et non plus simplement visible) que, en partenariat avec la Bibliothèque municipale de Grenoble, des chercheurs de l'Université Stendhal – Grenoble 3 se sont lancés dans la conception d'un site commun Ville / Université, couplé avec une base documentaire, CLELIA (Corpus littéraire et linguistique assisté par des outils d'intelligence artificielle). Le projet « Manuscrits de Stendhal » s'appuie sur une collaboration fructueuse et inédite entre des littéraires de l'équipe Traverses 19-21 et des informaticiens et linguistes du laboratoire LIDILEM de l'Université Stendhal Grenoble 3.

Le principe est de donner à voir les pages numérisées des manuscrits de Stendhal, mais aussi leur transcription et différentes informations sur leur contenu textuel par le biais d'un moteur de recherche, en fournissant des modes d'accès et de représentation variés aux utilisateurs. La plateforme CLELIA a été en effet conçue en visant le plus large public possible, du « grand public » aux spécialistes de Stendhal ou de la littérature du XIXème siècle.

Les premiers utilisateurs de la base sont nécessairement les transcripteurs littéraires qui vont l'alimenter progressivement. Ils doivent pouvoir saisir toutes les informations qui leur semblent pertinentes pour l'analyse des pages, et c'est pour cette raison que l'outil doit être adaptable et évolutif. Plus généralement, tous les chercheurs stendhaliens et spécialistes du XIXème siècle sont des utilisateurs potentiels de la base. Pour ne prendre qu' un exemple, un chercheur travaillant sur le rôle du souligné et des traits en marge au crayon chez Stendhal (c'est le cas de Christopher Thompson, qui a montré combien cette pratique est intéressante et révélatrice sur la réutilisation par Stendhal d'extraits de ses textes dans d'autres textes, transcendant ainsi les genres littéraires traditionnels) doit pouvoir trouver ces informations dans la base.

Au niveau microscopique, celui de la page, l'outil doit permettre de reconstituer autant que faire se peut la genèse de la page en identifiant les traces et les strates d'écriture, les ratures, variantes, soulignés, traits en

2 Almuth Grésillon, « Méthodes de lecture », Les manuscrits des écrivains, Paris, Hachette CNRS éditions, sous la direction de Louis Hay, 1993, p. 138-161 (la citation se trouve page 143).

marge, interlignes, ajouts, notes... autant d'éléments qui peuvent apporter des informations essentielles aux chercheurs sur le travail d'écriture et d'auto-relecture de Stendhal[3].

Au niveau macroscopique, celui des ensembles de pages, il s'agit d'identifier et de représenter de façon rigoureuse les documents et ensembles documentaires qui ont souvent été déplacés, par Stendhal ou par les bibliothécaires au moment de la reliure des manuscrits, voire qui ont été « désossés » lors des éditions, comme s'ils appartenaient à des corpus différents, et sans tenir compte de l'unité du support. Pour ce faire, le travail sur le fonds a amené les chercheurs à rationaliser l'analyse codicologique[4] par un inventaire systématique dont les informations (dimensions des papiers, trous de couture permettant d'identifier des cahiers, des liasses, etc.) seront introduites dans la base pour permettre des regroupements de documents présentant les mêmes caractéristiques. Toutes ces informations sont essentielles pour envisager un reclassement virtuel de ces ensembles désorganisés. Des analyses littéraires sont ainsi rendues possibles par les requêtes effectuées, qui viennent infirmer ou confirmer de façon rationnelle les intuitions des chercheurs, ou peuvent même les amener à formuler de nouvelles hypothèses.

La deuxième fonction de l'outil mis en place pour ce public que constituent les transcripteurs est de permettre de produire des éditions papier à la demande, en s'appuyant réellement sur les manuscrits.

L'équipe littéraire a des exigences en termes de contenu et de mise en forme. Il est important en effet autant que possible d'être fidèle à la mise en page stendhalienne, qui a le plus souvent une signification. Ainsi il convient de conserver le statut et la présentation des titres, qui donnent souvent une dimension solennelle à un début de cahier. Autre exemple, les notes de bas de page et les marginales n'ont en général pas la même fonction, les premières contiennent les références bibliographiques et les données chiffrées, tandis que les secondes peuvent être soit un commentaire du texte en regard duquel elles se trouvent, soit une simple notation diariste de l'état physique et mental de Stendhal et de ses activités et observations du moment.

Enfin, il s'est avéré pertinent de signaler les réclames et contre-réclames[5] pour respecter la mise en page voulue par Stendhal mais aussi pour préparer le travail sur l'édition papier. En effet, il faudra procéder à la

3 Stendhal étant coutumier de l'annotation a posteriori de ses propres écrits, nous avons ainsi créé un corpus « Notes de relecture de l'année XXXX » pour permettre des regroupements virtuels et identifier ainsi de façon rigoureuse ses centres d'intérêt selon les époques de sa vie, ses périodes de relecture active, etc.

4 L'analyse codicologique est l'analyse des caractéristiques du papier du document, voire du feuillet (identification de traces de couture, de piqûres d'épingle, etc.)

5 La réclame est un mot ou groupe de mots que le scripteur écrit sur la dernière ligne de la page en l' (les) alignant à droite, et qu'il répète éventuellement au début de la première ligne de la page suivante (contre-réclame). Cette pratique était courante dans les manuscrits et œuvres publiées jusqu'au début du XIXème siècle.

suppression automatique de tous les mots désignés comme contre-réclames pour éviter la répétition de ces mots).

Le **deuxième type d'utilisateurs regroupe les linguistes** pour lesquels les manuscrits constituent un corpus inédit. En effet, les manuscrits de Stendhal représentent 40 années d'écriture. Sous leur forme papier, les manuscrits sont difficilement utilisables pour le chercheur en linguistique. Il en est de même pour les numérisations qui présentent le défaut de lisibilité déjà évoqué plus haut.
En revanche, transcrits et annotés rigoureusement, les manuscrits représentent un matériau langagier unique : des milliers de pages d'écriture, appartenant à différent styles (diariste, ébauches et critiques littéraires et théâtrales…), dont tous les composants sont délimités et identifiés, quelque soit le grain hiérarchique (du bloc de texte au mot biffé en passant par lignes, paragraphes, ajouts en marge…).
Ainsi structuré, l'ensemble des pages de manuscrits forme non seulement un ensemble de corpus dans le sens littéraire du terme, mais aussi un corpus au sens linguistique du terme. Parmi les objets d'études linguistiques, nous envisageons notamment la caractérisation du sabir[6], de la dysgraphie (ou paragraphie[7]), la description linguistique des phénomènes de réécriture assimilables à des formes de disfluences écrites…

Les professeurs de lycées et leurs élèves constituent le troisième type d'utilisateurs visés, sachant que l'étude de la genèse des œuvres littéraires est au programme de français en seconde, afin de permettre aux élèves de mieux comprendre le processus de création chez les écrivains. Des parcours pédagogiques simples d'accès et d'utilisation, éventuellement téléchargeables, doivent donc être prévus, en gardant à l'esprit qu'il s'agit de distinguer les manuscrits de leur image souvent un peu poussiéreuse et ennuyeuse, en montrant aux jeunes générations à quel point l'analyse de la genèse d'une œuvre peut prendre des dimension inattendues d'enquête à partir d'indices. On mettra ainsi à disposition des enseignants des dossiers portant par exemple sur les ensembles significatifs de pages illustrant la démarche de création des personnages de romans chez Stendhal, sur les pratiques d'écriture autobiographique, ou sur le plagiat, en laissant bien sûr à l'enseignant, voire à ses élèves, la possibilité de se constituer des dossiers personnalisés.

6 Le sabir est une pratique récurrente chez Stendhal qui intègre dans ses écrits des séquences en langue étrangère (Bordas, 2007), cette pratique variant dans le temps, atteignant son paroxysme à la fin de sa vie dans « Earline ».

7 Les termes de dysgraphie et de paragraphie sont tous deux utilisés pour désigner les variantes orthographiques par rapport à la norme. Seule l'étude approfondie de ces phénomènes nous permettra de déterminer lequel des deux termes est le plus adéquat.

Enfin, dans ce souci légitime d'exhaustivité et de rigueur scientifique, il ne faut pas, oublier un public essentiel : **les amateurs éclairés et les simples curieux**. De fait, la valorisation du patrimoine culturel et scientifique doit se faire en ayant en tête un souci de vulgarisation pour que la diffusion de ces informations culturelles ne concerne pas qu'une élite. D'où la nécessité de mettre en place des parcours guidés ludiques et interactifs dans le fonds des manuscrits. Le principe sera ainsi, entre autres exemples, de donner à voir et à entendre par le biais d'hyperliens (faire voir la reproduction d'un tableau dont parle Stendhal, ou faire entendre un extrait d'un morceau de musique évoqué), de faire écouter un commentaire oral sur une page de manuscrit, de permettre un affichage dynamique de la page en cours d'écriture ou de correction. Il s'agit donc de montrer un auteur vivant à travers son œuvre.
Les recueils physiques sont reproduits numériquement, mais l'utilisateur peut aussi accéder aux manuscrits par le biais de parcours guidés, de regroupements par types d'écrits mais aussi par recherche simple ou avancée. La quantité et l'affinement des annotations sont adaptés au type d'utilisateur pour ne pas surcharger l'affichage inutilement. La plateforme permet par ailleurs aux utilisateurs de constituer leurs propres recueils de pages manuscrites. A terme, cette fonctionnalité permettra des éditions numériques et papier à la demande.

2 Une description « sémantique » des pages manuscrites

La page de manuscrit est un objet complexe à décrire. Y intégrer des informations paratextuelles, d'ordre scientifique ou didactique, augmente et complexifie la tâche de description.
Il existe de nombreuses normes de description et d'encodage de textes, parfois contradictoires ou antagonistes, représentant des points de vues différents sur l'objet textuel de manière général. La TEI (*Text Encoding Initiative*) joue un rôle majeur dans cet univers depuis plus de deux décennies[8], plus encore depuis la plus récente création de *guidelines* pour la transcription des manuscrits[9].
Toutefois, dans le cadre de ce projet, nous devions répondre à trois impératifs : 1) un impératif scientifique de description précise, 2) un impératif économique dû aux faibles ressources humaines pour le développement logiciel et 3) un impératif d'accessibilité à des utilisateurs peu formés aux outils et formalismes informatiques. Afin de répondre à ces trois exigences, nous avons opté pour une grammaire de description

8 Text Encoding Initiative – History : http://www.tei-c.org/About/history.xml

9 Guideline for « Manuscript description » : http://www.tei-c.org/release/doc/tei-p5-doc/html/MS.html

(DTD) conçue dans un dialogue interdisciplinaire qui permette de nommer les objets et leurs propriétés dans une terminologie accessible aux transcripteurs. Cette grammaire est accompagnée d'une feuille de style permettant une visualisation approximative de la transcription dans un logiciel d'édition de fichiers XML libre de droit. Ainsi, les transcripteurs sont guidés et contraints dans leur tâche par la DTD sur un outil ne nécessitant que peu d'apprentissage, et disposent d'un rendu visuel s'approchant du rendu envisagé en ligne.
A l'image de nombreux formalismes de description, la grammaire développée pour les manuscrits de Stendhal se décompose en deux parties principales : une en-tête de méta-données permettant pour le référencement et les renseignements sur la page incluant un commentaire du transcripteur ; et un corps contenant la transcription elle-même. L'équipe ne souhaitait pas une transcription hyper-diplomatique, par conséquent, la page est décomposée en 9 cadrans pour le positionnement des unités textuelles et graphiques : la zone centrale principale, les quatre coins de la page, les quatre cotés (marges lattérales, supérieure et inférieure – voir figure 1 page suivante). Chacune de ces zones peut contenir des éléments textuels (blocs de texte, blocs de citation, paginations, foliotations, marginales, notes…) qui se décomposent en entités identifiables visuellement (paragraphes, lignes, interlignes, figures, tableaux…) et en entités de mises en forme (biffe, calligraphie…). Tous les éléments peuvent être enrichis d'annotations d'ordre critique (commentaires pour le grand public, pour les spécialistes, pour les membres de l'équipe, identification du scripteur, datation, localisation géographique…). Enfin, l'ensemble est complété d'un système de pointage et de références. En effet, il arrive qu'un ajout soit effectué en interligne au dessus de son point d'ancrage, puis s'enchaîne faute de place en dessous de ce point d'ancrage pour se terminer en marge. La description des éléments textuels est faite dans une représentation pseudo-diplomatique (à l'image de la page), la représentation linéarisée (à l'image de la résultante de la tâche scripturale) est calculée notamment grâce à ce système de pointeurs.

Fondée sur XML, la grammaire de description en hérite les qualités et les défauts. Ces derniers sont reconnus, notamment la contrainte d'imbrication des balises. Certes, des solutions sont proposées, telle celle proposée par (Portier, 2009) qui dissocie le contenu de son annotation, ou LMNL (Caton, 2005) qui permet l'enchevêtrement de balises ouvrantes et fermantes. Ces méthodes et techniques présentent néanmoins le défaut d'être peu outillées et peu intuitives pour l'utilisateur peu formé en informatique. Par ailleurs, les plateformes de partages de données textuelles, telles Pinakes 3 (Scotti, 2006), se fonde sur un encodage XML voire sur la TEI. C'est pourquoi nous avons planifié au sein du projet le

développement de modules de conversion des données vers la TEI. Cette prospective a certes influencé les principes d'encodages tout en laissant à l'équipe une grande liberté sur ses choix méthodologiques.

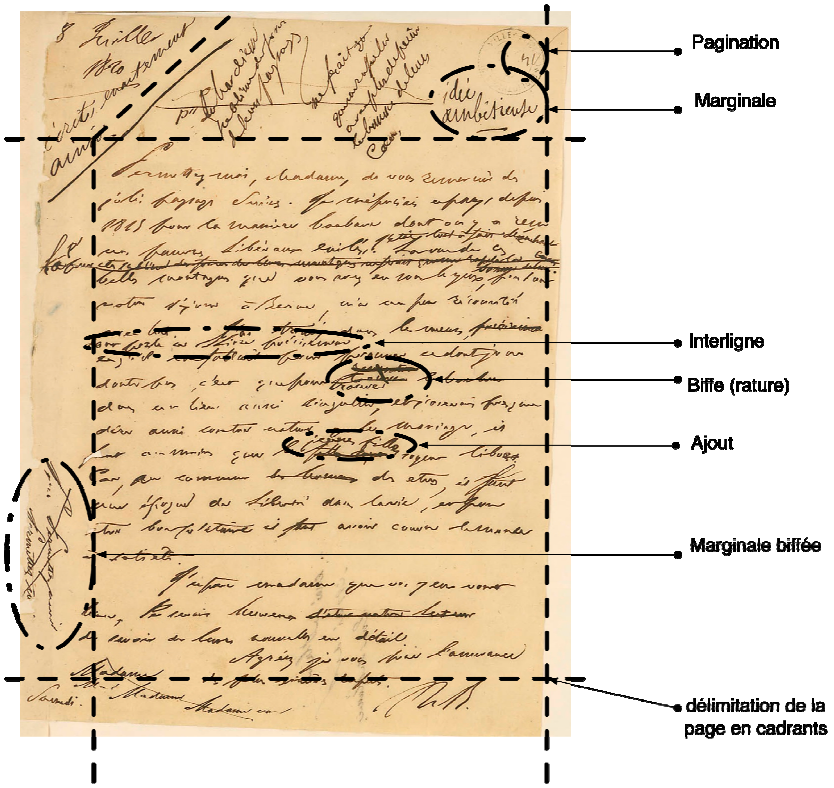


Figure 1 : délimitation de la page de manuscrit en cadrants et dénomination des éléments (R. 5896, volume 1, feuillet 71 recto, image propriété de la BmG).

Les transcripteurs disposent d'un outil hors-ligne pour effectuer les transcriptions. La plateforme CLELIA leur permet de déposer les transcriptions XML et de les visualiser telles qu'elles apparaîtront aux différents types d'utilisateurs et ainsi de corriger les transcriptions le cas échéant (la mise en ligne est assistée par une analyse du fichier mettant en évidence les erreurs et incohérences des données). Un processus de relecture puis de validation par les pairs, accompagné d'un code déontologique, permet de garantir la qualité scientifique des transcriptions mises à la disposition du public.
Du point de vue de l'utilisateur, l'accès aux manuscrits se fait selon trois méthodes différentes :

- par les registres physiques (à l'image des recueils conservés à la BmG) ou par regroupement cohérents d'un point de vue littéraire (corpus) ;
- par des regroupement artificiels générés automatiquement (ex : « les pages contenant des dessins de la main de Stendhal ») ou conçus par les spécialistes (ex : « les plus belles pages ») ou par des utilisateurs ;
- par recherche de mots-clés en plein texte.

La consultation des manuscrits correspondant au choix de l'utilisateur peut alors se faire selon trois modes d'affichage, l'utilisateur ayant toute liberté de basculer d'un mode à l'autre :

- par « planche contact » des pages numérisées, à l'image de leurs homonymes photographiques, permettant d'identifier rapidement la ou les pages pertinentes ;
- par « feuilletage », tel un livre dont on tourne les pages ;
- par vis-à-vis de la page et de sa transcription donnant ainsi une aide à la lecture et à l'analyse (par le biais d'infobulles).

La mise à disposition des manuscrits et de leurs transcriptions sur Internet n'est toutefois qu'un des aspects (certes majeurs) du projet et de la plateforme. Les transcriptions ainsi formalisées et enrichies constituent une donnée structurée qui permet de se défaire de la page en tant qu'objet physique,

2.1 De la transcription à l'ontologie des documents et des usages

L'objet transcrit et affiché est la page. Il est décrit non seulement par des propriétés physiques telles que le format, le type de page, le scripteur, la date de rédaction, etc., mais aussi par son contenu. Le contenu de la page est l'ensemble des éléments textuels la constituant (pagination, foliotation, marginales, note de bas de page, ajout en interligne, paragraphe, titre, etc.). Chaque bloc de texte dispose d'un ensemble de propriétés (corpus, scripteur, emplacement dans la page, type d'écriture, etc.) où chaque élément textuel peut contenir à son tour d'autres types de blocs de texte qui héritent de ses propriétés de façon implicite ; ces blocs peuvent également disposer de propriétés différentes, spécifiées explicitement.

Une page peut être considérée comme un objet dont les éléments sont organisés hiérarchiquement. L'observation de ces éléments peut se faire aux différents niveaux de la hiérarchie, correspondant à autant de niveaux de granularité. L'objectif de CLELIA et de permettre aux utilisateurs d'interroger l'ensemble documentaire des manuscrits afin de reconstruire virtuellement des objets textuels à tous les niveaux de granularité. Au niveau macroscopique, l'outil permet de remettre de l'ordre des pages, par exemple de reconstruire le « *Journal de mon 3ème voyage à Paris,*

1804-1805 », ensemble de cahiers aujourd'hui physiquement en désordre. Au niveau microscopique, d'une part les littéraires pourraient, par exemple consulter *« les pages de manuscrits de type diariste, contenant des figures, de la main de Stendhal, figure représentant des plans, classées selon les lieux décrits, puis par date de rédaction»***.** D'autre part, les linguistes chercheront un corpus correspondant à des critères tels que *« paragraphes comportant des biffes et des ajouts, classés par type d'écrits, puis par ordre chronologique »* pour étudier le phénomène de réécriture.

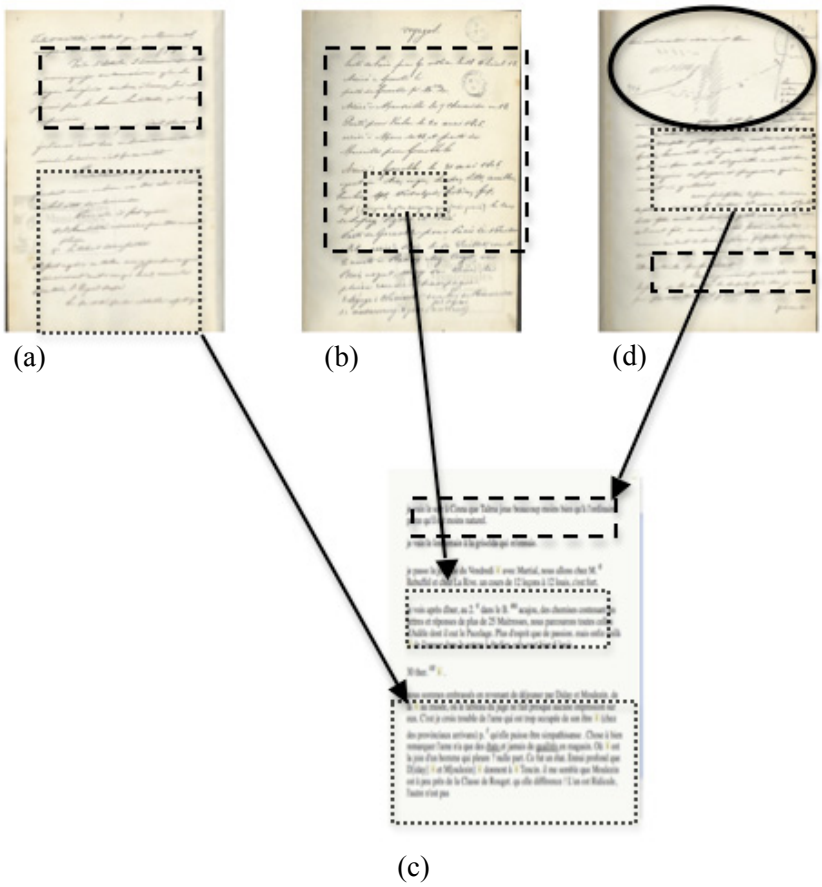


Figure 2. Exemple construit de restructuration dynamique

Comme nous l'avons cité plus haut**,** le formalise XML conçu pour les transcriptions permet de décrire fonctionnellement et sémantiquement les éléments textuels de la page manuscrite et d'en donner une représentation

hiérarchique. Pour répondre aux diverses requêtes, l'ensemble des éléments textuels de l'ensemble des pages est indexé, leur contenu et leurs propriétés sont ainsi enregistrées en conservant l'organisation hiérarchique. Cette indexation nous permet par la suite comme l'illustre la figure ci-dessous (Figure 1), de décomposer les éléments textuels constituants la page, pour nous donner une plus grande flexibilité de manipulation de ces objets (éléments textuels), les objets, quels qu'ils soient – de la page au mot -, peuvent être sélectionnés, en fonction de leurs propriétés, puis classé selon d'autres. Dans la figure 2, le document résultant (c) de la composition de fragments textuels présents dans trois pages manuscrites (a, b et c), est fait dynamiquement grâce aux propriétés qui caractérisent ces éléments et qui correspondent aux critères de sélection données par l'utilisateur.

La restructuration dynamique de corpus nous conduit à nous interroger sur l'accès au contenu des Manuscrits, et sur la navigation au sein des bases documentaires. Les publics de CLELIA sont nombreux, et ont des pratiques documentaires variées, que ce soit entre individus ou entre groupes. Pour permettre à chaque utilisateur de parcourir l'ensemble documentaire selon ses besoins propres, nous avons intégré à CLELIA un système de recherche d'information (SRI), couplé à une aide contextualisée. Comme l'expriment (Labiche & Holzeim, 2009) : « Nous avons estimé nécessaire de réfléchir à la conception de systèmes pour lesquels les interactions avec ces utilisateurs-concepteurs sont essentiels ».
Devant la masse de données et les différences entre utilisateurs, il nous semble important d'aider l'utilisateur de façon personnalisée. Les conditions d'interprétation d'une information sont différentes pour chacun ; La construction du sens est un acte dynamique et individuel. Afin de rendre la navigation pertinente pour chaque utilisateur, il nous apparaît nécessaire de nous inscrire dans une démarche centrée utilisateurs.
Le système de recherche d'information greffé à CLELIA fonctionne de manière simple. Il comprend une interface d'interrogation, par laquelle l'utilisateur exprime son besoin, ainsi qu'une interface de réponse, regroupant les résultats que l'utilisateur doit interpréter. L'articulation de ces deux étapes consiste en la mise en relation de l'expression du besoin de l'utilisateur et des données contenues dans la base documentaire, modélisées au préalable, pour sélection et affichage.
Dans l'interface d'interrogation, l'utilisateur exprime son besoin par une requête composée de listes de termes (une ou plusieurs, composées d'au moins un terme chacune). Par observation empirique[10], nous pouvons

10 Expérience menée sur quatre utilisatrices dans le cadre de travaux de Master 2 : (Faure, 2008)

affirmer que les utilisateurs construisent une liste comme un ensemble lexical définissant un « thème » de recherche (soit un angle de lecture). La construction de plusieurs listes vise ainsi à caractériser ou opposer plusieurs thèmes. Nous laissons ainsi l'utilisateur couvrir des champs lexicaux les plus larges possibles.
La sélection des documents du fonds correspondant à la recherche est effectuée par l'intermédiaire d'un index, qui contient l'ensemble du vocabulaire des Manuscrits. Lorsqu'un terme contenu dans une requête d'utilisateur est indexé, la mise en relation avec le(s) document(s) dont le terme est issu est opéré. Les documents modélisés le sont donc par l'intégralité de leur contenu, et non par une liste de mots-clés issus d'une sélection (humaine ou automatique). Il n'y a donc pas d'étape intermédiaire entre le document et la formulation du besoin de l'utilisateur ; l'absence d'interprétation donnée par une tierce personne va dans le sens d'un SRI centré utilisateur.
Le résultat est ensuite affiché sous la forme d'une cartographie, représentation graphique de l'ensemble documentaire sur laquelle est projetée la requête de l'utilisateur. Cette cartographie est un support de navigation plus éloquent qu'une liste de liens vers les documents sélectionnés. Elle constitue un point d'entrée dans l'ensemble documentaire ainsi que dans le document lui-même. Elle constitue également un premier support d'interprétation pour l'utilisateur, qui peut observer les différences ou ressemblances entre documents au regard de sa recherche. Il peut décider de consulter un document, ou de réajuster sa recherche. La cartographie est donc le point de départ de la navigation dans l'espace documentaire.

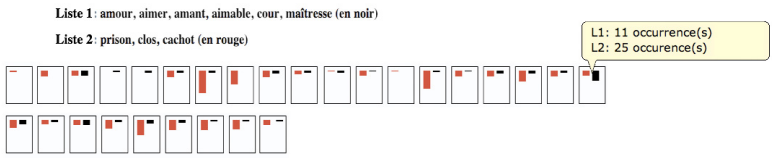


Figure 3 : Cartographie selon deux listes de termes (quantifiées)

Afin d'apporter une aide contextualisée à l'utilisateur construisant son parcours documentaire, c'est-à-dire centrée sur ses besoins et ses pratiques, nous proposons la construction d'une ressource terminologique qui lui soit propre. Cette ressource lui est proposée quand il construit sa requête de recherche. Lorsqu'un terme ajouté à la liste en construction existe dans la ressource terminologique, ses termes frères[11] sont proposés à l'utilisateur. Dans la perspective du sens construit dans l'interprétation,

11 Par termes frères au sein de la RTO, nous entendons les termes coprésents avec le terme de la requête courante

il semblerait incohérent de proposer une seule et même ressource lexicale à tous les utilisateurs, tel un dictionnaire des synonymes, ou un panel « à l'aveugle » des dix termes les plus employés dans les Manuscrits ou dans les requêtes d'utilisateurs. Nous souhaitons proposer des ressources légères individuelles, qui soient pertinentes en contexte et de taille raisonnable[12].

Dans (Roy & Beust, 2006), l'utilisateur exprime ses connaissances sur un domaine en organisant des lexies par regroupement ou par opposition. Nous choisissons pour notre part de construire la ressource de manière automatique, en l'implémentant à chaque nouvelle requête d'utilisateur. Les traces d'interaction entre l'utilisateur et le SRI viennent ainsi compléter la ressource terminologique de l'usager, en tenant compte des termes existant au préalable et des liens qu'ils entretiennent[13].

La ressource construite traduit les centres d'intérêt et les pratiques de l'utilisateur. En effet, les termes régulièrement employés seront liés à un vocabulaire large, tandis que les termes n'ayant pas satisfait la recherche de l'utilisateur tomberont en désuétude. Par ailleurs, la navigation inter et intra-textuelle que mène l'utilisateur lui permet de découvrir de nouveaux termes appartenant au vocabulaire de Stendhal mais pas forcément au sien. L'intégration de ces termes dans les requêtes de recherche suivantes conduit du même coup à leur intégration dans la ressource terminologique.

De nombreux travaux se sont penchés sur l'extraction automatique de termes pour construire des ressources terminologiques, mettant au jour un vocabulaire propre aux auteurs des documents de référence. Cette perspective ne correspond pas à notre démarche centrée utilisateur. Nous privilégions une ressource issue des seules pratiques de l'utilisateur, contenant exclusivement des termes qu'il choisit d'employer en conscience. Une telle ressource modélise donc à la fois ses centres d'intérêts et ses usages, face à l'outil et à la navigation dans l'ensemble documentaire : il s'agit de modéliser son parcours interprétatif. Il devient ainsi possible d'apporter une aide plus concrète à l'utilisateur dans sa recherche, par l'accès à cette ressource durant le parcours. Ces ressources sont également mutualisées, pour générer une ressource globale afin d'émettre de nouvelles pistes interprétatives. La comparaison de traces d'utilisateurs différents devrait permettre d'esquisser des petits groupes d'utilisateurs dont les processus cognitifs et les usages sont proches, dans un objectif de modélisation.

12 Plusieurs travaux sur la création de ressources terminologiques personnelles ont inspiré cette proposition : citons (Perlerin, 2004) ou encore (Roy, 2007).

13 Pour une étude des liens entre termes d'une ressource terminologique personnelle, consulter (Faure & Lebarbé, 2009)

3 Conclusion

Les méthodes et outils que nous présentons ont été développés autour du projet des manuscrits de Stendhal mais dans une perspective plus généraliste afin d'être adaptés en tant que plateforme ou en tant qu'outils indépendants à d'autres formes de ressources textuelles. Il est d'ores et déjà prévu d'étendre le champ d'application méthodologique et informatique aux Carnets des Canuts, propriété de la Ville de Lyon, autre patrimoine de l'histoire de la France. Conçus dans un dialogue interdisciplinaire permanent, ils offrent aux spécialistes comme au grand public de nouvelles approches documentaires du patrimoine culturel écrit ainsi enrichi, permettant de concevoir simultanément et complémentairement des éditions numériques en ligne et des éditions papier.
Valoriser le patrimoine apparaît ainsi comme le moyen de construire également des supports de réflexion méthodologiques pluridisciplinaires.

4 Références bibliographiques

J.-L. Bouraoui, Ph. Boissière, M. Mojahid, N. Vigouroux, A. Lagarrigue, F. Vella, J.-L. Nespoulous. Problématique d'analyse et de modélisation des erreurs en production écrite. Approche interdisciplinaire. Actes de TALN 2009.

B. Boie. L'écrivain et ses manuscrits, Les manuscrits des écrivains, Paris, Hachette CNRS éditions, sous la direction de Louis Hay, 1993, p. 34-53.

E. Bordas. Le Babeylisme scriptural de Stendhal, ou le style comme langue étrangère, Stendhal à Cosmopolis, ELLUG, sous la direction de Marie-Rose Corredor, 2007, p. 219-231.

P. Caton. « LMNL Matters », Extreme Markup Languages 2005, Montréal, Quebec

A. Faure. Cartographies et OntologieS des Manuscrits et Œuvres de Stendhal (COSMOS). Mémoire de Master 2, Université Stendhal Grenoble 3, 2008.

A. Grésillon. Méthodes de lecture , Les manuscrits des écrivains, Paris, Hachette CNRS éditions, sous la direction de Louis Hay, 1993, p. 138-161.

J. Labiche. M. Holzem. Couplage et perturbation versus boîte noire et entrée-sortie. Journées de Rochebrune 2009, 2009.

V. Perlerin. Sémantique légère pour le document. Thèse de doctorat en Informatique, Université de Caen / Basse-Normandie, 2004.

P.-E. Portier, S. Calabretto. Modélisation des connaissances dans le cadre de bibliothèques numériques spécialisées. Extraction et Gestion des Connaissances (EGC) 2009, Strasbourg.

F. Rastier. Sémantique interprétative, Presses Universitaires de Frances, 1987.

T. Roy et P. Beust. Ressources termino-ontologiques différentielles personnelles : construction et projection sur corpus. Revue I3E, Hors-série, 2006.

T. Roy. Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires. Thèse de Doctorat, Université de Caen/Basse-Normandie, 2007.

A. Scotti. Postgres & Java in the cultural heritage research: the Pinakes 3.0 Project. In WorldWide PG Day, Prato, Juin 2006.

Diversité de l'Information dans les Sites de Presse

Cyril Laitang, Elöd Egyed-Zsigmond, Sylvie Calabretto

Université de Lyon
LIRIS, INSA de Lyon

Mots-clés : catégorisation, web, sémantique, presse, flux RSS, thématique, ontologie

Keywords: RSS feeds, web, semantic, categorization, press, ontology

Résumé : La multiplication des acteurs de communication sur internet devrait logiquement entrainer une plus grande diversité dans les sources d'information disponibles. Toutefois de plus en plus d'études récentes mettent en doute cette hypothèse. Le projet IPRI a été créé afin de faire le point sur ces croyances et déterminer les schémas de propagation des actualités sur ce media. Notre principal objectif est de fournir une aide à l'analyse des informations de presse sur des évènements d'actualités. Nous assistons la catégorisation en fournissant un outil de classification dynamique des thématiques et des sujets extraits de flux RSS au moyen de ressources du Web sémantique. Notre solution lie la catégorisation sur des domaines non spécialisés à des sources et concepts sémantiques tel que des ontologies et des thésaurus.

Abstract: As internet communication actors grow in number it is widely admitted that informational offer diversity over this media should increase as well. However we recently came over more and more studies that seem to go against that assumption. The IPRI project was created in order to asset these believes and to figure out news events spreading schemes over the web. Our main objective is to help press information analysis over current events. We assist categorization by providing dynamic thematic and subject sorting tool over RSS feeds by the mean of semantic web resources. Our solution links information categorization over unspecialized data sources and various fields with semantic concepts like thesaurus and ontologies.

1 Introduction

1.1 Problématique

Considéré comme le socle de la démocratie, le pluralisme de l'information est l'objet d'une régulation dans les médias écrits et audiovisuels. Sur Internet en revanche il est permis de penser que la multiplicité des sources accessibles garantit naturellement cette diversité de l'information.
Qu'en est-il réellement? C'est pour répondre à cette problématique qu'a été créé le projet IPRI [1]. En partenariat avec des sociologues et des journalistes nous cherchons à aider la catégorisation des articles de presse depuis un échantillon représentatif des grandes familles de type de publication. Ceci afin de déterminer le niveau de répétitivité et le schéma de propagation de l'information.
Dans ce papier nous décrirons comment l'ajout et l'utilisation d'informations sémantiques peut fournir de l'aide à la fois à la catégorisation thématique et au rapprochement des sujets traités sur un intervalle temporel défini.

1.2 Objectifs

Notre problématique tente de répondre à deux besoins de catégorisation dont le sens propre au journalisme mérite une redéfinition :
La thématique, statique elle concerne le domaine dans lequel se classe l'article. Sport, politique, etc. Elle est nécessaire à toute catégorisation ultérieure dans la mesure où elle donne des informations utiles au regroupement par sujet. Il est à noter également que malgré sa normalisation par de grands organes de presse elle n'est peu ou pas appliquée à la publication de la grande majorité de nos sources (blogs, agrégateurs). Exemple: thématique sport, politique, etc.
Le sujet, ce dernier apparaît, évolue et disparaît au cours du temps. Cela implique deux problématiques : le choix de la distance temporelle entre deux analyses et son regroupement. Exemples : élection américaine, tremblement de terre, nouvelles lois.
Les deux types de catégorisations mentionnées ci-dessus sont, à l'heure actuelle, effectuées manuellement : par une annotation de l'article à sa publication pour la thématique; et par une analyse subjective de journalistes, sociologues, ou analystes de presse pour le regroupement par sujet. L'exemple de l'étude sur les sujets de presse [7] illustre les besoins existants dans ces domaines. En effet, cette analyse a demandé trois jours

[1] Internet, Pluralisme et Redondance de l'Information. Projet soutenu par la Maison des Sciences de l'Homme-Paris Nord

d'annotation manuelle, durée que notre solution permettrait de réduire considérablement.
Notre solution se caractérise par une série d'apports et la combinaison d'approches ayant déjà fait leurs preuves dans les domaines concernant à la fois la catégorisation et l'enrichissement sémantique. C'est pourquoi, dans la suite de ce papier, nous commencerons par un bref état de l'art des connaissances actuelles en RI[2] orienté catégorisation, avant de poursuivre par une série de définitions accompagnées d'exemples sur les sources sémantiques utilisées par notre solution. Par la suite nous décrirons les apports de notre solution pour l'aide à la catégorisation et à l'analyse des dépêches de presse. Enfin nous présenterons brièvement l'état actuel de notre prototype et des résultats obtenus avant de conclure sur les perspectives ouvertes à notre solution et les implémentations à venir.

2 État de l'art

Du fait du sujet même de notre problématique de recherche (la catégorisation assistée des dépêches de presse) il nous semble important dans un premier temps de rappeler et d'expliciter les concepts fondamentaux en RI orienté sur ce domaine. Nous définirons donc dans un premier temps les concepts de RI associés à notre projet avant d'approfondir la catégorisation à proprement parler. Par la suite, du fait de l'intégration de multiples sources sémantiques, il nous semble nécessaire de fournir un descriptif succinct ainsi qu'une présentation des sources utilisées.

2.1 Recherche d'information et catégorisation

On définit la Recherche d'Information par « l'ensemble des techniques permettant de sélectionner à partir d'une collection de documents ceux qui sont susceptibles de répondre aux besoins de l'utilisateur » [1] [2]
La catégorisation consistant en un rapprochement d'éléments similaires, il existe une forte corrélation entre les techniques dites de RI et celles associées à notre domaine d'étude. En effet, dans les deux cas il s'agira de déterminer la pertinence d'un ensemble, soit par rapport à une requête, soit par rapport à des documents voisins. C'est pourquoi le premier des deux processus clef de RI, à savoir le processus de « représentation », également appelé processus d'indexation. [1] [2] présente un fort intérêt pour notre système de catégorisation.

[2] Recherche d'Information

2.1.1 Indexation

Les processus d'indexation consistent en la description d'un document à l'aide de la représentation du contenu de celui ci. [1] [2] Autrement dit ils se caractérisent par la sélection puis par la pondération des termes pertinents.
La sélection des descripteurs se fait au moyen d'un équilibrage entre deux méthodes que sont la discrimination (distinction avec le reste du corpus de document) et la représentation (caractérisée par le contenu, modélisation du sujet dont traite le document.)

2.1.2 Extraction

Les processus d'extraction comportent deux principales familles d'approche :

- *L'extraction par cooccurrence*: est l'utilisation de groupes de mots pour définir un concept, elle mesure le nombre d'apparitions d'un mot sémantiquement non vide et son positionnement dans le document.
- *L'extraction par analyse linguistique*: (détection de patrons de mots NOM PREP. NOM) qui donne des informations sur la syntaxe et l'importance du terme. (un nom sera plus important qu'un adjectif par exemple).

Au regard des études sur le sujet et des solutions développées jusqu'ici l'approche dite mixte, c'est à dire prenant en compte à la fois le comptage des occurrences du terme dans le document et sa forme lexicale semble la plus appropriée à notre projet. Nous utilisons *TreeTagger*[3] [12] qui permet à la fois l'extraction, le traitement (transformation des verbes à l'infinitif, élimination des formes plurielles et autres problèmes pouvant nuire à la performance des comparaisons de termes) et l'identification de la famille grammaticale du terme.

2.1.3 Distance sémantique

Afin de répondre à notre problématique nous cherchons à effectuer deux types de rapprochements : un rapprochement entre les flux et les thématiques statiques (explicitées par notre thésaurus extrait des catégories normalisées de l'*IPTC*[4] , mais nous y reviendrons dans la section contribution) et un rapprochement sémantique entre les articles de presse eux même. Il nous semble donc que les mesures dites de « *similarité sémantique* » sont les plus adaptées à notre problématique.

[3] "TreeTagger" http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[4] Information Press Telecomunication Council, (voir état de l'art thésaurus).

Il convient donc de rappeler brièvement ce que sont ces mesures. En premier lieu il faut faire les distinctions entre la similarité sémantique et la proximité sémantique [1]. La similarité peut se calculer soit par la distance entre les arcs d'une représentation arborescente des termes, soit par un comparatif entre l'occurrence des termes. La proximité quant à elle prend en compte les relations entre les éléments que nous représentons dans notre solution sous forme de graphes, les relations étant les arcs et les éléments les nœuds.
Du fait de la structure même de nos sources de données au sein de notre prototype (se reporter à la section prototype de ce papier) nous utiliserons à la fois un système de calcul basé sur l'occurrence des termes et la distance entre les concepts en suivant le plus court chemin qui les sépare.

2.2 Les ressources sémantiques

Les travaux récents en recherche d'information et en catégorisation tendent vers la prise en compte de la sémantique. Pour rappel, la sémantique est l'étude du sens des mots et des relations qui les lient. L'analyse de corpus peut présenter de nombreux problèmes auxquels l'ajout de sources sémantiques peut pallier efficacement par des procédés tels que :

- *La désambiguisation* : des mots structurellement proches peuvent voir le sens qui leur est attribué changer du tout au tout (on parle alors de mots appartenant à plusieurs catégories syntaxiques). Ainsi par l'utilisation d'ontologies et des procédés de calcul de distance sémantique on peut ré-identifier le sens de ce mot.

- *L'enrichissement* : en associant les synonymes et les termes sémantiquement proches on précise le domaine de la recherche de correspondance entre nos termes et leurs catégories.

La distinction entre *Ontologies*, *Thésaurus* et *Bases de Données lexicales* est ténue dans la littérature [4]et parfois même dans la définition qu'en donnent leurs créateurs. Nous avons donc tenté d'en redonner une définition simplifiée dans la suite de ce papier tout en les illustrant par certains de leurs représentants utilisés dans notre prototype.

2.2.1 Thésaurus

Les thésaurus peuvent être définis comme des « *Ensembles hiérarchiques de termes clés représentant des concepts d'un domaine particulier.* »[5] Généralement organisés de façon thématique, les éléments de leur vocabulaire sont liés entre eux par des liens sémantiques qui peuvent être

[5] "Thesaurus." Dicodunet http://www.dicodunet.com/definitions/moteurs-de-recherche/thesaurus.htm

: la synonymie, l'équivalence, la spécificité (lien vers un concept de sens plus précis), ou la généralisation (lien vers un concept de sens plus large).
Dans le cadre de notre projet nous avons effectué l'étude d'un large échantillon de thesauri francophones librement exploitables tels qu' *Agrovoc*, *Jurivoc, Hurivoc, Dadi, Delphe, GeoEthno* ou encore *Eurovoc*.
Après l'analyse de soixante-quinze d'entre eux nous avons découvert qu'environ la moitié mettent à disposition leur ontologie en téléchargement et que sur cette moitié restante seulement cinq l'étaient sous la forme d'un ensemble de documents normalisés et réellement exploitable. La majorité se limitant à une série de *pdf* commentés. Concernant le volume lui même oscille entre quelques dizaines de milliers et plusieurs centaines de milliers.
La spécificité même des thésaurus associés aux particularités de notre problématique (documents francophones, domaines variés) font que l'intégration à notre base des seuls thésaurus disponibles risquait de parasiter le processus de calcul de la thématique. Nous avons donc fait le choix de n'intégrer que le thésaurus traduit par *Raphael Troncy* [4] des catégorisations de dépêches de presse de l' *IPTC*[6] et de l'utiliser comme base de référence thématique statique.
Pour information, le thésaurus de l'*IPTC* contient 1400 catégories, réparties sur trois niveaux d'abstraction et décrites succinctement.

2.2.2 Ontologies

On définit les ontologies comme des « Ensembles structurés des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances». La différence la plus intéressante par rapport à un thésaurus est qu'une ontologie offre un plus large champ de relations entre ses concepts que la simple hiérarchie comme par exemple la possibilité de retournement soit la navigabilité bidirectionnelle des arcs.
Parmi les ontologies disponibles nous avons porté notre choix et nos analyses sur deux d'entre elles:

- *DBPedia* : ontologie multi-domaine et multilingue créée entre autre à partir des « *infobox* » de *Wikipedia*. Son utilisation à l'heure actuelle reste peut répandue, même si elle a déjà été adoptée comme source de données sémantiques dans plusieurs projets de recherche et d'ingénierie. [4] [6]. Pour information l'ontologie *DBPedia* regroupe 2,6 millions de concepts dans 36 langues.

- *Yago* [10][11] : une ontologie généraliste anglophone basée sur *Wordnet* et *Wikipedia*. Son niveau de généralisation le plus élevé est établi d'après le système de classification de *Wikipedia*. Son intérêt

6 The IPTC-NAA standards."
http://www.iptc.org/cms/site/index.html?channel=CH0086

principal résidant en son système de représentation des liens entre concepts sous la forme {Sujet}-{Relation}-{Sujet}. *Yago* regroupe 95% des concepts *Wikipedia*.

Ces ontologies sont librement exploitables au format RDF, par « *endpoint* » (requête sur service web) et interrogeable via *SPARQL*. Il est à noter que *Yago* fournit un outil java de traitement et de conversion que nous avons utilisé pour franciser ses concepts par requête *SPARQL*.

2.2.3 Bases de données lexicales

Nous définissons les bases de données lexicales comme des ensembles de termes liés par synonymie ou proximité sémantique et organisés hiérarchiquement.
Au contraire des thésaurus, les bases de données lexicales sont généralistes. Et à la différence des ontologies les concepts exprimés sont simplifiés. Bien que considéré par beaucoup comme un thésaurus, *WordNet*[7] est l'illustration parfaite de ce que nous qualifierions de base de données lexicale.
Le problème de ces sources de données sémantiques est que leur représentant le plus significatif est en anglais, et du coup non exploitable par notre solution. Nous sommes donc en phase de recherche d'une base de données lexicale francophone récente et complète.

3 Contribution

Afin d'étudier le pluralisme des informations publiées sur le web, nous voulions analyser les informations publiées par les différents sites d'information, qu'ils soient des sites web de journaux classiques, des webzines (journaux exclusivement numériques), des blogs, des sites participatifs, des portails, des agrégateurs d'information ou des agences de presse. Une première idée était de prendre en compte le contenu complet de ces pages web, mais devant l'ampleur de la tâche face aux ressources disponibles, nous avons limité notre champ aux sites dotés d'un flux RSS. Un tel choix présente l'avantage de traiter des données homogènes issues de sources différentes.
Nous avons recueilli 89 flux RSS répartis dans les catégories ci-dessus. Nous les avons enregistré en continu pendant plusieurs semaines. Pour chaque item émis par un flux RSS, nous avons recueilli son titre, la date et l'heure d'émission, …
Nous proposons de raffiner le modèle statistique de calcul de la proximité des catégories de sujet au moyen de thésaurus, d'ontologies et de

[7] WordNet - Princeton University Cognitive Science Laboratory." http://wordnet.princeton.edu/

dictionnaires lexicaux. Parallèlement nous proposons des méthodes d'identification, d'extraction et de pondération des termes significatifs. Enfin nous ouvrons des perspectives de développement de notre projet quant à la catégorisation temporelle des dépêches de presse alternativement nommée « *Catégorisation de sujets* ».
Le schéma ci-dessous représente la structure de la base de données dont il sera question dans les sections suivantes. A titre d'information (même si il en sera question plus longuement dans les sections consacrées à l'intégration des sources sémantiques et à l'agrégation) : les flux RSS agrégés sont répartis sur les tables *lemmes*, *rss_item* ; les thématiques et leur relation hiérarchique sont contenues dans les tables *thema_lemmes*, *thematique*, *relation*, et *thesaurus*. Les ressources *DBPedia* dans *dbpedia_ressources* ; et les informations sur les sources de nos flux dans *tags*, *fluxrss_tags*, *fluxrss*.

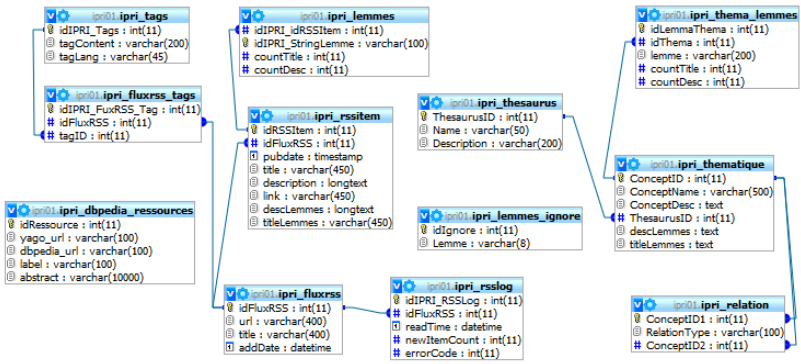


Figure 1 Structure de la base de données du projet

3.1 Intégration des sources sémantiques

De notre section état de l'art nous avons choisi d'intégrer à l'heure actuelle trois sources sémantiques, à savoir : le thésaurus de l'*ITPC*, l'ontologie *DBPedia* et l'ontologie *Yago*. Il est à noter également que nous interrogeons *DBPedia* de deux manières sur les trois possibles (voir section 3.2.1)

3.1.1 IPTC

Du fait de leur nature hiérarchique et afin de formaliser l'ajout d'autres sources sémantiques aux liens plus complexes nous avons choisi de représenter les concepts sous la forme d'une table thématique et d'une

table association décrivant le type de relation ,la thématique étant associée à une clef étrangère décrivant le type de source. Nous séparons les concepts et thésaurus en quatre tables :

- *La thématique* : chargée de stocker le titre et la description du concept ainsi que de stocker l'ensemble des lemmes qui en sont issus.
- *Les relations* : qui peuvent se résumer à une table association.
- *Les thésaurus* : contenant les noms et descriptifs de la famille associée.
- *Les lemmes* : contenant les lemmes stockés un par un et associés à une thématique ou concept.

Cette représentation présente un triple intérêt. Nous conservons d'une part par une forme abstraite une modélisation souple applicable à différentes sources de données. Nous facilitons ensuite le travail de représentation sous forme de graphes explicités dans la section consacrée au prototype; enfin, nous facilitons le travail de pondération et de calcul de distance des mesures de similarité sémantique.

3.1.2 DBPedia

DBPedia permet trois types d'interrogation sur sa base. Ils mettent à disposition des « *dumps* » de leur base, un service par « *endpoint* » est accessible comme service Web et interrogeable par requête *SPARQL*, enfin des fichiers sous forme *XML* sont fournis pour une correspondance avec *Yago*. De ces méthodes d'extraction sus cités nous utilisons les deux dernières.
Du fait de la spécificité linguistique de notre projet nous nous intéressons aux sources disponibles dans la langue française. C'est pourquoi nous avons peuplé notre base automatiquement de quelques deux millions de concepts à partir d'un fichier XML référençant les ressources et au moyen d'une requête *SPARQL* filtrant sur les titres et les descriptions en français.
Nous interrogeons également l'« *endpoint* » de *DBPedia* pour établir une correspondance entre les termes que nous avons détectés comme ayant une forte probabilité d'être des noms propres et les concepts correspondants. (voir section 3.2).

3.1.3 Yago

Nous proposons d'utiliser l'ontologie *Yago* au travers d'un logiciel librement distribué en Java [12]. Pour ce faire nous utilisons à la fois le service de conversion des sources en SQL et le service de conversion en XML.

Le premier nous permet d'obtenir une table de lien de la forme {Concept}-{Relation}-{Concept} tel que {Einstein}-{a gagné}-{Prix Nobel} nous permettant par la suite d'utiliser les liens de relations dans notre mesure de similarité.

Le second nous permet d'obtenir un fichier de quelques deux millions de liens vers les ressources de *DBPedia* que nous injectons par la suite dans nos algorithmes de requête SPARQL pour en extraire une francisation des ressources sous la forme label et commentaires. Nous préconisons à ce sujet de ne pas utiliser la section résumé du fait de sa taille trop élevée qui atténuera la détection et la pondération des termes importants.
Ces deux utilisations nous permettent donc de faire le lien entre les ressources de *DBPedia* et de *Yago* tout en corrigeant le problème de la langue, « *Yago* » étant une ontologie anglophone et nos dépêches étant francophones.

3.2 Traitement des flux

L'actualité est par nature changeante, événementielle, ainsi sont donc nos flux. Toutefois nous proposons une série de traitements et de règles générales pouvant s'y appliquer et permettant le calcul de la proximité sémantique.
Nous procédons à l'agrégation de flux RSS dont la structure se compose d'un titre et d'une description que nous lemmatisons pour analyse et traitement. Ainsi on relèvera que les lemmes de titre sont par nature plus significatifs que les lemmes de description. (Le titre est une accroche qui représente l'essence de l'article). Un poids plus important doit donc être accordé aux lemmes associés aux titres par rapport aux lemmes associés à la description du contenu de l'article.
Basé sur l'article de J.Savoy [5] nous avons pu identifier un certain nombre de termes à la signification sémantique nulle (70 termes parmi lesquels des propositions, des noms, des verbes, etc.). On peut la rapprocher de la technique dite de discrimination dans le sens où ces termes n'apporteront pas de valeur ajoutée à notre algorithme de catégorisation. Nous proposons de filtrer et d'éliminer ces termes à la lemmatisation de nos articles.
Du fait même de la structure de nos thématiques statiques il semble important de prendre en compte le niveau hiérarchique. En effet, après analyse, nous observons une répétition entre les différents niveaux. Ainsi « M*usique Classique »* de niveau rappel dans son contenu les caractéristiques de son parent *Musique* de niveau deux. Ces deux spécialisations étant des dérivées de la super-catégorie *Art et spectacle* de niveau un et représentant des formes d'expression artistique. On retrouvera donc une occurrence de termes communs qui perdent de leur pertinence à la montée progressive du niveau d'abstraction.

Nous proposons donc une double approche sur la pondération des termes:

- *Déplier l'ensemble des sous concepts* : autrement dit, pour la détection du premier niveau, inclure dans notre recherche l'ensemble des lemmes des sous-ensembles pour le calcul de la probabilité de proximité sémantique.
- *Pondérer au fur et à mesure des itérations* : lorsque l'on « accroche » un nœud, réduire le poids des lemmes de la super-catégorie et augmenter ceux de la sous-catégorie. Cela nous permet à la fois d'obtenir une indication sur la justesse du chemin choisi par la conservation des propriétés acquises et de réduire l'influence des termes redondants.

Basé sur l'analyse manuelle des articles de presse [7] nous assignons comme prioritaire pour la détermination d'une catégorie de presse les noms propres et géographiques utilisés. En effet, un article ayant, par exemple, pour titre le nom d'un sportif ou un club de foot, verra ses probabilités d'être associé à la thématique « sport » beaucoup plus élevées.

Nous établissons donc une fonction de détection et de reconnaissance des noms propres couplés à l'ontologie *DBPedia* [4][8]. Pour ce faire nous proposons d'utiliser les propriétés d'analyse lexicale de *TreeTagger*. Les candidats à la catégorie que nous qualifions de nom propre ayant en commun la caractéristique de ne justement pas appartenir à un ensemble de catégorie lexicale. Après analyse, si un nom ou un adjectif est détectable comme tel, un nom propre n'a lui aucune des caractéristiques de ce dernier (nous reviendrons sur les résultats dans la section expérimentation).

Enfin en ce qui concerne la mesure de proximité sémantique nous proposons un approche mixe prenant en compte à la fois le poids des termes sus cité et les relations pouvant exister entre ces termes.

3.3 Calcul de proximité

L'ensemble de nos catégories sont représentés sous formes de graphes au sein de notre prototype. Cette classe abstraite contient une liste de ses nœuds parents et enfants ainsi qu'un ensemble de lemmes prétraités par *TreeTagger*. C'est sur cette représentation que notre système de calcul de proximité sémantique effectue une série d'opération permettant la détermination du nœud thématique le plus proche de l'article soumis.

[8] "DBPedia" http://dbpedia.org/About

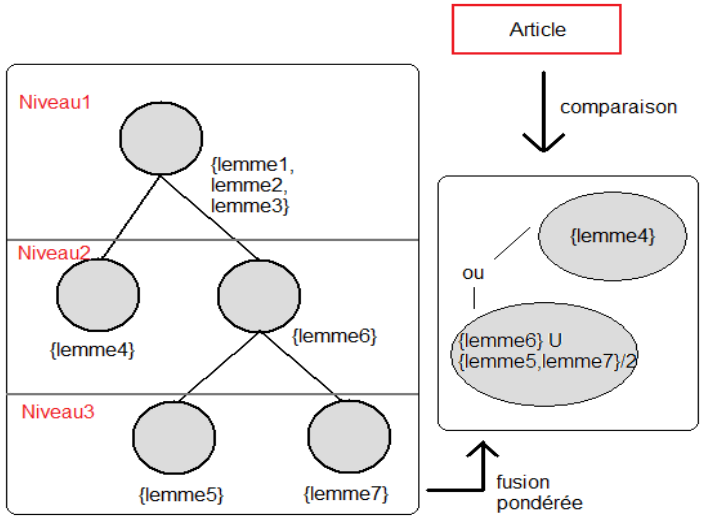


Figure 2 Processus de transformation par fusion pondérée des lemmes enfants

```
Entrée
          Niveaux à comparer
          Item RSS d'une dépêche
          L'ensemble des thématiques IPTC
Début
          Récupérer les lemmes de l'item RSS
          Pour chaque thématique de premier niveau
            récupérer les lemmes
            récupérer les nœuds enfants
            Si niveau >1 pour chaque nœud enfant
            récupérer les lemmes
              Si niveau > 2 pour chaque enfant des enfants
                  récupérer les lemmes.
          Comparer et retourner fréquence d'occurrence des lemmes de
          l'article   avec les lemmes des nœuds en divisant par deux
          pour chaque niveau d'éloignement du nœud à joindre.
Fin
```

Algorithme 1 Algorithme de comparaison

Pour ce faire notre algorithme récupère, des nœuds de plus haut degrés, les lemmes associés aux nœuds enfants et après avoir pondéré (on divise par deux la valeur pour chaque lemme de niveau inférieur), effectue un calcul de cooccurrence avec les lemmes de l'article. (Figure 2). De là il détermine le candidat le plus apte et répète l'opération sur les nœuds enfants.

L'originalité de l'algorithme tient principalement en sa procédure de « dépliage » des nœuds de niveaux inférieurs (nous pensons appliquer cette approche à des systèmes non hiérarchiques à l'avenir). Nous augmentons ainsi les chances de trouver des lemmes candidats à l'union tout en réduisant leur influence par pondération et par division de leur nombre.

4 Expérimentation et validation

Notre prototype écrit en Java dispose déjà de plusieurs fonctionnalités : l'agrégation de flux RSS d'un échantillon large, représentatif des différentes politiques de publication des sources de presse internet; la conversion et l'intégration de thésaurus; la lemmatisation et la pondération des termes liés à leur niveau hiérarchique à la fois pour les flux RSS et pour les thésaurus; l'identification des termes candidats au marquage en nom propre au moyen de la fonctionnalité d'analyse lexicale de *TreeTagger;* et enfin un algorithme de calcul de la proximité sémantique.

4.1 Agrégation

S'intéresser au pluralisme de l'information sur Internet implique d'intégrer les particularités de la publication de contenus sur le web, en comparaison des supports antérieurs comme l'imprimé ou l'audiovisuel. Car le processus de publication sur le web, ne se limite pas au modèle classique de diffusion mass-médiatique (sites de presse en ligne) : il comprend aussi le registre de l'auto publication (blogs), de la publication distribuée, et le niveau méta-éditorial (agrégateurs). Nous avons choisi 105 sources différentes représentatives de cette diversité:

- Des sites de presse en ligne : Le Monde, Libération, Le Figaro, etc.
- Des blogs : Plume de presse, Jean-Michel Apathie, etc.
- Des agences : Ria Novosti, etc.
- Des portails de news : Actualités Orange, MSN Actualités, etc.
- Des sites d'information participatifs : Rue89, etc.
- Des agrégateurs : Google Actualité, Wikio, etc.

Notre prototype propose au choix soit l'analyse des sorties du flux RSS d'une source sélectionnée, soit l'agrégation et la lemmatisation de l'ensemble. Nous convertissons le flux et l'insérons dans la base en en conservant la date, l'heure de publication (pour le regroupement

temporel), le titre, la description, et les termes extraits après lemmatisation.

4.2 Lemmatisation et filtrage

La lemmatisation se fait au travers d'un script utilisant *TreeTagger* qui permet accessoirement l'extraction linguistique ou le traitement automatique de la langue en convertissant notamment les verbes conjugués en leur forme infinitif. Notre prototype l'utilise pour trois de ses fonctionnalités, à savoir la lemmatisation des flux RSS, la lemmatisation des thésaurus (conversion linguistique) ainsi que la détection et le marquage des noms propres.

Comme explicité dans la section consacrée à notre contribution nous nous basons également sur le papier de J.Savoy [5] et éliminons soixante-trois lemmes vides de sens réduisant ainsi en moyenne le nombre de lemmes associé aux articles, définition et titre compris de 7%. Cette étape nous a permis de réduire légèrement les temps de calcul de proximité sémantique et d'alléger le poids de la base. Il est à noter que les lemmes en questions sont localisés à 80% dans les descriptions du fait de leurs natures plus verbeuses que les titres.

4.3 Proximités et résultats

Afin de tester la première étape d'implémentation de notre proposition de calcul de distance sémantique et d'aide à la catégorisation thématique des dépêches de presse nous avons sélectionné quarante articles aléatoirement sur un échantillon large de nos sources puis nous les avons catégorisés manuellement et avons récupéré en sortie les catégorisations proposées par le prototype selon que nous choisissions la fusion pondérée des nœuds enfants ou non.

Il est à noter que l'actualité récente (la grippe mexicaine) domine avec un tiers des dépêches orientés sur la santé.

Le Tableau 1présente les résultats obtenus. Les colonnes *sortie1, sortie2 et sortie3* représentent la catégorie retournée par l'algorithme de catégorisation en tenant compte des lemmes respectivement de niveau un, un et deux et des trois niveaux ensemble.

On observe dès les premières analyses que le niveau trois (à savoir prendre en compte toutes les lemmes de tous les descendants de chaque catégorie de niveau un et deux) donnent des résultats plus en adéquation avec nos attentes.

manuel	sortie1	sortie2	sortie3
Social	Police et justice	Police et justice	Police et justice
Santé	Alertes	Santé	Santé
Santé	Alertes	Santé	Santé
Politique	Santé	Santé	Santé
Santé	Police et justice	Police et justice	Police et justice
Santé	Alertes	Bulletins	Politique
Economie et finances	Société	Société	Economie et finances
Désastres et accidents	Politique	Politique	Désastres et accidents
Santé		Social	Politique
Politique	Police et justice	Police et justice	Politique
Société	Société	Société	Guerres et conflits
Sport	Sport	Sport	Sport
Police et justice	Politique	Politique	Sport
Police et justice	Police et justice	Police et justice	Police et justice
Santé	Social	Social	Politique
Santé	Politique	Politique	Sport
Police et justice	Politique	Politique	Sport
Politique	Politique	Politique	Politique
Economie et finances	rien	Social	Social
Economie et finances	Politique	Politique	Politique
Sport		Social	Sport
Santé	Santé	Santé	Santé
Politique	Politique	Politique	Politique
Politique	Politique	Politique	Politique
Social	Social	Social	Social
Police et justice	Politique	Politique	Sport
Sport	rien	Désastres et accidents	Sport
Santé	Désastres et accidents	Politique	Sport
Social	Vie quotidienne et loisirs	Vie quotidienne et loisirs	Social
Santé	rien	Santé	Santé
Guerres et conflits	Politique	Politique	Politique
Social	Vie quotidienne et loisirs	Vie quotidienne et loisirs	Social
Science et technologie	Gens animaux insolite	Gens animaux insolite	Social
Santé	Science et technologie	Politique	Politique
Environnement	Environnement	Police et justice	Environnement
Sport	Politique	Politique	Sport
Désastres et accidents	Désastres et accidents	Désastres et accidents	Désastres et accidents
Politique	Politique	Politique	Politique
Politique	Politique	Politique	Police et justice
Gens animaux insolite	Statistiques	Police et justice	Social

Tableau 1 Résultats attendus et résultats obtenus

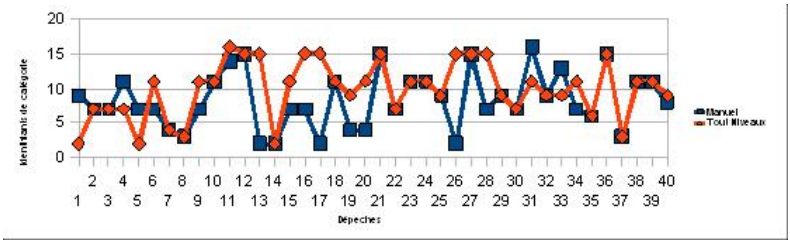


Figure 3 Correspondance avec la méthode de fusion pondérée

Une chose est à noter toutefois, notre catégorisation de référence est subjective. Bien que plus juste que celle obtenue par le prototype elle n'est pas exempte d'ambiguïté. En effet la frontière entre un mouvement social et une action politique est ténue et l'on trouvera dans l'article des références à ces deux thématiques.
Une dernière représentation booléenne de nos résultats (Figure 4 : somme des valeurs correctes en choisissant vrai pour thématique équivalente et faux dans le cas contraire divisé par le nombre total d'articles) vient toutefois confirmer notre analyse quant au taux de rappel entre la thématique attendue et les résultats de l'implémentation.

Methode	1	2	3
Satisfaction	30.00%	35.00%	55.00%

Figure 4 Taux de correspondance

Nous tirons trois conclusions de ces résultats :

- La prise en compte des nœuds enfants dans le choix des thématiques est une approche qui améliore significativement nos résultats.
- La quantité de lemmes contenus dans l'article influence la performance de l'appareillage. En effet la moitié des dépêches mal ou non catégorisées contenaient une description plus courte que la moyenne.
- Un apport sémantique est indispensable au bon fonctionnement de notre solution. Ce qui nous renforce dans notre idée d'enrichir le contenu des articles par des liens à *DBPedia*. Il semble clair que par la détection des concepts associés aux noms propres nous augmenterions considérablement la taille de l'ensemble de termes pertinents pour la catégorisation.

5 Conclusions et perspectives

Nous avons, au travers de ce papier, présenté notre approche d'aide à la catégorisation thématique des dépêches de presse francophones depuis des flux RSS, catégorisation basée sur un enrichissement sémantique permis d'une part par des sources externes (ontologies, thésaurus) et d'autre part par l'utilisation d'une mesure de calcul de proximité sémantique, mixte entre les méthodes par groupes de termes et les méthodes de calcul de chemin dans un arbre.

Les différents traitements successifs auxquels nous soumettons nos sources (lemmatisation, filtrage, analyse lexicale) nous permettent déjà d'obtenir un ensemble de propositions de rapprochement par catégorie thématique.

Nous projetons dans un prochain temps de fournir une réponse à la deuxième problématique de notre sujet à savoir de fournir une extension à la catégorisation temporelle des sujets. Au vu de l'état actuel de notre solution et des conclusions que nous avons tirées de nos recherches nous avons d'ores et déjà identifié trois facteurs de rapprochement, à savoir :
Le facteur thématique : On suppose que la probabilité de rapprochement d'articles à la thématique proche sera plus élevée que celle d'article à la thématique éloignée. Autrement dit la longueur du chemin qui sépare deux articles est un facteur déterminant de leurs proximités.
Le facteur temporel : du fait même de la nature événementielle des sujets de presse il apparaît essentiel dans une optique de catégorisation temporelle de déterminer un intervalle. De même la détection de ce dernier voit la probabilité du rapprochement de l'article à une famille dynamique de sujet augmenter. (Si l'on prend comme exemple le cadre des élections américaines, si l'on crée la section Obama [7] sous cette dernière les probabilités qu'un article contenant « *Obama* », « *américain* », « *élection* » ou n'importe quel lemme de ce type soient à rapprocher de cette catégorie s'en voient considérablement augmentées.)
Le facteur sémantique : La proximité et l'importance de certains lemmes utilisés pour la catégorisation thématique est un facteur de détermination et d'association au sujet. Dans le cadre même de regroupement au sein d'un nouveau sujet ces lemmes seraient choisis comme titre.

Nous projetons également une francisation et une normalisation finale des liens entre concepts intégrés à notre base, travail en cours qui devrait permettre de résoudre les problèmes de rapidité de traitement de correspondance actuel dû au fait du nombre élevé de tuples que représentent nos ontologies généralistes.

Enfin nous pensons ouvrir la voix à une implémentation d'une fonction d'apprentissage des termes importants relatifs aux évènements qui nous permettraient d'optimiser nos coûts en détection et en traitement.

6 Références bibliographiques

[1] H. Zargayouna. I*ndexation sémantique de documents XML,* Thèse section état de l'art, 2004

[2] M. Baziz, *Indexation Conceptuelle guidée par Ontologie pour la recherche d'Information*, Thèse pages 34-95, 2004

[3] E. Marty, F. Rebillard, N. Smyrnaios, A. Touboul, *Pluralisme et redondance ce l'information sur l'internet,* Revue Mot.

[4] R. Troncy, *Explorer des actualités multimédia dans le Web de Données*, IC 2009.

[5] J. Savoy, *Indexation et représentation comparative: Application au discours électoral*. CORIA 2009

[6] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, R. LeeMedia, *Meets Semantic Web - How the BBC uses DBpedia and Linked Data to make Connections*, 2009.

[7] A. Touboul, *Synthèse de l'étape « Analyse qualitative du traitement d'un sujet : Élection de Barack Obama» 2008.*

[8] Y. Matar, E. Egyed-Zsigmond, S. Lamji, *KWSim: Concepts Similarity Measure* CORIA 2008.

[9] A. Formica, *Concept similarity by evaluating information contents and feature vectors: a combined approach*. Commununications of the ACM 52 (2009) 145-149.

[10] F M. Suchanek, *Automated Construction and Growth of a Large Ontology* 2008.

[11] Fabian M. Suchanek, Gjergji Kasneci, GerhardWeikum, *YAGO:ALarge Ontology from Wikipedia andWordNet*, 2009.

[12] H. Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees* 2007.

Connaissances prescrites ou connaissances décrites ? L'apport de la sémantique des textes

Monique Slodzian (1), Mathieu Valette (2)

CRIM-ERTIM (EA 2520) INaLCO, Paris (1)
ATILF (UMR 7118) CNRS, Nancy (2)

Mots-clés : Connaissances prescrites, Vérité forte/vérité faible, Systèmes d'organisation des connaissances, Sémantique des textes, Parcours interprétatif, Planification de l'information, Forme sémantique, Thématisation, Lexicalisation

Keywords: prescriptive knowledge, Strong/weak truth, Knowledge Organisation Systems, Text Semantics, Interpretative path, Information planification, Semantic form, thematisation, lexicalisation.

Résumé : L'article vise à montrer que le modèle collaboratif de communication des connaissances revendiqué par le Web 2.0 ne rompt pas de manière significative avec le modèle épistémologique antérieur, issu du positivisme logique, notamment par son primat référentialiste prescriptif. En postulant *in fine* l'existence de concepts primitifs partagés, il est conduit à reproduire les mêmes limites que le Web sémantique fondé sur un socle de métadonnées réputées universelles. Par ailleurs, une acceptabilité indiscutée des connaissances de vérité faible pose des problèmes de fiabilité et de garantie susceptibles de compromettre le succès du modèle. L'article entend démontrer dans une deuxième partie en quoi la sémantique des textes peut contribuer à objectiver les connaissances par la description de parcours interprétatifs. Considérant que les textes relèvent d'une planification de l'information, l'article explicite la notion de *forme sémantique*, entre le texte et le concept, et envisage la possibilité de faire émerger des *préconnaissances* non encore lexicalisées. Cette proposition théorique est illustrée à partir de discours de prévention contre le tabagisme issus du Web.

Abstract : This paper aims at showing that the collaborative communication model upheld by Web 2.0 doesn't significantly break with the previous epistemological model, stemming from logical positivism, mainly due to its prescritive referentialist primacy. By assuming the existence of shared primitive concepts, it is finally led to reproduce the same drawbacks as the Semantic Web founded on a bunch of metadata given as universal. Moreover, accepting non-expert knowledge without debate raises the issue of reliability and expertise, exposing Web 2.0 to a fatal risk of misinformation. In its second part, the paper

argues that textual semantics is able to contribute to objectivating knowledge by describing interpretative scenarios. Considering that texts come under a planified pattern of information, the authors clarify the notion of semantic form, between text and concept, and consider the possibility of eliciting pre-knowlede elements, not yet lexicalized. As a concrete framework to this theoretical proposal, arguments will be supported by anti-smoking texts trawled from the Web.

1 Introduction

La masse de données qui constitue le Web ne relève pas pour l'essentiel de champs de connaissances homogènes et discrétisables en ontologies. En effet, ce qu'on appelle faute de mieux « connaissances générales » ou « connaissances vulgarisées » s'opposent en premier lieu aux connaissances historiquement encadrées et présumées garanties par les domaines scientifiques et techniques que des documentalistes ont organisées en classifications au fur et à mesure de leur constitution.
L'hypothèse que l'on puisse accorder de la valeur à des connaissances de vérité faible coïncide avec l'avènement du Web et de ses contenus surabondants offerts à tous. Le propre du tout-venant du web consiste en effet à demeurer hors du contrôle rationnel régi par les communautés d'experts et à défier ainsi toute dichotomie de type vulgarisé *vs* scientifique, voire scientifique *vs* pseudo-scientifique. Ceci par opposition aux notions d'arbre de connaissance ou d'ontologie qui supposent une finitude des données et une production raisonnée, restrictive et contrôlée des connaissances. Les théories et systèmes de classification produits tout au long du XXème siècle concevaient en effet « la connaissance » comme une structure rigoureuse (mathématique ou logique), se présentant comme un système formel. La science étant définie comme une structure logique commune à tous, le but de la connaissance consistait à orienter parmi les objets et à prédire leur comportement : on était censé y parvenir en découvrant leur *ordre* et en assignant à chacun d'eux la place qui est la sienne au sein de la structure du monde. Carnap donnera la théorisation la plus aboutie de ce modèle partagé par les positivistes logiques (Carnap, 1928). Sans nous attarder davantage sur la dichotomie entre connaissance intuitive vs connaissance scientifique comme genre supérieur de connaissance, nous retiendrons ce moment précis où le mode énonciatif de la science passe impérativement par des formes d'expression et de relations fortement *prescriptives*, au point que la conformité de la forme garantit la valeur du contenu (Carnap, 1934, Schlick, 1932). Il nous semble être le point de référence indispensable pour mesurer l'ampleur des ruptures épistémologiques accomplies et pour appréhender la question du texte et de sa description, souvent oubliée dans les plis de la philosophie de la connaissance.

Nous entendons montrer dans cet article, que les textes présentent une matérialité qui se prête à l'analyse et à l'objectivation. Par objectivation nous n'entendons pas une extraction de connaissances déliées des textes et des interprétations possibles, comme le font les approches prescriptives, mais au contraire une prise en compte des conditions de production et d'interprétation des documents. L'enjeu est de faire émerger les connaissances des textes et de les caractériser en tenant compte des conditions de leur production et de leur interprétation, de façon à évaluer leur pertinence par rapport à une tâche donnée. Notre proposition méthodologique, s'inspirant de la sémantique des textes, aura pour objectif de faire émerger des informations susceptibles d'être constituées en connaissances.

2 Problématique

2.1 L'information est une relation

La construction formelle que présupposait la science unitaire des positivistes logiques reposait notamment sur le concept de *relation***,** au point que l'on pourrait suivre Barlow (Barlow, 1994, p.13) et poser que "l'information est une relation". Les controverses sur la taxinomie et le nombre de relations logiques depuis l'élaboration de la Classification décimale universelle (DCU) montre bien le succès de cette formule lapidaire: d'une vingtaine pour Coates (Coates,1960), on passe à quelque 5000 relations dans le système CYC [1]. La polémique sur le statut épistémique des métadonnées proposées par le Dublin Core Metadata Initiative trouve ici sa place [2]. En effet, les relations classiques qui structuraient les systèmes de classification bibliographiques relevaient de l'empirisme logique et de l'heuristique de la vérité scientifique, tandis que les relations de CYC concernent des connaissances de sens commun et, plus paradoxal encore, les métadonnées du Dublin Core s'adressent aux connaissances illimitées du Web sémantique. Les folksonomies ne seraient-elles pas en dernière instance le symptôme de l'impuissance intrinsèque de tout système de métadonnées à structurer une masse illimitée de données?

En fin de compte, la difficulté théorique et pratique de concevoir des ontologies générales institutionnellement prédéfinies pour classer la masse d'informations hétérogènes aboutit à une substitution de paradigme : on passe du paradigme de l'information à celui de la

[1] Base de connaissances issues du sens commun ayant pour but le développement d'un système intelligent.

[2] Référence sur ladite polémique.

communication (Web 2.0). La notion de web horizontal exprime ce renversement en invitant les usagers à indexer eux-mêmes le web à l'aune de leurs intérêts propres. Classification traditionnelle et folksonomie correspondraient ainsi à deux options philosophiques opposées. La première répondrait à une structure hiérarchique *top-down* adossée à l'objectivité supposée des technosciences, la seconde assumerait la subjectivité d'un étiquetage *sui generis* issu d'un filtrage collectif.

2.2 Entre l'expert et la sagesse des foules, quelle place pour l'objectivité ?

Cependant, l'opposition entre « stockage de l'information hiérarchique » et « filtrage collaboratif » (Origgi, 2008) mérite d'y regarder de plus près. S'ils se distinguent par leur intentionnalité (informer vs communiquer), il reste à prouver qu'ils procèdent d'épistémologies différentes et qu'ils se situent de part et d'autre de la ligne référentiel /non référentiel, autrement dit qu'ils ne procèdent pas également de l'empirisme logique. L'activité de connaissance présupposée dans les deux cas n'est-elle pas conçue comme le transcodage d'un langage-objet en un métalangage et réciproquement (Rastier, 1995) ? C'est en tout cas ce que suggère l'idée de monde comme « catalogue » (Olivier Ertzscheid, 2008) qui laisse entendre que l'on peut accéder directement aux objets, l'univers de référence étant induit par l'objet lui-même. Il s'agit de saisir les mots par la dénotation et donc de représenter la relation des objets au monde. Cette vision néo-platonicienne du rapport des objets –virtuels et réels- au monde à travers le médium technologique est théorisée sous le nom d' « Internet des objets » et certains de ses tenants se risquent à présenter l'informatique comme le socle de la structure du monde (Wolfram, 2001). Leurs propositions viennent en droite ligne des travaux des positivistes, Russel, Whitehead et en particulier Carnap, dont *La Construction logique du Monde* peut être considérée à cet égard comme préfigurant la philosophie du Web.

On se demandera, par ailleurs, si le filtrage collaboratif à l'origine des folksonomies échappe à un modèle communicationnel inférant un lexique mental qui s'enracinerait lui-même dans des universaux cognitifs donnant accès à une vérité-évidence. La « sagesse des foules » dont procéderait le PageRank de Google signifierait que, selon Adam Bosworth cité par Ertzscheid, pour tout élément donné (texte, image, document), on aurait une série de mots et de termes composant le plus petit lexique commun (expression d'un consensus conceptuel) permettant de décrire l'objet ou le document. Ainsi, l'indexation sociale, dont FlikR par exemple serait le parangon, consiste à rechercher le consensus sur des valeurs phénoménologiques (opinions consensuelles) qui présupposent qu'on tient le monde pour un « pan-catalogue » d'objets. Peut-on être plus référentialiste ?

D'une certain manière, cette vérité-évidence, qu'elle soit dictée par des experts ou par la sagesse des foules, demeure l'affectation de valeurs subjectives à des contenus et procède en conséquence du prescriptif, quand même ses prescriptions sont de nature différentes, fortes lorsqu'elles sont émises par une autorité experte et faibles lorsqu'elle sont induites par des stéréotypes partagés. Ainsi, on rapporte le jugement d'individus ou de collectivité d'individus sur les contenus plus qu'on ne les décrit dans leurs contextes sociaux et culturels.

3 Faire parler les textes

3.1 La pertinence en jeu

La dichotomie objectivité *vs* subjectivité qui présuppose l'existence de « normes scientifiques » actualisées par des méthodes, des standards et des pratiques devient à son tour un critère déterminant de démarcation entre « bonne » et « mauvaise » science. Au-delà des enjeux juridiques et économiques sous-jacents à ce débat, nous nous intéressons à sa dimension épistémologique. Cette dernière est en effet déterminante si l'on considère les textes comme lieux de production de l'information. Plus particulièrement, la catégorisation des genres textuels (par exemple scientifique *vs* vulgarisé) pose directement la question de la possibilité de discriminer les textes scientifiques et pseudo-scientifiques. Autrement dit, y a-t-il des caractéristiques formelles stables et généralisables qui permettent de distinguer un texte scientifique d'un texte pseudo-scientifique? A priori, la présence de tableaux statistiques ou d'indices de quantification et de bibliographie (parmi d'autres traits) semble caractérisante de textes présentant une valeur de vérité forte. Or, la fabrication d'une argumentation pseudo-scientifique consistera précisément à exhiber ces indices, parmi d'autres, de telle sorte qu'il sera impossible de trancher tant la conformité à la forme attendue est confondante. La question du vrai/faux, qu'on la considère comme pastiche ou sorte de spam, invite à prendre la textualité au sérieux. Le cas limite du « faux » – problème général posé aujourd'hui au Web – impose que l'on s'appuie sur une sémantique des textes élaborée, tant il est vrai qu'une liste finie de mots clés (concepts homologués du domaine) et de procédés rhétoriques externes (figures de style obligées) ne suffisent pas pour produire une analyse des textes suffisamment pertinente.
S'il est vrai, comme le suggère Gloria Origgi, que « la vérification directe de l'information n'est tout simplement pas possible à des coûts raisonnables », ce passage à une ère d'informations de vérité faible est porteur de risques socioculturels incommensurables. Face à la crise annoncée, des outils opératoires nouveaux doivent être proposés, faisant

appel à des approches transdisciplinaires demeurées à la lisière des travaux sur l'ingénierie des connaissances. En proposant la description de parcours interprétatifs assignant un ou plusieurs sens à un texte, la sémantique des textes, ouverte au document dans la perspective du numérique (RTP.DOC, 2006), affirme sa capacité à tracer et hiérarchiser les subjectivités qui traversent les textes et, en cela, à assumer leur part d'objectivation.
Par objectivation nous ne supposons pas une extraction immédiate de connaissances déliées des textes et de leurs interprétations possibles, comme le suggèrent les approches prescriptives en produisant des listes de mots censés livrer sans médiation les connaissances d'un texte. Nous posons au contraire la nécessité de passer par des procédures d'analyse pour faire émerger et caractériser les connaissances d'un texte en tenant compte de ses conditions de production et d'interprétation (ordre herméneutique), si l'on veut assurer leur pertinence par rapport à une tâche donnée.
Cette approche impliquant l'ordre herméneutique est incompatible avec la philosophie sous-jacente à l'Internet des objets qui se réduit à l'ordre référentiel ou, au mieux, à l'ordre communicationnel. Il y a là un débat de fond à mener.

3.2 La sémantique du document dans les SOC

La notion de document, défini comme "une artefact médiateur à dominante sémiotique inséré dans des flux transactionnels" qui nous vient des STIC (Zacklad *et al.*, 2007) s'accompagne d'une vision ouverte de l'ingénierie des systèmes d'information à partir d'une réflexion nouvelle sur le processus de documentarisation. La théorie du document qui en émane met en avant « la recherche d'une complémentarité entre SOC hétérogènes, impliquant un rapprochement plus grand entre champs et secteurs différents ». On y trouve une invitation à construire une approche unifiée des espaces sémiotiques ouverts par les TIC, à partir de la notion de co-production sémiotique.
Le processus de documentarisation ainsi décrit propose un couplage texte/document où les approches de la sémantique interprétative peuvent trouver leur légitimité, en même temps qu'elles s'y verront confrontées à une dimension sémiotique nouvelle susceptible de renouveler le concept de texte. Il s'agira en particulier de voir comment des approches relevant respectivement d'une sémiotique du document et d'une sémantique du texte peuvent converger.
Nous tenterons maintenant de démontrer la possibilité de cette convergence en soumettant quelques propositions méthodologiques susceptibles d'intéresser ceux qui, dans la communauté STIC, partagent avec nous une vision « constructiviste » des connaissances et confèrent au

texte/document un statut herméneutique en rupture avec les descriptions strictement référentielles.

4 Le texte comme système d'organisation des connaissances ?

Dire que le texte est un SOC introduit un débat entre linguistique et ingénierie des connaissances. En effet, si la pratique de l'extraction de terminologies ou d'ontologies à partir de textes donne à penser que le texte est un espace de collecte privilégié, il serait faux de le considérer seulement comme le terrain d'actualisation des concepts : les concepts ne préexistent pas aux textes, ils sont des îlots, des zones stables de sens construits, élaborés dans les textes et par les textes. C'est pourquoi la textualité exerce des contraintes fortes sur l'élaboration des concepts.

D'une manière générale, la production et l'interprétation des textes sont soumises à des contraintes tant linguistiques que socioculturelles. Ainsi, les discours et les genres textuels configurent les textes en constituant des ensembles de règles de production et d'interprétation acquises ou apprises, parfois de manière inconsciente.

Par exemple, les chercheurs en médecine, eux-mêmes médecins, sont susceptibles de produire, à partir du même contenu informationnel, différents discours : le discours scientifique (à l'attention des chercheurs) ; le discours de la presse médicale (à l'attention des praticiens) et le discours de prévention (à l'attention des patients). Ainsi, au syntagme subtantival *« prise de poids »*, on opposera dans certains textes institutionnels « la forme verbale *« grossir »*. Plutôt que « *surcharge pondérale »*, on lira par exemple sur un forum de discussion *« être ronde »*. En bref, les genres textuels organisent différemment la connaissance et à chaque pratique correspondent des genres particuliers. La prévention contre le tabagisme est fortement médicalisée dans les textes institutionnels, elle ne l'est que marginalement dans les forums de discussion dont l'objectif est pourtant identique[3].

Dans les textes spécialisés, le genre choisi sélectionne les concepts et les organise en fonction de contraintes textuelles précises. D'une certaine manière, il décide de son niveau de spécialisation en éliminant certains concepts et en en privilégiant d'autres. Par exemple, un texte médical sur le tabagisme utilisera le concept hyperonymique *tabac* pour *« cigarette »*, *« pipe »*, *« cigare »*, *« narghilé »*, etc. tandis qu'un texte de vulgarisation privilégiera les hyponymes en fonction de leur cible (« *cigarette »*,

[3] Ces observations proviennent d'études réalisées dans le cadre du projet ANR-07-MDCO-002 C-MANTIC destiné à élaborer des méthodologies et des outils pour l'application de la sémantique de corpus au filtrage des masses documentaires.

« tabac à rouler ») et de l'ethos du lecteur supposé (un fumeur de cigarette n'est pas un fumeur de cigare). D'une manière générale, des analyses statistiques révèlent que le texte institutionnel construit un discours distancié, intellectuel quand le texte informel est davantage *incarné* ; en forçant le trait, on peut dire qu'il faut de la *volonté* pour s'arrêter dans un texte institutionnel (c'est-à-dire une faculté intellectuelle) et du *courage* dans un texte informel (c'est-à-dire une actualisation sensible de la volonté)
De ces exemples rapides, on conclura que les textes relèvent d'une planification de l'information. Cette planification est différentielle dans la mesure où les textes explicitent et organisent des connaissances apparentées de manières différentes. On peut en conséquence se risquer à allouer au textuel le statut de *système d'organisation des connaissances*.

4.1 Textes, informations et connaissances différentielles

Pour illustrer notre propos, nous proposons d'étudier brièvement différents discours de prévention contre le tabagisme. Le projet général vise notamment les tabacologues et a pour objectif de mieux connaître les pratiques tabagiques. Pour cela, nous étudierons ici un corpus composé de deux ensembles : (a) un discours institutionnel composé de sites médicaux (*Ligue contre le cancer*), de sites de lobbying (*OFT*) et de site de prévention du tabagisme (*Pataclope*, qui s'adresse aux adolescents) ou d'aide au sevrage (*OFT*) et (b) un discours informel, constitué de blog et de forums contre le tabac, sur le sevrage tabagique (*Atoute*).
Sans entrer dans le détail d'une analyse textométrique qui n'est pas ici notre propos, nous tâcherons dans les paragraphes suivant de proposer des grilles interprétatives générales destinées à mieux circonscrire d'un point de vue linguistique les différences de traitement de l'information et d'organisation ou de production des connaissances, pour une thématique semblable, dans ces deux types de discours. Nous aborderons tour à tour les statuts macroscopiques du texte, de l'information et de la connaissance.

4.1.1 Le statut du texte

	Sites institutionnels	**Blogs et forums**
Statut	Objectif (*« les fumeurs »*)	Subjectif (*« moi je »*)
Zone anthropique	Distal (*« le tabac »*)	Identitaire et proximal (*« une cigarette »*)
Fonction	Exposition	Construction

Figure 1. Statut différentiel des textes des deux sous-corpus

On observe que les sites institutionnels adoptent une perspective qui se présente comme objective. Ils mettent à distance l'objet. Ainsi, les différents actants des textes sont par exemple *« les fumeurs », « le fumeur », « le tabac », « la nicotine »*, des entités abstraites correspondant éventuellement à des positions ontologiques. A l'inverse, les forums privilégient la subjectivité. On y relève de nombreux marqueurs identitaires et de coordonnées spatiotemporelles, tels que les pronoms personnels, des déictiques (*« moi », « je »*). Le tabac ou les substances sont peu actualisées, on leur préfère les objets *« clope »* ou *« cigarette »* qui correspondent à une pratique concrète (après le repas, on fume *une* cigarette, pas *du* tabac). La cigarette, enfin, est un objet personnel qui relève de l'identitaire ou, dans le cas du tabagisme socialisant, du proximal (la relation de soi à l'autre). Le tabac est à l'inverse vu comme une plante, une substance, sa conceptualisation est scientifique donc distale – elle ressortit à une mise à distance.
Enfin, la fonction des textes institutionnels est exposante. Ils exposent les risques liés au tabagisme sur un mode dysphorique (*« cancer »*, *« maladie »*, etc.) tandis que les textes informels sont davantage dans la construction d'un savoir, l'élaboration d'une connaissance à partager.

4.1.2 Le statut de l'information

	Sites institutionnels	**Blogs et forums**
Statut	Sanctionnée Haut niveau	Débattue Bas niveau

Figure 2. Statut différentiel de l'information dans les deux sous-corpus

L'information des sites institutionnels est sanctionnée par le corps médical, elle est dite de « haut niveau », les produits de substitution présentés sont par exemple ceux valider par la recherche médicale (ou pharmaceutique) : *« substitut nicotinique »*, *« patch »* ; *« aide médicale »*, *« consultation tabacologique »*, etc. Dans les forums et les blogues, l'information est débattue, dialectisée, les classes sémantiques produites sont davantage liées à des pratiques de sevrage qu'à des catégories générales. On pourrait y trouver pêle-mêle *« chewing-gum »*, *« coup de fil à une copine »*, *« verre d'eau », « footing »*, etc.). L'information n'est pas sanctionnée et peut-être considérée comme de bas niveau (par exemple, un internaute rapporte avoir recouru au cannabis comme substitut nicotinique).

4.1.3 Le statut des connaissances

	Sites institutionnels	**Blogs et forums**
Statut	Exposées	Produites
Modèle	Ontologique	Praxéologique
Représentation des connaissances	Mots-clés, concepts	Passages-clés, *formes sémantiques*

Figure 3. Statut différentiel des connaissances dans les deux sous-corpus

Dans les textes institutionnels, les connaissances existent en amont de la production textuelles, elles relèvent de savoirs scientifiques, médicaux voire encyclopédiques construit dans d'autres situations énonciatives, d'autres pratiques sociales (par exemple, dans des articles scientifiques). Les connaissances sont des concepts déjà lexicalisés, des termes (qui correspondent à des mots-clés) et la textualité a pour fonction l'exposition et la mise en relation de ces connaissances. Les textes institutionnels « déploient » des ontologies, ou des mondes conceptuels similaires aux connaissances ontologiques. Dans les textes informels, blogs et forums, les connaissances sont *produites* par le texte. Elles n'existent pas préalablement aux textes mais résultent de l'élaboration ou de la collaboration des auteurs qui construisent des connaissances partagées En rendant compte de pratiques tabagiques et de scénarii de sevrage par exemple, les textes relèvent d'une praxéologie. Les connaissances dès lors ne sont pas données comme préalables à la mise en texte, elles sont élaborées par la textualité et n'accèdent pas à proprement parler au statut de concepts, mais de préconcepts ou de connaissances préconceptuelles suivant des modalités textuelles particulières que nous décrirons dans le paragraphe suivant.

4.2 Entre le texte et le concept, la *forme sémantique*

La tradition linguistique et terminologique privilégie le lexique, et plus particulièrement les groupes nominaux, dans la détermination des concepts. Or, la linguistique, depuis Saussure, pose que le versant psychique d'un signe, le *signifié*, ne se confond pas avec le concept. Un concept, au sens linguistique proposé ici, n'est donc pas systématiquement lié à un signe particulier, il peut s'actualiser dans une *forme sémantique*, c'est-à-dire un ensemble de valeurs sémantiques systématiquement cooccurrentes et groupées dans différents textes, relativement stabilisées, mais non nécessairement lexicalisé.
Par exemple, si les mots *« tabac »* et *« choix »* sont en cooccurrence dans un texte et *« fumer »* et *« liberté »* dans un autre, on peut avoir deux fois la même forme sémantique composée minimalement des traits

sémantiques /fumer/ et /liberté/ (à considérer que ces traits sémantiques sont contenus dans les signifiés de ces différentes unités lexicales). Ainsi dans les deux extraits ci-dessous, la forme sémantique /fumer/+/liberté/ est actualisée de façon différentes :

- Opter pour la consommation du tabac$^{/fumer/}$ relève du choix$^{/liberté/}$ personnel de l'individu. (http://www.orinfor.gov.rw/DOCS/Sante47.htm)
- Le citoyen est libre$^{/liberté/}$ de fumer$^{/fumer/}$ ou de ne pas fumer, de manger de la salade si ça lui chante et des rillettes s'il en a envie (http://www.le-tigre.net/Fumer-ne-tue-pas.html)

On pourrait évidemment enrichir cette forme sémantique des traits /personne/ ou /humain/ mais cette cooccurrence simple est en soi suffisante. Il ne s'agit pas d'un concept à proprement parler mais d'une connaissance préconceptuelle non lexicalisée susceptible de se stabiliser. Cette stabilisation peut mener à un figement lexical (par exemple le syntagme *« liberté de fumer »*) ou à la constitution de connaissances communes partagées.

Selon (Rastier 2008), ce que nous appelons ici connaissance préconceptuelle (ou *préconnaissance*) constitue une connaissance :

> Une connaissance est un ensemble de passages de textes (éventuellement multimédia) : dans leurs récurrences, le contenu de ces passages (les fragments) et leurs expressions (les extraits) sont en relation de transformation, ne serait-ce que par changement de position.
>
> Résultant de figements et de réductions de syntagmes, les mots sont une sorte très particulière de ces passages, et comme les autres passages, ils restent impossibles à interpréter sans recontextualisation.
>
> En somme, la connaissance est issue d'une décontextualisation de certaines formes sémantiques saillantes et des expressions qui leur correspondent.

On peut donc avancer qu'un concept est une forme sémantique lexicalisée. Mais un concept, au sens textuel que nous défendons ici, ne correspond pas forcément à une unité lexicale. La lexicalisation d'une forme sémantique, qui aboutit à la formation du concept, ne doit pas être envisagée exclusivement comme sa naissance, ni même comme l'aboutissement de la conceptualisation. Elle s'apparente davantage à un état de stabilisation provisoire, correspondant à un usage circonscrit d'un point de vue socioculturel et temporel.

5 Conclusion

En tant que Système d'Organisation des Connaissances, les textes ne permettent pas de niveler les connaissances ni des les laisser indifférenciées. Les ontologies, en globalisant les connaissances, les désituent, au détriment de la pertinence.

Par ailleurs, le passage du texte au document renforce la dimension sémiotique. C'est la congruence des dimensions sémantique et sémiotique qui est à même de susciter de nouveaux débats féconds sur des méthodes alternatives de constitution de connaissances à partir des textes.
Jusqu'ici, l'ingénierie des connaissances et la terminologie textuelle notamment se sont plus particulièrement consacrées à l'extraction de candidats termes dans les textes pour les expertiser et les valider ou non comme concepts ou termes. Elles se sont peu intéressées en revanche à l'émergence de ces concepts dans les textes dès lors que l'on admet que les textes en sont les lieux de production et pas seulement ceux de leur exposition. Nous faisons l'hypothèse qu'avant d'accéder au statut de signes dont les signifiés sont normés (les termes), les concepts émergents se manifestent dans les textes comme formes sémantiques qui se coaguleront ou non en unités lexicales nouvelles et en termes. On peut dès lors considérer que ces formes sémantiques ont valeur de préconcepts. L'enjeu pour la linguistique est d'être capable de décrire et de formaliser ces formes sémantiques et de les requalifier en zones de pertinence. Ce processus d'émergence intéresse d'un point de vue théorique la terminologie et, au plan pratique, l'identification et de détection pour la veille et la constitution de terminologies.

Remerciements : Nous remercions toute l'équipe du projet C-MANTIC, ainsi que Manuel Zacklad et Alain Giboin qui ont sollicité ce débat en organisant l'atelier *Modèles, méthodes, pratiques pour la conception de logiciels basés sur des SOC* à la plateforme AFIA 2009.

6 Références bibliographiques

Barlow, J.P., 1994, *A taxonomy of information*, in Bulletin of the American Society for Information Science, 20, 13-17

Coates, E.J., 1978, *Classification in Information Retrieval : Headings and Structure.* London, Library Association

Carnap, R., 1928, *La Construction logique du monde*, trad. fr. Elisabeth Schwarz et Thierry Rivain. Paris : J. Vrin, 2002

Carnap, R., 1934, *La syntaxe logique du langage*, trad.

Ertzscheid, O., 2008, *Indexation sociale et folksonomies: le monde comme catalogue*, Journées ABES.Montpellier, 20 et 21 mai 2008, http://wwwaffordance.info

Origgi, G., 2008, *Le sens des autres. L'ontogenèse de la confiance épistémique* in Raisons Pratiques n°17 – L'épistémologie sociale, Paris, ed.CNRS

Rastier, F., 1995, *Le terme : entre ontologie et linguistique*, in La banque des mots

Rastier, F., 2002, « Anthropologie linguistique et sémiotique des cultures », in Rastier, F. et Bouquet, S. (éds.), *Une introduction aux sciences de la culture*, Paris, PUF.

Rastier, F., 2007, « Passages », *Interprétation, contextes, codage*, B. Pincemin (éd.), Corpus, 6, 125-152.

Rastier, F., 2008, « Sémantique du Web vs Semantic Web ? Le problème de la pertinence », *Textes, documents numériques, corpus. Pour une science des textes instrumentée,* M. Valette (éd.), *Syntaxe & Sémantique*, n°9, 15-35.

Schlick, 1932, *Forme et contenu. Une introduction à la pensée philosophique*, Agone, Paris.

Valette, M., Slodzian, M., 2008, « Sémantique des textes et Recherche d'information », *Extraction d'information : l'apport de la linguistique*, A. Condamines & Th. Poibeau (éd.), *Revue Française de Linguistique Appliquée*, volume XIII-1 – juin 2008, 119-133.

Wolfram, S., 2001, *A new kind of Science*, Wolfram Media Inc.

Zacklad et al., 2007, Hypertopic: une métasémiotique et un protocole pour le Web socio-sémantique, Francky Trichet (Eds), Cépaduès

Une approche générale pour l'extraction de lignes des documents Arabes anciens multi-orientés

Nazih Ouwayed, Abdel Belaïd

Université Nancy 2, LORIA, Équipe READ, Vandœuvre-Lès-Nancy, France

Mots-clés : documents Arabes manuscrits, extraction de lignes, estimation de l'orientation, Snake, distribution de Wigner-Ville, chevauchement et connexion de lignes.

Keywords: handwritten Arabic documents, text line extraction, orientation estimation, Snake, Wigner-Ville distribution, overlapping and touching lines.

Résumé : Dans cet article, nous présentons une nouvelle approche pour l'extraction de lignes des documents Arabes anciens multi-orientés. En raison de la multi-orientation de lignes et de leur dispersion dans l'image, nous utilisons un maillage automatique de l'image qui nous permet de déterminer progressivement et localement les lignes. Le maillage est initialisé avec une petite fenêtre où sa taille est corrigée par extension jusqu'à ce que suffisamment de lignes et de composantes connexes ont été trouvées. Nous utilisons le Snake pour l'extraction de lignes. Une fois le document est divisé en fenêtres, l'orientation est déterminée en utilisant la distribution de Wigner Ville (DWV) sur l'histogramme de projection. Ensuite, cette orientation locale est élargie pour limiter l'orientation dans les fenêtres voisines. Ensuite, les lignes de texte sont extraites localement dans chaque zone en se basant sur le suivi des lignes de base et la proximité des composantes connexes. Enfin, les composantes connexes qui se chevauchent et se connectent dans les lignes adjacentes sont séparées en considérant la morphologie des lettres terminales des mots Arabes. L'approche proposée a été expérimentée sur 100 documents atteignant une précision d'environ 97.6%.

Abstract: This paper presents a novel approach for the multi-oriented text line extraction from historical handwritten Arabic documents. Because of the multi-orientation of lines and their dispersion in the page, we use an image paving allowing us to progressively and locally determine the lines. The paving is initialized with a small window and then its size is corrected by extension until enough lines and connected components were found. We use the Snake for line extraction. Once the paving is established, the orientation is determined using the Wigner-Ville distribution (WVD) on the histogram projection profile. This local

orientation is then enlarged to limit the orientation in the neighbourhood. Afterwards, the text lines are extracted locally in each zone basing on the follow-up of the baselines and the proximity of connected components. Finally, the connected components that overlap and touch in adjacent lines are separated. The morphology analysis of the terminal letters of Arabic words is here considered. The proposed approach has been experimented on 100 documents reaching an accuracy of about 97.6%.

1 Introduction

La segmentation du texte en lignes est vue comme une étape nécessaire dans le domaine d'analyse de document avant la reconnaissance des mots. La difficulté de cette tâche vient des caractéristiques des documents manuscrits et particulièrement lorsqu'ils sont anciens (voir Figure 1).
Ces documents présentent des espacements irréguliers entre les lignes et des fluctuations de la ligne de base de l'écriture par rapport à l'horizontale. Les lignes ont des longueurs différentes et peuvent se chevaucher ou se connecter lorsque leurs hampes et leurs jambages appartiennent à deux lignes consécutives. En outre, les lignes sous forme d'annotations peuvent exister dans les marges. Ces lignes sont en général obliques en raison de la réduction de l'espace qui constitue de nouvelles orientations. Par exemple, la Figure 1.a contient 4 annotations. La présence massive des points diacritiques complique en plus cette tâche. Également, l'écriture manuscrite arabe présente de grandes variations, dans les formes des lettres ou des mots, et dans la mise en page.

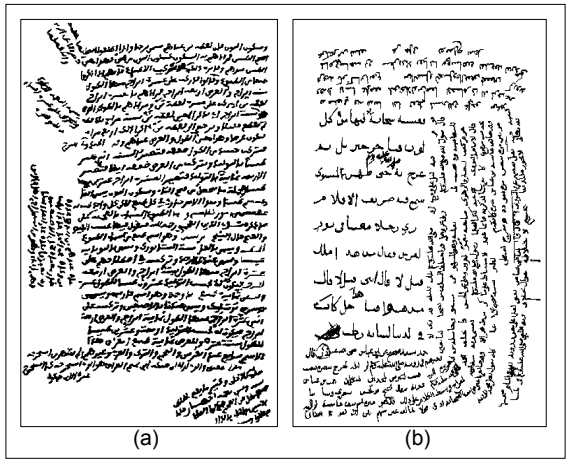


Figure 1: Exemples des documents multi-orientées.

L'article est organisé comme indiqué ci-après. Dans la Section 2, nous présentons les techniques existantes pour l'extraction des lignes. Les différentes étapes de notre approche pour l'extraction de lignes multi-orientées dans les documents arabes manuscrits sont détaillées dans la Section 3. Nous présentons dans la Section 4 nos résultats expérimentaux et une conclusion dans Section 5.

2 État de l'art

Dans la littérature, deux classes d'approches existent pour l'extraction de lignes : descendant et ascendant. Les approches descendantes sont essentiellement basées sur la technique de la projection. Dans [9, 16, 23], la projection horizontale est appliquée au document entier ou à une bande (horizontale ou verticale) de celui-ci. Ensuite, les maxima et les minima sont déterminés et les composantes connexes entre deux minima consécutifs sont recherchées. Ces composantes connexes forment la ligne.

Les approches ascendantes sont basées sur le bas niveau (pixel ou composante connexe). Dans cette catégorie, nous trouvons la classification par k-ppv (plus proches voisins), la transformée de Hough, la technique du lissage (smearing), et la technique de répulsif-attractif. Dans la classification par k-ppv [12], les alignements sont détectés en choisissant les composantes connexes, prolongées dans des directions spécifiques. Puis, les lignes sont extraites en groupant ces alignements selon trois critères : proximité, similitude et continuité de direction. Dans la transformée de Hough [13, 20], les points votants sont les centres de gravités des composantes connexes. Un ensemble de points alignés dans l'image ayant un pique dans la transformée de Hough, représente une ligne. Dans la technique de lissage [8, 21], RLS (Run-Length Smoothing), les pixels noirs consécutifs sur la direction horizontale sont lissés. Les boîtes qui englobent les composantes connexes dans l'image lissée forment les lignes. Dans la technique de répulsif-attractif [18], les lignes de base qui représentent les forces attrayantes, sont construites une par une à partir du fond de l'image jusqu'au début. Dans l'image les pixels représentent les forces répulsives. Les forces répulsives-attrayantes entre deux lignes de bases consécutives sont étudiées pour déterminer les lignes du document. D'autres méthodes utilisent le principe de regroupement par l'arbre couvrant minimal (MST : Minimum Spanning Tree). Dans [5], Yin et Liu ont proposé une méthode fondée sur le regroupement par MST avec l'apprentissage de distance. Compte tenu de la distance métrique, les composantes connexes de l'image sont

regroupées dans une structure arborescente. Les lignes de texte sont extraites de l'arbre en coupant ses bords avec une fonction objective. Pour une description plus détaillée de ces méthodes les lecteurs sont invités à lire [14].

Toutes les approches citées ci-dessus sont soit trop globales, procédant par projection ou par recherche d'alignements, soit trop locales, opérant par suivi de composantes connexes. Elles trouvent ici leurs limites face à la mauvaise qualité et à la multi-orientation des documents. La méthode que nous présentons dans ce papier se base sur le maillage automatique de l'image, la détection des zones multi-inclinées et l'extraction locale des lignes.

3 Approche proposée

L'écriture dans les documents arabes manuscrits anciens est multi-orientées, tordue et bruitée. Pour cette raison, nous avons décidé de nous concentrer sur une méthode locale qui divise d'abord l'image en plusieurs fenêtres. Puis, elle estime l'orientation dans chaque fenêtre. Ensuite, elle applique des règles de correction et d'extension de l'orientation afin de trouver toutes les zones (une zone est un ensemble des fenêtres qui ont la même orientation) multi-inclinées. Enfin, elle suit les lignes de base pour extraire les lignes du document. Les lignes qui se chevauchent et se connectent sont séparées après dans une étape de post-traitement.

3.1 Maillage automatique

Dans cette étape, le document est partitionné en petites fenêtres de taille (w×h). Cette taille est automatiquement généré en se basant sur l'idée qu'une fenêtre doit contenir environ 3 lignes pour être capable de produire un histogramme de projection représentatif de l'orientation. Cela fait suite à plusieurs étapes. Tout d'abord, une première fenêtre de taille arbitraire (15×15 pixels) est placée au milieu de l'image (voir Figure 2.a). Ensuite, l'approche du Snake est appliquée pour calculer les lignes (voir les explications ci-après). La largeur de la fenêtre est agrandie jusqu'à ce que le Snake se donne au moins 3 lignes. Une fois les lignes sont trouvées, la hauteur moyenne $\bar{h}$ est estimée ainsi que la distance moyenne entre les lignes $\bar{g}$. La distance entre les lignes est estimée en utilisant l'enveloppe convexe des lignes [15] (voir Figure 2.b). La fenêtre finale a une taille qui est égale à (w×h), où $w = h = 3 \times \bar{h} + 2 \times \bar{g}$ (voir Figure 2.c).

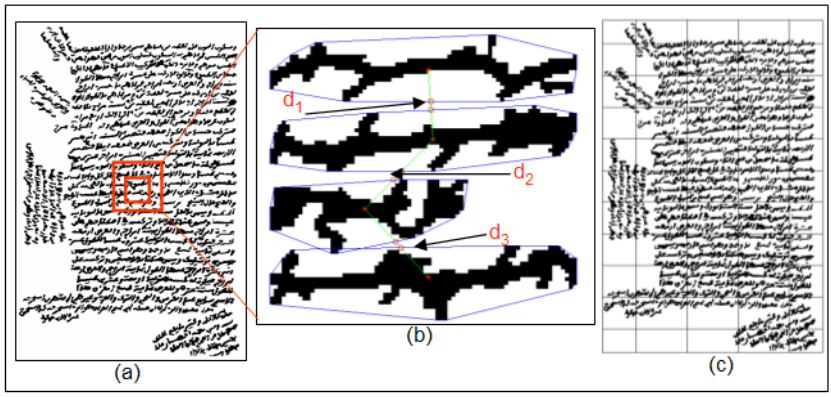


Figure 2: Algorithme de Maillage automatique (d1, d2 et d3 sont les distances entre les lignes).

Modèle de contour actif (Snake) : Le Snake est défini comme un contour virtuel qui va être emmené vers un contour réel dans l'image en utilisant un mécanisme de minimization d'énergie avec une méthode itérative qui le déforme [11]. Le Snake est un ensemble de points v(s)=[x(s), y(s)], où s $\in$[0,1]. Le contour final peut être obtenir en minimisant la fonction d'énergie suivante :

$$E_{snake} = \int_0^1 E_{\text{int}}(v(s))ds + \int_0^1 E_{ext}(v(s))ds + \int_0^1 E_{cont}(v(s))ds \qquad (1)$$

où E_{int} est l'énergie interne du Snake qui sert à la régularization du Snake (forme, convexe, etc..), E_{ext} est l'énergie externe de l'image qui pousse le Snake vers les lignes, les contours des objets qui se trouvent dans l'image et E_{cont} est l'énergie de contexte qui exprime certaines contraintes supplémentaires qui peuvent être imposées par l'utilisateur vu le Snake qu'il veut obtenir.

L'énergie externe traditionnelle est située sur les contours externes. Cela oblige à initialiser le Snake à proximité du contour cible. En outre, les valeurs du gradient ont des sens inverse sur les deux côtés du même contour, qui empêche le Snake d'entrer dans les concavités. Pour cette raison, Xu et al. [22] ont développé un nouveau type d'énergie externe qui permet d'initialiser le Snake loin du contour cible et d'aller vers les concavités. Cette énergie est nommé GVF (Gradient Vector Flow). GVF est défini comme un vecteur V(x, y) = (u(x, y), v(x, y)) qui minimise la fonction d'énergie suivante :

$$\varepsilon = \iint \mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + |\nabla f|^2 |V - \nabla f|^2 \, dxdy \qquad (2)$$

V peut être trouvé par les deux équations d'Euler :

$$\mu\nabla^2 u - (u - f_x)(f_x^2 + f_y^2) = 0$$

$$\mu\nabla^2 v - (v - f_y)(f_x^2 + f_y^2) = 0$$

où μ est un paramètre de réglage de bruit. Pour trouver *u* et *v,* Xu et al. proposent dans [22] plusieurs implémentations numériques. Dans notre application, nous avons utilisées la GVF sur l'axe majeure de la composante connexe dans chaque fenêtre (voir Figure 3).

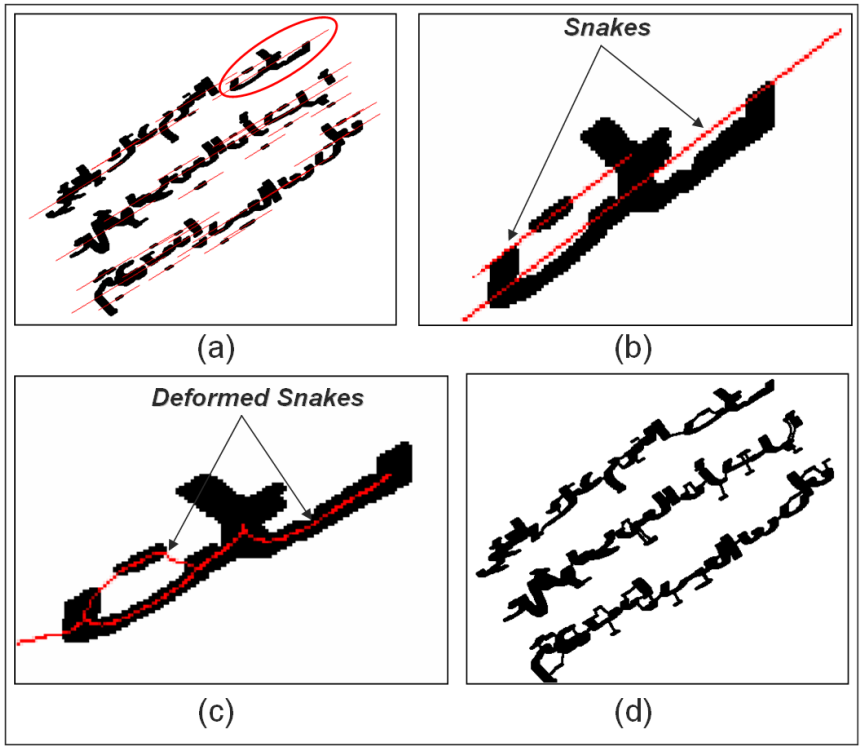


Figure 3. Application du Snake pour la détection de lignes, (a) L'axe majeur tracé pour chaque composante connexe des lignes. L'ellipse englobe la première composante connexe de (b), (c) montre la déformation de Snake (b), (d) donne le résultat final indiquant les composantes connexes groupées dans chaque ligne.

3.2 Détection des zones multi-orientées

Cette étape consiste à extraire la première orientation dans les fenêtres et les étendre vers les fenêtres voisines. Ces deux étapes se déroulent comme suit.

3.2.1 Estimation de l'orientation dans la fenêtre

Traditionnellement, l'estimation de l'orientation est faite en analysant l'histogramme de projection [9, 16, 23] ou par une autre méthode comme la transformée de Fourier [19]. Cependant, ce découpage en petites fenêtres peut parfois favoriser les orientations individuelles des mots à celle des lignes. Nous corrigeons ce biais en utilisant une distribution énergétique de Cohen [6, 7, 10] sur le profil de projection. Cette distribution réagit mieux que la projection aux pics engendrés par les lignes en traduisant leur présence en forte énergie. En effet, l'histogramme de projection crée des faux maxima qui perturbent l'extraction de l'orientation. Nous nous sommes limités à la distribution de Wigner-Ville (DWV) dont les propriétés permettent d'être plus réactive à ces présences de pics que les autres distributions de la classe de Cohen [4].
La DWV est définie comme la transformée de Fourier du signal représentant le profil de projection. Ce signal est donné par : $x(t+\tau/2)\, x^*(t-\tau/2)$ où τ exprime le retard et montre que la distribution est invariante aux translations temporelles et fréquentielles, ce qui peut tolérer le décalage du profil dans les fenêtres. La valeur de DWV est :

$$W_x(t,\nu) = \int_{-\infty}^{+\infty} x(t+\tau/2)x^*(t-\tau/2)e^{-j2\pi\nu\tau}\, d\tau \tag{3}$$

La DWV a quelques propriétés intéressantes par rapport aux autres distributions, comme la production de valeur réelle, facilitant son interprétation :

$$\int\int_{-\infty}^{+\infty} W_x(t,\nu)\, dtd\nu = \int_{-\infty}^{+\infty} |x(t)|^2\, dt \tag{4}$$

Pour estimer l'angle d'inclinaison, on projette la fenêtre suivant de multiples angles [-75° ; +90°] avec un pas de +15° donnant à chaque fois un signal x(t). Ensuite, la DWV est appliquée à chaque x(t) pour déterminer l'intensité d'énergie. L'angle correspondant au profil de projection ayant l'intensité d'énergie la plus élevée est choisi comme angle d'orientation (voir Figure 4, l'angle exact est +15° qui correspond au profil de projection qui a les pics les plus aigus et les vallées le plus creusées). Figure 7.b montre la première estimation de l'orientation pour le document dans la Figure 1.a.

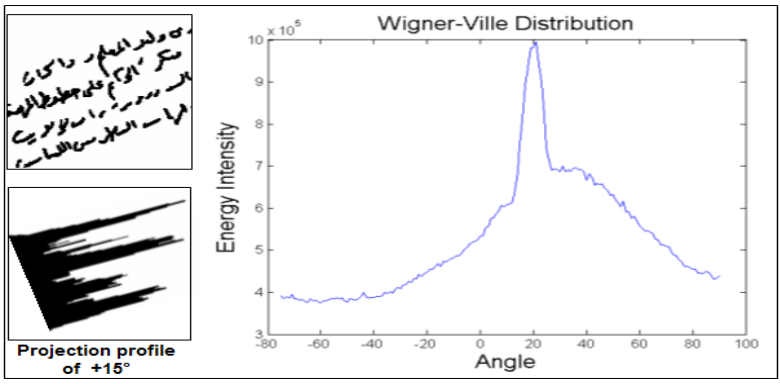


Figure 4: Distribution des valeurs d'énergie de la fenêtre à gauche.

3.2.2 Correction et extension de l'orientation

Deux cas peuvent arriver après l'estimation de l'orientation. Le premier cas arrive lorsque l'orientation au sein d'une fenêtre n'est pas franche, alors la fenêtre est découpée suivant la vallée de projection la plus profonde. Ensuite, chaque orientation de la fenêtre est revue en fonction de l'orientation des fenêtres voisines, prises suivant les règles d'écriture de l'Arabe : Est-Ouest, Est-NordOuest, Est-SudOuest, Sud-Nord et Nord-Sud. Soit, l'orientation est confirmée et étendue aux deux fenêtres, soit gardée identique dans chaque fenêtre. Cette opération est suivie d'une étape de correction du fenêtrage qui consiste à décaler les limites de fenêtres suivant l'orientation principale, de manière à ne pas intercepter les lignes d'écriture (voir Figure 7.b).

3.3 Extraction de lignes

En nous basant sur l'orientation dans les fenêtres, nous recalculons la projection par rapport à cette orientation, puis on procède à la recherche de nouveaux maxima (voir Figure 5.a). Chaque maxima représente le point de départ d'une ligne Ps, à partir duquel nous suivons la ligne de base blj en respectant l'angle de l'orientation. Le suivi commence dans la première fenêtre dans le coin droite du document. Le point d'arrivé Pe de la ligne de base est calculé en utilisant, l'angle, la largeur et la longueur de chaque fenêtre (voir Figure 5.b). La ligne de base blj est calculée en se fondant sur les deux points (Ps, Pe) et sur l'orientation de la fenêtre.
Pendant ce suivi, les composantes connexes qui appartiennent à une ligne de base sont recherchées pour former les lignes dans chaque fenêtre (voir Figure 5.c).

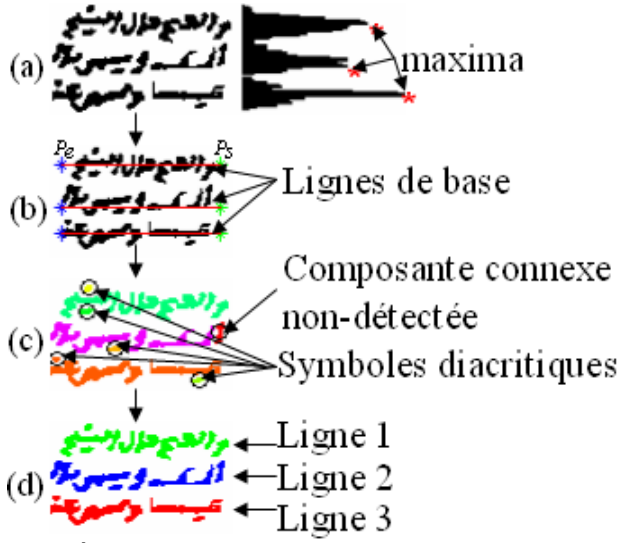


Figure 5: Étapes de la détection d'une ligne dans une fenêtre.

Une étape de correction de la détection suit cette étape pour attribuer les composantes connexes non détectées et les symboles diacritiques à la ligne appropriée (voir Figure 5.c et Figure 5.d). Une méthode de distance est utilisée pour résoudre ce problème. D'abord, la distance entre le centre de gravité de la composante non détectée ou symbole diacritique (C_i) et la ligne est calculée. C_i est attribué à la ligne l_j si $d_{ci,lj} < d_{ci,lj+1}$ sinon à l_{j+1} (voir Figure 6).

Pour chaque zone, les lignes sont groupées pour former les lignes de zone. Les relations entre les lignes de zones sont alors étudiées pour composer les lignes de document (voir Figure 7.c). Pour cela, nous regardons si une composante connexe appartient à deux lignes dans deux zones différentes. Si c'est le cas, nous fusionnons les deux lignes en une ligne.

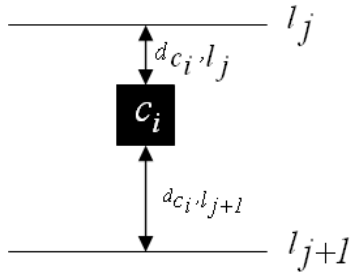


Figure 6: Attribution des composantes connexes non détectées et des symboles diacritiques.

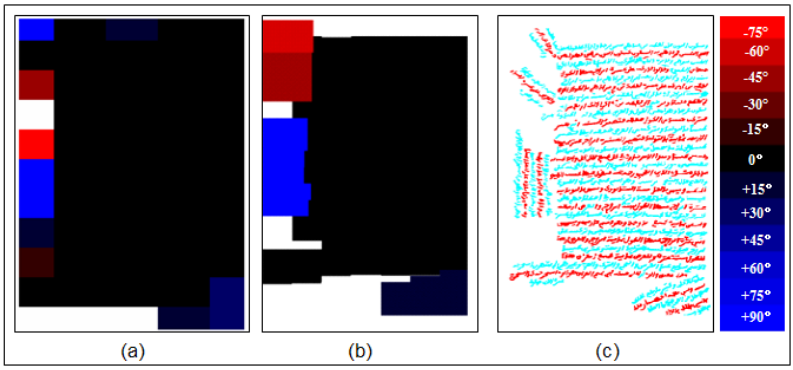


Figure 7: Etapes d'extraction des lignes.

3.4 Séparation de lignes connectées

La dernière étape de la méthode est relative à la séparation des lignes adjacentes. La connexion est relativement fréquente dans les documents manuscrits Arabe souvent favorisée par l'extension des lettres terminales de la ligne supérieure avec les hampes des lettres de la ligne inférieure (voir Figure 10.a).

Là encore, nous avons privilégié une technique prenant en compte la morphologie de l'écriture Arabe. Deux configurations de ligatures sont d'abord identifiées suivant l'orientation des boucles des lettres terminales : boucle ouverte vers la gauche (cas du RA, WAW, NOUN, etc.) et boucle orientée vers la droite (cas du HA, KHA, AIN, etc.) (voir Figure 8). Chacune peut se ligaturer soit avec un allographe vertical (cas du ALEF, LEM, KEF, etc.), soit avec un allographe avec boucle (cas du TA, SA, HA, etc.). Nous cherchons à reconstituer les terminales en suivant les tracés ayant une continuité, i.e. la même variance angulaire.

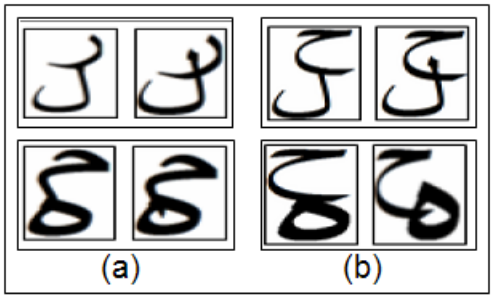


Figure 8: Configurations des ligatures dans l'Arabe.

La technique démarre des points d'intersection trouvés entre les lignes d'écriture adjacentes. L'analyse se concentre dans une fenêtre autour de

chaque point (Sp). On suit ensuite le tracé à partir du point le plus haut de la fenêtre (Bp), et on le prolonge au-delà du point d'intersection avec le tracé qui a la même variance angulaire que lui. Dans le cas de la Figure 10, le tracé d'étiquette 1 se poursuit avec celui d'étiquette 3, correspondant à la lettre RA, puis le 2 va avec le 4 : cas du ALEF. La Figure 10.b montre les lignes séparées suivant des couleurs différentes (pour des détails [17]).

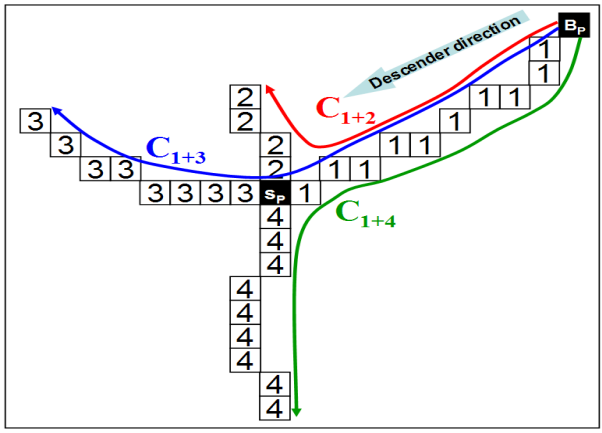


Figure 9 : Suivi des terminales des lettres ligaturées : le RA qui intercepte la lettre ALEF.

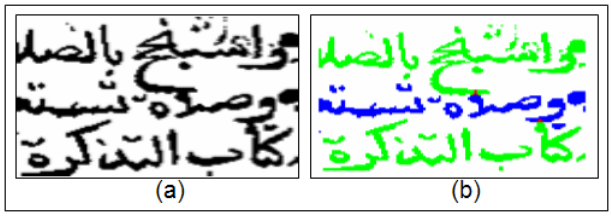


Figure 10 : Résultas de l'algorithme du séparation des lignes connectées.

4 Résultats expérimentaux et discussion

Pour étudier l'efficacité de notre approche, nous l'avons testée sur 100 documents arabes manuscrits anciens qui contiennent 2500 lignes. Ce sont des manuscrits de la bibliothèque nationale de Tunisie [1], la bibliothèque nationale de médicine aux États-Unis [2] et la bibliothèque et archives de l'Égypte [3]. Les essais ont été préparés après un calcul

manuel de zones et de lignes de chaque document, l'angle de rotation examiné pendant ces expériences varie du -75° jusqu'au +90°. Le temps d'exécution est mesuré à partir du maillage jusqu'à la séparation de lignes. Il dépend de la taille du document et de la taille de la fenêtre du maillage. Les essais ont été effectués sur un PC équipé d'un microprocesseur Pentium M de 1.4 GHz et une mémoire cache de 1 GO sous Windows XP. L'application a été développée avec MATLAB R2009b. Dans la détection de zones multi-inclinées, nous avons obtenu un pourcentage de détection autour de 96%, qui passe à 98% si nous ne prenons pas en considération les petites zones non-détectées. Le taux d'erreur de 2% est dû au maillage et à la fausse inclinaison. Dans le niveau de la segmentation en lignes, nous avons eu un taux d'extraction de 97.6%. Le 1.5% de lignes non détectées est dû au l'algorithme de la détection de zones. Le taux d'erreur de 0.9% est dû à la présence des symboles diacritiques dans le début de lignes qui créent des faux maxima. La Figure 11 illustre l'efficacité de notre algorithme sur un échantillon de 4 documents choisis arbitrairement parmi les 100 documents traités. Pour identifier les lignes, chaque paire consécutive de lignes sont présentées par deux couleurs différentes. Le Tableau 1 décrit les résultats de notre méthode.

	Extraites	**Non-Extraites**	**Erreur**
Zones	96 %	2 %	2 %
Lignes	97.6 %	1.5 %	0.9 %

Tableau 1: Résultats de notre méthode.

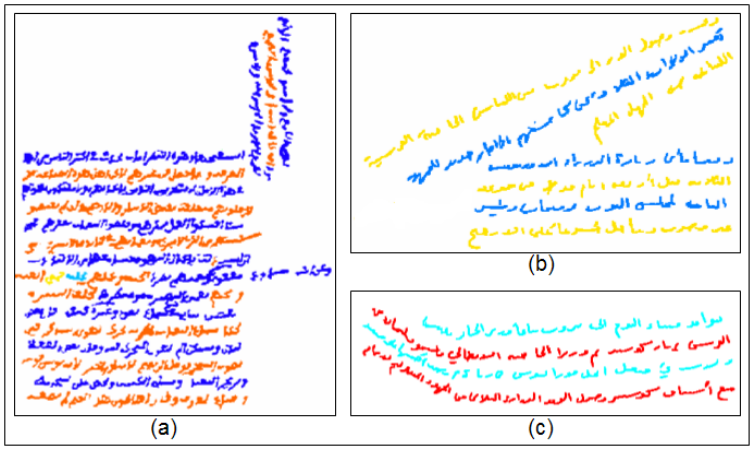


Figure 11 : Quelques résultats de l'approche de l'extraction de lignes multi-orientées.

5 Conclusion et perspectives

Nous avons proposé dans cet article une approche originale, qui vise à extraire les lignes multi-orientées dans les documents arabes manuscrits anciens. Au début, les zones multi-inclinées sont détectées en utilisant un maillage automatique du document. Après, l'orientation est estimée, corrigée et étendue afin de trouver toutes les orientations locales en utilisant la distribution de Wigner-Ville sur l'histogramme de projection. Ensuite, les lignes sont extraites en se basant sur l'orientation et les lignes de base de chaque fenêtre. Enfin, les lignes adjacentes connectées sont séparées en utilisant des informations statistiques sur la morphologie des lettres terminales Arabes. Le taux d'extraction de 97.6% montre l'efficacité et la performance de notre approche. La prochaine étape dans notre travail sera la généralisation de l'approche pour d'autres scripts comme (Latin, Urdu etc.) et l'application sur des documents contient un mélange des textes et des images.

Remerciements
Nous tenons à remercier la Bibliothèque Nationale de Tunis,la Bibliothèque Nationale de médecine aux U.S.A, l'Archives et Bibliothèque Nationales d'Egypte qui ont bien voulu mettre à disposition du public leurs documents et qui ont constitués notre ressource essentielle de données.

6 Références bibliographiques

[1] http://www.bibliotheque.nat.tn/

[2] http://www.nlm.nih.gov/hmd/arabic/welcome.html

[3] http://portal.unesco.org/ci/photos/showgallery.php/cat/559.

[4] L. Cohen. Generalized phase-space distribution functions. J. Math. Phys., 7(5):781-786, 1966.

[5] F. Yin, C. Liu. Handwritten text line segmentation by clustering with distance metric learning. In Proc. 11th ICFHR, pages 229-234, 2008.

[6] P. Flandrin. Time-Frequency/Time-Scale Analysis. Academic Press, San Diego, CA, 1999.

[7] P. Flandrin and W. Martin. A general class of estimators for the Wigner-Ville spectrum of nonstationary processes, a. bensoussan, j. l. eds. lions. systems Analysis and Optimization of Systems, Lecture Notes in Control and Information Sciences, 62:15-23, 1984.

[8] B. Gatos, A. Antonacopoulos, and N. Stamatopoulos. Handwriting segmentation contest. In ICDAR '07: Proceedings of the Ninth

International Conference on Document Analysis and Recognition, pages 1284-1288, 2007.

[9] A. Hashizume, P. S. Yeh, and A. Rosenfeld. A method of detection the orientation of aligned components. Pattern Recognit. Lett., 4:125-132, 1986.

[10] F. Hlawatsch and G. F. Boudreaux-Bartels. Linear and quadratic time-frequency signal representation. IEEE Signal Process. Mag., 9(2) :21-67, 1992.

[11] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. Proc. 1st ICCV, pages 259-268, June 1987.

[12]L. Likforman-Sulem and C. Faure. Extracting lines on handwritten documents by perceptual grouping, in advances in handwiting and drawing: multidisciplinary approach. C. Faure, P. Keuss, G. Lorette, A. Winter (Eds), pages 21-38, 1994.

[13] L. Likforman-Sulem, A. Hanimyan, and C. Faure. A Hough based algorithm for extracting text lines in handwritten document. In Proc. of ICDAR'95, pages 774-777, 1995.

[14]L. Likforman-Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: a survey. 9(2) :123-138, 2007.

[15]U. Mahadevan and R. C. Nagabushnam. Gap metrics for word separation in handwritten lines. In ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1), page 124-127, 1995.

[16]G. Nagy, S. Seth, and M. Viswanathan. A prototype document image analysis system for technical journals. Computer, 25:10-22, 1992.

[17]N. Ouwayed and A. Belaïd. Separation of overlapping and touching lines within handwritten Arabic documents. In the 13th International Conference on Computer Analysis of Images and Patterns (CAIP'2009), to appear in September 2009.

[18]E. Oztop, A. Y. Mulayim, V. Atalay, and F. Y. Vural. Repulsive attractive network for baseline extraction on document images. Signal Processing, 75 :1-10, 1999.

[19]W. Postl. Detection of linear oblique structures and skew scan in digitized documents. In Proceedings of the Eighth International Conference on Pattern Recognition, IEEE CS Press, Los Alamitos, CA, pages 687-689, 1986.

[20]Y. Pu and Z. Shi. A natural learning algorithm based on hough transform for text lines extraction in handwritten document. pages 637-646, 1998.

[21]Z. Shi and V. Govindaraju. Line separation for complex document images using fuzzy run length. In Int. Workshop on Document Image Analysis for Librairies, 2004.

[22]C. Xu and J. L. Prince. Gradient vector Flow: A new external force for snakes. Proc. IEEE Conf. on Comp. Vis. Patt. Recog. (CVPR), pages 66-71, June 1997.

[23]A. Zahour, L. Likforman-Sulem, W. Boussellaa, and B. Taconet. Text line segmentation of historical Arabic documents. In 9th Int. Conf. on Document Analysis and Recognition, pages 138-142, 2007.

Intertextual semantics generation for structured documents: a complete implementation in XSLT

Yves MARCOUX

GRDS – EBSI – Université de Montréal

Mots-clés : sémantique intertextuelle, documents structurés, langages de balisage, XML, XSLT, descriptions formelles de jeux de balises

Keywords: intertextual semantics, structured documents, markup languages, XML, XSLT, formal tag-set descriptions

Résumé : La sémantique intertextuelle (SI) [1][4] attribue aux documents balisés un sens en *langue naturelle*. Alors que les sémantiques formelles visent une représentation du sens des documents pour la machine, la SI vise l'humain. Dans la forme actuelle de l'approche, la SI d'un modèle (DTD, schéma) est donnée par deux *péritextes* associés à chaque élément: un *texte-avant* et un *texte-après*. La SI d'un document est la concaténation des péritextes et des contenus d'élément dans l'ordre du document. Nous présentons une implantation complète, en XSLT 1.0, de la génération de SI. L'implantation traite les attributs tel que décrit dans [2], et les hyperliens et éléments locaux tel que décrit dans [1]. Elle indente aussi l'extrant pour une meilleure lisibilité tel que suggéré dans [3] et gère les exceptions que sont les éléments et attributs inconnus.

Abstract : Intertextual semantics (IS) [1][4] is a framework in which the meaning of marked-up documents is given in *natural language*. While formal semantics aims at conveying the meaning of documents to machines, IS aims at conveying it to humans. In the current framework, the IS of a model (DTD, schema) is expressed by *peritexts* associated with each element: a *text-before* segment and a *text-after* one. The IS of a document is obtained by concatenating peritexts and element contents in document order. We present a complete implementation, in XSLT 1.0, of the IS generation mechanism. The implementation handles attributes as described in [2], and hyperlinks and local element definitions as described in [1]. It also indents its output for increased readability as suggested in [3] and handles unknown element or attribute exceptions.

1 Introduction

In a structured document (XML, SGML, etc.), what is the "meaning" of the various tags (the *markup*) present in the document? How is the meaning of the document augmented—or otherwise affected—by the presence of markup?

Fundamentally, there are two possible avenues to give an answer to that question: the formal one and the informal one. One can devise a framework in which the meaning of a marked-up document is represented by a set of *formal* statements, for example in first-order logic. Or, one can seek a framework in which the meaning of a marked-up document is represented by a set of sentences in an *informal* language, for example a natural language.

If automatic inferencing (through an inference engine) is aimed at, then a formal approach probably has a leading edge. However, if some other use of the "meaning" of the document is envisioned, which for example involves showing that meaning to humans, then the situation may be reversed.

Formal Tag-Set Descriptions (see for example [6], [7] and [8]) are an example of the approaches along the formal avenue. *Intertextual semantics* [1][2][4] is an approach along the informal avenue. In intertextual semantics (IS), the meaning of a marked-up document is entirely and exclusively represented *in natural language*.

The intertextual semantics (IS) approach is based on the hypothesis (of which traces can be found in, among other places, the works of Wirzbicka [9], Smedslund [5], and even Wittgenstein [10]) that humans ultimately "make sense" of artefacts through the use of *natural language* (NL), and that in designing artefacts, one should be preoccupied by how, and how easily and with how much ambiguity (or unambiguity), humans can derive NL from those artefacts. No matter how useful intermediate formal representations of meaning (including marked-up documents) may be for conciseness, machine processing, etc., they must ultimately be translatable (not necessarily translat*ed*) to NL, and are ever only as "meaningful" as such NL expressions of them are.

In the realm of structured (i.e., marked-up) documents, IS suggests that the creators of tag-sets (modelers) should be preoccupied by how markup can be translated to NL. Even if "end users" never see any marked-up document, some other humans, for example, processing software developers, or archivists, will have to deal with them directly or indirectly, unless the documents are totally pointless. One might say it is even more important to be preoccupied by that translation as the number of intermediate representations increases, because there are then more opportunities for misinterpretations.

IS proposes a mechanism by which NL passages (or whole documents) are generated from marked-up documents, according to an *IS specification* for the tag-set. So far, only very weak NL generation mechanisms have been explored, *and it is extremely important that those mechanisms be weak*, because too powerful mechanisms would "hide under the carpet" inherent interpretation complications which IS, in contrast, seeks to uncover. In the current state of the IS framework, an IS specification takes the form of a table giving, for each element type two NL segments: a "text-before" segment and a "text-after" segment, generically called "peritexts."

Attributes require special attention, but a way of handling them in keeping with the spirit of IS is presented in [2]. They are handled through the possibility of including in the peritexts "guarded segments," segments guarded by an attribute name, that are only included if the corresponding attribute is specified on the element, and that can refer to the attribute value. "Local" elements (in the sense of W3C schemas) are supported, so that different peritexts can be assigned depending on the ancestors of the element.

The IS generation process is akin to styling the document with the peritexts, concatenating peritexts *and* element contents as the document tree is traversed depth-first. The *IS*, or *IS-meaning*, *of the document* is the resulting character string. It is important to stress that, in spite of the similarity between styling and the generation of the IS of a document, the preoccupations of IS are absolutely not at the *presentational* level, but really at the *semantic* level.

In this article, we present a complete implementation, in XSLT 1.0, of the intertextual semantics generation mechanism. The transformation is *model-independent* in that it reads the peritexts from an XML document encoding the IS specification for a given model. It implements attribute handling as defined in [2]; hyperlinks in peritexts or as attribute or element content, as described in [1]; and local element definitions, also as described in [1]. In addition, it performs indentation of the output (in the same line as [3], but more elaborate), for increased readability, and handles *exceptions*, elements for which no peritext exists in the IS specification and unexpected attributes.

2 General approach

As mentioned earlier, the implementation is model-independent: the same XSLT stylesheet is used to process any document. In principle, the association between elements and peritexts is determined by the model (DTD or schema) to which the document conforms. Knowing the model,

the generic stylesheet can read an IS specification (ISS) file, giving the peritexts for all elements, and compute the IS of the instance.
In theory, namespace or schema-location information could be used to identify the appropriate ISS file applicable to a document. However, complications would arise from the fact that the same document can conform to different schemas, and contain elements of different namespaces. Thus, it seems simpler to determine the model of a document for IS purposes independently from namespace and schema-location information. One possibility would be to point explicitly to the ISS file through a processing instruction in the document. We chose a more implicit approach, requiring no model-specific addition to the documents, and which proved flexible enough: the generic ID of the document element of the instance is used to form the filename of the ISS file. More specifically, the file named `genID.iss.xml` in the same directory as the generic stylesheet is used as an ISS file, where `genID` is the generic ID of the top-level element of the document.
The processing performed by the generic stylesheet is one-pass, i.e., it takes as input the document instance and directly generates its IS. Thus, no pipelining environment is necessary. Any current browser with an XSLT 1.0 processor can be used to view the IS of documents directly, provided a link to the generic stylesheet is included in the documents, for example:

```
<?xml-stylesheet type="text/xsl" href="ISG.xsl" ?>
```

3 IS specifications

3.1 Overview

Essentially, an ISS file gives the peritexts (text-before and text-after segments) for all the "elements" in the model. Remember, however, that local elements (in the sense of W3C schemas) are possible [1], so that different peritexts can be assigned to elements with the same generic ID but different ancestral lines. Thus, in effect, a peritext is not assigned to some fixed generic ID, but to a *path*, and will be applicable to elements *matched* by that path. A path P is said to *match* an element E iff E's ancestral line ends with P, i.e., iff P is a suffix of E's ancestral line. For example, a peritext assigned to path `title` is applicable to all elements with generic ID `title`, regardless of their ancestral line, but a peritext assigned to path `titleStmt/title` is applicable only to elements with generic ID `title` that are children of elements with generic ID

`titleStmt`. A peritext assigned to path `/TEI` is applicable only to document (top-level) elements with generic ID `TEI`.[1]
For convenience, peritexts can be assigned simultaneously to more than one path, and peritexts are in fact assigned to paths in *pairs*, consisting each of one text-before segment and one text-after segment.
Peritexts can contain certain delimiters for handling attributes as described in [2] and allowing hyperlinks in the resulting IS as described in [1]. The hyperlink delimiters in [1] were [ ]; however, in [2] and here, those are used for attributes. Thus, we will use {{ }} for hyperlinks.

3.2 ISS files

An ISS file is an XML document. All its elements and attributes belong to the specific namespace:

`http://grds.ebsi.umontreal.ca/ns/ISS/`

Its top-level element is an `iss` element. The content of that element is one or more `rule` elements. Each `rule` element is empty and has three mandatory attributes: `paths`, `text-before`, and `text-after`. The effect of a `rule` element is to assign the pair of peritexts `text-before` and `text-after` to the path or space-delimited paths given in `paths`.
A rule is applicable to an element iff one of its paths matches the element. If more than one rule applies to an element, the one with the most specific (longest) matching path is chosen; if more than one rule has that longest matching path, the first one (in ISS file order) is applied.
The sequences {{ and }} *in peritexts* are hyperlink delimiters, i.e., what is between them is interpreted as a URL and converted to a hyperlink in the IS. It is possible to have {{ in a text-before and }} in the corresponding text-after, but this will only work when the element contains neither sub-elements nor }} character sequences (which would be unusual in a URL).
Peritexts can contain passages "guarded" by an attribute name, such as:

```
@attribName[Some text containing exactly one @.]
```

Such guarded passages in peritexts are included in the resulting IS only if the guarding attribute is present on the element to which the peritext is applied. Otherwise, the entire guarded passage is omitted. When the passage *is* included, the actual value of the attribute is inserted in place of the `@`.
It is possible to use `xmlns` as an attribute to refer to the `namespace-uri` of an element. The guarded passage is then included only if the element belongs to a namespace.

1 Paths play a role similar to match attributes of xsl:template elements in XSLT. However, paths are deliberately much less flexible and cannot, for example, contain wildcards or refer to arbitrary XPath axes.

3.3 Examples

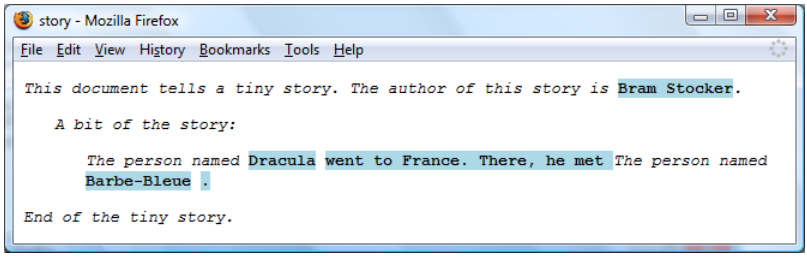
Here is the ISS file used for our examples. It is intended for a top-level element of `story`, and should thus be named `story.iss.xml` and reside in the same directory as the generic stylesheet:

```
<?xml version="1.0"?>
<iss xmlns="http://grds.ebsi.umontreal.ca/ns/ISS/">
<rule paths="story"
  text-before="This document tells a tiny story.@xmlns[ The
     document belongs to the XML namespace &quot;@&quot; (if
     you are not familiar with XML namespaces, you can read
     about them at {{http://www.w3.org/TR/REC-xml-
     names/}}).]@author[ The author of this story is @.]"
  text-after="End of the tiny story."/>
<rule paths="para" text-before="A bit of the story: " text-
     after=""/>
<rule paths="person" text-before="The person named "
  text-after=" @key[{{http://en.wikipedia.org/wiki/@}} ]"/>
<rule paths="place" text-before="The place named "
  text-after=" @key[{{http://en.wikipedia.org/wiki/@}} ]"/>
</iss>
```

Example 1 is the following XML document:

```
<?xml version="1.0" ?>
<?xml-stylesheet type="text/xsl" href="ISG.xsl" ?>
<story author="Bram Stocker">
  <para><person>Dracula</person> went to France. There, he
     met
      <person>Barbe-Bleue</person>.</para>
</story>
```

Note that `ISG.xsl` is the generic stylesheet and it is here assumed to be in the same directory as the document. The resulting IS, as can be viewed in any XSLT 1.0-compliant web browser, is:

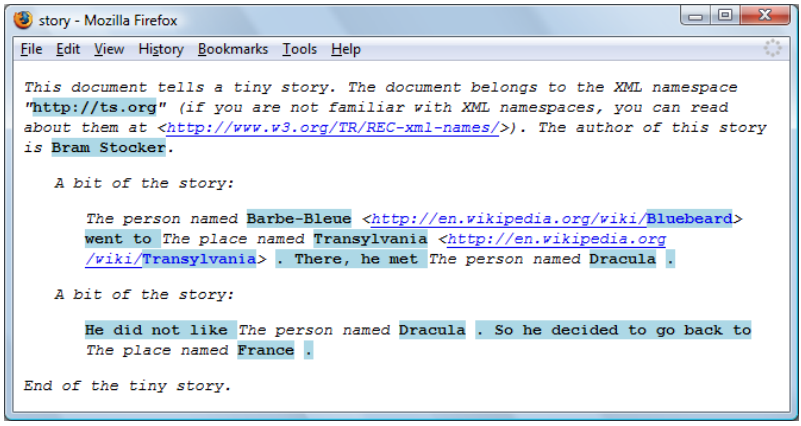


Note that the text contributed by peritexts is typeset in italics and the text contributed by the document is typeset in normal font on blue

background. This is in keeping with the philosophy of IS, which demands that the origin of all text in the IS of a document be clearly identifiable. Example 2 uses more of the richness allowed by the model:

```
<?xml version="1.0" ?>
<?xml-stylesheet type="text/xsl" href="ISG.xsl" ?>
<story author="Bram Stocker" xmlns="http://ts.org">
  <para>
    <person key="Bluebeard">Barbe-Bleue</person> went to
    <place key="Transylvania">Transylvania</place>. There,
     he met
    <person>Dracula</person>.</para>
  <para>He did not like <person>Dracula</person>. So he
     decided
    to go back to <place>France</place>.</para>
</story>
```

This time, the resulting IS is:



Note that the `key` attributes (of both `person` and `place`) become part of a clickable hyperlink in the IS.

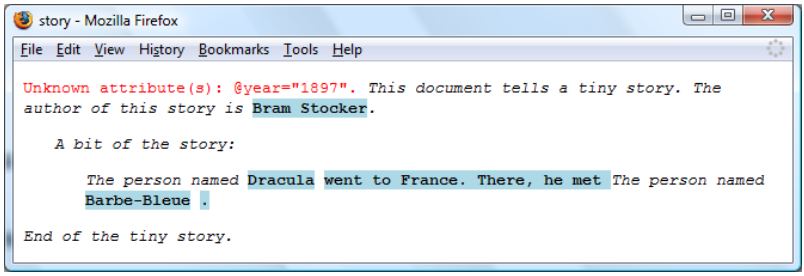
Both Example 1 and Example 2 exhibit a block structure with some parts indented. This feature, globally referred to as "automatic indentation," is not controlled by the ISS file, but rather realized automatically through heuristics. It will be discussed in more detail later.

Example 3 will illustrate exception handling. In the document of Example 1, we add to some element (**`story`**) an attribute (**`year`**) that is not mentioned anywhere in the peritexts of that element. This is considered to be an "unknown attribute:"

```
<?xml version="1.0" ?>
<?xml-stylesheet type="text/xsl" href="ISG.xsl" ?>
<story author="Bram Stocker" year="1897">
```

```
  <para><person>Dracula</person> went to France. There, he
     met
      <person>Barbe-Bleue</person>.</para>
</story>
```

The resulting IS is:



Note that exceptions are reported in red, so as to be easily differentiated from "normal" IS text. An element to which no rule in the ISS file applies is also an exception, of type "unknown element."

4 Structure of the generic stylesheet

The general structure of the generic stylesheet (`ISG.xsl` in the above examples) is as follows:

1. Global variables initialization.
2. Overall HTML structure template, matching /.
3. Templates for text-only elements.
4. Templates for all other elements.
5. Templates for text nodes (PCDATA).
6. Called templates for finding the best matching rule.
7. Called templates for processing attributes.
8. Called templates for processing hyperlinks.
9. Called templates for exception handling.

The rules contained in the model-specific ISS file *genID*`.iss.xml`, where *genID* is the generic ID of the top-level element of the document, are read in Part 1 and placed in a global variable. Part 2 produces the overall HTML structure of the output, including an internal CSS stylesheet.
Parts 3, 4, and 5 are actually pairs of templates: one for block formatting and one for flowed formatting. The *indentation heuristics*, mentioned earlier, is realized by the templates in Part 4 and determines how blocks are indented relative to one another and at which level the formatting should be changed from block to flowed.

The templates in Part 6 are used to determine the rule, in the ISS file, that "best" matches an element. As sketched earlier, a rule is a best match for an element E iff one of its paths matches E (i.e., is a suffix of E's ancestral line) and no other rule specifies a longer path matching E. If two rules or more are best matches for an element, the first one (in ISS file order) is chosen.
The templates in Parts 7 and 8 process attributes and hyperlinks, respectively. Processing consists essentially in recursive search-replace of various delimiters and placeholders. The templates in Part 9 are called by those of Parts 3 and 4 to verify the presence of exceptions and, if needed, include the appropriate warnings in the produced IS.

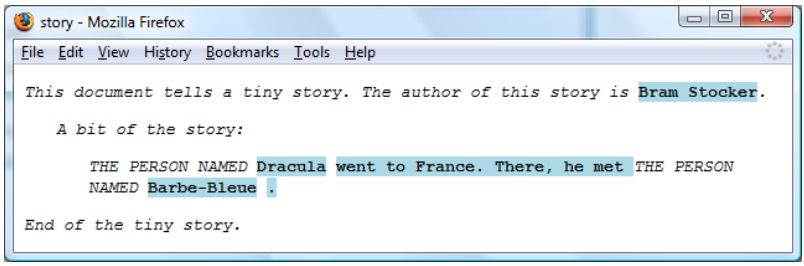
5 Discussion

The examples illustrate that peritexts can be very long. It is an essential feature of IS that there be no limit on their length. They should not be constrained lexically either. However, this is not entirely the case in the current implementation. Indeed, it is currently not possible to include some of the delimiters and placeholders for attribute guarded-passages and hyperlinks as data in the peritexts (and, in a few cases, in element content). One possible improvement would thus be to define and implement conventions for allowing all delimiters and placeholders to be included as data in peritexts (and element content).
A related issue is the validation of the syntax used for attribute guarded-passages and hyperlinks in peritexts. At the moment, no validation or error detection is performed. While this will never cause the abnormal termination of the transformation, it could yield unexpected results. Another possible improvement would thus be to implement full syntactic validation of the peritexts.
Certain delimiters are used internally as placeholders in text variables during processing. Their use relies implicitly on certain character sequences not occurring as textual content in the processed document. This should be replaced by more robust mechanisms.
One of the challenges of developing a generic stylesheet in XSLT 1.0 is to maintain a one-pass approach. Solving the above-mentioned weaknesses would without doubt make this challenge even bigger. Switching to a two-pass approach may thus be an attractive avenue for future developments.
Adopting a two-pass approach can be done in essentially two ways: an external pipelining mechanism (such as XProc <http://www.w3.org/TR/xproc/>) can be used with XSLT 1.0, or multi-passes can be handled internally in XSLT 2.0, through the `node-set` function, which allows some pipeline-like processing. In both cases,

browser integration could be non-trivial. One interesting way to exploit an external pipelining mechanism would be to have a generic stylesheet generate a model-specific stylesheet from the IS specification, then apply the generated stylesheet to the document instance to generate its IS.
Another functionality that could benefit from the enhanced possibilities of multi-pass / XSLT 2.0 processing is the automatic indentation of the output. Right now, the heuristics used is fairly simple, and it can break down even on simple cases. A more sophisticated and robust heuristics should thus be developed, and this would likely be easier with multi-pass / XSLT 2.0 processing.
A question that needs to be investigated through experimentation is that of determining how much of the indentation should be automatic. In [1], indentation was specified explicitly in a conventional manner in the peritexts.
Let us now consider the foreseeable evolutions of the IS framework and how they could impact the IS generation mechanism. Consider the output of Example 1. As textual content, it includes the following passage:

```
There, he met The person named Barbe-Bleue
```

Note that the article `The` has been capitalized. Why? The answer is that it comes from a text-before segment that is sometimes located at the beginning of a sentence, where capitalization is appropriate. But capitalization in the middle of a sentence (as in Example 1) is inappropriate. The source of the problem is that, in the current framework, the same text-before segment must be used consistently, regardless of its position in a sentence.
Remember that in IS, the focus is not on presentation but on meaning. Since the problem at hand only affects presentation and does not hinder comprehension, it must not be considered major. Moreover, it can be alleviated by various devices, such as writing the peritext all in capitals. With Example 1, this gives the following output, which, though still unusual, is not as strange-looking as the original output:



Thus, it is not clear that the framework needs to be modified to accommodate peritexts that vary according to their position in a sentence.

Other possible extensions in the same line would include peritexts that vary with the position of the element relative to its siblings, with the number of children of the element, and with the grammatical gender of a word or expression in the content of some element or attribute. Clearly, adding any such extension to IS would complicate the IS-generation mechanism. For one thing, it might, require the inclusion of additional peritexts in the IS specification of a model. Then, those additional peritexts would have to be appropriately processed during IS-generation.
We believe that, in all cases, experimentation should be used to determine whether an extension is truly necessary or if some workaround without extension is possible. We think extreme parsimony is of utmost importance for the evolution of IS, because the inclusion of too powerful mechanisms could severely impair the explanatory power of the approach.

6 Conclusion

In this article, we presented a complete implementation, in XSLT 1.0, of the intertextual semantics (IS) generation mechanism for XML documents. The implementation is *model-independent*, in that a generic XSLT stylesheet reads the peritexts from an *IS specification file*, an XML document giving the IS specification (ISS) applicable to the document being processed. The implementation handles attributes as described in [2], hyperlinks (in peritexts or as attribute or element content) as described in [1], and local element definitions (in the sense of W3C schemas), also as described in [1]. In addition, it performs indentation of the IS produced (in the same line as [3], but more elaborate), for increased readability, and handles *exceptions*, elements for which no peritexts are given in the IS specification or attributes unexpected by the peritexts.
After describing the format adopted for ISS files, we gave examples illustrating the functionalities of the implementation, then outlined the structure of the generic stylesheet. Finally, we discussed various aspects of the implementation, possible improvements, and the impact that foreseeable generalizations of the IS framework might have on the IS generation mechanism.
The current version of the stylesheet is available through <http://grds.ebsi.umontreal.ca/>. It is published under the Creative Commons “Attribution-Noncommercial-Share Alike 2.5 Canada” license <http://creativecommons .org/licenses/by-nc-sa/2.5/ca/>. We warmly encourage readers to experiment with it, look at the examples, write IS specifications for their models, either extant or under development, and send comments and suggestions.

7 References

[1] Marcoux, Yves. “A natural-language approach to modeling: Why is some XML so difficult to write?” *Proceedings of Extreme Markup Languages 2006*.

[2] Marcoux, Yves; Rizkallah, Élias. “Exploring intertextual semantics: a reflection on attributes and optionality.” *Proceedings of Extreme Markup Languages 2007*.

[3] Marcoux, Yves; Rizkallah, Élias. “Experience with the use of peritexts to support modeler-author communication in a structured-document system.” *Proceedings of SIGDOC 2007*.

[4] Marcoux, Yves; Rizkallah, Élias. “Intertextual semantics: a semantics for information design.” *Journal of the American Society for Information Science & Technology*, Perspectives issue on design. September 2009, *in Press*.

[5] Smedslund, J. *Dialogues about a new psychology*. Chagrin Falls, Ohio: Taos Institute. 2004.

[6] Sperberg-McQueen, C. M., Huitfeldt, C., & Renear, A. “Meaning and interpretation of markup.” Markup Languages: Theory and Practice 2, 3 (2000), 215–234.

[7] Sperberg-McQueen, C. M., Dubin, D., Huitfeldt, C., & Renear, A. “Drawing inferences on the basis of markup.” In Proceedings of Extreme Markup Languages 2002 (Montreal, Canada, August 2002), B. T. Usdin and S. R. Newcomb, Eds.

[8] Sperberg-McQueen, C. M. & Miller, E. “On mapping from colloquial XML to RDF using XSLT.” Proceedings of Extreme Markup Languages 2004.

[9] Wierzbicka, A. *Semantics, culture, and cognition : universal human concepts in culture-specific configurations*. Oxford University Press. 1992.

[10] Wittgenstein, L. *Philosophical investigations*. Oxford: Blackwell. 1953.

Quelques techniques de visualisations de contrôle pour la numérisation massive

Rodrigo Almeida, Pierre Cubaud

Centre d'études et de recherche en informatique (CEDRIC).
Conservatoire national des arts et métiers (CNAM)

Mots-clés : numérisation massive, visualisation d'images, contrôle qualité, bibliothèques numériques

Keywords: massive digitization programs, visual quality control, digital libraries, image browsing and visualization

Résumé : Les programmes de numérisation de masse ont besoin de nouvelle techniques de visualisation adaptées pour le contrôle qualité. Nous décrivons quelques prototypes fonctionnels d'interfaces fluides pour un logiciel permettant l'inspection rapide de la conformité de grands lots de numérisations d'ouvrages.

Abstract : Massive digitization programs need massive visualization techniques for quality control. We describe some functional prototypes of a fluid interactive environment enabling a rapid inspection of pages conformity for large batches of digitalized books.

1 Introduction

En matière de bibliothèques numérisées, une dimension nouvelle a été atteinte fin 2005 avec l'annonce de projets très ambitieux : l'accord de Google avec plusieurs très grandes bibliothèques américaines (en particulier avec l'Université du Michigan qui atteint le million de titres en ligne début 2008) et la mise en place d'Europeana, une fédération de bibliothèques européennes (6 millions de titres prévus en 5 ans). Chez les sous-traitants qui effectuent la numérisation, un opérateur peut atteindre la vitesse de 700 p./h. Ainsi, pour un centre de numérisation d'une dizaine de numériseurs, la productivité est de l'ordre d'un millier de volumes traités par semaine. Avec de telles charges, il est clair que le processus de numérisation doit être soigneusement mis au point. Malheureusement, la productivité chute pour les collections patrimoniales du fait de leur hétérogénéité (reliures pincées, existence d'hors-textes, pagination erratique, tables défectueuses, etc.). De fait, la dissimilitude entre les

volumes et les particularités de chacun imposent un plus grand nombre d'intervention humaine pour le paramétrage du scanneur. Le projet DEMAT-FACTORY, une collaboration regroupant pour trois ans des entreprises de numérisation (A2IA, Banctech, SAFIG, Themis) et des laboratoires (CNAM, ESIEE, Univ. Paris 6), s'est donné pour but d'étudier ces problèmes.
Une des pistes explorées consiste à évaluer l'apport de techniques récentes de visualisation d'information pour les interfaces de contrôle qualité. Ce domaine n'a, à notre connaissance, fait l'objet d'aucune étude systématique. Les logiciels actuellement en usage au sein du consortium sont basés sur les techniques usuelles d'association de vignettes et de vues de page en gros plan (fig. 1). Ces applications, à la fois conçues selon la logique de la visualisation de photos et non destinées à la visualisation de milliers d'images, n'offrent que peu de modalités d'affichage prenant en compte les particularités graphiques des pages des documents.
Nous allons d'abord présenter la problématique du contrôle de qualité visuelle des images. Ensuite, nous indiquerons quelques aspects des techniques de visualisation traditionnelle qui les rendent peu adaptées à ce type de tâche. Puis, nous présenterons quatre techniques de visualisation conçues pour ce contexte de travail que nous avons explorées.

2 Contrôle qualité d'un lot de pages numérisées

Lorsque les images sont livrées par les prestataires de numérisation, on peut identifier différents types d'erreur, quelques-uns assez imprévues. De fait, lorsqu'on numérise des ouvrages rares et difficilement accessibles ailleurs, le contrôle qualité devrait mériter encore plus d'attention afin que les documents numérisés soient suffisamment fiables pour être consultés à la place des originaux. Certains lecteurs trouveraient peu utilisables des documents numérisés dont quelques pages ou planches sont absentes. De plus, la ré-insertion de ce volume dans un nouveau train peut paraître moins intéressante que la numérisation d'un titre pas encore numérisé. La phase de contrôle qualité est donc essentielle afin de découvrir des images manquantes, les pages absentes dans le volume original (et non détectées avant l'envoi du volume), des prises de vue inexploitables, etc. Cette phase de contrôle occupe de fait une place importante dans les chaînes de traitement des projets de numérisation [1].
Certains aspects du contrôle qualité peuvent être validés de façon automatique ou quasi-automatique. Par exemple, les metadonnées de chaque volume, les noms donnés aux fichiers image, l'intégrité de ces fichiers, la résolution des images, etc. [2]. Il est cependant recommandable de procéder à un contrôle visuel des images livrées [2].

Lors du contrôle visuel, on cherche des erreurs qui ne sont pas identifiables par les procédures automatiques. Ces erreurs peuvent être classées selon le moment où ils sont survenues : avant, durant ou après la prise de vue. Le tableau 1 présente une liste non-exhaustive des problèmes que l'on peut trouver lors d'un contrôle visuel des pages numérisées.

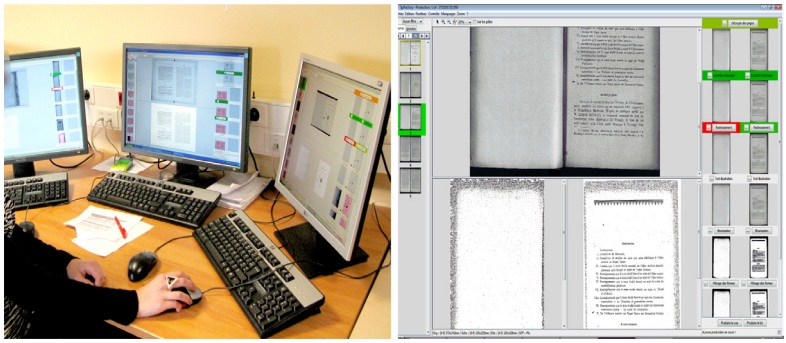


Figure 1. Contrôle qualité courant dans Demat-Factory : espace de travail et interface.

Phase	**Type de problème**
avant	- pages manquantes - pages abîmées (pliées, déchirées, tachées) - mauvais volume (mauvaises informations sur la reliure)
durant	- volume peu ouvert - volume mal orienté - pages mal tournées - planches non dépliées - doigts ou d'autres objets couvrant la page - images floues - bordures du *setup* trop larges - artefacts numériques - variation de la couleur parmi les pages d'un même volume - illisibilité des caractères/dessins les plus petits
après	- images « binarisées » trop foncées ou trop claires - mélange d'images issues de volumes différents - images répétées - images manquantes - incohérence entre le n° de la page imprimé et celui déclaré dans le fichier - incohérence entre le n° du vol. imprimé et celui déclaré dans le répertoire

Tableau 1 : Type de problème de numérisation en fonction du moment où ils apparaissent.

Quelques-uns des erreurs décrits dans le tableau 1 (pages répétées, incohérence entre le numéro imprimé sur la page et le numéro dans le nom du fichier, page mal orientée) pourraient éventuellement êtres détectés avec l'aide des techniques de vision par ordinateur. Cette approche automatique est plutôt complémentaire à l'approche visuelle. Tandis que la première permet de détecter vite des erreurs « modélisables », un avantage clair de la seconde est de permettre que l'on décèle des erreurs qui n'étaient pas « prévues » ou qui sont plus facilement détectées par l'œil humain.
Dans un contrôle qualité traditionnel, un petit échantillon de pages numérisées est examiné et on cherche des problèmes récurrents. Les vignettes sont rapidement balayées et quelques-unes sont chargées en taille normale. Par ailleurs, dans une chaîne non de numérisation non-massive, le volume physique original peut être facilement accessible par le personnel de la bibliothèque numérique, qui connaît ses particularités, et qui peut être confronté à son équivalent numérique. En revanche, dans un programme de masse, des connaissances sur le document numérisé, ou en cours de numérisation, sont réduites, voire inexistantes, à l'exception de celles explicitement déclarées dans le cahier des charges. De même la comparaison des images numérisées avec l'original physique est infaisable.
Nous considérons ainsi que, pour compenser les manques d'indices plus riches sur la pertinence de la numérisation, des chaînes de numérisation massive devraient adopter une stratégie de *visualisation exhaustive*, et non pas *par échantillon*. On devrait pouvoir rapidement balayer *toutes les images* et pouvoir passer sans effort à une observation détaillée de toutes celles qui attirent son attention.

3 Vue globale et vue détaillée

Les applications traditionnelles d'affichage d'images sont principalement conçues pour la gestion et la visualisation de photos personnelles, plutôt que pour la visualisation de pages numérisées. L'utilisation des pixels de l'écran (« *screen real estate* ») n'est pas optimale. De plus, passer d'une photo à l'autre peut prendre du temps lorsqu'on souhaite en voir un grand nombre. On accorde plus d'importance à la notion d'élément individuel qu'à la notion d'un ensemble d'éléments disposant de traits en commun (comme c'est le cas des pages numérisées issues d'un même volume). Or, dans la visualisation de pages, l'important est de voir chaque image dans son « contexte », c'est-à-dire entourée par les pages « voisines » et à côté des pages issues des volumes d'un même lot, et de pouvoir en visualiser un grand nombre avec un minimum d'encombrement. Dans un bon nombre d'applications, une interface pour la visualisation et la navigation

des vignettes existe, bien qu'elle soit souvent présentée comme un mode de consultation auxiliaire. Dans ce type d'interface, la priorité est donnée pour la présentation de la page dans sa totalité.
L'interface *Space-filling Thumbnails* (SFT) propose d'exploiter davantage la présentation de vignettes qui miniaturisent les pages [3]. Cette interface présente toutes les pages d'un livre distribuées de façon matricielle où chaque page a une position fixe par rapport aux pages voisines. La matrice a toujours la même configuration (la réduction de la fenêtre réduit la taille de toutes les cellules sans changer les nombres de lignes et de colonnes), ce qui permet à l'utilisateur de mémoriser la position relative d'une vignette et de la revisiter facilement. Par ailleurs, cette interface ne dispose d'aucun mécanisme de défilement. De ce fait, toutes les pages (d'un même document) sont toujours simultanément visibles sur la fenêtre. Pour cela, la taille de chaque vignette décroît à proportion que le nombre total de vignettes s'accroît. Le désavantage de ce changement d'échelle automatique est le fait que les vignettes deviennent microscopiques lorsque le volume est doté d'un grand nombre de pages.
Outre la visualisation globale via un affichage « miniaturisé » du volume, le passage à un affichage de la page en haute résolution est une fonctionnalité également importante. Certains problèmes (comme la lisibilité des caractères du texte ou la netteté des illustrations) ne sont identifiables que lorsqu'on visualise l'image en « taille réelle ». Les mécanismes de zoom sont variés : le zoom peut agrandir toute la page qui occupera toute fenêtre (comme dans *Adobe Photoshop*), il peut agrandir toutes les photos présentées dans la fenêtre en même temps (comme dans *Picasa*) ou il peut agrandir temporairement qu'une photo en conservant au fond les autres images (comme le zoom contextuel de SFT). La transition entre les différents niveaux d'échelle peut aussi être soit discrète (comme dans *Adobe Acrobat Reader*), soit continue (comme dans *Photomesa* [4]). Les applications, dites « zoomables » comme *Picasa* et *Photomesa* présentent un bon compromis entre les vues globale et détaillée.

4 Visualisation par réduction extrême

Nous avons commencé à explorer les apports de la visualisation par vignettes en produisant des miniatures pour les périodiques de « La Nature »[1]. Ce périodique est le plus visité de la bibliothèque numérique du CNAM (Conervatoire Numérique, CNUM) et l'un des plus volumineux. Il compte 32.500 pages, avec environ une gravure toutes les deux pages ; actuellement 32 années (soit 1696 fascicules reliés dans 65 volumes) de cette publication sont disponibles en ligne. À l'aide de la

1 http://cnum.cnam.fr/redir?4KY28

bibliothèque de programmes *NetPBM*. [2] , nous avons produit des miniatures, chacune fait 10 par 16 pixels, pour chaque page de tous les volumes de cette collection.

Figure 2 : Les 32.500 pages de La Nature affichées (chaque page fait 10 x 16 pixels) sur une même image (8000 x 600 pixels). Chaque ligne correspond à un volume de fascicules reliés. Détail sur la fin des volumes dans les années 1890.

Nous avons ensuite produit un script Processing[3]. pour que les pages d'un même volume soient affichées sur une même ligne (fig. 2). Cette configuration fournit des pistes visuelles très intéressantes. On peut, par exemple, observer : la longueur relative des volumes les uns par rapport aux autres, les pages typiques qui apparaissent en fin de volume et celles qui apparaissent en début de volume, la regularité des pages illustrées, si un volume est plus illustré que les autres. Nous pouvons aussi voir que dans le volume 34, les suppléments qui apparaissent en fin du volume sont visiblement manquants (ce que nous ignorions avant cette expérience !). Toutes ces indications sont utiles pour repérer des lacunes dans les volume, qu'elles soient dues à la numérisation ou aux volumes physiques fournis à la numérisation.
Cette visualisation statique met en évidence les points forts d'avoir une vue globale sur une collection numérisée, principalement pour vérifier son intégrité et comprendre quels sont les types de page qui la composent. La visualisation interactive d'un ensemble de telle grandeur, proche du giga-pixel, est actuellement hors de portée de la puissance de calcul des ordinateurs standard. Nous avons commencé à explorer des techniques interactives, sur des ensembles de pages plus réduits.

5 Interface Grille de détails

Nous avons développé l'interface « Grille de détails » pour que les pages numérisées puissent être vérifiées de façon *exhaustive* et non pas par échantillons. Il s'agit d'une interface Web, utilisée au sein du projet CNUM, qui affiche les vignettes des pages numérisées dans une

2 http://netpbm.sourceforge.net/

3 http://processing.org/

disposition matricielle (fig. 3). Dans cette interface, hormis l'espace d'un pixel entre les vignettes, tous les pixels de la fenêtre sont utilisés pour afficher des zones des pages numérisées. Quelques aspects visuels de toutes les pages d'un volume de mille pages peuvent ainsi être vérifiés en quelques minutes. En outre, les vignettes sont partiellement recoupées afin de prioriser la zone de la page contenant du texte ou de l'illustration. L'affichage résultant est donc dense en informations visuelles. Enfin, le fait que les pages soient serrées les unes contre les autres, sans beaucoup d'espace entre elles, facilite la détection d'anormalités. Par exemple, une page avec du texte plus foncé que le texte des pages voisines saute aux yeux.

Modalités d'affichage. Les vignettes des pages utilisées par la Grille de détails sont générées lorsqu'on charge les images dans le serveur du CNUM. Deux types de vignette sont fabriqués à partir des images master : une miniature de la page et un détail du centre de la page. Ce détail a la même résolution de l'image qui sera diffusée dans le site.[4] Pour la miniature, on recoupe un carré dont le côté fait la dimension la plus courte de la page et on l'échantillonne ensuite à 128 x 128 pixels. Pour le détail, on échantillonne d'abord l'image à la résolution du format de diffusion ; ensuite on recoupe, au centre cette image échantillonnée, un carré de 128 x 128 pixels. Pour la miniature des images brutes, les vignettes font 256 x 128 pixels (ces images sont recadrées le moins possible). Ces vignettes sont aussi fabriquées par lot à l'aide des programmes *NetPBM*. La grille de détails affiche les vignettes dans une page générée par du code *PHP* dont des fonctions *JavaScript* offrent quelques micro interactions.

Inspection des images de diffusion. Ce mode de visualisation est actuellement utilisé dans la vérification des images qui seront mises en ligne. Les vignettes de chaque page sont affichées dans une matrice qui se sert de tout l'espace de la fenêtre. L'utilisateur décide si, par défaut, toutes les vignettes affichent la vue « détaillée » ou la vue « miniature » de la page. Lorsqu'on passe le curseur sur une vignette, celle-ci bascule vers le mode de présentation qui n'est pas celui par défaut. En cliquant sur la vignette, on charge la page dans le format GIF de diffusion. La vue de détail permet de vérifier, par exemple, si le texte est lisible (si les caractères ne sont ni trop rongés ni trop petits) ou si le résultat du tramage des gravures est satisfaisant. Ce changement rapide et sans clic entre la vue miniature et la vue détail fonctionne comme un zoom contextuel et discret (d'un seul pas). Il permet de vérifier simultanément des indices visuels de tout un volume avant de charger les pages en taille réelle.

4 Il faudrait à terme ajouter une étape de reconnaissance du contenu de la page pour que le détail recoupé encadre forcément du texte ou de l'illustration et non pas du vide (comme il arrive dans certaines pages).

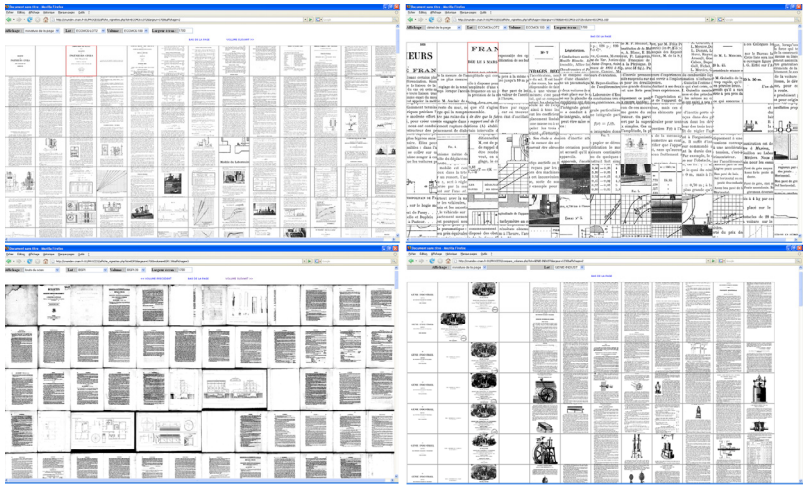


Figure 3. Copies d'écran de l'interface par Grille de détails. En haut à gauche, les vignettes affichent les pages en miniature ; à droite on voit le détail du centre des pages. En bas à gauche, les miniatures des bruts de scan ; à droite les dix premières pages des volumes d'une même collection.

Inspection des images de préservation. Cette deuxième interface présente les miniatures des images brutes en double page. Le fait d'afficher les pages impaires à côté des pages paires permet de voir si une page est manquante dans le volume original ou si le fichier image a été « perdu » lors des transferts ou des traitements des images brutes. Par ailleurs, les pages apparaissent dans l'ordre dans lequel elles ont été numérisées, ce qui facilite la vérification de l'emplacement des planches ou d'autres éléments sans pagination. Enfin, les tonalités de gris et d'autres éléments visuels, comme des traits ou des taches, communiquent une certaine homogénéité des pages et peuvent éclairer l'utilisateur sur des pages « étrangères » ayant été placées par erreur dans le répertoire du volume.

Visualisation comparative de plusieurs volumes. Cette troisième méthode n'est pas faite pour vérifier les images elles-mêmes mais pour vérifier la cohérence entre les images des volumes appartenant à une même collection. On se sert des mêmes vignettes (miniature et détail) utilisées dans la première interface et les images de chaque volume sont affichées dans une ligne.[5] Ainsi, en balayant les lignes de haut en bas, on

5 Comme il serait trop lourd d'afficher toutes les images de chaque volume (certains lots font 24.000 pages), seules les 15 premières vignettes de chaque volume sont affichées.

peut voir si des différences importantes entre les volumes (ou des similarités indiquant des volumes répétés) sautent aux yeux. Cette présentation est spécialement intéressante dans la visualisation des revues qui se servent d'une même structure au fil de plusieurs volumes (comme, par exemple, La Nature). On peut identifier des volumes répétés ou des pages manquantes.
Comme le montre le tableau 2, on peut considérer ces différentes options d'affichage comme des modalités d'échantillonnage auquel l'ensemble d'un volume est soumis. Vu la difficulté pratique de visualiser *tous les pixels de toutes les pages du volume*, on échantillonne soit le nombre de pages, soit la zone visualisée de chaque page, soit la résolution de la page.

Type d'affichage	Nombre de pages	Zone observée	Résolution
Page entière	2%	100%	100%
Miniature	100%	90%	10%
Détail du centre	100%	10%	100%

Tableau 2 : Résumé des caractéristiques de l'« échantillonnage visuel » procuré par chaque méthode.

6 Mur de pages

La *grille de détails* augmente le débit de pages que nous pouvons visualiser mais elle ne fournit qu'un niveau de zoom pour chaque page. On sacrifie le paramétrage de la visualisation au profit de la simplicité d'interaction (ainsi qu'au profit d'une technologie Web facile à déployer). Lorsque l'utilisateur souhaite plus de détail, il va devoir charger l'image entière. Il s'agit donc d'une *interaction discrète* à trois pas : (1) miniature, (2) zoom serré et (3) chargement de l'image en taille réelle.
Pensant à ces limitations, nous avons développé un prototype qui pourrait restituer les détails de l'image de façon fluide et progressive. Comme dans la grille de détails, les pages sont disposées de façon serrée dans une matrice. Grâce à cette disposition, l'ensemble de pages est perçu comme une grande image. Utilisant la technologie 3D, les pages sont disposées dans un environnement où le zoom et la navigation deviennent des opérations très naturelles. Le zoom en arrière réduit la grille progressivement jusqu'à ce que l'intégralité des pages soit visible en même temps. Par le zoom en avant, l'utilisateur peut rapidement trouver le niveau de détail qui lui convient. Il peut alors défiler les pages à une

très haute vitesse grâce au dispositif isométrique utilisé pour contrôler la caméra.

6.1 Stratégies de navigation

Deux stratégies de navigation sont importantes dans le contexte d'une tâche de contrôle de qualité : nous les dénommons la « navigation global-détail » et le « balayage séquentiel ».
Dans la *navigation global-détail*, l'utilisateur peut visualiser la totalité des pages d'un coup d'œil. Cette vue globale est importante pour qu'il puisse estimer la proportion entre des pages de texte, d'illustration et des planches. Cet affichage permet aussi de voir s'il y a des pages dont les aspects visuels (niveaux de gris, taille des caractères, bordure) sont différents de la majorité des pages. Puis, une fois identifiée des « zones du mur » qui méritent d'être observées avec plus d'attention, l'utilisateur doit pouvoir rapidement s'en approcher allant, si nécessaire, jusqu'à la trame du papier.
Pour la stratégie *balayage séquentiel*, l'utilisateur peut « balayer » toutes les pages à une résolution et à une vitesse qui lui conviennent. En voyant les pages dans l'ordre dans laquelle elles ont été numérisées, il pourra plus aisément avoir des indications sur des pages manquantes (par exemple, un chapitre qui finit dans une page et la page suivante commence par du texte). De même, si on souhaite vérifier la qualité des illustrations dans un volume qui en possède beaucoup, il sera plus facile de faire défiler toutes les pages plutôt que de faire des allers-retours de zoom sur chaque illustration.
C'est pour cette raison aussi que ce défilement doit pouvoir se faire à une grande vitesse : pour que l'utilisateur puisse « sauter » les zones visuelles méritant peu d'attention. Enfin, il est important que l'utilisateur puisse positionner la caméra à une résolution donnée (affichant ou bien la totalité de chaque page, ou bien uniquement le détail de chaque page) et qu'il puisse en suite avancer en gardant fixe cette distance. C'est pensant à ces deux stratégies que nous avons créé et adapté la scène et la navigation du *Mur de pages*.

6.2 Scène et contrôle de la caméra

Les pages sont plaquées sur la face interne d'un hémicylindre (fig. 4, gauche). Lorsque la caméra se trouve au centre de la scène, toutes les pages sont visibles : le rayon du cylindre est calculé en fonction du numéro de pages à afficher et de l'amplitude du champ de vision de la caméra virtuelle. Les limites du mouvement de la caméra : en zoomant, elle peut aller jusqu'au détail de la page ; en reculant, elle doit s'arrêter au centre du cylindre. Cette disposition permet que les trajectoires rectilignes du centre du cylindre vers n'importe quelle page aient la même

longueur. En revanche, si le mur était plat, il serait plus coûteux de déplacer la caméra jusqu'aux pages qui se trouvent dans un point extrême de cette surface.

Navigation cylindrique. Dans un premier prototype du Mur de pages, il était possible de contrôler la caméra dans quatre degrés-de-liberté (DDL), à savoir, les déplacements le long des trois axes et la rotation autour de l'axe vertical. Cependant, nous avons constaté que les utilisateurs n'arrivaient pas à maîtriser facilement ces quatre DDL et cela les retardait pour atteindre les positions qui leur convenaient. La rotation de la caméra autour de l'axe Y et son déplacement le long de l'axe X jouent, dans ce type de tâche, des rôles similaires. Nous avons ainsi décidé de restreindre la navigation horizontale à la rotation autour de l'axe vertical. Alors, bien que le déplacement le long de X et de Z soit interdit, toutes les zones du mur restent accessibles et visibles (dans une topologie convexe comme celle-ci). Comme le déplacement de la caméra suit des coordonnées cylindriques, elle reste toujours parallèle au mur. L'utilisateur voit ainsi les images sans aucune distorsion de perspective. Contrairement à d'autres interfaces qui se servent des lignes de fuite pour montrer la « périphérie » d'un document [5], la non-distorsion du *Mur de pages* réduit les chances que des erreurs de numérisation soient confondues avec des effets de perspective.

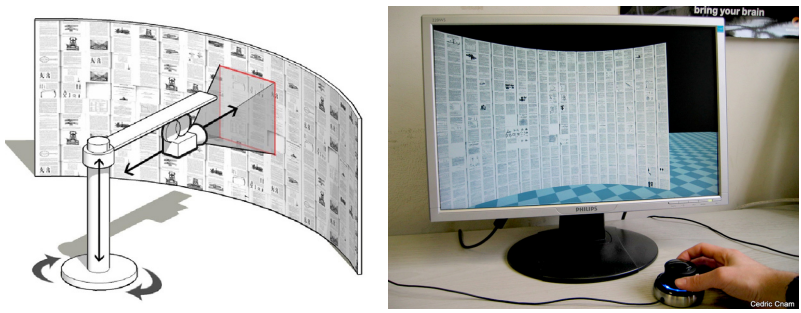


Figure 4. Interface Mur de pages*. Gauche : Principe de navigation. Droite : utilisation.*

Navigation sur l'axe vertical. Dans le *Mur de pages*, les pages sont disposées sur deux dimensions pour tirer un maximum de profit de l'espace de la fenêtre. Puisque les pages sont ordonnées le long des rangées, parcourir l'axe vertical procure une « visualisation aléatoire » des pages, ce qui est intéressant dans le contexte du contrôle qualité. Par exemple, si l'utilisateur zoome sur la page 7 et monte la caméra, il verra passer la page 36, puis 65, 104 et ainsi de suite. Cette navigation lui permet de visualiser rapidement un échantillon de pages du volume.

Balayage séquentiel sans rupture. L'axe horizontal, privilégié par la disposition des pages, est consacré au balayage séquentiel. Son effet est similaire à l'action de défiler rapidement un document dans une application comme *Adobe Acrobat Reader*. Cependant, l'utilisateur doit passer à la ligne du dessus lorsqu'il atteint la fin de la rangée de pages où il se trouve. Ce mouvement ponctuel de monter la caméra d'une ligne introduit une interruption dans le ballayage. C'est pourquoi nous avons considéré d'autres topologies de scène qui éviteraient cette opération. L'idéal serait que l'utilisateur puisse parcourir toutes les pages avec un même geste de contrôle pour que le contrôle porte uniquement sur la vitesse du balayage et l'attention sur l'identification des problèmes.
Dans cette perspective, un cylindre complet, où les pages sont disposées de façon hélicoïdale, produirait une bande d'images continue de la première à la dernière page [6]. Cependant, dans cette configuration il n'y a aucune point où la caméra puisse voir toute le scène (il y aurait toujours une partie du mur derrière la caméra). Par ailleurs, l'environnement serait trop homogène, sans points de repère, ce qui pourrait rendre plus difficile à l'utilisateur de retourner à des pages déjà visitées.
Comportements pour passer d'une rangée de pages à l'autre. Nous étudions deux techniques qui ont pour but de combler cette lacune de notre scène 3D. Un test d'ergonomie pourra montrer si elles sont efficaces et naturelles. Le comportement *monte-page* serait déclenché dans des conditions précises : en réalisant un mouvement panoramique à la proximité du mur de pages (une rangée de pages occupe plus de la moitié de la hauteur du *viewport*), on atteint la fin de cette rangé de pages. Le *monte-page* transporte alors la caméra à la rangée de pages d'au-dessus. Bien que cette solution impose un changement dans le sens du mouvement appliqué au dispositif d'entrée qui contrôla la caméra, elle a l'avantage de permettre que les pages soient disposées dans un ordre facile à suivre visuellement. Ainsi, il est inutile de tourner le regard vers le début de la ligne lorsqu'on atteint la fin d'une rangée. Le deuxième comportement, le *vide-de-pages*, fait que la caméra soit transportée d'un côté à l'autre du hémicylindre lorsqu'elle atteint la fin d'une ligne. L'avantage de ce deuxième comportement est de permettre que les pages soient disposées toutes dans le même ordre (croissant de gauche à droite, quelle que soit la ligne).

6.3 Implémentation du prototype

Un *SpaceNavigator* a été utilisé comme périphérique d'entrée pour le contrôle de la caméra. Il s'agit d'un sort de joystick isométrique à six DDL fabriqué par la société *3D Connexion*. Il est formé par une base lourde sur laquelle un petit cylindre (la manette) est monté. On saisit cette manette par les bouts des doigts. Les dispositifs isométriques procurent un contrôle plus efficace pour une tâche de navigation car, par une simple

modulation de la pression appliquée, on peut faire défiler toutes les pages [7]. En relâchant la manette, la caméra s'arrête.
L'application prototype a été développée en *C* et *OpenGL*. Les images master en niveaux de gris ont été légèrement recadrées pour que leurs dimensions aient un ratio de 2:1. Elles ont été ensuite six fois échantillonnées en avance dans des versions plus petites. Toutes ces versions avaient des dimensions qui étaient des puissances de deux (de 2048x1024 pixels jusqu'à 32x16 pixels). Des images de ces dimensions tirent profit de la mémoire de la carte graphique lorsqu'elles sont chargées en tant que textures. Le jeu de versions réduites de chaque image alimente le filtrage par *MIP mapping*, ce qui procure un zoom plus fluide et plus net.
Les GIFs sont chargées dans la mémoire de la carte graphique utilisant une fonctionnalité de compression de textures lors de l'initialisation de l'application. Comme les textures sont toutes résidentes, il n'y a pas de délai dû à la pagination de la mémoire virtuelle et l'animation n'est pas saccadée. En effet, dans un ordinateur doté d'une carte GeForce 9400M, nous avons chargé 500 pages et l'utilisateur peut zoomer sur n'importe quelle page en moins de deux secondes (fig. 4, droite).

6.4 Retour des utilisateurs

Plusieurs utilisateurs (20 personnes parmi bibliothécaires, professionnels de la numérisation, utilisateurs débutants et avancés) ont déjà essayé notre prototype. Il a été présenté dans un salon des techniques de la numérisation (FAN'2008, Paris) où il a été très apprécié par les personnes qui l'ont utilisé. Ceux qui travaillent dans la numérisation (ou dans la visualisation d'images en général) ont trouvé que cette interface apporte une facilité qu'ils ne connaissent pas dans d'autres applications. Certains utilisateurs, surtout ceux qui ne sont pas habitués aux dispositifs 3D, ont eu des difficultés pour contrôler les mouvements de zoom de la caméra. La forme du dispositif et son faible feed-back élastique paraissent troubler les utilisateurs qui finissent pour réaliser des mouvements imprécis. Le mouvement vertical, par exemple, est assez difficile puisque la prise pour ce type de mouvement n'est pas confortable.
Ajustement de la vitesse du panorama. Igarashi et Hinckley évoquent le problème du « débordement visuel » généré par le défilement accéléré d'un document (par exemple, une longue page Web) [8]. Cet effet est bien similaire à celui expérimenté par les utilisateurs qui ont essayé le *Mur de pages* lorsque la caméra se trouve proche des pages. Dans notre application, nous avons adopté une « navigation cylindrique ». De ce fait, la vitesse angulaire parcourue par la caméra, qu'elle soit proche du centre ou dans l'extrémité du cylindre, est toujours la même. En revanche, la longueur parcourue sera toutefois beaucoup plus grande lorsque la caméra se trouve proche au mur (c'est-à-dire loin du centre du cylindre).

Nous étudions à cette fin un *mapping* variable pour réduire la vitesse de rotation de la camera en fonction de sa distance par rapport au mur. Cela ferait en sorte que l'effet visuel généré par la rotation soit toujours le même. Un deuxième *mapping* que nous étudions est une adaptation directe de la technique de Igarashi et Hinckley, c'est-à-dire reculer la caméra (vers le centre du cylindre) lorsqu'on la tourne rapidement la caméra et la rapprocher lorsque que le mouvement se ralenti ou s'arrête.

7 Panorama de détails

Comme nous avons vu dans la *Grille de détails*, il peut être avantageux visualiser les détails de plusieurs pages en même temps. On se sert des indices disponibles dans le détail affiché pour vérifier la qualité de l'image. Dans l'interface utilisée actuellement, on pré-fabrique ces détails en recoupant le centre de la page. Il s'agit d'un « cache graphique ». Ce cache graphique est nécessaire pour que l'interface soit utilisable pour les lots de toutes tailles et aussi pour qu'on la consulte via des navigateurs différents et sans *plugin*. Cependant, la grille de détails ne nous permet pas de changer dynamiquement la position de la fenêtre de détail. On peut vouloir balayer les détails du haut de la page, pour vérifier si des notes ont été prises. Ou bien, on peut vouloir aussi voir une zone de détail un peu plus grande de chaque page. Ces changements en temps réel ne pourraient pas être pris en compte par la plate-forme de travail actuelle.
Nous avons donc exploré une première maquette d'une « grille de détail paramétrable », que nous appelons « Panorama de détails ». Il s'agit d'une fenêtre similaire à la Grille de détails, mais où chaque carré correspond à une « sous-fenêtre ». On peut imaginer que l'image-détail de chaque page est fournie par une caméra qui plane au-dessus et parallèle à la page correspondante. Dans cette maquette, l'utilisateur contrôle la position et le zoom d'une « camera-abstraite », celle-ci est associée à toutes les caméras qui planent sur chaque page de la grille. Il est ainsi possible de faire un panorama multiple qui s'affiche dans chaque sous-fenêtre de l'interface.

7.1 Paramètres de navigation

Outre le mouvement de panorama sur les pages, notre maquette permet aussi de changer la taille de la fenêtre de visualisation de chaque page. L'utilisateur peut ainsi, par exemple, voir une partie plus grande des pages à l'échelle 100%. Cependant, l'augmentation de la taille des sous-fenêtres réduit le nombre de sous-fenêtres qui peuvent être simultanément affichées sur la fenêtre principale. La solution est de considérer la fenêtre principale elle aussi comme une « super-caméra » qui plane sur une grille. On peut approcher ou écarter la super-caméra des sous-fenêtres.

Comme le montre la figure 5, on peut contrôler les mouvements de la camera-abstraite et les mouvements de la super-camera.

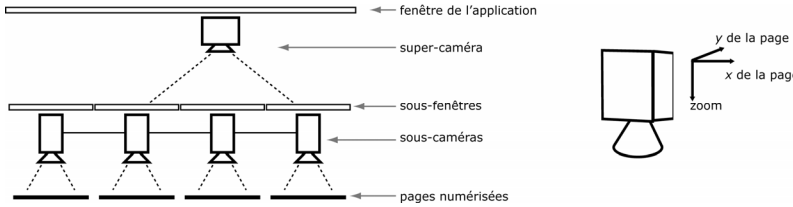


Figure 5. A gauche, schéma avec les caméras virtuelles de la maquette du Panorama de détails*. Les sous-caméras sont reliées les unes aux autres et exécutent toujours les mêmes mouvements. A droite, les DDL le long desquels les caméras virtuelles se déplacent. Les sous-caméras restent toujours parallèles aux images des pages numérisées. La super-caméra reste toujours parallèle au plan des sous-fenêtres.*

7.2 Interface de navigation

L'obstacle que nous avons rencontré dans cette première maquette est le grand nombre de paramètres de visualisation à contrôler, à savoir le zoom et le déplacement panoramique à la fois de la caméra-abstraite et de la super-caméra (trois DDL pour chaque caméra). Il s'agit ainsi d'une interface à six DDL. Dans cette maquette, nous avons choisi de représenter les viewports des deux caméras par des icônes qui fonctionnent comme interface de navigation. Ces représentations, à l'instar du *widget Navigator* dans *Photoshop*, peuvent être glissées et redimensionnées directement avec la souris ou par des touches du clavier. Elle a l'avantage d'indiquer quels sont la position et le niveau de zoom de la caméra par rapport à la page.

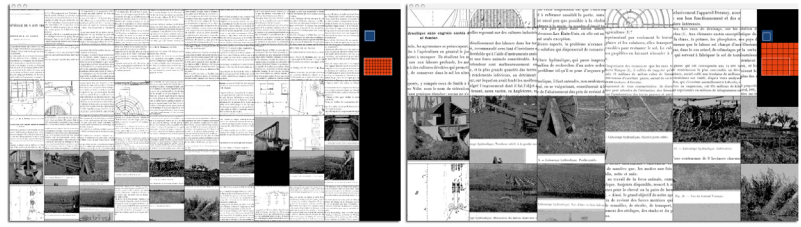


Figure 6. A gauche, schéma avec les caméras virtuelles du Panorama de détails*. Les sous-caméras sont reliées les unes aux autres et exécutent toujours les mêmes mouvements. A droite, les DDL le long desquels les caméras virtuelles se déplacent. Les sous-caméras restent toujours parallèles aux images des pages numérisées. La super-caméra reste toujours parallèle au plan des sous-fenêtres.*

Le contrôle de six DDL est naturel s'il s'agit des six variables qui définissent la position et l'orientation d'un objet dans l'espace, c'est-à-dire (X, Y, Z) et (R_X, R_Y, R_Z). Un joystick isométrique pourrait offrir une solution bureautique pour cette manipulation. Nous ne voyons cependant pas comment associer un mouvement six DDL à la métaphore de caméra que nous avons décrite. Une piste à explorer serait de se servir d'une manipulation bi-manuelle. La main gauche pourrait contrôler la super-caméra tandis que la main droite manipule la caméra-abstraite.

8 Conclusion et perspectives

Les bibliothèques numériques massives ont (ou vont avoir) besoin de procédés de visualisation également massifs. Les premières expériences décrites ci-dessus sont encourageantes dans la mesure où elles montrent que les systèmes graphiques actuels ont la puissance nécessaire pour de telles visualisations interactives. Nous avons décrit quatre techniques de visualisation, en différents niveaux de développement, qui ont pour but de favoriser la compréhension d'un lot de pages numérisées. L'interface *Grille de détails* est maintenant utilisée de manière systématique pour la revue des lots du CNUM.
Le prototype de *Mur de pages* a, quant à lui, reçu un accueil positif de la part des personnes qui l'ont testé. Ce prototype, bien qu'il soit techniquement bien avancé, ne procure pas encore de fonctions qui nous permettraient de l'utiliser dans un cadre réel de contrôle qualité. Par exemple, il n'est pas encore possible de récupérer des informations additionnelles sur une page donnée (nom du fichier, date de numérisation, dimensions en pixel, colorimétrie) et nous ne disposons pas non plus d'une interface d'annotation intégrée. Outre le développement de ces fonctions, nous souhaitons conduire une étude ergonomique qui nous aiderait à choisir un *mapping* plus efficient et confortable pour la caméra. Deux idées concernant le *mappings* ont déjà été proposées et sont en cours d'implémentation : le passage semi-automatique d'une ligne à l'autre lors d'un balayage séquentiel et l'adaptation « écologique » de la vitesse de rotation (en fonction de la distance qui sépare la caméra des pages).
D'autre part, nous souhaitons rendre le prototype robuste pour la visualisation d'un ensemble de dimensions très importantes (plus de 2000 pages). Une architecture qui prenne en compte le chargement et le déchargement dynamique des textures doit être étudiée [9]. Ils nous semble aussi que les travaux liées à la visualisation des images giga-pixel peuvent nous fournir des pistes [10]. D'autant plus que ces techniques évoluent rapidement grâce à leur popularisation via les outils Web de consultation de cartes et photos satellitaires. Un gestionnaire de

chargement adapté aux parcours caractéristiques de la tâche de contrôle de qualité pourra donc être développé.
Nous allons ensuite approfondir l'analyse du travail des opérateurs de contrôle de qualité des partenaires de DEMAT-FACTORY, et continuer le développement avec une démarche de design participatif. Nous procéderons ensuite à une évaluation quantitative. Au-delà, il conviendrait d'étudier en quoi ces techniques seraient utilisables pour faciliter la navigation des utilisateurs finaux des grands corpus numérisés.

Remerciements : Nous remercions M. Emile Prior, directeur technique de SAFIG, ainsi que le personnel du centre de La Châtre. Merci également à Alexandre Topol et Pedro Alessio pour leur aide dans l'expérience du mur de page.

9 Références bibliographiques

[1] Site Web du projet de numérisation de la bibliothèque de l'Université de Michigan, "http://www.lib.umich.edu/news/millionth.html"

[2] J. Riley et K. Whitsel. Practical Quality Control Procedures for Digital Imaging Projects. *OCLC Systems & Services* 21(1), pp. 40--48 (2005)

[3] A. Cockburn, C. Gutwin et J. Alexander. Faster Document Navigation with Space-Filling Thumbnails. *Proc. of CHI '06*, pp. 1--10. ACM Press, New York (2006)

[4] B. Bederson. PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps, *Proc. of UIST '01*, pp. 71--80. ACM Press, New York (2001)

[5] Y. Guiard, O. Chapuis, Y. Du et M. Beaudouin-Lafon. Allowing Camera Tilts for Document Navigation in the Standard GUI: A Discussion and an Experiment. *Proc. of AVI '06*, pp. 241--244. ACM Press, New York (2008)

[6] A. Topol. Interaction 3D pour les paysages informationnels. Thèse de doctorat en informatique, Conservatoire national des arts et métiers, Paris, France (2002)

[7] S. Zhai, B. A. Smith et T. Selker. Improving Browsing Performance : A Study of Four Input Devices for Scrolling and Pointing Tasks. *Proc. of INTERACT '97*, pp. 286--293, Chapman & Hall, London, UK (1997)

[8] T. Igarashi et K. Hinckley. Speed-Dependent Automatic Zooming for Browsing Large Documents. *Proc. of UIST '00*, pp. 139--148. ACM Press, New York (2000)

[9] D. Cline et P. K. Egbert. Interactive Display of Very Large Textures. *Proc. of VIS '98*, pp. 343--350, IEEE Press, Los Alamitos, CA, USA (1998)

[10] J. Kopf, M. Uyttendaele, O. Deussen et M. F. Cohen. Capturing and viewing gigapixel images. *ACM Trans. Graph.*, 26(3) : 93, ACM Press, New York (2007)

Les nouveaux enjeux de la mise en valeur du Patrimoine scientifique et technique de la recherche dans l'espace Francophone

Rachel Kamga, Khaldoun Zreik

Université paris 8, Laboratoire Paragraphe

1 Introduction

La réflexion proposée dans cette communication se veut une incitation à une réflexion globale sur le repérage, l'inventorisation, la documentation, la sauvegarde et la valorisation du patrimoine scientifique et technique francophone de la recherche par le biais des technologies d'information et de la communication. En abordant la question du recensement des éléments de ce patrimoine et celle de leur valorisation, on ne peut manquer de poser la question de l'histoire et particulièrement celle de l'histoire des sciences et des techniques. Faut-il rappeler l'importance de ce riche patrimoine, à la fois témoin de l'histoire de la recherche scientifique, de l'évolution et de l'apport des sciences et des technologies et support incontestable de la culture scientifique et technique dans une société de la connaissance[1] ? En effet, aussi bien la science, la technique et l'industrie que les régions elles-mêmes ont un passé qui se matérialise dans un patrimoine et qui reste sous-jacent à maintes problématiques du présent[2]. A cette fin, on analysera d'abord le concept, du matériel scientifique obsolète, témoin non seulement d'une époque, mais de pratiques scientifiques ou pédagogiques, faisant partie de notre histoire. À ce titre, il doit être évalué et préservé. L'intérêt pour la préservation des archives et instruments scientifiques et techniques s'est de plus en développé et se caractérise par des actions coordonnées comme celui de

1 Colloque Patrimoine scientifique et technique, culture et société ? Du 13 Au 14 mars 2008 - Musée des arts et métiers, Paris, http://www.ocim.fr/Colloque-Patrimoine-scientifique

2 Robert Halleux, Directeur de recherche du FNRS, « L'histoire et le patrimoine dans la culture scientifique, technique et industrielle », http://www.embarcaderedusavoir.ulg.ac.be/journeeshubertcurien/actes/JHCurien-RHalleux.pdf

l'Union Internationale d'Histoire et de Philosophie des Sciences[3]. La mise en valeur de ce patrimoine contribue également au développement de la culture scientifique et technique, permettant aux publics de se familiariser avec les savoirs, les techniques et les innovations.

Comme le souligne les entretiens de Jacques Cartier[4], les francophones bénéficient d'un patrimoine scientifique et culturel dont nous pouvons être fiers et qu'il faut faire partager. Seule une recherche de qualité garantira l'existence d'un espace scientifique francophone. Il s'agit d'accélérer les échanges, de multiplier les rapprochements pour que des projets scientifiques de qualité voient le jour. C'est l'identité francophone qui est en jeu[5]. La rétroaction de la communauté scientifique et technique est très positive à ce sujet, il est opportun de faire le point sur de nombreuses initiatives engagées dans le monde francophone. D'exemples actions (internationales, nationales, régionales ou institutionnelles) menées ou en cours au sein de divers organismes (musées, ministères, universités, collectivités, associations, PME, associations …) francophones, constituent un levier significatif pour la sauvegarde et de valorisation du patrimoine scientifique et technique contemporain, particulièrement pour la recherche. Comme le projet de numérisation des 200 000 « *manuscrits de Tombouctou* », témoins de la grandeur de la ville entre le XIV° et le XIX° siècle ; Conservatoire National des Arts et Métiers à Paris, pionnière en la matière par la création du conservatoire numérique des arts et métiers suite à une mission nationale de sauvegarde du patrimoine scientifique et technique contemporain confiée à son directeur. Le CNUAM couvre le domaine de la recherche et de l'enseignement en histoire des sciences et des techniques, en épistémologie et en didactique. Deux autres exemples de projet de grande envergure sont : le Centre d'Histoire des Sciences et des Techniques de l'Université de Liège, et Bibliotheca Alexandrina.

Les notions de collecte, conservation, sauvegarde et valorisation font ressurgir aussi bien les enjeux que les problèmes méthodologiques n'ont résolus. Comment décider qu'un objet plus qu'un autre sera classé patrimoine ? Les procédures d'évaluation des intérêts de ces objets restent floues ou peu comprises du grand public. Car les processus de sélection, conservation, valorisation des objets font appels à des critères aussi bien objectifs que subjectifs ce qui complexifie la démarche

3 R.W. HOME, P. HARPER, O. WELFELÉ (eds), Arhives of Contemporary Science. Proceedings of the Symposium organised by the Commission on Bibliography and Documentation, Liège,20-26 July 1997, Liège, DHS Secretariat, 1998

4 Créés en 1987, les Entretiens Jacques Cartier s'articulent autour de quatre axes : des colloques pointus initiés et pris en charge par les différents pôles d'excellence de la région Rhône-Alpes, des colloques sur des grands problèmes de société d'aujourd'hui et de demain, des espaces de dialogue consacrés à l'économie, des échanges et des rencontres culturels.

5 LES ENTRETIENS JACQUES CARTIER, http://cjc.univ-lyon2.fr/article.php3?id_article=45

méthodologique. Nous pensons que les outils virtuels peuvent être ici d'un grand secours, tel le laboratoire virtuel de Fabio Bevilacqua mettant en œuvre les instruments électriques de Pavie, et les « histoires d'instruments » sur CD élaborés par Catherine Cuenca[6].

2 Patrimoine scientifique et technique francophone : Un potentiel patrimonial et intellectuel d'exception?

La région francophone peut être fière de ce son riche potentiel patrimonial scientifique et technique comprenant, les établissements d'enseignement supérieur, les organismes de recherche, les réseaux de compétence et de nombreuses initiatives territoriales ou individuelles. Ces organismes et actions offrent un tissu riche pour l'exploration méthodique de ce patrimoine. Avant d'aller plus loin, une remarque préliminaire sur la pertinence de l'association du «*patrimoine scientifique* » au « *patrimoine technique* » s'impose. En effet, associer le patrimoine scientifique au patrimoine technique se justifie par le lien de proximité qui existe entre ces deux types patrimoines. D'une part, dans le domaine de la connaissance, le lien entre science et technique est fort, tout à fait évident dans le Monde moderne et contemporain. Et d'autre part, le patrimoine technique est majoritairement constitué d'instruments obsolètes utilisés par les chercheurs qui malheureusement échappent à notre considération. L'instrumentation scientifique est pourtant un élément essentiel du patrimoine car témoin et conséquence d'au moins deux Révolutions technoscientifiques : celle de la microélectronique (qui a entraîné celles de la Microinformatique et de l'automatisation) et celle de la microbiologie (biologie moléculaire puis la génomique)[7]. Les instruments scientifiques et appareils témoins de la recherche et de l'innovation disparaissent progressivement des laboratoires, remplacés par d'autres, tandis qu'un grand nombre de chercheurs et d'ingénieurs qui ont participé vers 1960-1970 à la création de ces laboratoires quittent la vie professionnelle, alors qu'ils constituent « *une mémoire irremplaçable de ce demi-siècle d'évolution technique et scientifique* ». Une « *mémoire vivante* » qu'il faut conserver et transmettre dans les meilleures conditions aux générations futures[8].

6 Robert Halleux, Directeur de recherche du FNRS, « L'histoire et le patrimoine dans la culture scientifique, technique et industrielle », http://www.embarcaderedusavoir.ulg.ac.be/journeeshubertcurien/actes/JHCurien-RHalleux.pdf

7 Yves THOMAS, directeur de la valorisation de la recherche à l'Université de Nantes, et Professeur à l'Université de Nantes, « Un exemple de sauvegarde du patrimoine scientifique et technologique contemporain »

8 Serge Chambaud, Directeur du Musée des arts et métiers, « La protection du patrimoine scientifique et technique », Le 15ème Cahier des Soirées scientifiques est paru (décembre 2008).

De nombreuses Institutions se sont spécialisées dans la préservation du patrimoine scientifique et technique, mais pour préserver, encore faut-il connaître et inventorier ? La démarche d'inventorisation de cette mémoire scientifique comporte bien évidemment une dimension intellectuelle mais on ne peut réduire à ce seul plan le souci patrimonial. Faut-il rappeler que la région Francophone regorge d'un riche patrimoine scientifique et technique ?
Le patrimoine universitaire et le patrimoine mobilier scientifique et technique représentent un potentiel intellectuel et culturel unique. Pourtant ses infrastructures ne sont cependant que partiellement accessibles au public car cette catégorie de patrimoine est aujourd'hui sous-représentée sur la liste du patrimoine mondial[9]. Les laboratoires doivent alors se montrer, aller vers les citoyens. Et l'objet scientifique prendra alors une dimension démonstrative nouvelle. Car la population connaît mal ces chercheurs arc-boutés sur leurs microscopes, leurs lasers [10] ou leurs archives. En effet, peu de biens reconnaissent explicitement cette dimension comme prépondérante ou même simplement présente dans l'analyse de leur valeur universelle exceptionnelle.

En quelques décennies, les sciences et les technologies ont connu une rapide évolution, le patrimoine scientifique se qui jadis se matérialisait sous forme de « papiers » : cahiers, articles, thèses, actes, etc., se dématérialise peu à peu et se présente sous une forme entièrement numérique transmissible instantanément dans le monde entier constitue une véritable révolution culturelle. Même si le support papier reste évidemment très utilisé, l'essentiel est désormais dématérialisé sur les disques durs des stations, des portables ou des serveurs[11], cette nouvelle forme vient grossir les rangs du patrimoine immatériel[12]. Ce patrimoine immatériel scientifique soulève d'autres problématiques liées à leur organisation et à leur accès (disque dur, supports multimédias, clé USB,

9 Michel Cotte, Conseiller de l'ICOMOS pour le Patrimoine mondial « Le patrimoine scientifique : quelques remarques introductives »

10 Alain Beltran, « Introduction », La Revue pour l'histoire du CNRS, N°14 - Mai 2006, [En ligne], mis en ligne le 3 mai 2008. URL : http://histoire-cnrs.revues.org/document1748.html. Consulté le 29 juillet 2009.

11 Jean-Michel Trio, Adjoint du Délégué Régional Alsace du CNRS « La protection du patrimoine scientifique : une démarche globale »

12 Le patrimoine immatériel est une catégorie adoptée officiellement par l'Unesco pour désigner des pratiques, des représentations, des expressions, des connaissances et des savoir-faire ainsi que les objets, les espaces et les groupes qui leur sont associés. Une convention pour la sauvegarde du patrimoine culturel immatériel a été adoptée par l'Unesco en octobre 2003 et entre actuellement en vigueur. Elle a été signée par la France en juillet 2006. Les critères permettant de définir un patrimoine comme immatériel, l'étendue du domaine qu'il recouvre, les formes prises par sa protection, sont autant de questions qui se posent aujourd'hui à la Direction du Patrimoine et auxquelles la mission à l'ethnologie et le Lahic peuvent contribuer à répondre. La notion de patrimoine immatériel rejoint en effet en grande partie celle de patrimoine ethnologique qui était justement caractérisée par son immatérialité.

messagerie, fichiers joints, …). Dans cette articulation l'inventorisation dudit patrimoine doit porter aussi bien sur leur aspect matériel qu'immatériel. Et pour valoriser ce patrimoine, certains projets de grande envergure dont nous détaillerons quelques exemples dans la partie suivante ont vu le jour.

3 Valorisation du patrimoine scientifique et technique par les TIC : Exemples de projets de préservation et valorisation ?

3.1 La Mission nationale de sauvegarde du patrimoine scientifique et technique

Sur le plan national, la Mission nationale de sauvegarde du patrimoine scientifique et technique confiée au Musée des arts et métiers du Cnam par le Ministre de la Recherche, a pour objectifs principaux de sensibiliser les organismes d'enseignement supérieur et de recherche, ainsi que les entreprises, à la sauvegarde du patrimoine scientifique et technique, de développer un réseau national de sauvegarde et de valorisation du patrimoine de la recherche et de l'industrie, en assurant un rôle de conseil et d'expertise dans le domaine, en suscitant des initiatives régionales, en accompagnant la mise en œuvre du programme dans les régions.

3.2 Projet de numérisation des manuscrits arabo-islamiques[13]

« Les apports arabo-islamiques dans le domaine des sciences médicales » est le premier volume d'une série intitulée « Contribution de la civilisation arabo-islamique aux sciences » récemment publié en coopération par le Bureau de l'UNESCO au Caire, le Centre national égyptien pour la documentation sur le patrimoine culturel et naturel (CULTNAT) et la Bibliothèque nationale d'Egypte.[14]

« Ce projet pilote est la première phase d'un effort encyclopédique de longue haleine pour utiliser les TIC à fin d'accès et d'archivage des trésors uniques des civilisations arabo-islamiques »[15]. Le principal résultat de cette première phase du projet a été la numérisation de 2.000 manuscrits destinés à tomber dans le domaine public, au bénéfice des chercheurs, analystes et de toutes les personnes intéressées. Les manuscrits numérisés font partie d'une collection de 1.084 manuscrits

13 http://portal.unesco.org/ci/fr/ev.php-URL_ID=5322&URL_DO=DO_TOPIC&URL_SECTION=201.html

14 © UNESCO 1995 - 2009

15 explique Tarek Shawki, du Bureau de l'UNESCO au Caire

d'origine arabe, turque ou persane relatifs aux sciences médicales et détenus par la Bibliothèque nationale égyptienne du Caire. Le catalogue sur support papier fournit, quant à lui, la description détaillée de 31 d'entre eux. La totalité de la collection numérisée sera publiée sur un CD-Rom trilingue (arabe, français et anglais) et sera publiée sur internet en fin d'année 2009.

3.3 Valoriser le patrimoine de "la Cité de l'écrit" : Tombouctou, la Renaissance par le Numérique[16]

Le projet de numérisation des manuscrits de Tombouctou entre dans le cadre de la coopération décentralisée entre la région Rhône-Alpes et la région de Tombouctou, région située au nord du Mali, entre le fleuve Niger et les frontières algériennes et mauritaniennes.

Ce vaste projet, dont l'ambition première est de redonner à Tombouctou son rôle historique de « *Capitale du Savoir et de la Culture* », consiste d'une part, à inventorier les manuscrits, à élaborer le catalogue, à

16 http://www.manuscritsdetombouctou.org/fr/introduction.php

informatiser le réseau des bibliothèques, et d'autre part à contribuer au mouvement de sauvegarde physique des manuscrits à travers la numérisation de 50 000 manuscrits (4 millions de pages). Il permettra aussi de créer des conditions de stockage pérenne des données numérisées, d'élaborer une bibliothèque numérique et de créer un pôle de compétence regroupant un centre d'études supérieures de l'écrit et des arts graphiques, un musée, des ateliers de copie.
Ce vaste chantier s'inscrit comme une suite sociale et économique logique du cœur du projet intégré à la stratégie de développement touristique de Tombouctou et sa région. Environ 180 à 200 000 manuscrits anciens, provenant de l'Espagne Arabo- Andalouse, de la vallée du Niger, du Sahara central, de la Mauritanie, du Maroc... sont conservés dans la région de Tombouctou. Certains sont universellement connus et reconnus pour leur intérêt historique et relatent le passé d'une société aux activités commerciales intenses, muée par une pensée africaine et musulmane « humaniste » très originale. Ces manuscrits sont actuellement gardés en divers lieux et entretenus de diverses manières : au CEDRAB-IHERIAB créé par la volonté de l'Etat malien en 1970, dans des bibliothèques privées entretenues par les descendants des fondateurs, ouvertes ou non au public, mais aussi dans des demeures familiales. Face aux problèmes de dégradations physiques, de trafics et de difficultés de conservation de ces manuscrits dans un pays pauvre frappé par une succession d'années sèches et une période de guerre, de nombreuses réflexions sur l'inventaire, la restauration et le stockage de ces manuscrits ont vu le jour ces dernières années.

3.4 Projet à l'échelle francophone[17]

RFBNN Réseau francophone des bibliothèques nationales numériques

Le RFBNN a pour mission de réunir au sein d'une instance coopérative ouverte les grandes institutions documentaires de la Francophonie déjà engagées dans des programmes de numérisation patrimoniale, ou développant des projets dans ce domaine. Il entend offrir à ces institutions un forum d'échanges permettant de mener de façon concertée les initiatives lancées dans le champ de la numérisation patrimoniale, afin d'éviter tout dédoublement d'effort et d'assurer le partage des meilleures pratiques.

17 http://www.rfbnn.org/html/Pages/mission.htm

Le RFBNN contribuera ainsi à la préservation à long terme et à la diffusion auprès d'un large public d'un patrimoine précieux et souvent menacé de disparition, faute de conditions adéquates de conservation.
Le RFBNN assurera également un transfert de savoir-faire auprès d'un nombre croissant d'institutions documentaires de la Francophonie par l'organisation de stages de formation, l'élaboration d'outils didactiques et l'échange permanent d'information entre ses membres.
Pour l'instant seuls 14 pays francophones participent à ce projet de grande ampleur, il s'agit :
Belgique, Canada, Cambodge, Égypte, France, Haïti, Luxembourg, Madagascar, Mali, Maroc, Québec, Sénégal, Suisse, Tunisie, Vietnam.

3.5 La Bibliothèque numérique mondiale (BNM)[18]

L'UNESCO en collaboration avec 32 institutions partenaires lance la Bibliothèque numérique mondiale, une plate-forme promouvant pour la libre circulation de l'information, la solidarité internationale, la célébration de la diversité culturelle et l'édification de sociétés du savoir. Cet hypermédia propose un éventail unique de matériels culturels provenant de bibliothèques et d'archives d'un peu partout dans le monde. Le site offre manuscrits, cartes, livres rares, films, enregistrements sonores, illustrations et photographies. L'accès à ces ressources est libre et gratuit. L'objectif visé est de réduire la fracture numérique, à promouvoir la compréhension et à renforcer la diversité linguistique et culturelle. Outre la promotion de la compréhension internationale, le projet vise à augmenter la quantité et la diversité des contenus culturels sur internet, à fournir des matériels aux éducateurs, aux élèves et au grand public, mais aussi à réduire la fracture numérique au sein et entre les pays, en renforçant les capacités dans les pays partenaires. La BNM offrira des fonctions de recherche et de navigation en sept langues (anglais, arabe, chinois, espagnol, français, portugais et russe) et proposera des contenus dans plus de quarante langues. Parmi les autres trésors figurant dans la BNM, on trouve des manuscrits scientifiques arabes provenant de la Bibliothèque nationale et des Archives d'Egypte ; d'anciennes photographies d'Amérique latine

18 La BNM a été développée par une équipe de la Bibliothèque du Congrès. Une aide technique a été fournie par la Bibliotheca Alexandrina (Alexandrie, Egypte). Parmi les institutions ayant contribué à la BNM, on compte des bibliothèques nationales et des institutions culturelles ou éducatives d'Afrique du Sud, d'Arabie saoudite, du Brésil, de Chine, d'Egypte, des Etats-Unis, de France, d'Iraq, d'Israël, du Japon, du Mali, du Maroc, du Mexique, d'Ouganda, des Pays-Bas, du Qatar, du Royaume-Uni, de la Fédération de Russie, de Serbie, de Slovaquie et de Suède.
Site web officiel : http://www.wdl.org/fr

fournies par la Bibliothèque nationale brésilienne ; le Hyakumanto darani, un parchemin datant de l'an 764 détenu par la Bibliothèque du Parlement japonais ; la fameuse Bible du diable, du XIIIème siècle qui se trouve à la Bibliothèque royale de Stockholm ; des calligraphies arabes, persanes et turques provenant de la Bibliothèque du Congrès.

4 Méthodologie de valorisation du patrimoine scientifique et technique proposée dans le cadre du projet EFRARD

Nous constatons fort agréablement l'émergence de projets patrimoniaux ayant recours aux TIC dans une approche de numérisation, d'archivage et de conservation, mais également à des fins d'attractivité et de valorisation touristique de ces territoires. Le projet EFRARD[19] s'organise en réseaux scientifiques et d'expertises par spécialisation et thématique sous la forme d'une structure matricielle au sein de laquelle coopèrent étroitement les principaux acteurs de la R&DI (chercheurs, experts, administrateurs, politiques et institutions francophones) pour partager leurs expériences et savoir-faire. La méthodologie développée dans le cadre du projet EFRARD[20] vise à la fédération des différents acteurs francophones de la sauvegarde, de la conservation, de la valorisation et de la recherche dans le domaine du patrimoine scientifique et technique. Les collectivités territoriales ont un rôle primordial dans cette valorisation, puisqu'elles possèdent chacune des outils de culture scientifiques intéressants mais qui gagneraient à renforcer leur coopération avec la communauté scientifique. Par exemple, par la mise en commun du matériel archivistique par des programmes de numérisation. Le premier point central de la problématique des politiques, programmes et actions tient à la question de la collaboration et à la coordination des objectifs communs. Il paraît essentiel de s'interroger sur la notion de patrimoine dans le contexte actuel de révolution technologique. Comment associer inventorisation et la valorisation du patrimoine scientifique et technique aux mutations technologiques actuelles ? Une transformation majeure s'est opérée avec la révolution numérique et la mondialisation, nous

19 Espace Francophone pour la Recherche, le Développement et l'Innovation http://www.efrardwiki.org/plate-forme/index.php?title=Accueil

20 L'espace francophone pour la Recherche, le Développement et l'Innovation (EFRARD) est un Consortium Francophone de Recherche et Développement durable initialement créé par un groupe de chercheurs de l'équipe de recherche C.I.T.U à l'université de Paris 8. L'objectif d'EFRARD est d'aider à la mise en place d'une communauté internationale majoritairement francophone de chercheurs, experts et institutions travaillant sur les enjeux et modalités du renforcement des coopérations entre les acteurs de la Recherche, du Développement et de l'Innovation (R&DI) pour le développement durable et le changement social

obligeant à considérer les mutations technologiques comme une opportunité pour constituer des espaces dans lequel cohabitent immatériel et le matériel. Seules les technologies de l'information et de la communication permettent d'accélérer les échanges, de multiplier les interactions et collaborations pour que des projets scientifiques de qualité voient le jour.

Les objectifs de cette méthode proposée sont de sensibiliser les acteurs de la recherche à la protection de leur patrimoine scientifique (articles, archives, base de données, thèse, instruments,…), la première étape doit consister à faire un état des lieux du patrimoine scientifique et technique reconnus ou non dans les pays Francophones. Il s'agit là d'un vaste chantier qui requiert l'implication de tous les acteurs concernés comme nous l'avons souligné plus haut. Notons que le patrimoine scientifique est souvent volumineux, complexe, délicat et nécessite des investissements importants depuis l'inventorisation jusqu'à la valorisation. Les entités productrices n'ont pas toujours une suffisante sensibilité au sauvetage. Le public lui-même demande à être sensibilisé par des mises en scène scénographiques innovantes et pédagogiques. Une fois le patrimoine inventorier s'en suit un processus de restauration et de préservation, processus qui sera poursuivi par une dissémination de ce patrimoine dans l'espace francophone par le biais de TIC, musées virtuels, portails dédiés et des actions de formation ad hoc. La mise en place d'un dispositif de formation et de collaboration sur la valorisation du patrimoine y compris le personnel privé et public des institutions impliquées dans la préservation, la restauration et la valorisation s'avère indispensable. On comprend donc que la sauvegarde du patrimoine scientifique et technique relève d'une démarche globale, sans cesse renouvelée, à caractère sociétal.

L'universalisme scientifique et la diversité culturelle qui caractérise les pays francophones peuvent être considérés comme un vecteur de richesse et de développement pour la mise en valeur de ce patrimoine matériel à qui s'attache un patrimoine immatériel : les connaissances de chacun de ces instruments. L'un est indissociable de l'autre et une réflexion doit être menée pour que cette dualité puisse être conservée au mieux et immatériel[21]. La coopération entre les chercheurs, les institutions en charge de la gestion du patrimoine, les acteurs politiques et économiques s'en trouve nécessairement renforcée surtout en ce concerne leur recensement mais aussi des possibilités de stratégies de développement durable. Les TIC sont d'un grand secours dans ce type de collaboration car ils permettent non seulement le travail en commun malgré la distance géographique (équipes chargées de recenser, sauvegarde, documenter) mais également la mise en visibilité de ce patrimoine d'un pays

21 Culture Scientifique et Technique, « Préservation du patrimoine », http://cst.univ-pau.fr/live/Patrimoine?isPdf=1

francophone à un autre en mettant en valeur des sites, des musées, des documents, des hommes. Dans cette conception, un inventaire du patrimoine scientifique et technique francophone doit être réalisé, suivi de sa préservation, restauration, conservation et valorisation.
Cette mise en valeur par les TIC, peut se faire à travers un espace virtuel communautaire. Utiliser le terme musée virtuel serait réducteur, car cet espace en plus des fonctionnalités d'un musée virtuel doit également intégrer les moyens de coopération pour un double enjeu : servir la communauté et son développement une fois constituée et, par ailleurs, aider à construire la communauté au préalable. Une itération parait indispensable.

5 Conclusion

La valorisation du patrimoine scientifique et technique est un processus critique dans toute démarche de gestion du patrimoine. En effet, de nombreux travaux et études [22] ont abordé la question du patrimoine scientifique et technique en termes de protection et de valorisation. Si on part du postulat que patrimoine scientifique et technique est un enjeu majeur de visibilité, à des fins de vulgarisation scientifique, d'attractivité touristique et de développement des territoires en question, la valorisation dudit patrimoine peut-être envisagée comme vecteur du développement urbain durable. Le concept du Collaboratoire » représente un potentiel important en ce qui concerne le recensement et la préservation du patrimoine scientifique et technique de recherche, comme moyen efficace qui favorise davantage le travail en commun. La sauvegarde et la mise en valeur du patrimoine scientifique et technique ne sont pas encore une cause gagnée. Notre prochaine réflexion s'attachera à illustrer que le

22 A) Jean-Pierre Dalbera est conseiller du directeur pour la recherche et la technologie au musée des civilisations de l'Europe et de la Méditerranée, avec lequel il a créé un nouveau portail consacré aux recherches ethnologiques. Il est chercheur associé au laboratoire LEDEN de l'université Paris VIII et collabore avec le Centre Georges Pompidou, l'Ecole du Louvre, le Museum national d'histoire naturelle, le CNRS pour susciter de nouvelles formes de communication et d'échange avec le public dans le monde scientifique et culturel. Il a piloté le premier plan national de numérisation du patrimoine et produit d'importantes publications multimédias sur les grands sites archéologiques et les Célébrations nationales, primées dans des festivals internationaux.

B) Etude réalisée en juillet-novembre 2000 par Catherine Roth, sur une commande de la Mission à l'ethnologie de la direction du patrimoine et de l'architecture du ministère de la Culture (responsable de la mission à la date de janvier 2001 ; les enjeux de la mémoire scientifique », 2000

C) Dominique Lecourt professeur l'université de Denis Diderot - Paris VII, L'enseignement de la philosophie des sciences, Rapport au ministre de l'éducation nationale, de la Recherche et de la Technologie.

D) L'enquête sur le patrimoine des sociétés savantes lors du colloque de 1997 du Comité des Travaux Historiques et Scientifiques (CTHS),

F) Les travaux de Jean-Marc Levy-Lebond, notamment « Patrimoine scientifique et recherche », Le patrimoine écrit scientifique et techniques : définition, usages et accessibilité, ou encore « Défisciences

patrimoine et les nouvelles technologies peuvent s'enrichir mutuellement. L'intérêt pour la communauté francophone est énorme puisque l'idée est basée sur un projet de reconstruction 3/4D du patrimoine scientifique et technique francophone. Pour ce faire, un appel à manifestation d'intérêt sur ce projet permettra aux différents acteurs[23] d'exprimer leur intérêt vis-à-vis du projet mais éventuellement des projets réalisés, en cours, en gestation proche de ce projet. Cet appel à manifestation d'intérêt devra servir de base à la constitution non seulement d'une équipe de volontaires mais aussi de pays ou régions francophones volontaires à un travail collaboratif et collectif sur ce projet de reconstruction 3/4D du patrimoine scientifique et technique francophone.

6 Références bibliographiques

Ali Mahmoud Radwan, « Trésors de collections inconnues : exemples issus du site d'Helwan », dans MUSEUM International, no 225-226 : Paysages du patrimoine en Égypte, mai 2005, p. 87-91 (ISSN 0304-3002)

Callon, Michel (sous la dir.), La science et ses réseaux - genèse et circulation des faits scientifiques, Paris, Éditions La Découverte, Conseil de l'Europe, UNESCO, coll. Textes à l'appui, 1988, 215 p.

Callon, Michel ; Latour, Bruno (sous la dir.), La science telle qu'elle se fait, Paris, Éditions La Découverte, coll. Textes à l'appui, 1990, 391 p.

Colonna, Jean-François, « Du réel au virtuel « , Le patrimoine écrit scientifique et technique, Actes du colloque de Roanne « Mois du patrimoine écrit », 5-6 octobre 1993, Paris, coédition FFCB, ARALD, Roanne - bibliothèque municipale, 1994, pp. 43-51.

Claude FORRIERES, Jean-Luc REMY, « Le rôle des savoir-faire techniques et la protection du patrimoine industriel. Le cas du métal et des machines mécaniques », décembre 1992.

Christian Leblanc, « Recherche, valorisation et gestion du patrimoine sur la rive gauche du Nil: autour du Ramesseum », Museum international, ISSN 0304-3002, Nº. 225-226, 2005 , pages. 79-86

Dalbéra, Jean-Pierre, « Patrimoine culturel et société de l'information », http://www.culture.fr/culture/mrt/numerisation/fr/intro_generale.htm

D. FERIOT, M. LOURENÇO, « De l'utilité des musées et collections des universités », La lettre de l'OCIM, 93 (mai-juin 2004), p. 4-16.

Fathi Saleh, Hala N. Barakat, « Le village planétaire du patrimoine: la contribution du Centre de documentation sur le patrimoine culturel et

23 Collectivités territoriales, chercheurs, architectes, institutions gouvernementales, établissement d'enseignement supérieur, institutions de recherche, acteurs locaux, aux pouvoirs publics ….

naturel (CULTNAT) », Museum international, ISSN 0304-3002, Nº. 225-226, 2005, pages. 73-78.

G. EMPTOZ, D. WORONOFF, « L'histoire des techniques en France. Bilan et perspectives ». Rapport à la D.B.M.I.S.T. et à la Mission Scientifique de la Direction de l'Enseignement Supérieur et de la Recherche, avril 1982.

LAMARD, « Le patrimoine industriel comme vecteur de reconquête économique », Lavauzelle, 2007.

Lévy, Pierre, « Qu'est-ce que le virtuel ? », Paris, Éditions La Découverte, coll. Essais, 159 p.

Poulot, Dominique, « La représentation du patrimoine des bibliothèques, XVIe-XXe siècle », Le patrimoine, histoire, pratiques et perspectives, Oddos, Jean-Paul (sous la dir.), Paris, Éditions du Cercle de la Librairie, coll. Bibliothèques, 1997, pp. 17-41.

Welfelé-Capy, Odile, « Quel archivage pour l'information scientifique en ligne ? », Patrimoine et multimédia : le rôle du conservateur, Colloque 23-25 octobre 1996, Paris, La Documentation française, 1997, pp. 243-254

Yasser Mansour, « Le projet du Grand Musée égyptien : architecture et muséographie », MUSEUM International, no 225-226 : Paysages du patrimoine en Égypte, mai 2005 (ISSN 0304-3002)

Zahi Hawass , « Une nouvelle ère pour les musées égyptiens », ISSN 0304-3002, , pages. 7-23

Un nouveau dictionnaire électronique structuré et évolutif de la langue arabe : DESELA

Abd El Salam AL HAJJAR (1)(2), Mohammad HAJJAR (1), Khaldoun ZREIK (2)

(1)Institut Universitaire de Technologie, Université Libanaise, Liban
(2)Laboratoire Paragraphe, Université de Paris 8 - Vincennes - Saint-Denis, France

Mots-clés : Langue arabe, Corpus, Dictionnaire, Extraction d'information, Racine.

Keywords: Arabic Language, Corpus, Dictionary, Information Extraction, Root.

Résumé : Dans cet article, nous proposons un nouveau dictionnaire électronique structuré et évolutif de la langue arabe (DESELA) qui peut être présenté sous la forme d'une base de données relationnelle ou d'un document XML et qui est facilement exploitable à l'aide des langages de requêtes appropriés. En effet, on trouve beaucoup de dictionnaires arabes mais qui ne sont pas directement exploitables puisqu'ils sont sous forme des fichiers texte plats. DESELA contient essentiellement les racines, les préfixes, les suffixes, les infixes, les modèles et les mots dérivés. De plus, pour un mot donné, il fournit les liens avec sa racine, avec les affixes associés et avec son modèle éventuel. DESELA est alimenté automatiquement à partir d'un ou de plusieurs dictionnaires textuels classiques et est enrichi en permanence avec des corpus textuels arabe quelconques grâce à un système que nous avons construit. Ce système est composé d'un parseur, d'un classifieur, d'un comparateur et d'un analyseur. Le parseur permet de transformer une source textuelle (dictionnaire ou corpus textuel) en un ensemble de mots. Le classifieur permet de classer un mot donné et de l'ajouter au DESELA en tant qu'une racine ou en tant qu'un mot dérivé. L'analyseur permet d'extraire les affixes et le modèle à partir d'un mot dérivé et de sa racine. Le comparateur permet d'éviter d'avoir des doublons, à tous les niveaux, dans DESELA. Ce dictionnaire peut être utilisé pour évaluer ces méthodes d'extraction d'information à partir d'un document arabe. Étant donné que le vocabulaire de la langue arabe est essentiellement construit à partir des racines, un mot arabe est construit à partir de sa racine en y ajoutant des affixes (préfixe, infixe, ou suffixe) selon un modèle précis. La plupart des méthodes d'extraction d'information à partir d'un document arabe procèdent inversement en extrayant la racine à partir du mot.

Abstract: In this article, we propose a new structured and progressive electronic dictionary for the Arab language (DESELA) which can be presented in the form of a relational database or in the form of an XML document which can be easily exploitable using suitable query languages. Indeed, many Arab dictionaries are found but are not directly exploitable since they are in flat textual files form. DESELA contains the roots, the prefixes, the suffixes, the infixes, the patterns and the derived words. Moreover, for a given word, it provides links to its root, to their associated affixes, and to its possible pattern. DESELA is supplied automatically from one or several traditional textual dictionaries and is enriched permanently with any Arab textual corpus using system that we built. This system is composed of a parser, a classifier, a comparator and an analyzer. The parser allows transforming a textual source (dictionary or textual corpus) into a set of words. The classifier allows to classify a given word and to add it to DESELA as a root or a derived word. The analyzer allows extracting the affixes and the model from a derived word and of its root. The comparator permits to avoid duplication of roots, affixes or patterns in DESELA. This dictionary can be used to evaluate the information extraction methods from an Arab document, given that; the vocabulary of the Arab language is essentially built from the roots. In general, an Arab word is built from its root while adding to it the affixes (prefix, infix, or suffix) according to a precise pattern. Most methods of information extraction starting from an Arab document proceed conversely by extracting the root from the mot.

1 Introduction

Les performances des systèmes d'extraction d'information en langue arabe restent très problématiques et ceci pour plusieurs raisons [1], [3], [6]. L'une des raisons principales est due au fait que le vocabulaire de la langue arabe est essentiellement construit à partir des racines. En effet, la langue arabe possède cinq à sept milles racines distincts. Un mot arabe est construit à partir de sa racine en y ajoutant des affixes (préfixe, infixe, ou suffixe) selon un modèle précis [2], [3], [8]. Ces modèles sont au nombre de cent vingt, environ. Les méthodes d'extraction d'information à partir d'un document arabe procèdent inversement en extrayant la racine à partir du mot. Dans ce domaine, plusieurs méthodes ont été proposées [1], [4], [6], [8] , [11], [14], [15], [17], [21] , [23], [24]. Ces méthodes sont soient basées sur les caractéristiques morphologiques de la langue arabe soient sur des calculs statistiques. Pour évaluer ces méthodes, nous avons développé un système d'évaluation et nous avons construit un corpus limité à vingt racines et à deux milles mots. Pour valider ces résultats, il faut bien sur un corpus plus important, un dictionnaire par exemple.

En effet, on trouve beaucoup de dictionnaires arabes comme Lisan Al-Arab, Al Qamous Al Mouhit, Al Wasit, Al Mouhit, Mouhit Al Mouhit, Al Ghani et d'autres [26], [27], [28], [29]. Bien que ces dictionnaires indiquent la racine, la définition, l'orthographe, les sens et les modes d'utilisation d'un mot donné, ils ne sont pas directement exploitables informatiquement puisqu'ils sont aux formats textuels non structurés (fichiers texte plats). Donc, l'absence d'un tel dictionnaire nous a poussé à construire un dictionnaire électronique structuré et informatiquement exploitable pour l'utiliser dans l'évaluation des méthodes d'extraction d'information à partir des documents arabes.
Dans cet article, nous proposons un dictionnaire électronique structuré et évolutif de la langue arabe (DESELA). Ce nouveau dictionnaire peut être présenté sous la forme d'une base de données relationnelle [18], [19] ou d'un document XML [22] facilement exploitable à l'aide des langages de requêtes appropriés. Ce nouveau dictionnaire contient les racines, les préfixes, les suffixes, les infixes et les modèles, en plus des informations fournies par un dictionnaire classique. De plus, il fournit les liens d'un mot donné avec sa racine, avec les affixes associés et avec son modèle éventuel. Pour atteindre cet objectif, nous avons construit un système automatique qui permet d'alimenter DESELA à partir d'un ou de plusieurs dictionnaires textuels classiques. Ce système permet aussi d'enrichir DESELA, en permanence, à partir d'un corpus textuel arabe quelconque d'où l'évolutivité de notre dictionnaire.

2 Architecture

La figure 1 présente l'architecture générale de notre système qui permet d'alimenter et d'enrichir automatiquement DESELA à partir de plusieurs dictionnaires textuels classiques et des corpus textuels arabe quelconques. Ce système est composé de plusieurs modules qui sont le parseur, le classifieur, le comparateur et l'analyseur.

- Le parseur permet de transformer une source textuelle (dictionnaire ou corpus textuel) en un ensemble de mots.
- Le classifieur permet de classer un mot donné et de l'ajouter au DESELA en tant qu'une racine ou en tant qu'un mot dérivé. L'analyseur permet d'extraire les affixes et le modèle à partir d'un mot dérivé et de sa racine.
- Le comparateur permet d'éviter d'avoir des doublons dans DESELA, à tous les niveaux. Une remarque d'ordre générale pour cette partie est que les transcriptions de tous les mots arabes utilisés (racine, affixes, modèles, mots) dans ce document sont données dans la table 1.

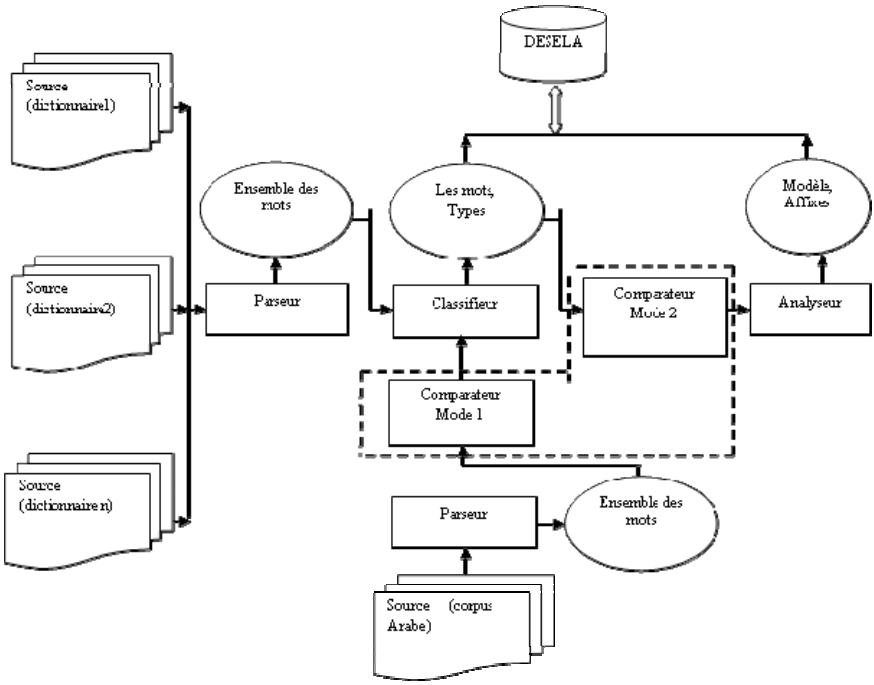


Figure 1 : Architecture générale du système d'alimentation et d'enrichissement automatiquement de DESELA.

1.1 Le parseur

Le parseur constitue le point d'entrée de notre système. L'objectif de ce composant est de transformer une source textuelle en un ensemble de mots. Il peut bien s'agir d'un dictionnaire sous format des fichiers texte plats d'ou d'un corpus textuel quelconque. Le parsing d'une source textuel est effectué en plusieurs étapes. La première sert à déterminer les délimiteurs qui séparent les mots. Ces délimiteurs peuvent être des espaces, des symboles particuliers ou d'autres selon le document à traiter. La deuxième étape consiste à fournir un premier ensemble des mots bruts à partir du document source. La dernière étape dans ce module sert à nettoyer l'ensemble de mots bruts ainsi obtenus. Cette étape consiste en plusieurs phases. La première sert à éliminer les non-caractères, les chiffres et les symboles de l'ensemble des mots bruts. La deuxième sert à en supprimer les mots parasites, ou des mots courts (من, إلى,...) pour les ajouter à DESELA. Donc, la sortie de cette dernière étape, et du cout du parseur, est un ensemble de mots qui sont, soient des racines, soient des mots dérivés des racines.

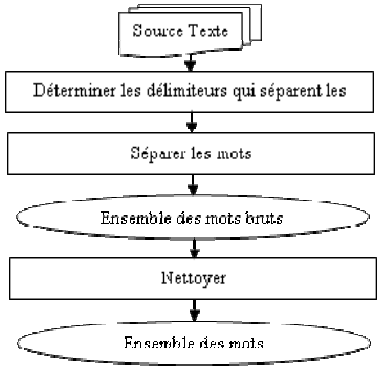


Figure 2 : Le parseur.

1.2 Le classifieur

Les entrées de ce composant peuvent être fournies soit par le parseur, soit par le comparateur. L'objectif de ce composant est de décider si un mot est une racine ou non. Le classifieur permet de classer un mot et de l'ajouter au DESELA. Trois classes sont possibles : racines, mots dérivé d'une racine, mot isolé. S'il est une racine, il l'ajoute au DESELA en tant qu'une racine. Dans le cas contraire, il détermine la racine de laquelle il dérive, l'ajoute au DESELA en tant que mot dérivé et établit le lien avec sa racine. S'il n'a pas de racine, il est isolé, dans ce cas aucun lien n'est établit avec les racines.

La question primordiale à résoudre dans ce composant est : comment déterminer si un mot est une racine ? Pour répondre à cette question, plusieurs cas se présentent (Figure 3).

Dans le cas d'un dictionnaire, les racines sont, en générale, encadrées par des séparateurs spéciaux et les mots, qui sont situés après cette racine et avant la racine suivante, dérivent de la première. Le faite de valider un mot avec sa racine est dû au faite que certains mots qui se trouvent après une racine peuvent ne pas dériver d'elle. Ce type des mots est à ne pas considérer dans DESELA. Pour déterminer ce type des mots, nous utilisons l'une des méthodes d'extraction de la racine d'un mot arabe [24]. Par exemple, dans le cas du dictionnaire Lissan Al Arabe [26], [27]. Chaque racine est précédée par le symbole « @ » et suivit par le symbole « : » (Figure 4), la plupart des mots qui sont situés après une racine et avant la racine suivante dérivent de la première. Considérons l'exemple de la racine أكل donné dans la figure 4. Dans cet exemple, tous les mots qui sont situés entre les deux racines أكل et غرب sont validés par une méthode d'extraction de la racine arabe, en l'occurrence « Arabic Stemming without a root dictionary » [8]. Par contre le mot تقول, qui situe entre les deux racines أكل et غرب ne dérive pas de la première racine أكل.

Dans ce cas, le mot تقول, qui n'est pas validé par rapport à la racine أكل, est à ne pas considérer dans DESELA.
La méthode « Arabic Stemming without a root dictionary » est basée sur l'élimination de plusieurs ensembles de diacritiques et d'affixes et sur l'application de plusieurs modèles qui ont déjà définit [8]. Nous avons choisis cette méthode pour déterminer si un mot est u ne racine ou pour valider les mots par rapport à une racine parce qu'elle n'utilise aucun dictionnaire pour extraire la racine arabe.

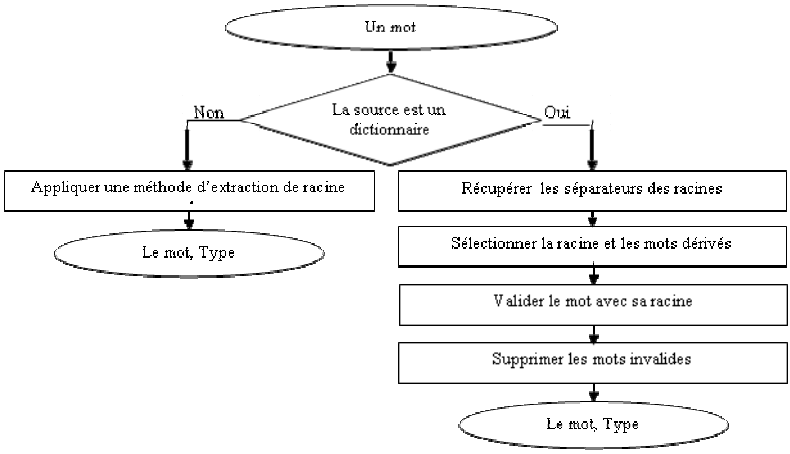


Figure 3 : Le classifieur.

Par contre, dans le cas d'un corpus quelconque, nous utilisons l'une des méthodes d'extraction de la racine d'un mot arabe [24] pour décider si un mot est une racine ou pour déterminer la racine de laquelle elle dérive.

Figure 4 : Cas du dictionnaire Lisan Al Arabe

2.1 1.3 L'analyseur

En générale, un mot arabe est dérivé à partir de sa racine en y ajoutant des affixes (préfixe, infixe, ou suffixe) selon un modèle précis. L'analyseur permet d'extraire les affixes et le modèle à partir d'un mot dérivé et de sa racine. Ce composant prend en entrée un couple {Mot dérivé, Racine} et produit en sortie les préfixes, les suffixes et les infixes éventuels ainsi que le modèle selon lequel le mot est dérivé. Pour ce faire, Nous commençons par repérer les positions des lettres constituant la racine dans le mot dérivé. L'étape suivante consiste à déterminer les lettres appartenant au mot dérivé et ne faisant pas partie de la racine. Ainsi, les lettres qui précédent la première lettre de la racine, si elles existent, dans le mot dérivé constituent les préfixes. De même, les lettres qui suivent la dernière lettre de la racine, si elles existent, dans le mot dérivé constituent les suffixes. En suite, les lettres qui sont situés entre la première lettre et la dernière lettre de la racine, si elles existent, dans le mot dérivé et qui ne font pas parties de la racine constituent les infixes. L'étape suivante consiste à déduire le modèle dans le mot dérivé, le modèle est déduit, selon les positions des lettres constituant la racine dans le mot dérivé. La première étape consiste à supprimer les suffixes, la deuxième consiste à supprimer les préfixes s'ils n'appartiennent pas à l'ensemble {س, ا, م, ت}, la troisième étape consiste à transformer les lettres après les préfixes de l'ensemble {س, ا, م, ت} s'ils y existent de la racine dans l'ordre où la première lettre permute en "ف", la deuxième permute en "ع" et la troisième permute en "ل". Les infixes sont repris tel qu'ils sont.

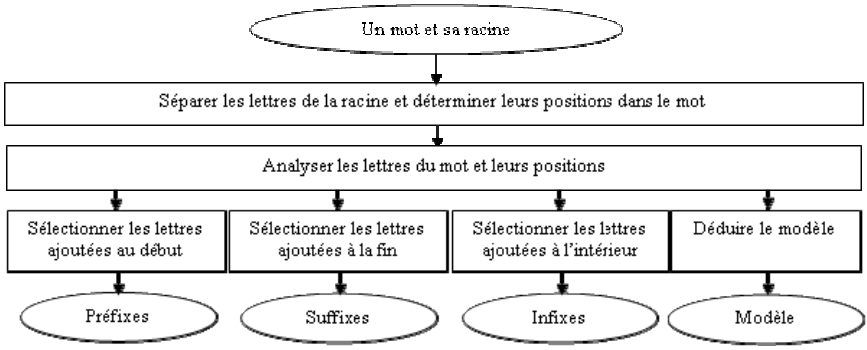


Figure 5 : L'analyseur

Considérons l'exemple du couple {Mot dérivé = المأكولات, Racine = أكل}, la phase du repérage des lettres de la racine (en rouge) dans le mot dérivé

donne ا ل م أ ك و ل ا ت. En suite, les lettres qui précédent la première lettre de la racine الم constituent les préfixes (en vert). Les lettres qui suivent la dernière lettre de la racine ات constituent les suffixes (en jaune). Les lettres qui sont situés entre la première lettre et la dernière lettre de la racine et qui n'en font pas parties و constituent les infixes (en bleu). Le modèle est déduit comme لوعفم, en partant des lettres du mot ا ل م أ ك و ل ا ت, nous supprimons le suffixe ا ت et le préfixe ا ل, car م appartient à {س, ت, م, ل}nous obtiendrons le mot م أ ك و ل, ensuite nous permutons respectivement أ ك ل en ف ع ل. L'infixe و est repris tel qu'il est pour obtenir le modèle لوعفم.

1.4 Les comparateurs

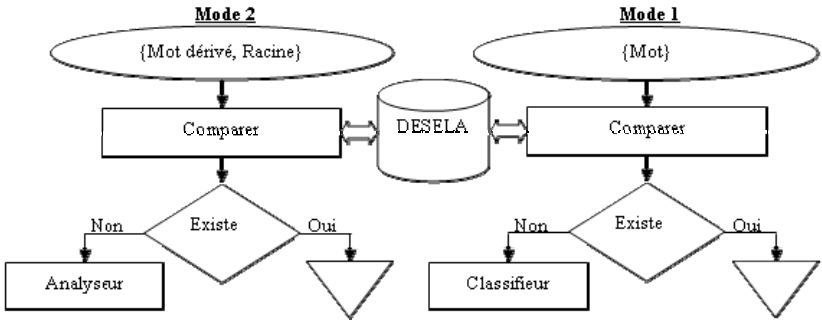


Figure 6 : Le comparateur

Le comparateur permet d'éviter d'avoir des doublons dans DESELA aux niveaux des mots, des racines, des préfixes, des infixes, des suffixes et des modèles. Ce composant est sollicité seulement en phase d'enrichissement. Cette phase d'enrichissement présente le problème suivant : comment enrichir vraiment notre dictionnaire et non pas ajouter des doublons à tous les niveaux. Donc, le rôle du comparateur est de filtrer les mots avant de les ajouter au DESELA. Ce comparateur possède deux modes de fonctionnement. Le premier quand il reçoit en entrée un ensemble de mots fournit par le parseur. Dans ce cas, pour chaque mot il va vérifier s'il existe dans DESELA, s'il n'y est pas il le passe au classifieur pour le traiter. Le second mode de fonctionnement du comparateur est quand il reçoit en entrée un couple {Mot dérivé, Racine} fournit par le classifieur. Ce couple n'est ajouté au DESELA avec le lien entre la racine et le mot dérivé que s'il n'y était pas. Si seule la racine y était, ce couple est passé à l'analyseur pour extraire les affixes et le modèle et si ceux-là n'existent pas dans DESELA, ce mot est ajouté au DESELA avec le lien avec la racine, ils y sont ajoutés.

Mot/Lettre	Transcription
إ	Alef avec Hamza au dessus
أ	Alef avec Hamza on dessous
آ	Alef avec Maada
ب	Baa
ة	Taa Marbouta
ت	Taa
ث	Tha
ج	Jeem
ح	H'a
خ	Khaa
ر	Raa
ز	Thal
س	Seen
ش	Cheen
ص	Saad
ض	Daad
ط	T'aa
ظ	Zha
ع	Ain
غ	Ghain
ف	Faa
ق	Qaf
ك	Kaf
ل	Lam
م	Meem
ن	Noon
ه	Haa
و	Waw
ؤ	Waw avec Hamza
ى	Alif Makzora
ي	Yaa
ئ	Hamza avec Hamza
من	Min
إلى	Ila
أكل	Akala
الأكل	Alakil
المأكلة	Almaekala
المأكول	Almaekol
إيكالا	ikalan
المأكولات	Almaekolate
تقول	takole

أكلت	Akalte
أكل	Akale
أكلا	Aklan
أكلة	Aklah
أكلها	Akalaha
تأكلا	Taeakolan
غرب	Garaba

Table 1 : Les transcriptions des lettres et des mots arabes utilisés dans ce document.

3 Résultat

Le résultat principal de ce travail est le nouveau dictionnaire électronique structuré et évolutif de la langue arabe (DESELA) qui peut être présenté sous la forme d'une base de données relationnelle ou d'un document XML. DESELA contient essentiellement les racines, les préfixes, les suffixes, les infixes, les modèles et les mots dérivés. De plus, pour un mot donné, il fournit les liens avec sa racine, avec les affixes associés et avec son modèle éventuel.
Un deuxième résultat est le système qui permet d'alimenter DESELA automatiquement à partir d'un ou de plusieurs dictionnaires textuels classiques et de l'enrichir en permanence avec des corpus textuels arabe quelconques. Ce système est composé d'un parseur, d'un classifieur, d'un comparateur et d'un analyseur. Le parseur permet de transformer une source textuelle (dictionnaire ou corpus textuel) en un ensemble de mots bruts. Le classifieur permet de classer un mot donné et de l'ajouter au DESELA en tant qu'une racine ou en tant qu'un mot dérivé. L'analyseur permet d'extraire les affixes et le modèle à partir d'un mot dérivé et de sa racine. Le comparateur permet d'éviter d'avoir des doublons, à tous les niveaux, dans DESELA.

4 Conclusion

Dans cet article, nous avons présenté DESELA le nouveau dictionnaire électronique structuré et évolutif de la langue arabe. Ce nouveau dictionnaire peut être présenté sous la forme d'une base de données relationnelle ou d'un document XML facilement exploitable à l'aide des langages de requêtes appropriés. Ce nouveau dictionnaire contient les racines, les préfixes, les suffixes, les infixes et les modèles, en plus des informations fournies par un dictionnaire classique. De plus, il fournit les liens d'un mot donné avec sa racine, avec les affixes associés et avec son

modèle éventuel [2], [3], [6]. Nous avons présenté aussi le système automatique qui permet d'alimenter et d'enrichir DESELA à partir d'un ou de plusieurs dictionnaires textuels classiques et des corpus textuels arabe quelconque.
Notre dictionnaire électronique évolutif et structuré comble un besoin au niveau du patrimoine électronique arabe. Ce dictionnaire peut être utilisé pour évaluer les méthodes d'extraction d'information à partir d'un document arabe cette évaluation contribue sans doute à améliorer les méthodes existantes d'extraction d'information à partir des documents arabes. L'originalité de notre dictionnaire réside dans le fait qu'il s'agit d'un dictionnaire évolutif qui contribue aussi à l'évolution de la langue arabe.

5 Perspective

La prochaine étape est de doter DESELA d'une dimension sémantique en ajoutant des relations sémantiques entre les mots. Pour établir les relations sémantiques entre les mots il faut que nous exploitions les caractéristiques des dictionnaires classiques. En général, un dictionnaire classique fournit les mots avec leurs synonymes. Ces synonymes peuvent être des mots ou des racines. Donc, nous utiliserons le classifieur et l'analyseur pour établir les relations sémantiques aux niveaux des mots et des racines. Ainsi, pour déterminer les relations sémantiques entre deux mots on pourra passer par leurs racines.

Remerciements : Ce travail est effectué dans le cadre des projets "Arabic Web Intelligence" financé par le Centre National de Recherche Scientifique Libanais (CNRSL) et « Recherche d'information Multimedia Multilingue Arabe » financé par le comité Franco-Libanaise (CEDRE).

6 Références bibliographique

[1] W. Adamson George, J. Boreham, *The use of an association measure based on character structure to identify semantically related pairs of words and document titles*, Information Storage and Retrieval, Vol. 10, pp 253-260, 1974.

[2] I. Al Kharashi, *A Web Search Engine for Indexing, Searching and Publishing Arabic Bibliographic Databases*, 1999.

[3] A. Chen, F. Gey, *Building an Arabic stemmer for information retrieval*.TREC-11 conference 2002.

[4] K. Darwish, *Building a Shallow Arabic Morphological Analyzer in One Day*. The ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, USA, 2002.

[5] L. S. Larkey, L. Ballesteros, M. E. Connel, *Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis*, Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275 – 282, 2002.

[6] H. Suleiman Mustafa, *Character contiguity in N-gram based word matching: the case for Arabic text searching* . Information Processing and Management.41 (4), 819-827, 2004.

[7] G. Kanaan, R. Al-Shalabi, J. Jaarn, M. Al-Kabi, A. Hasnah, *A New Stemming Algorithm to Extract Quadri-Literal Arabic Roots*, 2004.

[8] K. Taghva, R. Elkoury, J. Coombs, *Arabic Stemming without a root dictionary*, International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I pp. 152-157, 2005.

[9] H. Al Ameed, S. Al Ketbi, A. Al Kaabi, K. Al Shebli, N. Al Shamsi, N. Al Nuaimi, S. Al Muhairi, *Arabic Light Stemmer: A new Enhanced Approach* , The Second International Conference on Innovations in Information Technology (IIT'05), 2005.

[10] L. Larkey, L. Ballesteros, M. Connell, *Light Stemming for Arabic IR*, Arabic Computational Morphology: Knowledge-based and Empirical Methods, A. Soudi, A. Van Bosch, and G. Neumann Editors. Kluwer/Springer's series on Text, Speech, and Language Technology, 2005.

[11] F .Douzidia, G. Lapalme, *Un système de résumé de textes en arabe*, 2ème Congrès International sur l'Ingénierie de l'Arabe et l'Ingénierie de la langue, Alger, 2005.

[12] Y. Kadri, J. Nie, *Effective Stemming for Arabic Information Retrieval*, proceedings of theChallenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni, 2006.

[13] L. Khreisat, *Arabic Text Classification Using N-gram Frequency Statistics A Comparative Study*, The 2006 International Conference on Data Mining Part of the 2006 World Congress in Computer Sciences DMIN: 78-82, 2006.

[14] F. Ahmed, A. Nürnberger, *N-grams Conflation Approach for Arabic*, ACM SIGIR Conference, Amsterdam, 27 Juillet 2007.

[15] A. M. El-Halees, *Arabic Text Classification Using Maximum Entropy*, The Islamic University Journal (Series of Natural Studies and Engineering) Vol. 15, No.1, pp 157-167, ISSN 1726-6807, http//www.iugzaza.edu.ps/ara/research/, 2007.

[16] A. Khemakhem, B. Gargouri, A. Abdelwahed, G. Francopoulo, *Modélisation des paradigmes de flexion des verbes arabes selon la norme LMF - ISO 24613*, Traitement Automatique des Langues Naturelles, Toulouse, France, 5- 8 Juin 2007.

[17] M. Ben Abderrahmen, B. Gargouri, M Jmaiel, *LMF-QL: A graphical Tool to Query LMF databases*, Third Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland, 2007.

[18] E. Norbert, *Arabic Language Support in SQL Server, Microsoft corporation, SQL Server Technical Article*, http://msdn.microsoft.com/en-us/library/cc295829(SQL.90).aspx, 2008.

[19] C-A. Comes, L-D. Savu, I-O Spatacean, B. Stefan, A. Avram, *Universal Symbolic Translator for Procedural Language over SQL*, 7th WSEAS Int. Conf. on Applied Computer & Applied Computational Science (ACACOS '08), Hangzhou, Chine, 6-8 Avril, 2008

[20] J. Micher, C.Voss, *Buckwalter-based Lookup Tool as Language Resource for Arabic Language Learners Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 66–67, Columbus, Ohio, USA, June 2008.

[21] M. Sinane, M. Rammal, K. Zreik, *Arabic documents classification using N-gram*, Conférence ICHSL6, Toulouse, 2008.

[22] F. Baccar, A. Khemakhem, B. Gargouri, K. Haddar, A. Ben Hamadou, *Modélisation normalisée LMF des dictionnaires électroniques éditoriaux de l'arabe*, TALN 2008, Avignon, France, 9-13 juin 2008.

[23] G. Francopoulo, M. George, *Language resource management − Lexical markup framework (LMF)*, ISO/TC 37/SC 4 Rev.15, 2008.

[24] A. Al Hajjar, M. Hajjar, K. Zreik, *Classification of Arabic Information Extraction methods*, 2nd International Conference on Arabic Language Resources and Tools, Le Caire, Egypte, 21-23 Avril 2009.

[25] A. Al Hajjar, M. Hajjar, K. Zreik, *Un nouveau système d'évaluation des méthodes d'extraction de la racine des mots arabes*, (soumis), 2009.

[26] Ibn Manzour, *Lisan Al-Arab*. www.muhaddith.org, 2009.

[27] Sakher, *Lexicons: Lisan Al-Arab, Al Qamous Al Mouhit, Al Wasit, Al Mouhit, Mouhit Al Mouhit, Al Ghani, Taj Al Arous, Najaat Al Raed*, http://lexicons.sakhr.com, 2009.

[28] Academy of the Arabic Language, http://lexicons.sakhr.com/intro/intro.aspx?fileurl=intro01.asp, 2009.

[29] Islamic Library, *Arabic Dictionaries: Al Misbah Al Mounir, Al Qamous Al Mouhit, Moujam Makayys Al Lougha, Moukhtar Al Sihah,* http://www.islamweb.net/newlibrary/bookslist.php?subject= كتب اللغة العربية, 2009.

La préservation des connaissances : instrumentation de contenus et interprétation des vues documentaires

Bruno BACHIMONT (1), Stéphane CROZAT (2)

Université de Technologie de Compiègne, UMR CNRS Heudiasyc (1)
Université de Technologie de Compiègne, Unité Recherche Action Ingénierie des contenus et des savoirs (2)

Mots-clés : préservation, connaissance, gestion, OAIS, chaîne éditoriale, herméneutique, philologie

Keywords: preservation, knowledge, management, OAIS, editorial chain, hermeneutics, philology.

Résumé : La gestion des connaissances tente d'expliciter la connaissance pour la rendre capitalisable sour la formes de documents. La préservation des contenus vise à maintenir l'intégrité et l'authenticité des contenus au cours du temps et de la longue durée. Or, la gestion des connaissances doit prendre en compte le transmission dans le temps des contenus qui capitalisent la connaissance, et la préservation ne s'effectue qu'en entretenant les connaissances associées aux contenus. L'approche proposée ici consiste à croiser ces deux points de vue pour la préservation des connaissances.

Abstract: knowledge management aims at explicitating knowledge in order to make it usable by means of documents. Content preservation aims at ensuring content integrity and authenticity through time. While knowledge management should take into account document transmission, content preservation relies on knowledge and its maintenance through time. The proposed approach consists in merging these two points of view for knowledge preservation.

1 Introduction de l'article

Alors que l'on parle habituellement de gestion des connaissances, on évoque plutôt la préservation des contenus. En effet, la connaissance serait cet actif immatériel qu'il faudrait seulement gérer, alors que les contenus sont des objets matériels dont il faut assurer l'intégrité physique.

Mais les connaissances s'expriment à travers des inscriptions qui les matérialisent, et les contenus s'interprètent par des connaissances qui les actualisent. Les opposer est donc artificiel, les traiter séparément stérile.
Or, ces deux domaines restent étonnement indépendants, la gestion des connaissances ignorant la longue durée, la préservation la connaissance. La proposition de cet article est de les articuler, chacun pouvant apporter à l'autre ce qui lui manque. La préservation nous apprend que la connaissance ne s'entretient que par son usage et sa mise en pratique : l'usage permet la conservation, non l'inverse. La gestion nous apprend qu'il faut expliciter les différents savoirs dont on dispose pour permettre leur formulation et transmission.
Nous proposons donc une approche de la préservation par l'accès où les connaissances sont capitalisées, préservées et transmises à travers une pratique entretenue et permanente, reposant sur une instrumentation des contenus. Dans le contexte numérique contemporain, les contenus matérialisant les connaissances sont en effet soumis à des transformations permanentes, que ce soit pour les lire ou les exploiter, si bien que tout accès devient dès lors une transformation des contenus. La préservation par l'usage devient une herméneutique des avatars documentaires créés par l'usage.

2 Une approche patrimoniale pour la capitalisation

2.1 Gestion des connaissances et inscription documentaire

La gestion des connaissances se définit essentiellement aujourd'hui comme le fait d'expliciter, rassembler et diffuser les savoirs et savoir faire d'une organisation ou d'une communauté. Ses principales difficultés relèvent alors de la difficulté de formuler et formaliser ces savoirs et savoir faire pour pouvoir les collecter, et d'autre part de les adapter et les convertir pour les diffuser et partager. Formulée ainsi, la problématique se pose essentiellement en terme de transcription, de l'implicite à l'explicite, de l'informel au formel, du pratique au théorique, du spontané au rationalisé. Cela étant, la gestion des connaissances bute sur deux difficultés :

- L'ancrage documentaire des contenus rassemblés et formulés
- La gestion dans le temps de la vie de ces contenus, pour préserver tant leur intégrité que leur intelligibilité.

Le premier problème correspond au fait de pouvoir instrumenter les contenus de manière à permettre leur exploration, comparaison et exploitation. En particulier, il est nécessaire de pouvoir mettre en résonance des contenus différents mobilisant des formats hétérogènes

comme le texte, la vidéo, le son, etc. Une problématique générale du dossier de la connaissance émerge, où le lecteur / utilisateur se trouve confronté à une multiplicité de contenus devant lesquels il doit être capable d'abstraire une connaissance et de la mobiliser.
Le second problème correspond au fait que le numérique est un support par nature instable permettant la transformation et la manipulation des contenus. Comme en témoigne les ressources bureautiques, un même fichier se donne à lire de manière différente selon les traitements de texte et outils utilisés. La question est alors de savoir en quoi un contenu vu différemment selon différentes vues produites par différents outils reste le même ; en quoi peut-il garder son authenticité à travers les accès à chaque fois modifiés qu'on peut en avoir. De même, les outils numériques permettent de généraliser une logique du remploi, de la réutilisation, de la ré-éditorialisation. Nombre de contenus sont originaux par leur intention, moins par leur composition qui peut reprendre force d'éléments préexistants. Un nouveau problème émerge ainsi de suivre la généalogie des contenus à travers leur leur mutation documentaire. La problématique est donc la suivante:

- d'une part, on a différents contenus renvoyant à une synthèse à construire et abstraire;
- d'autre part, on a différentes vues de « mêmes » contenus qu'il faut confronter et discuter pour déterminer et étudier l'authenticité et l'identité des contenus.

Le parti pris de cet article est de mettre en avant l'utilisation et la réutilisation des contenus comme approche générale de la capitalisation et de la préservation. D'une part on ne capitalise que ce dont on a besoin et l'on ne préserve que ce que l'on utilise ; et d'autre part l'accès aux contenus pour leur usage est une garantie de leur gestion continue et de leur préservation.
Ainsi, il faut prendre acte et non le regretter que le numérique confère une mutabilité intrinsèque aux contenus. Le problème n'est plus de conserver un contenu identique à lui-même, mais de le suivre et le reconnaître à travers ses différents avatars et représentations qu'il adopte au cours du temps. L'utiliser, c'est donc le transformer, le laisser identique à lui-même, c'est l'oublier. Cette mutabilité se retrouve tant sur le plan de la forme que du fond :

- un contenu numérique doit posséder une lisibilité technique car il repose sur une instrumentation donnée pour être lu ;
- un contenu en général doit reposer sur une lisibilité culturelle car il s'inscrit dans un système de normes et de codes pour être compris et interprété.

Par conséquent, gérer la connaissance, capitaliser pour la transmettre et la réutiliser, sont des tâches qui se définissent sur le fond de cette mutabilité. Pour se souvenir, il faut transformer, pour comprendre il faut interpréter. La capitalisation des connaissances doit donc s'effectuer comme une herméneutique des avatars documentaires créés par l'usage.

2.2 La gestion patrimoniale des contenus

L'approche est ici de considérer ce qui a fait le succès de la transmission des connaissances, pratiques ou théoriques, dans le temps, voire la longue durée, pour l'adapter à la gestion des connaissances et de leur inscription documentaire. Le problème abordé sous l'angle de la transmission peut se résumer de la manière suivante :

- Des contenus documentaires expriment certaines connaissances, savoir faire ou pratiques ;
- Ces contenus, au moment de leur production, sont encore plongés dans un environnement, une tradition de lecture et d'interprétation, qui permet de recouvrer leur signification et d'accéder à leur mise en œuvre, pratique ou théorique.
- Avec le temps cependant, ces contenus se décontextualisent progressivement : un fossé d'intelligibilité se creuse si bien qu'ils deviennent illisibles et incompréhensibles.

Afin d'éviter le creusement de ce fossé, il faut donc que le contexte d'interprétation des contenus soit maintenu au fil du temps. Pour cela il faut que les contenus soient utilisés (lu, annotés, rééditorialisé, etc.) et que les modalités de cet usage soit contrôlé par un dispositif qui permettent et maintiennent les conditions de l'usage dans le temps. La solution élaborée par la tradition occidentale repose sur les éléments suivants :

- Une conservation des supports physiques des contenus. C'est le rôle traditionnel des bibliothèques, qui rassemblent, inventorient et préservent les contenus dont elles ont la charge.
- Une interprétation permanente et dynamique des contenus, qui maintient l'intelligibilité de ces derniers, au besoin en les enrichissant de commentaires, gloses, explicitations, venant faciliter la recontextualisation des contenus dans l'environnement social et intellectuel du moment. C'est traditionnellement le rôle de l'université.
- Quand les contenus ont une dimension performative, comme les outils techniques ou les instruments de musique, on ajoute un « conservatoire » qui a pour but de répéter et transmettre les pratiques associées aux instruments : elles répliquent en leur sein la

dichotomie ci-dessus, en conservant les outils et en transmettant les savoir faire. Ce sont les conservatoires des arts et métiers par exemple, ou les conservatoires de musique.

On voit ainsi que la capitalisation au cours du temps repose sur deux piliers fondamentaux : la conservation des contenus et le maintien d'une tradition d'usage et de lecture. Seul l'usage répété et systématique évite la rupture de la transmission du savoir. La conservation peut être plus ou moins instrumentée, de manière à permettre un accès décontextualisé et autoriser ainsi des relectures dans des contextes différents de l'origine, mais il reste toujours la nécessité de s'inscrire dans un cadre partagé par les compétences du lecteur d'une part et le contexte du document d'autre part. Si par exemple, on peut ajouter des glossaires pour expliquer des mots anciens, on supporte ainsi la perte de savoir, mais on repose sur la connaissance de la langue commune. La question est à présent de savoir comment instrumenter l'usage documentaire pour faciliter la préservation des connaissances.

3 Gestion patrimoniale des contenus et gestion documentaire

L'enjeu est de proposer une approche de la gestion documentaire des connaissances reprenant ces clefs du succès de la transmission patrimoniale des contenus. Cela pose plusieurs questions :

- Identifier les rôles et les actants : que sont les contenus, les interprètes, les institutions ? On s'aperçoit que, souvent, la capitalisation des connaissances constitue bien une sorte de bibliothèque, mais laisse ses documents dormants, sans une université venant constamment lire et actualiser le contenu.
- Instrumenter les rôles et les contenus : comment traduire de manière éditoriale la conservation des contenus de manière à rendre possible la dynamique interprétative et à intégrer dans la conservation et la transmission le résultat de ces interprétations ?

L'idée est alors de constituer une éditorialisation des contenus et de la glose comme la condition de la gestion documentaire, de manière à conserver son intelligibilité. Les propositions que l'on peut faire sont les suivantes :

- La gestion des contenus : pour une philologie documentaire et numérique. L'objectif est de revenir sur l'identité d'un contenu, son authenticité et intégrité, pour définir quelles sont ses différentes versions, variations et transformations. Ce problème est particulièrement aigu dans le contexte numérique où le simple fait

de lire un contenu à l'écran est d'emblée la transformation de la ressource conservée, où les reproductions se réalisent instantanément (copier/coller) et où un même contenu physique (un fichier par exemple) peut être mobilisé au sein de document différents (une image, une définition, etc.). La solution générale au problème de l'identité du contenu est le maintien d'une généalogie qui permette de connaître les liens entre une instance physique d'un contenu et ses parents (pères, frères, cousins, clones, etc.).

– L'interprétation des contenus : pour une herméneutique documentaire. L'objectif est de préciser comment constituer un apparat critique pour l'explicitation sédimentée d'un contenu, qui vient enrichir sa consultation. Le numérique permet une approche nouvelle de cette sédimentation, dans la mesure où l'ajout, le retrait, le référence, la variation devient une modification *intégrée* au contenu - en fait elle devient le contenu lui même et non seulement une glose au sens d'une marque « en marge ». La solution générale au problème de l'interprétation du contenu est l'enregistrement explicité des différentes couches de modification qui viennent enrichir un contenu au cours du temps et des usages.

Puisque l'approche est de *transformer* les contenus pour les conserver, transformation matérielle pour préserver leur accessibilité physique (changer de format, de support, etc.), et de les *enrichir* pour préserver leur accessibilité intellectuelle, à l'instar des contenus plus anciens comme ceux de la tradition manuscrite, il convient d'élaborer une philologie et herméneutique numérique pour rendre possible et gérer cette *transmanence* des contenus, la persistance patrimoniale du contenu à travers son évolution dans le temps.

4 Instrumentation éditoriale des contenus pour leur transmission

Nous avons posé comme stratégie de préservation une démarche fondée sur l'accès aux contenus, combiné à une gestion de leur identité et de leurs enrichissements. Dans le contexte du documentaire numérique, nous proposons de reformuler cela comme la mise au point d'un dispositif permettant l'instrumentation éditoriale - nous pourrions ajouter continue - des contenus.
L'hypothèse est que les usagers accèdent aux contenus pour les utiliser, et non pour les préserver, mais que cet usage est une source d'information riche pour la préservation. Or, utiliser un contenu, dans un contexte numérique signifie toujours le transformer (même pour le lire, surtout pour l'annoter ou le réutiliser). Ce que le numérique rend possible - voire

impose - le couple homme-machine ne le gère pas forcément. Ainsi nous copions, référençons, commentons à loisir, mais ces transformations ne sont que rarement exploitables a posteriori à des fins de préservation : l'on perd les liens entre les copies, l'on ne sait plus quel fragments est le dérivé de l'autre, etc. Ce qui nous incite à une gestion de l'instantané (le Web et ce qu'il me propose dans l'état où il me le propose maintenant) plutôt qu'à une gestion documentaire rigoureuse (tel contenu dans cette version, relié à tel autre contenu dans telle version, etc.). Notre proposition est donc de proposer aux usagers un dispositif éditorial reposant sur deux facettes :

- Il est l'instrument de l'usage au quotidien : il permet d'écrire, lire, annoter, commenter, compléter, fragmenter, réutiliser, augmenter, enrichir, etc.
- Il est l'instrument de la préservation dans le temps : il enregistre les actes des usagers et maintient la généalogie des contenus (fonction philologique) et l'actualisation des contenus (fonction herméneutique et d'accessibilité physique.)

L'approche présentée ici s'appuie sur plusieurs éléments :

- La chaine éditoriale Scenari, chaine permettant de créer des contenus multimédia structurés selon une modélisation documentaire donnée.
- La norme OAIS (Open Archive Information System) qui propose une organisation des informations et des processus pour gérer sur le temps long l'évolution d'une archive, c'est-à-dire de documents dont on veut conserver l'intelligibilité.
- L'approche Cyclops, élaborée pour la gestion patrimoniale de contenus artistiques et musicaux. Cette approche, fondée sur la préservation par l'accès, propose aux détenteurs de contenus de les déclarer et définir

5 Standards techniques et méthodes

5.1 La chaîne éditoriale Scenari

Scenari est un outil intégré de conception et de déploiement de chaînes éditoriales XML [1]. Il se compose de SCENARIbuilder et de SCENARIchain. SCENARIbuilder est un outil permettant de créer des *modèles documentaires*, c'est à dire des règles et programmes fixant la structure d'une famille de documents (schéma XML, règles de validation, etc.) et son comportement (logique d'édition, formats de publication, etc.). SCENARIchain est l'environnement permettant d'exécuter les

modèles créés avec SCENARIbuilder, il propose principalement un éditeur XML WYSIWYM[1], un gestionnaire d'intégrité et des interfaces de publication (HTML, Open Document, etc.).
Le principe de la chaîne éditoriale XML consiste à proposer un environnement permettant la création des contenus selon un langage XML qui représente la structure documentaire logique, plutôt - comme dans les outils bureautique - qu'un format de mise en forme. Dans un second temps l'outil permet de transformer ces contenus XML (dit canoniques) en des vues lisibles suivant des formats comme HTML pour le Web ou Open Document pour l'impression. Cette approche modifie assez profondément le rapport au document numérique, en ouvrant en particulier les voies du polymorphisme et de la rééditorialisation *sans recopie*.
Le *polymorphisme* consiste en la possibilité technique de disposer d'une *source unique (single sourcing* en anglais) de contenu et de la transformer à volonté selon les supports et mises en formes désirés. Le polymorphisme est un possible technologique qui reste limité dans la pratique : en effet il est rare que l'on souhaite présenter exactement la même information sous deux supports différents pour deux usages différents. Une nouvelle publication implique généralement la sélection du contenu (telle partie en plus, telle partie en moins), sa réorganisation (telle partie avant telle autre), sa remise en contexte (introduction, conclusions, transitions), etc.
L'idée est alors de profiter du découpage logique du contenu formalisé selon un langage XML métier, pour appliquer des césures physiques (découpage de fichiers XML et utilisation de liens par référence). Il devient alors possible de partager de mêmes fragments documentaires entre plusieurs documents, ce qui permet la réutilisation sans recopie. On appelle *ré-éditorialisation* (le terme anglais de *repurposing* étant encore plus adéquat) la remise en contexte de fragments issus d'un fonds documentaire, par leur réagencement au sein d'un nouveau document, leur augmentation par une création de contenus spécifiques et leur publication sur un nouveau support et/ou pour un nouveau public.
La chaîne éditoriale XML est donc un outil pour l'usage (il permet de créer, modifier, réutiliser, publier) et pour la préservation (il assure l'accessibilité technique par le langage XML technologiquement indépendant, il permet la réutilisation sans recopie, il facilite la mémorisation des différentiels entre versions, etc.).

5.2 Open Archive Information System

OAIS est une norme internationale issue du monde de l'astronomie qui a vocation à gérer les archives, en particulier numérique. Signifiant *Open*

[1]What you see is what you mean.

Archive Information System, OAIS n'a pas pour vocation de décrire les opérations technique de l'archivage ni de prescrire des processus particuliers, mais de fixer un cadre, une terminologie et un référentiel pour la gestion des archives. Elle propose un modèle d'information et un modèle de processus. Le modèle d'information repose, entre autres choses, sur l'idée fondamentale qu'un objet numérique n'est lisible que si l'on possède une information sur la manière de le lire. OAIS appelle cette information la *representation information* Par exemple, une suite de caractères aura pour information de représentation le fait qu'elle est exprimée en ASCII. Cette information a elle-même besoin d'être exprimée de manière numérique, et donc, en tant qu'information numérique, elle possède donc sa propre information de représentation. Par ailleurs, le modèle de processus prescrit que l'archive, comme institution, doit en permanence surveiller l'environnement pour détecter quand un écart, un fossé, se creuse entre ce que l'archive contient et ce que son environnement peut comprendre. En effet, une archive n'est pas gardée pour elle-même, mais pour une communauté désignée qui pratique la connaissance et les savoirs lui permettant de se saisir de l'archive. Quand un écart s'installe entre l'information numérique, son information de représentation et la communauté désignée, c'est qu'il faut enrichir l'archive de nouvelles informations de représentation.
A la base de nombreux projets de préservation, OAIS est une norme étonnement ouverte qui permet d'articuler l'usage de l'archive à sa conservation. Cependant, si elle intègre bien le fait de devoir surveiller l'exploitation par la communauté désignée et sa capacité de compréhension, elle ne prévoit pas d'articulation particulière pour intégrer de manière régulière les connaissances produites par cette communauté dans son utilisation de l'archive, faisant naturellement évoluer l'archive par son usage même. OAIS en reste encore à une vision centralisée, où l'archiviste contrôle et surveille l'usage pour faire évoluer le contenu, au lieu de suivre l'évolution du contenu par l'usage. C'est pourquoi nous proposons une méthodologie pour l'archivage reposant sur la préservation par l'accès, l'enjeu étant de pouvoir enrichir l'archive et la connaissance par le produit de son usage, conservant ainsi son actualité et son intelligibilité. Cette approche a été étudiée dans les arts médiatiques (voir section suivante).

6 Terrains

6.1 Contenus artistiques : préservation des arts médiatiques

Les arts médiatiques proposent un cadre particulièrement stimulant pour la capitalisation et la gestion des connaissances dans la mesure où les

objets conservés sont soumis à une double obsolescence technologique et culturelle. Exploitant des dispositifs hétérogènes, faisant appel à différents dispositifs et techniques, les œuvres médiatiques se conservent difficilement, le maintien de leur identité interdisant la possibilité de les rejouer ou de les réactiver et donc d'exprimer leur intelligibilité. De nombreuses initiatives préconisent de revenir aux propriétés significatives, ou les invariants de l'œuvre, dont on conserverait la description pour la reproduire plus tard en respectant ces descriptions. Autrement dit, on n'interdit pas l'usage au profit de l'identité, mais on accepte de réinventer l'œuvre dans sa totalité pour lui permettre de rester vivante, au prix de transformations à qui l'on impose de respecter certains invariants.

6.2 Expertises d'entreprises : capitalisation des expériences

L'UTC et la Caisse régionale de Crédit Agricole Brie Picardie collaborent dans le cadre du projet de recherche CAP-XP visant à créer une plate-forme d'échange de connaissances métier et d'expériences pour le réseau des commerciaux de la banque privée.
Le projet a mis à disposition d'un réseau d'une trentaine de commerciaux et directeurs d'agence un système éditorial leur permettant de consigner des comptes rendus d'expérience (une action de vente réussie, une action de conseil, etc.) et des éclaircissements techniques directement liés à leur métier (modification de loi, tendance économique, etc.). Le système permet (notamment) :

- l'édition XML des contenus en ligne (des contenus de type « histoire » et « note technique »)
- la publication pour la consultation Web directe, sous forme de fiches imprimables et sous forme de diaporamas
- la constitution de dossier regroupant des histoires ou notes techniques
- la « promotion » des histoires et notes techniques en (respectivement) cas et fiches techniques. Cette promotion signifie que les contenus sont copiés dans une nouvelle version qui va pouvoir être enrichie par la communauté, avec pour objectif d'obtenir des versions officielles finalisées (validées par la communauté d'usagers)

Ce dispositif rend possible deux axes perpendiculaires de sédimentation d'usage pour passer de la captation à la capitalisation : un axe vertical, le *raffinement* progressif et un axe horizontal, la réutilisation.
À la création du contenu, celui-ci est plutôt brut (selon le temps que le contributeur aura pu passer à l'élaborer) et plutôt personnel (selon la prise de recul du contributeur). Au fur et à mesure des usages de ce contenus

par les autres contributeurs (commentaires, corrections, co-élaboration lorsque les contenus sont promus au niveau de cas ou fiches) le contenu est amélioré pour devenir un contenu d'intérêt général au niveau de la communauté (tout les contenus n'auront bien entendu pas cet avenir).
Sur un axe complémentaires, et quelque soit le niveau de leur raffinement, les contenus peuvent être mobilisés au sein de différents dossiers, pour l'usage quotidien des membres de la communauté : fiche pour soi en vue d'un rendez-vous, dossier technique pour un client, diaporama de présentation pour les collaborateurs de l'agence, etc. Ce second axe est à la fois le moteur du premier (on améliore car on a un besoin) et son consommateur (on peut utiliser les contenus car ils sont de qualité). Le système a été installé en juin 2009 et donnera ses premiers résultats mesurables en septembre 2009, le colloque sera l'occasion de les présenter, et en particulier de valider l'hypothèse du couplage amélioration-utilisation.

7 Conclusion

Les approches proposées sont complémentaires, l'une apportant la méthodologie patrimoniale, l'autre l'approche documentaire, la dernière un environnement normatif. L'objectif consiste alors à proposer un modèle documentaire permettant l'édition structurée des différentes interprétations d'un contenu, intégrées avec sa version courante ainsi que ses versions antérieures. Dans tous les cas, il faut documenter tant la forme que le fond et construire une perspective temporelle permettant au lecteur d'accéder au contenu (la dernière version) et d'adopter s'il le désire une lecture critique qu'il peut approfondir aussi loin qu'il le veut grâce au modèle documentaire proposé.

8 Références bibliographiques

Bachimont, B. (2007). Ingénierie des connaissances et des contenus : le numérique entre ontologies et documents. — Hermès, Paris.

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS (2002) Reference Model for an Open Archival Information System (OAIS). Washington.

S. Crozat, Scenari la chaîne éditoriale libre, Eyrolles, Accès Libre, 2007

Caractériser l'information à partir des processus métiers : méthodes et enjeux

Noémie MUSNIK (1) (2), Manuel ZACKLAD (1), Philippe HAÏK (2), Benoît RICARD (2), Sylvain MAHÉ (2)

(1) CNAM – Laboratoire DICEN
(2) EDF-R&D – Département STEP
** Cette recherche a bénéficié du soutien de l'ANR MIIPA-Doc n°2008 CORD 014 03.*

Mots-clés : gestion des connaissances, classification à facettes, indexation participative, point de vue, système d'organisation des connaissances, recherche d'information.

Keywords: knowledge management, facetted classification, collaborative/participative indexing, point of view knowledge organisation system, information retrieval.

Résumé : Ce projet de recherche porte sur l'analyse et le traitement de contenus documentaires numériques volumineux et hétérogènes, la recherche pertinente et l'accès auxdits documents, situés dans un contexte opérationnel complexe. Cette étude a pour objectif d'apporter des éléments méthodologiques pour supporter la conception d'un outil de gestion de l'information, permettant de répondre au besoin informationnel "en situation", c'est-à-dire prenant en compte les problématiques des métiers mobilisés dans les activités d'exploitation et de maintenance d'une installation industrielle. L'analyse de la complémentarité de deux méthodes d'indexation est au centre de l'étude – l'une dite par « facettes » s'appuie sur l'analyse des processus organisationnels, l'autre dite « participative » se fonde sur la compréhension des activités métiers et introduit la dimension subjective de l'acteur en situation.

Abstract : This research project focuses on the analysis and processing of voluminous and heterogeneous digital documentary content, the information retrieval of and the relevant access to those documents, located in a complex operational context. This study aims to provide methodological framework to support the design of an information management tool, which could respond to the informational needs "in context", i.e. by taking into account the actor issues involved in the operations and maintenance activities of an industrial facility. The analysis of the complementarity of two different methods of indexing is the key issue of this study: one called "facetted classification", based on the analysis of organizational processes, the other called "participative", based on understanding the business activities which introduces the subjective dimension of the actors in situ.

1 Introduction

Confrontées à la nécessité d'une gestion efficace des connaissances, des entreprises industrielles, telle qu'EDF (Électricité de France), sont particulièrement intéressées par l'émergence et le déploiement des technologies de l'information et de la communication (TIC), qui ouvrent de nouvelles possibilités en termes d'accès à un nombre de sources d'information sans cesse croissant et de support aux démarches et processus de travail.
Définir une méthode d'analyse, de traitement et de caractérisation de l'information produite et utilisée dans un contexte opérationnel, en s'appuyant sur la compréhension des processus organisationnels d'une part, des activités et des connaissances métiers d'autre part, constitue une approche originale, qui permettrait de répondre au besoin informationnel des acteurs métiers et des opérationnels terrain impliqués dans des activités d'exploitation et de maintenance d'une installation industrielle complexe.
Dans une première partie nous décrirons le contexte général de l'étude ainsi que le cas d'application identifié, puis nous présenterons les pistes envisagées pour répondre aux problématiques de l'indexation et la recherche d'information en situation. Nous exposerons ensuite deux approches d'indexation : le modèle d'une infrastructure sémantique s'appuyant sur le principe de la classification à facettes, qui permet de caractériser les connaissances à partir des processus organisationnels, et l'indexation participative, qui, orientée sur les usages et la recherche d'information lors de leur utilisation par les métiers, viendra compléter la première approche. Nous introduirons par la suite les premières pistes d'une analyse portant sur l'impact des solutions considérées sur les pratiques métiers concernées.

2 Le contexte EDF et le cadre d'étude

2.1 Les problématiques caractéristiques d'EDF en regard de la recherche d'information

Le besoin de gestion des connaissances est particulièrement sensible à EDF, notamment dans le secteur de la production d'énergie nucléaire[1], dont les métiers sont très techniques et les exigences de sûreté, de disponibilité et de qualité de service importantes. La gestion des connaissances nécessaires à la conception, à l'exploitation, à la

1 80% de l'énergie produite par EDF provient des centrales nucléaires – on compte à ce jour 19 sites nucléaires avec 58 tranches.

maintenance et à la déconstruction des installations constitue en effet un enjeu majeur. Cette nécessité de préserver et de transmettre les connaissances et savoir-faire métiers est accrue par un contexte qui présente un certain nombre de problématiques caractéristiques [2]:

- L'organisation complexe et multi-métiers de l'entreprise ;
- La répartition des sites, dispersés sur l'ensemble du territoire français ;
- Un secteur d'activité à risque très réglementé : l'ensemble des activités et processus critiques suivent des procédures strictes et sont soumises à de fortes contraintes, liées à la sûreté nucléaire, à la radioprotection et à la réglementation environnementale, qui tendent à complexifier les activités.
- La complexité et la variété des systèmes techniques et des outils de production : différentes technologies sont utilisées et plusieurs générations d'outils sont exploitées de manière parallèle ;
- Le renouvellement important des effectifs attendu entre 2007 et 2015 et la forte mobilité professionnelle interne des agents ;
- La durée des cycles de vie industriels, plus importante que la période d'activité professionnelle d'un agent ;
- Les forts enjeux économiques ;
- La mutation des contextes technique et socio-économique dans lesquels l'entreprise évolue : entre autres, l'arrivée d'une nouvelle génération de réacteurs et l'ouverture du marché de l'énergie à la concurrence et la dérégularisation du marché de l'électricité en France et en Europe.

EDF produit, en grande quantité et dans un cadre très réglementé, des données et des documents de conception, d'exploitation et de maintenance de différentes natures et de divers formats, relatifs à ses centrales de production d'énergie nucléaire. Une part importante des activités repose sur une exploitation « intelligente » de documents (utiliser les bons documents dans le bon contexte) – qu'il s'agisse de documents internes (Programmes de Base de Maintenance Préventive, Procédures d'Exploitation, Référentiels Qualité, etc.) ou de documents externes (documentation de constructeurs, normes, documents provenant d'autres exploitants, ouvrages de référence, supports réglementaires, etc.).

Ce rôle central des documents (des informations et des connaissances qu'ils véhiculent) laisse entrevoir l'importance que peut revêtir l'efficacité de la recherche d'information. En effet, la problématique de la gestion des connaissances, des savoirs et des savoir-faire contenus dans les documents de l'entreprise, est considérée comme essentielle à la

performance de l'ensemble des activités de conception, d'exploitation, de maintenance et de déconstruction des outils de production du Parc nucléaire.

2.2 Le cas d'étude : la préparation et la gestion de l'arrêt de tranche pour la maintenance des centrales nucléaires

Nous centrons notre étude sur un contexte particulier : celui de l'arrêt de tranche (AT), c'est-à-dire des périodes d'arrêt d'exploitation programmées, pendant lesquelles une unité de production d'électricité nucléaire est mise à l'arrêt pour procéder au rechargement du combustible et à des opérations de contrôle, d'entretien et de maintenance. Il s'agit d'un contexte fortement contraint du point de vue de la sûreté des installations et de la sécurité des biens et des personnes [2], mais également du nombre important d'opérations à réaliser sur l'installation dans un temps limité (par exemple, de l'ordre de 4000 activités en une trentaine de jours). Dans cette étude, deux phases nous intéressent particulièrement :

- **La préparation de l'arrêt**. Cette phase s'étend sur quelques mois, durant lesquels différents acteurs identifient puis planifient les opérations qui devront être réalisées pendant l'arrêt, en anticipant les possibles difficultés et en optimisant l'utilisation des ressources et la gestion du temps (l'objectif d'un arrêt étant de ne pas excéder la durée cible retenue lors de l'élaboration des plannings nationaux d'arrêt, qui sont les garants de l'équilibre national entre production et consommation énergétiques) ;

- **La gestion des aléas lors de l'arrêt**. S'inscrivant dans la phase de conduite de l'AT, la gestion des aléas correspond à des situations, risquant de remettre en cause le planning prévisionnel de l'arrêt, au cours desquelles il s'agit de comprendre un événement inattendu, d'identifier les solutions à mettre en œuvre et les appliquer, en minimisant les allongements de délais et les surcoûts.

Une bonne préparation et une gestion efficiente des aléas reposent sur l'exploitation efficace de nombreuses informations. Ainsi, lors de ces deux étapes, les acteurs impliqués peuvent-ils avoir à effectuer diverses recherches d'information, entre autres, sur la documentation technique disponible, les retours d'expérience passés, les « bonnes pratiques », le cadrage et les procédures d'exploitation et de maintenance. En outre, les fortes contraintes temporelles qui pèsent sur la préparation et surtout la gestion de l'AT, ainsi que les multiples dépendances entre les tâches à réaliser au cours de celles-ci, mettent en évidence l'importance de l'efficacité de ces recherches d'information.

3 Quelques pistes pour améliorer l'efficacité de la recherche d'information (RI)

3.1 Les principaux écueils de la RI

Dans ce contexte, caractérisé par le nombre important de documents et d'entités, la complexité des circuits d'élaboration et de gestion de ces documents et l'existence d'une organisation complexe en appui aux activités de préparation et de gestion de l'arrêt de tranche, les problématiques relatives à l'accès aux documents [8] et à la recherche pertinente de l'information en situation, ainsi que celle de l'indexation [1], constituent des points clés. Les principaux écueils de la RI sont ce qu'on appelle classiquement le « bruit » et le « silence » :

- Le « bruit » correspond au cas d'une recherche renvoyant à un nombre trop élevé de sources d'information, dont la plupart ne sont pas pertinentes, ce qui rend difficile l'identification des « bons » documents, correspondant au besoin informationnel des opérationnels terrain lors de la préparation de l'arrêt ou au cours de la gestion d'un aléa. À titre d'exemple, retrouver le ReX (retour d'expérience) relatif à un événement similaire à celui rencontré lors de l'AT dans l'ensemble des documents disponibles est quasiment impossible avec les outils actuels de recherche d'information compte-tenu de la dissémination des informations dans un grand nombre de bases.

- Le « silence » correspond au cas où un document pertinent, disponible, n'est pas présenté dans les résultats proposés par le système suite à une recherche d'information. Par exemple, même si un événement a déjà été rencontré par le passé, il est assez rare que le ReX associé ait été formalisé : il est ainsi souvent plus facile de faire appel à la mémoire des acteurs que de se retourner vers les bases documentaires.

En outre, la grande variété des sources (en termes de format, de contenu et de structure), la diversité des bases dans lesquelles sont stockés les documents, ainsi que les contraintes liées à la plus ou moins grande souplesse des droits d'accès aux documents, sont autant de problématiques caractéristiques de notre cas d'étude, qui rendent la recherche d'information peu évidente et conduisent aujourd'hui à des résultats insatisfaisants.

3.2 Les pistes envisagées

Notre étude, centrée sur l'appui à la recherche et à la bonne gestion de documents, a conduit à identifier certains axes de recherche pour

répondre aux problématiques de la recherche d'information en situation. Ainsi, trois pistes ont été jusqu'à présent envisagées :

- **L'étude des contextes de production des documents**, de la structure et des processus organisationnels de l'entreprise, et de l'articulation entre les différentes entités et les métiers qui, en tant que ressources pour la modélisation d'un système de classification à facettes, constitueraient une première approche d'indexation des documents selon un point de vue institutionnel ;

- **L'analyse de la représentation de domaines métiers spécifiques**, *via* un système d'organisation des connaissances (SOC) [2] qui, en s'appuyant sur les représentations et le *sens* des notions manipulées, permettrait de faciliter et d'automatiser certaines étapes d'analyse lors de la classification des documents ;

- Une approche s'appuyant sur **l'intégration d'éléments caractérisant le contexte d'utilisation des documents et l'*utilité* des informations qui y sont contenues**, ainsi que sur les réseaux d'acteurs qui les manipulent. Cette approche permettrait d'introduire la dimension interprétative des acteurs en situation et d'intégrer, dans le système de gestion de l'information, la possibilité de rechercher une information selon des critères relatifs à la pertinence d'un document en fonction de son utilité lors de la gestion d'un aléa.

Nous envisageons ainsi d'étudier et d'expérimenter ces différentes approches de classification et d'indexation des connaissances, en partant de l'hypothèse que leur complémentarité, exploitée au sein d'un système de gestion électronique de documents, permettrait de rendre la recherche d'information en situation plus efficace.

3.3 Le projet de recherche

L'analyse des apports possibles de la complémentarité de deux démarches caractérisation des documents, relevant de l'idée d'exploiter à la fois la situation d'élaboration et de consultation des documents, est au centre de notre étude. Celle-ci s'inscrit dans le projet ANR MIIPA-Doc (Méthodes et services Intégrés Institutionnels et PArticipatifs pour la classification à facettes des contenus DOCumentaires complexes) qui vise à concevoir des méthodes de gestion de l'information pour l'organisation des contenus documentaires « complexes » et à développer l'architecture logicielle correspondante. Cette dernière a pour objectif de permettre un accès unifié à l'ensemble des ressources d'information

2 Les systèmes d'organisation des connaissances (SOC) sont des instruments conceptuels permettant de décrire et d'indexer des contenus documentaires, de représenter dans un formalisme opératoire les connaissances utiles à la méta-description des ressources numériques. Classification, ontologies, thésaurus, taxinomies, folksonomies, sont autant de SOC représentant un domaine de connaissances, utilisés pour diverses manipulations sémantiques aidant à interroger des ressources d'information

distribuées (bases notes, serveurs Intranet, outils de GED des différentes unités, etc.) et aux documents de natures variées de l'entreprise.
Afin de catégoriser les documents en lien avec les connaissances organisationnelles et institutionnelles de l'entreprise, les démarches d'indexation proposées par le modèle ISIS, présenté dans le point suivant, s'appuyant sur la classification à facettes, nous ont semblé pertinentes. Une approche participative – ou par points de vue – exploitant la dimension interprétative de l'acteur en situation, afin d'assurer une convergence entre indexation, pratiques et connaissances métiers, permettrait de compléter cette démarche afin de caractériser les connaissances métiers.

4 Deux approches pour caractériser les connaissances métiers

4.1 Caractériser l'information à partir des connaissances organisationnelles : la classification par facettes

4.1.1 Le modèle ISIS : une infrastructure sémantique pour la gestion des connaissances organisationnelles

Le modèle ISIS[3] (*Infor*mation Semantic Infrastructure Services) est une infrastructure sémantique qui propose de caractériser les documents en s'appuyant sur une méthodologie de classification à facettes [7]. Il permet de décrire à la fois le contenu de l'objet informationnel et le contexte dans lequel il a été créé, afin d'optimiser la précision de la requête de l'acteur, dans le cadre d'une opération de recherche d'information, et d'améliorer la gestion de l'information corporative à travers son cycle de vie. Ce modèle, illustré par la figure 1, constitue un jeu de facettes génériques, qui reflète l'organisation des entités et les objets qui les constituent. Ce jeu de facettes est construit et s'articule ainsi autour de la compréhension des structures organisationnelles et des processus de l'entreprise :

- Les facettes de contexte décrivent les liens entre le contenu du document numérique et son environnement administratif et/ou opérationnel, soit, les fonctions et les activités de gestion et d'exploitation auquel le document est rattaché, le type de contenu, le rôle et la position des acteurs impliqués.
- Les facettes de contenu s'attachent, quant à elles, à décrire « de

3 Le concept ISIS a été développé par la société canadienne Cogniva, à partir, notamment, des travaux de recherche portant sur la classification à facettes de Sabine Mas et de Michèle Hudon .

quoi parle le document », en précisant le corps même de l'information et les références au contexte de production de l'information (titre, auteur, thème, projets et affaires auxquels le document se réfère, etc.).

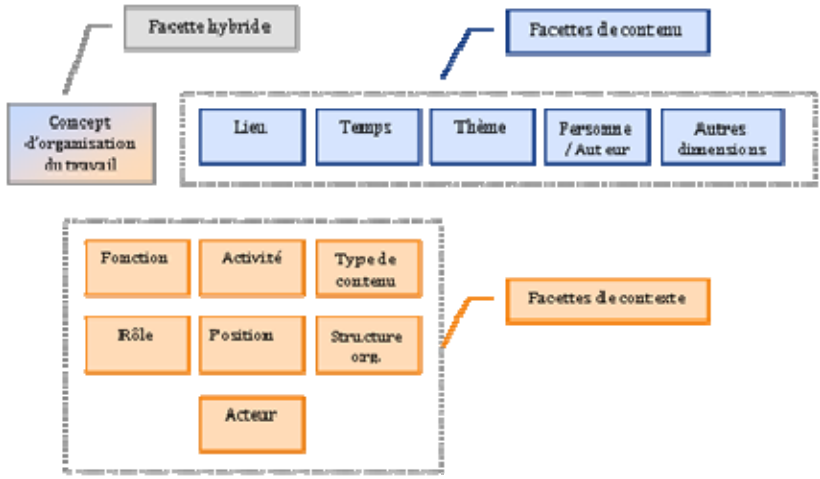


Figure 1 Les facettes du modèle ISIS [7]

Le modèle ISIS repose sur la création de relations sémantiques entre les valeurs de facettes. Ainsi, certains éléments du document peuvent être extraits automatiquement, en fonction du niveau de structuration de celui-ci, générant un certain « affinement » des valeurs proposées pour caractériser les autres facettes, tandis que d'autres sont analysés manuellement. Cette fonctionnalité permet de restreindre le nombre de valeurs présentes sous chaque facette, facilitant le processus de classification et de description du document par l'acteur.

Cette méthode d'indexation s'appuie sur les référentiels des processus métiers. Elle semble particulièrement adaptée à notre cas d'application, dans la mesure où elle se présente comme un compromis entre la rigueur imposée par une organisation du classement et la flexibilité offerte par le fonctionnement des facettes. Une première étape du travail consiste à examiner, à partir d'une étude du terrain visant à spécifier les différentes facettes, comment cette démarche peut être utilisée puis adaptée dans des contextes métiers spécifiques.

4.2 Caractériser l'information à partir des connaissances métiers : les approches par points de vue et l'indexation participative

Aussi s'agit-il de questionner les modalités actuelles de production de l'information et les usages de ces informations dans un contexte opérationnel, caractérisé par la variété des types de documents mobilisés (images, textes, plans, enregistrements audio, vidéos) et la diversité des bases dans lesquelles ils sont stockés, pour contribuer à la définition d'une méthode d'indexation. On envisage ainsi de compléter l'approche de classification par facettes par une approche d'indexation participative s'appuyant davantage sur les usages et la recherche de documents lors de leur utilisation par les métiers.

En suivant les développements du web socio-sémantique, les approches par points de vue (ontologies sémiotiques) et l'indexation participative (folksonomies, pratiques d'annotations libres) permettent, au moyen d'outils et d'interfaces, un « balisage » subjectivé, à vocation non pérenne, de l'information et plus globalement des connaissances.

Les approches par **points de vue** renvoient à une dimension d'analyse de la situation portée par un acteur, le plus souvent un expert reconnu, qui représente une pratique singulière sans que la dimension associée à cette pratique ne soit nécessairement explicitée ni qu'elle ait nécessairement vocation à s'inscrire dans un schéma d'ensemble [9]. Envisagées comme des points de vue systémiques génériques, les **ontologies sémiotiques** correspondent aux situations dans lesquelles l'organisation ou la communauté considérée cherche à définir un ensemble de dimensions explicites et communes sans qu'un consensus puisse être immédiatement établi [9]. Le processus de définition est progressif et il est même possible que l'on ne parvienne jamais à une situation totalement stable ; les ontologies sémiotiques se distinguent en ce sens des ontologies formelles qui proposent une modélisation *a priori* d'un domaine et font consensus.

Issues du croisement de deux phénomènes renvoyant à des techniques de recherche et de partage de documents sur le Web, les **folksonomies** [3] résultent d'un processus d'indexation participative par mots clés choisis librement par les utilisateurs.

Les approches par points de vue et l'indexation participative favorisent ainsi l'émergence de cercles sociaux intermédiaires, car les termes utilisés pour classer et caractériser les contenus documentaires sont destinés à se propager à l'échelle de la communauté de pratiques et des métiers, réunis autour de thématiques précises [4]. Elles ajoutent, de ce fait, de nouvelles fonctionnalités aux outils de classement et de partage des ressources documentaires et impliquent les utilisateurs dans la construction du système de gestion des connaissances. En les invitant à caractériser les documents au moment où ils les élaborent et au moment

où ils les manipulent, elles donnent aux utilisateurs la possibilité d'organiser leurs ressources et de caractériser les informations selon des termes qu'ils auront eux-mêmes choisis.

5 L'introduction des TIC dans les pratiques professionnelles

Ces problématiques de la recherche d'information en situation et de son indexation s'inscrivent dans le contexte plus général de l'introduction des technologies de l'information et de la communication (TIC) dans les pratiques professionnelles.

En effet, l'étude des différents modes de recherche d'information intervient dans un cadre dans lequel se produisent de profonds changements : d'une situation dans laquelle la documentation était matérialisée sur support papier – et son accès partait d'une démarche individuelle, les métiers doivent évoluer vers une situation dans laquelle la documentation prend des formes plus variées et est disponible en abondance sur des supports électroniques accessibles en réseau. L'objectif de l'introduction d'un outil de gestion de l'information, en support à la recherche d'information en situation opérationnelle, est de permettre aux acteurs métiers de se « construire » une représentation de leur environnement tant interne qu'externe, notamment en termes de ressources à disposition.

Les pistes envisagées pour répondre à ces problématiques de recherche d'information en situation mettent ainsi en exergue le passage d'une utilisation assez « linéaire » des outils informatiques (« *je questionne et je dépouille mes résultats* ») à une utilisation plus composite : par de multiples acteurs, jouant des rôles différents, parfois changeant au fil du temps, vis-à-vis des informations, et une utilisation participative desdits outils, s'appuyant sur un travail en réseau. Il convient également de souligner le changement du statut de l'information [8] dans ce nouveau contexte.

De ce fait, l'exploitation des facettes et l'intégration de logiques propres au web socio-sémantique renouvellent les agencements classificatoires dans les systèmes documentaires internes à l'organisation, en permettant l'intervention directe des usagers sur la description des contenus documentaires. Les développements méthodologiques dans le domaine doivent donc être associés à une analyse de l'impact possible des solutions considérées sur les pratiques des métiers concernés et, sans doute, à des démarches d'accompagnement adaptées.

6 Synthèse

L'étude présentée ici vise à évaluer, en situation opérationnelle, deux méthodes d'analyse et de traitement de contenus documentaires auprès d'un échantillon d'acteurs : la classification par facettes et l'indexation participative. L'objectif est d'analyser l'intérêt du déploiement d'un outil de gestion des connaissances en contexte, pour supporter les différentes phases de l'arrêt de tranche et optimiser la durée de la planification et de la conduite de l'ensemble des opérations. La démarche de l'étude consiste en l'observation des métiers, l'analyse de l'activité et l'expérimentation des approches auprès d'un échantillon d'opérationnels terrains, afin de prendre également en compte la dimension subjective de l'acteur en situation.
Soutenue par une expérimentation du modèle auprès d'un certain nombre d'acteurs, cette étude s'intègre dans un projet de recherche, dont l'ambition est de contribuer à la définition d'une méthodologie pour la mise en œuvre (et l'adaptation) d'un outil de gestion de l'information, orienté usages et usagers, permettant de répondre au *besoin informationnel* « en situation », c'est-à-dire prenant en compte les différentes problématiques des métiers mobilisés dans les activités d'exploitation et de maintenance d'une centrale de production d'énergie nucléaire.

7 Références bibliographiques

[1] M. Amar. *Les fondements théoriques de l'indexation : une approche linguistique*. Paris, ADBS Éditions. 2000.

[2] M. Bourrier. *Le nucléaire à l'épreuve de l'organisation*. Paris: PUF. 1999

[3] E. Broudoux. *Folksonomies et indexation. collaborative. Rôle des réseaux sociaux dans la fabrique de l'information*. 2006. Document en ligne sur DocForum: <http://www.docforum.tm.fr/documents/23&24nov06SavResPar06InterBroudouxE.pdf>

[4] O. Ertzscheid et G. Gallezot. Indexation sociale et continents documentaires. *Document numérique et société. Actes de la conférence DocSoc*. Paris, ADBS Éditions, pp. 291-305. 2006.

[5] A. Hatchuel et B. Weil. *L'expert et le système : gestion des savoirs et métamorphose des acteurs dans l'entreprise industrielle*. Paris: Economica. 1992.

[6] M. Hudon et S. Mas. *Analyse des facettes pour la classification des documents institutionnels au gouvernement du Québec.* Rapport

présenté pour le Groupe de travail en classification et indexation. Montréal : École de bibliothéconomie et des sciences de l'information, Université de Montréal. (Collection en ingénierie documentaire; 13). 2001. [consulté le 15.02.2009] <http://www.msg.gouv.qc.ca/fr/publications/enligne/administration/ingenierie/classification_analyse.pdf>

[7] Y. Marleau, S. Mas et M. Zacklad. Exploitation des facettes et des ontologies sémiotiques pour la gestion documentaire. In E. Broudoux et G. Chartron (dir.). *Traitements et pratiques documentaires : vers un changement de paradigme ?*. Paris, ADBS Éditions, p. 91-110. 2008.

[8] Roger T. Pedauque. *Le document : forme, signe, medium, les reformulations du numérique*. STIC-CNRS, Working Paper – 8 juillet 2003. http://archivesic.ccsd.cnrs.fr/sic_00000511.html

[9] M. Zacklad. Introduction aux ontologies sémiotiques dans le Web socio-sémantique, *Actes de la conférence Ingénierie des Connaissances 2005*, Nice. http://archivesic.ccsd.cnrs.fr_00001479.en.html

[10] M. Zacklad, Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information. Montréal, CAIS/ACSI, 2007. http://www.cais-acsi.ca

Une médiathèque virtuelle physique

Pedro ALESSIO, Boris GUILLOT, Alexandre TOPOL

Laboratoire CEDRIC / Conservatoire National des Arts et Métiers

Mots-clés : Bibliothèque numérique, interaction 3D, moteur physique

Keywords: Digital Library, 3D interaction, Physical Engine

Résumé : Un sujet d'étude particulier nécessite, de nos jours, d'accéder à des bases de connaissances diverses (vidéos, textes, objets 3D, sons, images 2D). Ces différentes bases de connaissances ne sont pas liées entre elles et nous devons les explorer séparément. Cela force l'utilisateur à ouvrir autant de fenêtre qu'il n'a de type de bases à interroger. Nous présentons dans cet article une interface en trois dimensions permettant de visualiser et manipuler conjointement ses informations hétérogènes. De plus, pour rendre leur manipulation plus naturelle, nous avons adjoint à notre environnement 3D des propriétés physiques pour le feuilletage des ouvrages et le déchirage des pages.

Abstract : A particular subject of study requires, nowadays, to reach various knowledge bases (videos, texts, 3D objects, sounds, 2D images). These various knowledge bases are independent and one must explore them separately. That forces the user to open as much window as there are bases to query. We present in this article an interface in three dimensions allowing to visualize and handle together this heterogeneous information. Moreover, to make their handling more natural, we associated with our 3D environment some physical properties to make it possible to turn and tear pages.

1 Introduction

Les techniques de visualisation et d'interaction 3D sont de plus en plus utilisées pour l'affichage et la gestion de documents numériques. Cette démocratisation de la 3D est directement liée au succès des jeux vidéos qui pousse le développement des architectures matérielles et logicielles 3D. Les aptitudes acquises au travers des jeux vidéos pour la manipulation des interfaces 3D ainsi que le bombardement de la 3D comme canon esthétique pour le grand public n'ont pas échappé aux concepteur des trois principaux systèmes d'exploitation (Linux, Microsoft Windows et MacOSX d'Apple). Les interactions 3D ont fait naturellement leur apparition dans les interfaces graphiques de ces trois

systèmes. Elles offrent maintenant des outils 3D pour la manipulation des fenêtres, des bureaux virtuels, voire même des documents.
On pourra regretter le côté gadget ou inutile des widgets 3D proposées. En particulier, le basculement des tâches sous Windows n'apporte véritablement rien de nouveau. Cependant, certaines autres interfaces, telles que la time machine ou le mur de signets de Safari sur MacOSX, exploitent véritablement la troisième dimension pour représenter les informations (respectivement, l'utilisation de la profondeur pour représenter une chronologie et la répartition 3D pour favoriser les heuristiques visuelles basées sur la mémoire spatiale). Par ailleurs, d'autres interfaces graphiques telles que Looking Glass de Sun et BumpTop [1] ont été développées pour pousser bien au-delà les expérimentations 3D et nous laissent entrevoir si ce n'est l'avenir de l'interaction homme-machine, tout du moins les atouts de la 3D dans ce domaine.
Une approche semblable mêlant documents 2D et interface 3D pour la lecture détaillée de livres numérisés a été étudiée dans différents travaux de recherche [3, 4, 5]. D'autres domaines, tels que la C.A.O., l'architecture, ou encore la numérisation d'objets pour les musées [2, 7] exploitent des technologies de visualisation 3D. Ces types d'applications rendent particulièrement nécessaire l'utilisation de méta-données textuelles pour enrichir le modèle 3D brut, mais leur présentation et utilisation reste complètement isolées de l'interface de navigation et de l'ensemble de la collection d'objets 3D. Ces informations sémantiques attachées à un objet 3D ne servent en général qu'aux moteurs d'indexation et de recherche.
Notre proposition de médiathèque numérique 3D multimédia se base sur ce constat. Le sujet central de cet article est la prise en compte de ces métadonnées pour enrichir une interface documentaire qui était à l'origine uniquement textuelle. Plus concrètement, nous proposons de lier sémantiquement différents médias dans une seule et unique interface 3D. Cela permet d'étudier de multiples ressources hétérogènes (textes, vidéos, sons, objets 3D) traitant d'un même sujet, sans avoir à effectuer une recherche de multiples fois dans différents moteurs de recherche ou applications. Ainsi on n'a plus à subir la lourdeur induite par le changement de paradigme de représentation graphique et de système recherche des informations et l'on y gagne par la mise en regard immédiate d'informations issues de bases différentes.

2 Le document en contexte

Il semble intéressant d'étudier la coexistence de documents textuels et d'objets 3D au sein d'un environnement de visualisation. Notre champ

d'application potentiel est celui des bibliothèques numériques sur l'histoire de la technologie, pour lesquelles on voudrait pouvoir associer les modèles des machines ou des dispositifs scientifiques et les travaux qui les décrivent. De manière générale, la coexistance de différents médias (textes, images 2D, objets 3D, vidéos, sons) et la possibilité de les consulter dans un même environnement est un usage intéressant pour une médiathèque. Nous nous sommes intéressés en particulier au cas où ce type d'association contextuelle entre différents médias n'est pas conçu a priori, par l'auteur d'un hypermédia, mais a posteriori par un lecteur qui a accès à plusieurs sources d'information. Il est à noter que ce genre d'activité est effectué par un nombre considérable d'utilisateurs du Web lorsqu'ils téléchargent et lisent d'un côté les notices descriptives et qu'ils visualisent de l'autre les représentations 2D ou 3D de marchandises qu'ils désirent acheter.

Figure 1. Exemple pour l'étude d'un polarimètre avec mise en situation réelle (haut gauche), numérique 2D (haut droite) et 3D (bas)

La Fig. 1 met en regard une situation réelle (en haut à gauche) et deux situations vituelles (2D en haut à droite et 3D en bas) pour la même sessions de travail. Il s'agit de l'étude d'un polarimètre et des écrits associés. Pour la situation réelle, les documents et les objets sont organisés sur un bureau en fonction des préférences de l'utilisateur. La

place dédiée à la lecture des documents nécessite une bonne organisation, une ergonomie appropriée voire même une esthétique adaptée à un travail efficace. D'un utilisateur à l'autre, on retrouvera ainsi des organisations spatiales différentes mettant à profit leur propre mémoire spatiale, leur propre habilité à organiser l'information et ceci sur la surface limitée du bureau. La contrepartie numérique repose en général sur des techniques classiques basées sur la métaphore WIMP (*Windows Icons Menus Pointer*) qui permet d'organiser les informations dans de multiples fenêtres recouvrantes que l'on peut déplacer et redimensionner à souhait. Cependant, la taille de ces bureaux virtuels est très insuffisante et ne permet pas d'atteindre un niveau de productivité comparable avec celui des bureaux traditionnels [8]. Les fenêtres recouvrantes qui habillent les documents mènent inexorablement à une superposition de documents qui est clairement opposée à une bonne organisation.
Nous avons décrit dans [5] comment des livres numériques peuvent être affichés avec des mouvements contraints sur un sol virtuel qui facilite leur manipulation. Concernant les objets 3D, la possibilité de les manipuler librement est la première fonctionnalité importante car ils doivent, comme tout autre document, être organisés spatialement. La seconde fonctionnalité est de permettre leur étude sous tous les angles tout comme de vrais objets peuvent être examinés en bougeant la main qui les porte. La contrainte de gravité qui s'applique à notre environnement est une obstruction à une telle métaphore d'examen des objets. Pour orienter et positionner un objet de la manière souhaitée il faudrait en effet prendre en compte pour chaque mouvement sa géométrie qui rentre en collision avec le sol à cause de la gravité. La solution retenue est de représenter un objet en lévitation au dessus du sol à l'intérieur d'une sphère englobante semi transparente qui elle seule subit la gravité. Cette sphère est posée sur un cylindre qui permet de positionner l'objet 3D. La sphère permet quant à elle de fournir une information de volume occupé par l'objet et y sont attachées les opérations d'orientation de l'objet. De cette manière, on peut laisser un objet au repos, dans la position et l'orientation qui conviennent à son étude.

3 La composante physique

Nos précédents travaux sur la manipulation 3D de documents ne mettent en œuvre que des techniques classiques. La manière d'interagir sur les objets y est basée sur les événements souris et sur leur répercussion visuelle à l'écran en terme de commandes de rendu 3D. De la même manière, les animations se font en calculant, par interpolation linéaire, une série de valeurs dans le temps. L'effet d'animation est alors obtenu

en attribuant ces valeurs à des transformations géométriques. D'un point de vue strictement esthétique, ces effets sont relativement pauvres par rapport à ce que l'on a l'habitude de voir dans les derniers jeux vidéos. Or, l'œil d'un utilisateur sur les interfaces 3D s'est affûté avec le temps et est devenu très critique à tel point que la prouesse technique est un argument de vente. Sans parler de réalisme qui n'est pas l'effet recherché dans notre interface, bien au contraire puisque l'on essaye de s'abstenir des contraintes réelles souvent anti-productives dans un univers virtuel, l'accueil et donc l'adhésion à une interface 3D par un large public passe par la prise en compte de certains critères technico-esthétiques. En s'inspirant du domaine vidéo ludique, une solution consiste à modeler physiquement notre environnement et de confier la gestion des animations à un moteur physique. Une deuxième partie de cet article portera donc plus précisément sur la partie technique d'un tel environnement physique et de sa valeur ajoutée par rapport aux méthodes classiques.

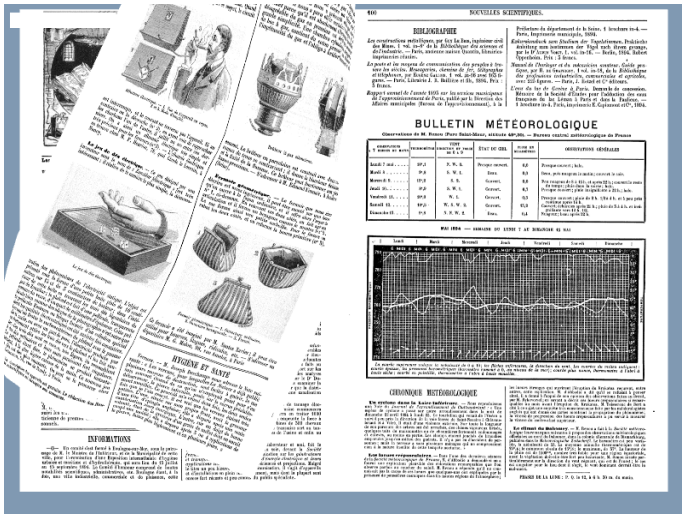


Figure 2. L'atelier physique

Pour étudier les apports d'un moteur physique dans notre contexte de consultation de documents numériques, nous avons développé deux applications. La première très dépouillée, composée d'un unique livre de 4 pages (fig. 2), nous permet de régler les paramètres du moteur physique pour qu'il réponde à nos besoins. Elle nous permet également de tester certaines idées de fonctionnalités et d'interactions. Parce que sa vocation est uniquement de proposer un cadre pour le prototypage des interactions physiques, nous appellerons cette application l'atelier physique. La deuxième application (fig. 3), exploitable pour la lecture multi documents, utilise comme socle de base notre version précédente de

l'atelier de lecture dans lequel les animations sont gérées par des transformations géométriques. Nous y avons donc remplacé ces dernières par certaines caractéristiques physiques validées techniquement dans l'atelier physique.

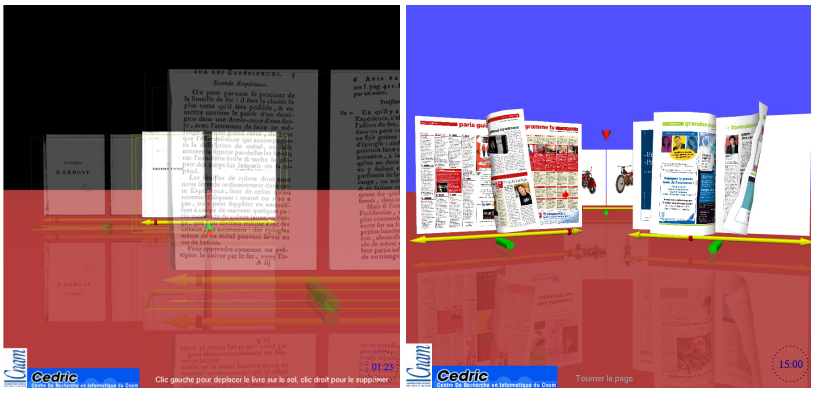


Figure 3. L'interface physique

En théorie, tout peut être géré par le moteur physique. Des problèmes apparaissent cependant avec une solution entièrement physique : le manque de contrôle fin dû à l'inertie des objets et les tremblements perceptibles lorsque le simulateur physique est en recherche de stabilité, sans parler des cas rares mais possibles de situations aberrantes comme l'interpénétration d'objets. Nous expliquons donc dans notre article les compromis proposés entre une gestion toute physique et une gestion toute géométrique.

3.1 Le trépied physique

Le principe de base de l'application de lecture reste le même que précédemment : l'organisation spatiale afin de permettre la lecture de multiples documents ouverts simultanément. L'utilisateur a donc la possibilité de gérer lui-même son espace de lecture en ouvrant plusieurs livres au format PDF et en les répartissant spatialement selon son bon vouloir. Pour ce faire, il translate et oriente indirectement les livres 3D par l'intermédiaire de la métaphore du trépied qui fait office de widget de manipulation. Ces opérations se font facilement à l'aide d'une souris standard à deux degrés de liberté car les trépieds sont posés sur un plan 2D représentant le sol. Dans la version préliminaire du poste de lecture, un trépied que l'utilisateur est en train de manipuler peut interpénétrer les autres éléments de la scène 3D afin de le positionner dernière. Cette fonctionnalité est rendue possible par une gestion de la transparence des trépieds inactifs.

Le premier apport évident de l'ajout de la physique est la prise en compte de la gravité pour gérer la contrainte du sol et des collisions entre objets. Plutôt que de calculer par des techniques de *picking* la position contrainte des trépieds sur le sol en fonction de la position de la souris, ce peut être réalisé par le moteur physique. Le déplacement proprement dit des trépieds se fait toujours en prenant les informations de position de la souris et non d'accélération. Ceci afin que l'utilisateur garde un contrôle optimal qui n'est pas aisé lorsque le moteur physique affecte une certaine inertie aux objets poussés.
Un deuxième aspect que l'on peut implémenter facilement avec un moteur physique est la gestion des collisions. Notre choix préliminaire de ne pas les gérer était délibéré et non induite par les algorithmes à mettre en place qui somme toute sont classiques. Nous pensons en effet que la manipulation des trépieds serait malaisée s'ils collisionnent les uns avec les autres. En particulier, lorsque l'atelier de lecture est fortement encombré, pour positionner un trépied derrière tous les autres, il devra se frayer un passage. Or, cela modifierait très certainement le positionnement ou l'orientation d'autres trépieds que l'utilisateur avait placé minutieusement pour travailler confortablement. Les tests facilités par le moteur physique nous ont permis de valider ce choix initial et de l'améliorer.
Trois options s'offrent à nous pour les collisions : ne pas les prendre en compte, les gérer finement (par l'algorithme CCD – *Continous Collision Detection*) ou les gérer grossièrement. Pour les raisons évoquées précédemment, qui se sont révélées exactes, la gestion exacte des collisions est écartée. La gestion grossière des collisions, en ne considérant que les contacts à un instant particulier et non les vecteurs de déplacement, permet un mode hybride intéressant. Lors de mouvement rapide, les trépieds s'interpénètrent car l'instant précis de la collision n'intervient pas au moment où on dessine la scène. De ce fait, ils agissent comme si les collisions n'étaient pas gérées. Par contre, lorsque le déplacement est lent, la collision est captée et les trépieds la subissent en conséquence. Ainsi, l'utilisateur pourra en connaissance de cause, pousser plusieurs trépieds en même temps.

3.2 La page physique

Le deuxième apport de la physique concerne plus particulièrement les pages d'un livre. Lors d'un précédent travail portant sur la numérisation 3D de livres par la technique de photogrammétrie, nous nous sommes intéressés à la problématique de l'aplatissement de la page acquise en utilisant un système masse-ressort. Il en est ressorti que ce genre de système permet de modéliser de manière très réaliste une page déformable. Certains moteurs physiques utilisent ces systèmes masse-ressort pour simuler des textiles. Nous avons donc tenté de détourner

cette gestion de vêtements pour que cela convienne pour nos pages de livres. Le but avoué était, dans un premier temps, de reproduire un feuilletage réaliste et, dans un second temps, de permettre des interactions riches sur les pages.
Après réduction du nombre de sommets et réglages des contraintes des ressorts, nous sommes arrivés à une page « textile » tout à fait exploitable. Les calculs nécessaires pour la déformation étant peu coûteux, nous avons pu introduire la notion de feuilletage multi-pages (fig. 3). Par ailleurs, le moteur physique utilisé permet de gérer le déchirage des textiles en spécifiant quels sommets du système masse-ressort peuvent être désolidarisés des autres. Ceci nous a permis de mettre en place un système de déchirage des pages. La possibilité de casser la structure préétablie d'un livre et d'en recréer d'autres à partir de pages déchirées est désormais devenue envisageable, et ceci de manière entièrement physique. Dans un deuxième temps, nous avons étudié une autre possibilité des systèmes masse-ressort : la rupture des ressorts liant les masses. Cela permet d'intégrer la fonctionnalité de déchirage réaliste de pages ou de parties de pages ; ceci afin de permettre à l'utilisateur de supprimer des éléments d'un document ou de constituer lui-même un livre composé de différentes sources d'information. Pour cela, nous avons adjoint aux images des livres manipulés une description XML des blocs composant chaque page.
Pour pousser plus en avant cette expérimentation physique, il nous a paru intéressant d'étudier également la physique du geste. Pour cela, nous avons intégré la prise en charge des interaction avec une wiimote. De simples gestes permettent de feuilleter les ouvrages, de les maximiser ou de les minimiser. Le point sensible concernait la sélection de l'ouvrage sur lequel ces opérations doivent être réalisées. Nous avons pour cela développé une technique baptisée sotalotre dont le but est de calculer, à partir d'un micro-mouvement (déterminé à l'aide d'un seuil temporel et spatial), le prochain trépied à sélectionner. En pratique, l'utilisateur n'a qu'à initier un mouvement à partir de la sélection courante et en direction du trépied suivant qu'il veut sélectionner.

4 Au delà du simple réalisme

Ces premiers travaux paraissent tout à fait prometteur et nous permettent d'envisager plusieurs autres applications. En particulier, nous pensons à intégrer dans notre représentation XML du contenu des pages, les différents systèmes de type popup décrits dans nos travaux précédents [6] pour qu'ils soient gérés de manière physique. De plus, nous voulons nous pencher sur un système d'annotation du type post-it. Chaque post-it se comportera comme une partie déchirable que l'utilisateur pourra

déplacer, enlever et recoller à souhait. Dans le même ordre d'idées, nous souhaitons fournir une description XML des interactions possibles sur les maillages 3D des objets afin que leur manipulation de soit pas nécessairement monolithique comme cela est fait dans [7].

5 Références bibliographiques

[1] Agarawala A., Balakrishnan R. Keepin' it Real: Pushing the Desktop Metaphor with Physics, Piles and the Pen. *Proceedings of CHI 2006 - the ACM Conference on Human Factors in Computing Systems*. p. 1283-1292. 2006.

[2] Alisi, T.B., Del Bimbo, A., Valli, A., Natural Interfaces to Enhance Visitors' Experiences, *IEEE MultiMedia*, vol. 12, no. 3, pp. 80-85, Jul-Sept, 2005

[3] Card, S. K., Hong, L., Mackinlay, J. D., and Chi, E. H. 3Book: a scalable 3D virtual book. *Proc. of ACM CHI '04.* ACM Press, New York, NY, 1095-1098.

[4] Chu, Y., Bainbridge, D., Jones, M., and Witten, I. H. Realistic books: a bizarre homage to an obsolete medium?. *Proc.of ACM/IEEE-CS JCDL '04.* ACM Press, New York, NY, 78-86.

[5] Cubaud, P., Stokowski, P., and Topol, A. Binding browsing and reading activities in a 3D digital library. *Proc. of ACM/IEEE-CS JCDL '02*. ACM Press, New York, NY, 281-282.

[6] Cubaud, P., Dupire, J., and Topol, A. Digitization and 3D modeling of movable books. *Proc. of ACM/IEEE-CS JCDL '05*. ACM Press, New York, NY, 244-245.

[7] Hemminger, B., Bolas, G., Schiff, D. Capturing content for virtual museums : from pieces to exhibits. *J. of Digital Information*, vol. 6(1), march 2005.

[8] Mackinlay J. D., Heer J., Royer C.: Wideband Visual Interfaces: Sensemaking on Multiple Monitors. *PARC Technical Report*, 2003.

Documents et Applications : CMS nouvelle génération

Jean-Marc LECARPENTIER, Hervé Le Crosnier, Jacques MADELAINE

GREYC – CNRS UMR 6072 – Université de Caen Basse-Normandie

Mots-clés : création document, document multimédia, document multilingue, FRBR - Functional Requirements for Bibliographic Records, CMS - Content Management System

Keywords: document creation, multimedia, multilingual documents, FRBR - Functional Requirements for Bibliographic Records, CMS - Content Management System

Résumé : L'internet est devenu la plate-forme de prédilection pour la création de documents via l'utilisation des CMS (Content Management System ou Système de Gestion de Contenu). Or, trop souvent les CMS sont conçus comme des outils de production de sites web. Nous imaginons la réalisation de nouvelles plate-formes de création et de gestion de documents qui épousent le web d'aujourd'hui : un web multilingue, multimédia, sémantique et social. L'objectif de cet article et de proposer une architecture de production et gestion des documents numériques basée sur l'expérience des bibliothèques avec le logiciel Sydonie (SYstème de gestion de DOcuments Numériques pour l'Internet et l'Édition).

Abstract : Most Web Content is nowadays published with Content Management Systems (CMS). However, most CMS consider each document as an independent entity which matches one web page. New systems need to consider today's web : multilingual, multimedia, semantic and social. Based on the experience of libraries, this article aims to propose an architecture to author and manage digital documents, with Sydonie (SYstème de gestion de DOcuments Numériques pour l'Internet et l'Édition).

1 Introduction

La notion de « page web » n'est plus synonyme de « document web » comme au début de l'Internet. Une page web est désormais un ensemble composite de documents, de données et d'applications. Les pages présentées sont souvent composées d'un ou plusieurs documents, qui sont

détenus en propre par le serveur, ou obtenus à partir de services distants. Nous avons décrit cette évolution dans un article antérieur [1]. Il s'agit maintenant d'en tirer des conclusions opérationnelles pour la gestion des documents numériques au travers du web.
L'internet est devenu la plate-forme de prédilection pour la création de documents, via l'utilisation des CMS (*Content Management System* ou Système de Gestion de Contenu). Or trop souvent les CMS sont conçus comme des outils de production de sites web, susceptibles d'intégrer des « documents » à l'intérieur des « pages » pré-organisées et conçues par un graphiste. Nous imaginons la réalisation de nouvelles plate-formes de création et de gestion de documents qui puissent épouser toute la complexité d'un réseau. En particulier, nous avons en ligne de mire trois ingrédients du web :

- un web multilingue : il s'agit de penser le document comme l'ensemble de ses expressions linguistiques disponibles, sur le serveur local (problématique de CMS traditionnel) comme entre serveurs distants (nécessité d'un schéma de numérotation global) ;
- un web multimédia, dans lequel les objets numériques ne sont pas de simples « fichiers numériques » destinés à un *player*, mais bien des documents à part entière, associant les métadonnées et annotations autour du fichier binaire proprement dit. Cette conception impose de voir tout document comme « composite », et de concevoir les objets numériques (images, vidéo, sons, animations,...) comme étant eux-mêmes des documents ;
- un web sémantique et social, dans lequel des données et informations sémantiques rendues disponibles en divers points de la toile peuvent être utilisées pour annoter les documents, et en sens inverse dans lesquels des documents, données et informations organisées peuvent être proposées aux autres acteurs du web sémantique, suivant la logique des « *Linked Data »* [2]

L'objectif de cet article est de proposer une architecture de production et de gestion des documents numériques répondant à ces contraintes.

2 CMS : création de documents et applications web

L'utilisation d'un CMS pour la création de documents est aujourd'hui fortement liée à la présentation du document créé *via* ce même CMS. Autrement dit, la création et la présentation sont intégrées au sein d'un même processus, visant à créer un document *pour* un logiciel et un site particuliers. Bien entendu des méthodes d'exportation des données du CMS existent en général, mais l'export (le plus souvent au format XML)

ne permet pas une exploitation immédiate du document, surtout s'il est composite.
Notre objectif est de séparer la création du document de sa publication (web ou autre). Dans cette optique, il nous semble nécessaire de penser un système composé de deux outils logiciels, actuellement en cours de développement :

- un outil de création de documents : **Sydonie** (SYstème de gestion de DOcuments Numériques pour l'Internet et l'Edition)
- un outil de création d'applications web (**Aglae**) permettant d'intégrer les documents créés, des applications externes (web services), voire d'autres documents respectant les protocoles d'échange XML de divers professions, tels PRISM (*Publishing Requirements for Industry Standard Metadata*[1]) ou newsML[2]

L'objectif de Sydonie est de permettre le stockage des documents produits avec des outils divers (en base de données, dans des fichiers XML, avec des fiches de métadonnées en RDF,...). Sydonie produit ensuite des formats physiques qui correspondent aux différentes méthodes de visualisation (ou audition) par les lecteurs humains (HTML, pdf, impression, synthèse vocale…) ou non-humains (représentations XML ou RDF selon les besoins, intégration des données dans les documents dans la logique de GRDDL).

3 Les frontières du document

La double contrainte de séparation entre le document et la page web d'une part et de conception d'un document composite multilingue d'autre part nous impose de préciser les « frontières d'un document ». Quelle relation particulière entretient une traduction avec l'original ? Le texte encodé en HTML et la version mise en page pour l'impression en pdf sont-ils un même ou bien deux documents différents ?

3.1 « Functional Requirements for Bibliographic Records"

Pour éclairer toutes ces questions, nous avons étudié la façon dont les bibliothèques approchent dorénavant la notion de « Document ». Alors que le catalogue traditionnel des bibliothèques ne connaissait que « l'exemplaire dans les mains du bibliothécaire », tendant à multiplier les fiches pour diverses éditions ou traductions d'une même œuvre, la tendance actuelle est l'équivalent d'une révolution copernicienne : au

1 PRISM, http://www.prismstandard.org/about/

2 NewsML, http://newsml.org

cœur de la nouvelle description documentaire est « L'Œuvre » (*Work*) dont les descriptions des ouvrages présents dans la bibliothèque vont dépendre. La logique catalographique reprend alors les pratiques des lecteurs : le moment de recherche globale d'une œuvre précède le choix d'une édition puis d'un exemplaire. Une place est accordée au travail intellectuel préalable à la réalisation d'une édition (une « manifestation » en FRBR-lingo), à la fois dans la définition du « travail » de création de l'original, mais aussi dans la conception de traductions ou les corrections liées aux diverses éditions d'une œuvre.
Cette transformation a été menée par une réflexion de l'IFLA[3] sous l'intitulé « FRBR – *Functional Requirements for Bibliographic Records* ». Le rapport final du groupe de travail des années 90 [3][4] définit un modèle pour les notices bibliographiques basé sur des relations entre entités. Trois groupes d'entités sont définis :

- le groupe 1 définit les entités *Work*, *Expression*, *Manifestation* et *Item*, qui représentent les différents aspects de ce qu'un utilisateur peut trouver dans les produits d'une activité intellectuelle ou artistique ;
- le groupe 2 définit les entités *Person* et *Corporate Body* qui représentent les personnes physiques ou morales qui ont la responsabilité du contenu intellectuel ou artistique, de la production matérielle et de la distribution, ou de la gestion juridique des entités du premier groupe ;
- le groupe 3 définit les entités *Concept*, *Object*, *Event* et *Place* qui représentent les langages documentaires et les contenus d'indexation (relation « a pour sujet »).

Nous nous sommes pour l'instant intéressés de plus près au groupe 1 (*Work*, *Expression*, *Manifestation* et *Item*). Notons dès à présent que nous avons volontairement choisi les dénominations anglaises de ces entités. En effet la traduction française produite par la BnF utilise le terme « Document » pour désigner l'entité *Item*, ce qui nous semble en contradiction avec la définition élargie de Roger T. Pédauque [5] de ce qu'est un document numérique. Pour notre part, nous désignons « Document » l'ensemble de l'arbre composite. La figure 1 (extraite du rapport FRBR) explicite les relations entre ces quatre entités.
Les entités *Work* et *Expression* expriment le contenu intellectuel de l'œuvre. Ils sont abstraits, et ne comportent donc que des métadonnées documentaires :

- L'entité *Work* (Œuvre en français) représente une création intellectuelle ou artistique déterminée ;

3 International Federation of Library Associations and Institutions, http://www.ifla.org

- Une *Expression* est la réalisation d'une œuvre. Par exemple les versions linguistiques, ou les divers enregistrements d'une même œuvre sonore, ou diverses éditions revues et corrigées.
- Les entités *Manifestation* et *Item* représentent la forme matérielle de l'œuvre :
- L'entité *Manifestation* exprime la représentation matérielle d'une *Expression* ;
- L'entité *Item* représente un exemplaire « physique » d'une *Manifestation.*

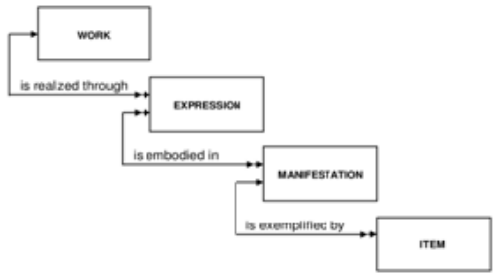


Figure 1 : Entités du groupe 1 et leurs relations

Par exemple le travail intellectuel *Madame Bovary* est représenté par une entité *Work*. L'*Expression* « originale » est la version française du roman. Une *Manifestation* pourrait être l'édition publiée par Flammarion, une autre *Manifestation* serait une édition de Poche, ou une édition électronique en PDF et une autre en format ePUB par exemple. Chaque *Manifestation* possède un ou plusieurs *Item* qui sont les exemplaires « physiques » présents sur les rayonnages, ou les sites de référence distribuant les versions électroniques. Une autre *Expression* serait par exemple une traduction en anglais, avec les métadonnées concernant la traduction (traducteur, année, etc.), et qui aurait elle-même plusieurs *Manifestations*...
FRBR permet donc de représenter des documents ayant différentes versions linguistiques, différents formats, mais plus intéressant encore, FRBR offre un point d'accès unique à une œuvre via l'entité *Work.*

3.2 FRBR et documents numériques

FRBR a été imaginé pour gérer des documents « physiques » présents sur des étagères de bibliothèque. Dans le cadre d'un Système de Gestion de Contenu ou d'une bibliothèque numérique, ce n'est manifestement pas le cas. De plus les catalogues de bibliothèques ne stockent que des métadonnées alors qu'un CMS doit stocker à la fois les métadonnées et le

contenu d'un document. Comment alors adapter ce modèle pour gérer des documents numériques présents sur le Web ?
Les entités *Work* et *Manifestation* représentant le travail intellectuel, ces deux entités peuvent être utilisées sans problème dans le cadre de documents numériques, les informations stockées étant des métadonnées. L'entité *Manifestation* exprime la représentation matérielle d'une *Expression.* Nous pouvons alors considérer que le représentation au format HTML (par exemple) est une *Manifestation* et que la représentation au format PDF en est une autre. Pour l'entité *Item*, qui représente un exemplaire avec un emplacement précis sur les étagères permettant de le trouver, nous utiliserons par analogie un URI qui permet de spécifier l'emplacement d'un document sur le web. Cette analogie permet alors à une *Manifestation* d'avoir plusieurs URIs associés, par exemple dans les cas suivants :

- document présent sur un serveur miroir (servant exactement les mêmes pages) ;
- document présenté avec un « environnement » différent, c'est-à-dire avec des informations péri-phériques dépendant du contexte d'affichage ;
- document publié par des sites tiers.

3.3 Gestion de documents composites et Sydonie

FRBR permet donc de modéliser des documents numériques en se concentrant sur l'aspect intellectuel du document via l'entité *Work.* Au travers des entités définies par FRBR, un document peut être représenté sous forme arborescente, comme illustré sur la Figure 2. Le modèle que nous choisirons est alors basé sur le même principe : nous considérons qu'un document est l'arbre complet tel que nous l'avons construit par analogie avec FRBR.

Figure 2 : Structure arborescente du Document

Ce choix permet à un document de connaître toutes ses versions linguistiques et formats de fichiers. Parmi les *Expressions* présentes, l'une d'entre elles a un rôle particulier puisqu'il s'agit de l'*Expression originale*,

à partir de laquelle les autres *Expressions* ont été créées. De même pour chaque *Expression*, l'une des *Manifestations* a la qualité de *Manifestation référence* à partir de laquelle les autres *Manifestations* seront créées (que ce soit automatiquement ou non).
Pour la mise en œuvre informatique, un document est une entité abstraite composée de plusieurs objets suivant le schéma présenté Figure 3.

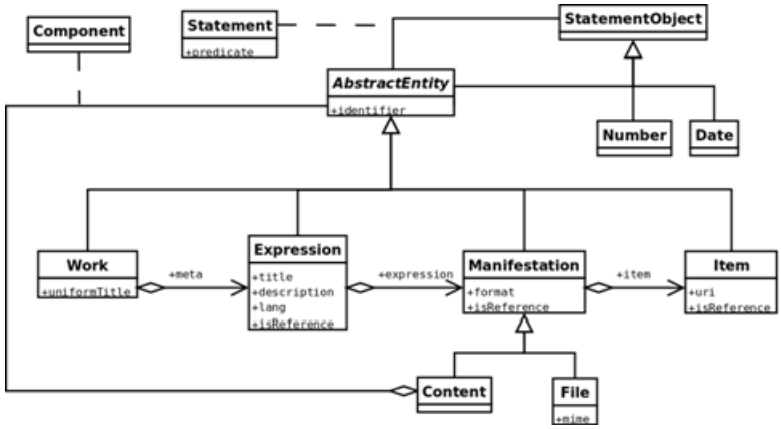


Figure 3 : Objets composant un arbre Document

Le logiciel de création de documents Sydonie en cours de développement utilise ce modèle pour stocker et gérer les documents créés. Nous devons y représenter les documents de façon générique (texte, image, vidéo, actualité, article, anthologie, etc.) grâce à un arbre en dérivant la classe abstraite *DocumentType* qui contient les propriétés et méthodes de base à tout type de document. Les spécificités d'un type de document sont gérées par un processus d'assertions indépendantes, ce qui favorise la modélisation en RDF des documents. Une telle conception du document permet d'approcher la notion de document composite. Par exemple, une image intégrée dans une page n'est plus considérée comme un simple lien vers un fichier numérique (modèle HTML) mais bien comme un document intégré, disposant de ses propres métadonnées, et pouvant éventuellement lui-même avoir plusieurs *Expressions*, cas d'un graphique avec du texte intégré, par exemple.

3.4 Publication Web

La séparation entre la création de documents via le web et la publication de documents impose l'utilisation de normes ouvertes afin de garantir l'interopérabilité entre les divers systèmes de publication. La norme PRISM [6] (Publishing Requirements for Industry Standard Metadata)

définit un vocabulaire XML permettant la gestion et le stockage de documents composites pour la publication de magazines et de journaux. Son utilisation permet de gérer les documents composites au sens de Sydonie, tout en garantissant la possibilité d'importation par d'autres systèmes. Une *Manifestation* (c'est-à-dire le contenu d'un document) peut donc être stockée sous ce format, permettant ainsi l'échange de données. Diverses transformations peuvent ainsi être appliquées pour construire des présentations dans les divers formats professionnels tout en conservant le même lot de métadonnées prduites par Sydonie.
L'accès aux documents peut se faire à tout niveau dans l'arbre représentant le document mais la vue qui sera servie à un agent utilisateur (ou *User-Agent*[4]) est toujours la vue d'une *Manifestation*. Un utilisateur final accède donc à une *Manifestation* via un URI modélisé par un *Item*. Un système de gestion et de déréférencement des usages portant sur ces divers URI devra être mis en place, autour d'une logique de *Handler,* à l'image de ARK[5] ou du DOI[6], avec une numérotation unique. Nous envisageons la mise en place d'un modèle et d'un système associé à Sydonie (GONG : Gestion des Objets Numérique Généralisée).
De plus, une stratégie doit être mise en œuvre pour définir la *Manifestation* qui sera présentée quand une requête porte sur l'accès à un document au niveau de l'entité *Work* ou *Expression*. Le rapport *Cool URIs for the semantic web* [7], publiée par le W3C en décembre 2008, donne des indications sur ce type de stratégie. Le protocole HTTP permet à deux machines d'entrer en négociation pour déterminer le contenu adapté à une requête. Le serveur peut donc utiliser les *Language-Negotiation* et *Content-Negotiation* pour adapter la langue et le format de document à fournir au client :

- si un nœud de type *Work* est demandé, alors le processus *Language-Negotiation* permet au serveur de déterminer quelle *Expression* utiliser (en fonction des préférences de langues du client et des langues disponibles pour le document). Puis un *Content-Negotiation* permet ensuite de déterminer quelle manifestation servir au client (cf. ci après) ;

- si un nœud de type *Expression* est demandé, alors *Language-Negotiation* n'est pas nécessaire et seul le *Content-Negotiation* aura lieu. Un exemple type sera le cas d'un navigateur qui recevra du XHTML (c'est le type de contenu que ce client déclare préférer) alors qu'un robot recevra du XML ou du RDF ;

4 Définition de User-Agent sur Wikipedia : http://fr.wikipedia.org/wiki/User-Agent

5 Archival Resource Key, http://www.cdlib.org/inside/diglib/ark/

6 Digital Object Identifier, http://www.doi.org/

- si un nœud de type *Manifestation* est demandé, alors le serveur peut renvoyer directement le contenu sans avoir à effectuer de négociation.

Le système devra toujours préciser si la *Manifestation* servie dérive de l'*Expression* originale et si elle est une référence (telle que définie plus haut), et si ce n'est pas le cas indiquer un URI de la *Manifestation* référence pour l'*Expression* originale. Ce processus de négociation est réalisé en chaîne pour tous les composants d'un document, ce qui permet par exemple de servir les versions linguistiques adaptés pour les images ou vidéos.

4 Perspectives et travail futur

La conception d'un document comme un objet représentant un arbre complet d'entités FRBR correspondant à un travail intellectuel est très attirante. Un tel modèle peut être implémenté dans un Système de Gestion de Contenus pour répondre aux besoins de rédacteurs de *Manifestations* et être modélisé avec XML pour la publication et l'échange de documents.
Le modèle proposé considère une *Manifestation* d'un document comme un composite avec des parties ou informations annexes qui sont elles-mêmes des documents basés sur le même modèle (c'est-à-dire eux-aussi représentés sous la forme d'un arbre d'entités FRBR). La mise en œuvre d'un tel modèle n'est pas simple, en particulier en termes d'ergonomie de l'interface de saisie. De notre point de vue, le modèle de saisie actuel de la plupart des CMS n'est pas adapté à cette conception. De nouvelles formes de saisies doivent être imaginées, combinant les possibilités des formulaires actuels, des éditeurs WYSIWYG en ligne (de type tinyMCE par exemple), des boîtes modales et des échanges Ajax avec le serveur.
Un mécanisme de saisie utilisant au mieux toutes ces possibilités devrait permettre de collecter des informations précises qui pourront être reprises dans des modèles de métadonnées comme RDF-a ou les microformats. La structure arborescente du modèle de document permet aussi d'éviter la duplication d'informations, lors de la saisie de nouvelles versions linguistiques par exemple, puisque les métadonnées au niveau *Work* sont déjà dans le système.
L'expérience des bibliothèques ouvrant de nouvelles perspectives dans la conception de systèmes de gestion de documents sur le web, le modèle présenté devra aussi intégrer les entités FRBR des groupes 2 (personnes) et 3 (sujets du document). Un groupe de travail sur l'harmonisation entre la classification FRBR des bibliothèques et la classification des objets de musée a publié un brouillon de $FRBR_{OO}$, une reformulation orientée objet du modèle entité-relations de FRBR sous la forme d'une ontologie [8]. Un

tel accord montre le caractère fécond de l'approche FRBR qui peut s'étendre au-delà des catalogues de bibliothèque, ce qui nous semble justifier notre approche par analogie pour le document numérique.
L'expérience des médias imprimés, synthétisée dans la spécification de PRISM, permet d'établir un format pivot entre l'édition et la publication. Le modèle proposé doit pouvoir être étendu à d'autres formats professionnels, comme NewsML par exemple.

5 Conclusion

L'avenir de la publication sur Internet réside principalement dans la « re-publication » de tout ou partie de travaux intellectuels dans divers médias, sites, blogs, etc. Une architecture globale qui collecte les divers URIs présentant un document ouvre la possibilité de statistiques sur les usages du document. Couplé à l'utilisation de licences adéquates, cela permet d'envisager des négociations et transactions pour la publication des travaux d'auteurs.
Ni *framework* généraliste, ni CMS fermé, le projet Sydonie tente de fournir des outils adaptés à l'édition et la publication de documents numériques composites dans un environnement ouvert.

6 Références bibliographiques

Jean-Marc Lecarpentier, H. Le Crosnier et J. Madelaine, Évolutions de l'architecture du web et des documents numériques, Traitements et Pratiques Documentaires. Vers un changement de paradigme? Actes de la deuxième conférence Document Numérique et Société, pages 13-30, ADBS éditions, Paris 2008

Chris Bizer, Tom Heath, Kingsley Idehen, Tim Berners-Lee, Linked Data on the Web (LDOW2008), Proceedings WWW2008, Beijing, China, 2008. http://www2008.org/papers/pdf/p1265-bizer.pdf

IFLA Study Group, Functional Requirements for Bibliographic Records, K . G. Saur, München, 1998

Groupe de travail IFLA, Spécifications fonctionnelles des notices bibliographique, BNF, Paris, 2001

Roger T. Pédauque, Le document à la lumière du numérique, C&F éditions, 2006

DEAlliance, PRISM : Publishing Requirements for Industry Standard Metadata, 2009, http://www.prismstandard.org/specifications/2.1/

W3C Interest Group, Cool URIs for the semantic web, W3C, 2008 http://www.w3.org/TR/cooluris

WG on FRBR and CIDOC CRM hamonization, FRBR object-oriented definition and mapping to FRBRER, http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/FRBR_oo_V0.9.pdf

Outil de butinage du contenu des documents de collections numériques

Lyne DA SYLVA

École de bibliothéconomie et des sciences de l'information, Université de Montréal

Mots-clés : indexation, collections numériques, index de livre, indexation automatique, accès à l'information, accès au contenu, aide à la lecture

Keywords: indexing, digital collections, back-of-the-book index, automatic indexing, access to information, access to contents, reading aid

Résumé : Cette recherche se veut une contribution à la recherche d'information dans les documents numériques, non pas pour le repérage de documents mais pour l'aide à la lecture et donc l'évaluation de la pertinence de documents repérés. L'introduction d'un outil de butinage est proposée pour accéder au contenu de documents des bibliothèques numériques, soit l'index de livre traditionnel. Celui-ci présente plusieurs avantages en tant qu'outil de navigation, bien que sa création automatique pose quelques difficultés. L'implémentation d'un outil de ce type est esquissée dans ses grandes lignes.

Abstract : Our research is a contribution to information search within digital documents, after the initial steps of document retrieval from a given digital library. We suggest introducing a type of browsing tool to aid in document perusal and thus to help in evaluating its relevance for the user's information need. The tool in question is the traditional back-of-the-book style index. We present its advantages as a browsing tool, some challenges posed by the automatic creation of this type of tool, and a sketch of our current implementation.

1 Introduction

Cette étude porte sur les collections numériques (ou bibliothèques numériques) de textes non structurés et sur les outils pour accéder à leur contenu. Nous proposons l'adjonction d'un certain outil de navigation dans les documents. Des outils puissants de description sont nécessaires pour permettre aux utilisateurs de repérer les documents pertinents à leurs besoins. Les outils de ce type s'arrêtent cependant à la tâche d'extraire un certain nombre de documents de la collection, n'aidant pas ou peu

l'utilisateur à prendre connaissance du contenu de ceux-ci afin d'évaluer leur pertinence réelle. Nous proposons d'ajouter un outil d'aide à la lecture du document afin de faciliter cette tâche de prise de connaissance du contenu.
La section 1 identifie les outils de recherche actuels et leurs lacunes pour accéder au contenu des documents. À la section 2, nous présentons un nouveau type d'outil, l'index de fin de livre, qui est bien connu pour faire des recherches dans des documents imprimés, mais pratiquement inutilisé pour les documents numériques. La section 3 esquisse une implémentation d'un tel outil et la section 4 discute des difficultés rencontrées lors de l'implémentation de cette approche, alors que la conclusion aborde des pistes de recherche futures.

2 Travaux précédents

Les outils de recherche, qui permettent de repérer des documents dans une collection, se déclinent en plusieurs variétés : moteurs de recherche généraux (Goole, Ask.com ou Yahoo!, etc.), moteurs spécialisés (Yahoo! Kids, Google Scholar, etc.) selon les utilisateurs, le domaine, le type de documents ou la région géolinguistique visés, méta-moteurs (Excite, Hotbot, Metacrawler, etc.) et autres. Leur fonction première est d'aider les utilisateurs à trouver un document qui peut répondre à leurs besoins d'information ; il reste à cet utilisateur à consulter le document en question (soit le lire, totalement ou partiellement), pour déterminer si son besoin d'information est satisfait.
Certains outils peuvent servir également à prendre connaissance (bien que sommairement) du contenu des documents. À cet effet, dans plusieurs cas, la liste des documents retournés en réponse à la requête contient un (très) court extrait de chaque document. De plus, ces outils reposent sur l'indexation préalable des documents, qui peut être faite en vocabulaire libre ou avec des vocabulaires contrôlés ; par exemple (Baca, 2003) : *Library of Congress Subject Headings*, *Art and Architecture Thesaurus*, *Thesaurus of Geographic Names*, *Library of Congress Thesaurus for Graphic Materials*. L'indexation fournit ainsi les métadonnées qui servent à décrire chaque document et à les apparier aux mots de la requête. Ces métadonnées peuvent également aider davantage l'utilisateur : elles peuvent aussi nourrir un système de visualisation de l'information, pour regrouper des documents semblables par exemple (comme dans les outils Metacrawler, Clusty, Grokker, etc.). Ceci peut aider l'utilisateur à se faire une idée de leur contenu. Certains systèmes de repérage (Davis, 2006) permettent à la fois la navigation dans une structure préétablie et une recherche par mots-clés ; la combinaison peut aider à mieux cerner la pertinence des documents repérés. Mais

l'utilisateur a peu de moyens, outre la lecture complète ou partielle du document, pour prendre connaissance de son contenu et évaluer la pertinence pour ses besoins. Un résumé peut alléger cette tâche, mais les résumés se font plutôt rares.
Certains chercheurs proposent des interfaces de navigation basées sur une analyse du contenu des documents (par exemple, Dakka et al., 2005). Elles servent alors à naviguer dans une collection de document, et non à l'intérieur d'un document donné. Quelques chercheurs (dont Hernandez et Grau, 2003) proposent un outil qui peut générer, pour un document, une structure semblable à une table des matières. Yaari et Gan (2000) font ceci à partir d'une analyse hiérarchique des sujets abordés et leur système construit aussi un index thématique, qui consiste essentiellement d'une liste de termes extraits d'une section donnée.
Nous cherchons à contribuer aux efforts de déploiement d'outils d'accès au contenu de documents numériques, qui permettraient aux utilisateurs de naviguer effectivement dans le document par le biais de son réseau conceptuel, et non de son organisation textuelle.

3 Un outil à considérer : l'index de livre

Nous voulons explorer un moyen, autre que le résumé ou la table des matières, qui aiderait l'utilisateur à prendre rapidement connaissance du contenu d'un document, en quelque sorte une aide à la lecture. Un outil à considérer pour l'accès au contenu des documents serait l'index de livre.

3.1 Structure

Un index de livre se présente comme une liste alphabétique d'entrées, chacune structurée en vedette principale et éventuellement de sous-vedettes, menant à une référence de page, par exemple :

> Température, 186-189 (Fenwick, 1997)
> du bain, 138, 141, 227
> de la chambre, 118, 121, 178
> fièvre, 180, 184, 186-188, 187
> pendant la grossesse, 38
> prise de la, 187
> urgence, 38, 174
> *voir aussi* Thermomètre

Chaque entrée représente un thème abordé dans le document ; les sous-vedettes le subdivisent en aspects secondaires, termes spécifiques, etc. Certaines entrées sont simples, constituées uniquement d'une vedette principale. La taille de l'index détermine sa couverture thématique par rapport au contenu global du document. Des renvois de type *voir aussi*

entre les entrées permettent d'établir des liens qui auraient pu échapper à l'utilisateur alors que les renvois de type *voir* (non illustré ici) mènent à des vedettes synonymes. Ce type d'outil est très familier aux utilisateurs de documents papier et il possède des caractéristiques différentes de celles offertes par les autres outils d'accès.

3.2 Comparaison avec autres outils d'accès au contenu

La structure de l'index est différente de celle de la table des matières : cette dernière reflète l'organisation textuelle alors que l'index est plutôt un inventaire de thèmes abordés. La table des matières aborde le contenu d'un document de manière séquentielle ; l'index permet d'y accéder de manière tabulaire (une longue réflexion sur des sujets reliés est présentée dans Vandendorpe, 1999).

L'index diffère d'un résumé en ce qu'il couvre davantage de thèmes, dans plus de détails, regroupant les mentions de ceux-ci qui seraient dispersées dans le document. Et bien sûr, le résumé est généralement un texte suivi (ou une grille textuelle), qui suit dans les grandes lignes l'organisation textuelle, alors que l'index est composé de courts termes juxtaposés et ordonnés (par ordre alphabétique ou un autre ordre systématique).

Par rapport à une fonction de recherche en texte intégral, l'index offre l'avantage de présenter ouvertement à l'utilisateur les thèmes du document ainsi que certaines des relations qui les unissent ; cela peut aider à mieux formuler une requête ultérieure. Une étude de Abdullah et Gibb (2009) a comparé trois types d'outils de navigation pour les livres électroniques (*e-books*) : index de livres, table des matières et fonction de recherche. L'index s'est révélé plus efficient que les autres en termes de rapidité pour trouver l'information, plus performant pour repérer correctement du contenu pertinent et plus convivial pour les utilisateurs.

Comme outil de navigation, il offre l'avantage de délimiter la couverture conceptuelle du document, proposant des termes pour une requête éventuelle. Il inclut des variantes des termes, pour repérer des thèmes peu importe la terminologie choisie. Et il peut représenter un outil additionnel, la recherche en texte intégral étant aussi souvent disponible. Enfin, Wathen and Burkell (2002) rapportent que les utilisateurs recherchent la familiarité de l'imprimé dans les environnements Web. L'inclusion d'un index de ce type, très connu des utilisateurs, présenterait donc de nombreux avantages.

La fonctionnalité de butinage qu'il incorpore est intéressante. Brown (1988) relève la distinction entre le butinage (ou navigation) et la recherche en termes de l'opposition entre ce que l'on recherche et l'endroit où il se trouve : pour le butinage, l'utilisateur procède de l'endroit vers l'information recherchée (*from where to what*) alors que pour la recherche, le mouvement est inverse, de l'information recherchée

à l'endroit où elle se trouve (*from what to where*). Pour la recherche, il faut donc savoir au départ ce que l'on veut trouver. Alors que la navigation permet une appropriation graduelle d'un contenu même si l'utilisateur n'a aucune connaissance préalable de celui-ci. Ertzscheid (2003) étudie les comportements différents qui caractérisent les deux activités liées au repérage d'information.

4 Construction automatique d'un index de livres

La construction d'un tel outil de manière automatique présente plusieurs obstacles. Les premiers prototypes (Artandi, 1963 ; Earl, 1970) n'ont guère eu de succès, étant limités par leur méthodologie qui consistait essentiellement à lister alphabétiquement les mots les plus fréquents du document. Or, une implémentation réussie doit contourner deux pièges : d'abord, éviter de confondre fréquence avec importance, ce qui implique de limiter les entrées d'index aux occurrences significatives d'un sujet ; ensuite, aborder le problème de la structuration des entrées, généralement faite (par les humains) sur la base de relations sémantiques, très difficiles à capter automatiquement.
Nous avons toutefois proposé une méthode inspirée de la méthodologie des indexeurs humains et qui permet d'atteindre de meilleurs résultats. Elle repose sur trois principes : (i) les indexeurs indexent des passages, et non des mots, et cette séparation en passages dicte le repérage des thèmes importants à l'intérieur de celui-ci ; (ii) à fréquence égale, les mots n'ont pas tous la même utilité dans l'index ; (iii) certains relations sémantiques entre les thématiques sont explicitées par les mots du passage. Les détails de l'implémentation sont présentés dans Da Sylva et Doll (2005) et sont esquissés ci-dessous. Mais il est utile d'approfondir d'abord la dernière idée, portant sur les relations sémantiques évoquées dans les entrées structurées de l'index.

4.1 Entrées complexes et relations utiles

Un index vise à aider l'utilisateur à repérer des passages utiles. Pour ce faire, l'index présente (en vedette principale) les concepts-clés du document. Lorsque plusieurs passages font référence à un même concept-clé, l'entrée énumère tous ces endroits en autant de numéros de page. Lorsque cette liste devient trop longue, elle est inutile à l'utilisateur, qui est confronté à un trop grand nombre de passages potentiellement intéressant pour sa recherche d'information. Il est préférable, alors, de distinguer chacune des références en introduisant des sous-vedettes qui explicitent l'aspect selon lequel le concept est envisagé dans chacun des passages. Cette présentation est souhaitable aussi bien dans un index créé manuellement que dans un autre construit automatiquement.

Il est alors important de déterminer quelle sous-vedette devrait être utilisée pour distinguer chaque référence. D'une part, pour simplifier la tâche du système, on peut supposer que la sous-vedette est présente explicitement dans le texte source (la dériver automatiquement exige des ressources sémantiques considérables) ; on voudra donc, dans l'implémentation, extraire au besoin des paires de termes, dont l'un sera finalement la vedette principale et l'autre sera la sous-vedette. D'autre part, on veut limiter le type de sous-vedettes utilisées (pour limiter le nombre de paires extraites), ce qui peut être fait en définissant les types de relations utiles dans un index.
Dans un travail précédent, nous avons analysé le type de relations entre la vedette principale et les sous-vedettes dans un certain nombre d'index créés par des indexeurs humains. Elles sont présentées dans Da Sylva (2004). Elles incluent entre autres la relation hyperonymique (voir la figure 1).

Type	Relation	Exemple
Syntagmatique	Mot – Terme avec ce mot	Grammaire - grammaire de dépendance
	Coordination	Café - et grossesse
Paradigmatique	Hyperonyme – Hyponyme	Mammifères – félins
	Tout – partie	Voiture – moteur
	Thème – Facette ou aspect	Robotique – développement

Figure 1. Relations principales observées dans les index de livres

4.2 Implémentation

Notre prototype d'indexation (Da Sylva et Doll, 2005) fonctionne de la manière suivante :

1. Segmentation du texte en segments thématiques. La méthodologie utilisée s'inspire de l'approche de Hearst (1997) et repose sur l'analyse de la cohésion lexicale : une coupure thématique est postulée entre deux segments quand le score calculé à partir d'indicateurs lexicaux (mots répétés, absence d'anaphores, etc.) chute. Nous avons modifié l'algorithme pour assurer la relative uniformité des segments. Cette segmentation sert à définir les passages auxquels les entrées d'index font référence.

2. Extraction des mots (et des suites de plusieurs mots, appelés multitermes) après lemmatisation et comptage des fréquences. Sur la base de la fréquence des mots (à l'intérieur des segments comme dans le document dans son ensemble), on déterminera la saillance d'un sujet dans un segment donné.

3. Identification, dans le texte, de paires de mots ou de multitermes qui pourront former des couples vedette principale/sous-vedette. Ces paires doivent relever de types précis, identifiés dans l'étude préalable (Da Sylva, 2004). Cette méthode permet de produire des entrées structurées comme celle donnée en exemple au début de cet article.
4. Pour chacun des éléments de la liste de candidats-termes (que sont les mots, termes et paires identifiés dans les étapes 2. et 3.), calcul d'un poids; pour chaque segment, on ne retiendra dans l'index que les candidats-termes dont les poids sont les plus élevés (au-delà d'un certain seuil).
5. Sélection des candidats-termes les plus saillants, regroupement sur la base des vedettes principales partagées et mise en ordre alphabétique.

Charlet et al. (2004) et Nazarenko et Aït El-Mekki (2005) présentent un outil très similaire à celui que nous avons développé de manière indépendante. L'introduction, dans le processus de construction de l'index, d'un auteur humain leur permet de contourner plusieurs problèmes liés à la limite de l'analyse automatique de la langue.
Notre originalité tient au traitement que nous accordons aux différents types de liens sémantiques qui peuvent tenir entre une vedette principale et une vedette secondaire. La figure 2 présente des exemples d'entrées d'index produites par notre prototype. Certaines entrées sont des mots simples, d'autres des multitermes, d'autres encore des paires de termes. Les numéros font référence aux segments obtenus par la segmentation automatique. Même dans ce court extrait, on voit que les concepts sont reliés dans l'index même quand ils sont disséminés dans le document.

béton béton armé, 5 limite, 5 dalles de béton ordinaire coulées, 4 renforcement du béton avec des fibres d'acier, 5 béton armé, 5 utilisation du béton précontraint, 4 béton ordinaire, 4	cheveu, 10 fibre, 9 fibres de noix de coco, 9 fines fibres, 3 renforcement du béton avec des fibres d'acier, 5 béton armé, 5

Notre prototype n'a pas encore fait l'objet d'une évaluation objective, sauf l'aspect segmentation de texte (Da Sylva, 2006), qui se compare favorablement à l'approche de Hearst (1997).

5 Quelques défis

On peut objecter qu'un index créé automatiquement doit offrir plus que simplement l'extraction des mots et expressions dans le texte, sinon il est d'une utilité limitée. Cependant, deux propriétés d'un index, même produit par pure extraction de termes du document, en justifie la création. D'abord, il offre un inventaire des concepts présents dans le document, explicitant du fait même la couverture conceptuelle aussi bien que lexicale ; c'est en quelque sorte une « photographie conceptuelle » de celui-ci. Et il indique également des relations entre les concepts (exprimées dans les entrées structurées en vedette principale et sous-vedettes). Ensuite, il restreint l'apparition des expressions à celles qui sont le plus importantes, alors qu'une fonction de recherche repérera chacune des occurrences.

Mais il est clair qu'il est préférable d'inclure dans l'index des expressions que l'on ne peut pas trouver directement dans le texte : par exemple, des synonymes ou des hyperonymes de termes du document. Si le document parle de « vélo », on voudrait trouver à l'index un renvoi « bicyclette, voir vélo » même si ce deuxième terme n'apparaît pas dans le texte. Également, un ouvrage qui parlerait de différents types de rongeurs, mais toujours dénotés par leur race spécifique (« souris », « rat », « écureuil », etc.), gagnerait à avoir une entrée « rongeurs » qui regrouperait chaque type. La difficulté réside alors à trouver des ressources lexicales externes au document qui contiennent ces informations. Nazarenko et Aït El-Mekki (2005) bénéficient d'une bonne solution à ce problème, ayant accès à une large base lexicale qui contient, pour chaque terme, des variantes aussi bien que des hyperonymes ou hyponymes. Un thésaurus général (disponible en format numérique) comme WordNet peut souvent fournir l'information nécessaire. En l'absence de ceci (par exemple, pour des langues pour lesquelles ces ressources n'existent pas), on doit imaginer d'autres stratégies. Comme par exemple l'analyse de grands corpus afin d'en extraire des généralisations pertinentes, parmi lesquelles pourront se trouver les relations qui nous intéresse (comme dans Ruiz-Casado, 2007, ou Hearst, 1992, par exemple). En plus, un thésaurus thématique serait avantageux : il contiendrait des connaissances disciplinaires spécialisées qui échappent aux thésaurus généraux.

On peut préférer une fonction de recherche en ce qu'elle nous amène directement à l'endroit dans le texte où l'objet de notre recherche apparaît. En contraste, l'entrée d'index nous amène normalement à une région textuelle (un passage, un paragraphe, une page) où il est du ressort de l'utilisateur de localiser l'endroit pertinent.

En outre, dans l'extraction des mots et termes du document, l'identification de ceux-ci se fait normalement sur la base de la chaîne

alphabétique, et non sur la base du sens du mot. Les ambiguïtés dues à la polysémie et à l'homographie amenuisent la performance du système.
En somme, cette tâche rencontre beaucoup des problèmes déjà identifiés dans d'autres applications de traitement automatique de la langue.

6 Conclusion

Pour faciliter l'accès au contenu de documents numériques, nous proposons un outil ancien, bien connu des utilisateurs et des indexeurs, l'index de livres. Cet outil serait particulièrement utile dans le cas de monographies assez importantes où la structure est peu apparente. Il offre un accès différent au texte, complémentaire à un résumé, à une fonction de recherche ou à une table des matières.
Nous avons proposé une implémentation qui tient compte, dans certains de ses aspects du moins, de la méthodologie des indexeurs humains et des propriétés attendues d'un index de qualité. Bien qu'une évaluation objective reste à faire, l'approche générale nous semble suffisamment motivée pour constituer un chantier de recherche intéressant.
Plusieurs pistes de recherche restent à explorer. Parmi celles-ci : l'ajout de ressources lexicales externes (comme un thésaurus général de la langue ou un thésaurus particulier au domaine de la monographie) ; l'évaluation par des utilisateurs et par des indexeurs professionnels ; la construction de différentes présentations de la structure conceptuelle définie par l'index, y compris des représentations en ontologies ou Topic Maps ; et le raffinement des types d'expressions qui constituent les entrées proposées pour l'index.

7 Références bibliographiques

[1] N. Abdullah et F. Gibb. Students Attitudes towards e-Books in a Scottish Higher Education Institute: Part 3 -- Search and Browse Tasks. *Library Review*, 58(1), 2009, 17-27.

[2] S. Artandi. *Book indexing by computer*. S.S. Artandi, New Brunswick, N.J. 1963.

[3] M Baca. Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information. *Cataloging & Classification Quarterly*, 36(3/4), 2003, 47-55.

[4] P. J. Brown. Linking and searching within hypertext. *Electronic Publishing*, 1(1), 1988, 45-53.

[5] J. Charlet, T. Aït el Mekki, D. Bourigault, A. Nazarenko, R. Teulier et B. Toledano. CEDERILIC : constitution d'un livre et d'un index

numériques. In : *Actes du Colloque International sur le Document Electronique (CIDE)*, 2004.

[6] W. Dakka, G. P. G. Ipeirotis et K.R. Wood. Automatic construction of multifaceted browsing interfaces. In *CIKM*, 2005, 768-775.

[7] L. Da Sylva. *Experiments in Proportional and Variable Automatic Text Segmentation* (poster). 19th Conference of the Canadian Society for Computational Studies of Intelligence (AI'06). 2006, Université Laval, Québec.

[8] L. Da Sylva et F. Doll. A Document Browsing Tool: Using Lexical Classes to Convey Information. In G. Lapalme et B. Kégl. *Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2005 (Proceedings)*, New York : Springer-Verlag, 2005, 307-318.

[9] L. Da Sylva. Relations sémantiques pour l'indexation automatique. Définition d'objectifs pour la détection automatique. *Document numérique,* 8, 3 (2004), 135-155.

[10] L. Davis. Designing a search user interface for a digital library. *Journal of the American Society for Information Science and Technology*, 57(6), 2006, 788-791.

[11] L. L. Earl. Experiments in automatic extraction and indexing. *Information Storage and Retrieval*, 6, 1970, 313-334.

[12] O. Ertzscheid. Comportements de navigation et documents électroniques : propositions d'invariants. In : C. Faure, J. Madelaine (réds), *Document électronique Dynamique. Actes du sixième colloque international sur le document électronique : CIDE.6*, Europia Productions, Paris, 2003.

[13] E. Fenwick. *Mon bébé, je l'attends, je l'élève* (traduction de *The Canadian Medical Association complete book of mother & baby care*). Reader's Digest Association, Montréal. 1992.

[14] M. Hearst. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1), 1997, 33-64.

[15] M. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, 1992, 539-545.

[16] N. Hernandez et B. Grau. What is this text about? Combining topic and meta descriptors for text structure presentation. In *Proceedings of the 21st annual international conference on Documentation (ACM SIGDOC)*, San Francisco, 12-15 Oct. 2003, 117-24.

[17] A. Nazarenko et T. Aït El Mekki. Building back-of-the-book indexes. *Terminology*, Special issue on Application-driven Terminology engineering, 11(11), 2005, 199-224.

[18] N. Hernandez et B. Grau. What is this text about? Combining topic and meta descriptors for text structure presentation. In: *Proceedings of the 21st annual international conference on Documentation (ACM SIGDOC*), San Francisco, 12-15 Oct. (2003), 117-124.

[19] M. Ruiz-Casado, E. Alfonseca et P. Castells. Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from Wikipedia. *Data & Knowledge Engineering*, 61(3), 484-99, 2007.

[20] Vandendorpe, C. Du papyrus à l'hypertexte: essai sur les mutations du texte et de la lecture, Boréal, Montréal, 1999.

[21] Y. Yaari et R. Gan. *NLP-assisted exploration of texts*. In In Proceedings RIAO'2000 Content-Based Multimedia Information Access, Paris, 2000, 2000.

Extraction de termes, reconnaissance et labellisation de relations dans un thésaurus – Vers une ontologie

Marie-Noelle BESSAGNET (1), Eric KERGOSIEN (2), Mauro GAIO (2)

UPPA, Laboratoire LIUPPA, IAE, Avenue du doyen Poplawski, 64012 PAU (1)
UPPA, Laboratoire LIUPPA, Faculté des Sciences, Département Informatique, 64000 PAU (2)

Mots-clés : Ingénierie des connaissances, Thésaurus, Représentation des connaissances, liste d'autorités, vedettes matière, ontologie

Keywords: Knowledge Engineering, thesaurus, Knowledge representation, Subject Headings, ontology

Résumé : Dans le domaine des systèmes de documentation, l'usage des thésaurus à des fins d'indexation puis de recherche d'information est courant voire obligatoire. Dans les bibliothèques et les médiathèques francophones, par exemple, les documents possèdent de par le travail effectué par les bibliothécaires de riches informations de description, sous la forme de notices descriptives, décrites sur la base du thésaurus RAMEAU. Nous exploitons ces deux types de ressources (documents et notices) afin de créer une première structure sémantique représentant le travail d'indexation des bibliothécaires pour élaborer le thésaurus TERRIDOC. Notre corpus de référence a une forte connotation territoriale. Nous nous intéressons également à la transformation de thésaurus en ontologie de domaine. En effet, nous souhaitons obtenir une ontologie de domaine offrant une représentation synthétique du territoire implicitement décrit par le fonds documentaire traité, en faisant appel à des ressources externes de type SIG.

Abstract : Within the documentary system domain, the integration of thesauri for indexing and retrieval information steps is usual. In libraries, documents own rich descriptive information made by librarians, under descriptive notice based on Rameau thesaurus. We exploit two kinds of information in order to create a first semantic structure. A step of conceptualization allows us to define the various modules used to automatically build the semantic structure of the indexation work. Our current work focuses on an approach that aims to define an ontology based on a thesaurus. We hope to integrate new knowledge

characterizing the territory of our structure (adding “toponyms” and links between concepts) thanks to a geographic information system (GIS).

1 Introduction

Les bibliothèques et les médiathèques, renferment des corpus documentaires de type patrimonial conséquents de plus en plus facilement disponibles pour le grand public grâce au format électronique (numérisation et OCRisation). Cependant l'accessibilité par le grand public à ces corpus reste encore problématique. Dans ces organismes de conservation, chaque document est associé à une notice descriptive établie par les bibliothécaires, elles sont construites sur la base d’un thésaurus faisant autorité dans le milieu, le thésaurus RAMEAU[1]. Nous proposons une exploitation automatique de ces deux types de ressources afin de créer une structure sémantique représentant le travail d’indexation des bibliothécaires.

Nous souhaitons, d’une part, proposer aux bibliothécaires des outils de visualisation et de parcours de cette structure afin de valider leur travail d’indexation. Cette approche se décompose en deux phases : la première étant d’identifier et de représenter l’information à l’aide des connaissances expertes extraites automatiquement des notices, la deuxième étant de donner la possibilité de naviguer dans le fonds documentaire via les connaissances identifiées pour faciliter la représentation du travail d’indexation. Actuellement, les relectures et éventuelles corrections sont réalisées manuellement notice par notice, rendant cette tâche fastidieuse. Nous pensons que la représentation sous forme de carte de connaissances extraites automatiquement du travail d’indexation apporte un premier élément de réponse à leurs attentes en leur offrant une synthèse exhaustive d’un état de l’indexation de la base documentaire. Nous nous intéressons, d’autre part, à la conceptualisation d'un sous-ensemble du thésaurus RAMEAU afin de produire une représentation ontologique de domaine mettant en avant un territoire. RAMEAU a été adopté dans le contexte d’informatisation des bibliothèques françaises dans les années 80. Cette liste d’autorités s’inspire largement du langage d’indexation RVM Laval (Canada), qui lui-même est issu d’un long travail de traduction à partir des vedettes-matières américaines tirées des LCSH (Library of Congress Subject Headings). Les thésaurus sont des vocabulaires contrôlés de termes représentant généralement un domaine particulier gérant des relations hiérarchiques, associatives et d’équivalence. On peut citer NML’s

[1] Répertoire d'autorité-matière encyclopédique et alphabétique unifié ; http://rameau.bnf.fr/. Thésaurus défini au sein de la Bibliothèque Nationale de France (BNF)

Medical Subject Headings (MeSH) dans le domaine médical pour indexer et rechercher des articles, le célèbre Wordnet, plus général, utilisé dans des travaux d'analyse sémantique. Dans le contexte de transformation de thésaurus en ontologie, le W3C travaille sur un méta schéma de référence, le SKOS (Simple Knowledge Organization System), basé sur les concepts [1].
Dans notre cas, l'objectif est d'expliciter la sémantique informelle du thésaurus autour des termes décrivant un territoire en se restreignant à l'aspect spatial. L'information du corpus (ici les notices descriptives) peut nous aider à spécifier des relations entre termes pouvant être ambigües dans un thésaurus afin de créer une première ontologie, L'analyse linguistique automatisée de ces notices doit ensuite nous permettre d'enrichir l'ontologie du domaine par de nouveaux concepts qualifiant un territoire.
Dans une première partie (&2), nous présenterons les problématiques et objectifs de notre travail de recherche. Nous développerons les travaux connexes dans le (&3) puis notre approche (&4) pour construire de manière automatique un thésaurus particulier : le thésaurus TERRIDOC. Enfin, nous expliciterons notre démarche pour passer d'un thésaurus particulier à une ontologie de domaine (&4) puis nous conclurons (&5).

2 Problématiques et objectifs

L'intérêt de disposer d'un fonds documentaire et de pouvoir ensuite proposer à des utilisateurs d'accéder aux informations nécessaires pour leur activité est primordial. Cela implique, d'une part, la nécessité d'identifier les informations pertinentes et d'autre part, la possibilité de fournir des moyens pour y accéder. Les possibilités pour organiser, classer et structurer un ensemble de documents sont nombreuses. Ainsi, afin d'offrir aux experts du domaine [2] un outil de validation de l'utilisation du langage contrôlé qu'ils ont mis en œuvre pour harmoniser leurs formulations de thèmes décrivant le contenu des documents, nous avons élaboré dans notre démarche deux phases préalables : (i) Extraction et Structuration des connaissances du domaine du fonds documentaire ; (ii) Navigation et interrogation du fonds documentaire en proposant une représentation sémantique de ce dernier. L'un de nos objectifs est la mise en place d'un processus pour passer d'un thésaurus classique à une base de connaissances. Ainsi, la première phase de notre démarche permet de créer automatiquement une structure représentant sous forme de thésaurus (le thésaurus TERRIDOC) le travail d'indexation des bibliothécaires en nous appuyant sur les notices descriptives pour

[2] Nous collaborons avec les bibliothécaires de la Médiathèque Intercommunale à Dimension Régionale (MIDR) de Pau)

identifier les termes et sur RAMEAU pour extraire les relations entre ces termes. Chacun des termes est ainsi enrichi par les relations de type « employé pour », « terme associé » et « terme générique » et par les termes RAMEAU se trouvant liés par ces relations. Ainsi, nous considérons chaque terme extrait des notices descriptives comme terme « de bas niveau » car rattaché directement à des documents et nous enrichissons le thésaurus avec les termes plus génériques du thésaurus RAMEAU. Le but visé par l'enrichissement du thésaurus via ces termes génériques est de permettre le regroupement en une seule structure des termes extraits. L'étape suivante consiste à enrichir cette première structure sémantique par des connaissances renseignant sur le territoire implicitement décrit par le fonds documentaire dans le but d'offrir aux utilisateurs un accès élargi à l'information. Ainsi, nous cherchons à exploiter dans nos ressources trois types d'informations : nous les qualifions d'entité thématique, d'entité spatiale (ES) [2] et d'entité temporelle (dans cet article, nous ne traiterons pas ce dernier type). Afin de capter ces entités et les relations existantes entre ces dernières, nous avons mis en place une chaîne de traitement sémantique automatisée, développée grâce à l'environnement Linguastream[3]. Elle est composée de quatre grandes phases [3]: (a) la lemmatisation pour segmenter les mots ; (b) l'analyse lexicale et morphologique pour la reconnaissance des mots ; (c) l'analyse syntaxique, basée sur des grammaires, afin de trouver les relations entre les mots ; (d) enfin l'analyse sémantique pour réaliser une interprétation plus spécifique sur les syntagmes retenus.
Afin de détecter ces entités, la partie extraction est découpée en étapes. La première (1) concerne la collecte d'ouvrages numérisés relatant d'un territoire. La seconde (2) supporte une analyse linguistique puis sémantique afin d'extraire les Entités précitées. La troisième (3) s'appuie d'une part sur des ressources géographiques (communes, lieux-dits, routes, pics, vallées, …) afin de valider les ES détectées à l'étape précédente et d'autre part sur la ressource RAMEAU afin de valider les Entités Thématiques. La dernière étape (4) propose la labellisation des relations entre ces diverses entités. Au vu de l'analyse de notre corpus, nous souhaitons nous intéresser à l'ensemble des relations binaires suivantes : Entité Thématique- Entité Spatiale et Entité Thématique- Entité Temporelle. Nous aborderons dans ce papier la relation Entité Thématique- Entité Spatiale. A cet effet, nous montrerons la démarche pour détecter des qualificatifs des toponymes ainsi que des relations d'approximation de sens avec les termes du thésaurus.

[3] http://www.linguastream.org/whitepaper.html

3 Travaux connexes

Transformer des thésaurus en ontologie fait l'objet de travaux de recherche récents. Depuis plusieurs années, les ontologies sont créées et utilisées dans le domaine de l'ingénierie des connaissances et notamment leur représentation. Le champ d'application est très large [4] : d'une manière générale dans l'indexation et la recherche d'information, et plus particulièrement dans le domaine médical, dans le domaine touristique, dans le domaine de l'éducation, dans le domaine de l'héritage culturel.
La conception automatique d'ontologies émerge comme un sous-domaine de l'ingénierie des connaissances. Afin de créer ces ontologies, il existe diverses approches et méthodes. Certains travaux reposent sur l'analyse de textes afin d'aider à la construction semi automatique des ontologies. D. Bourigault et al [5] décrivent les quatre étapes de la méthodologie de construction d'une ontologie à partir de textes (constitution du corpus à partir d'une analyse des besoins de l'application, étude linguistique afin d'identifier les termes et relations constituant la structure sémantique, normalisation sémantique définissant dans un langage formel les concepts et relations identifiées, validation de la formalisation par des spécialistes du domaine étudié). On peut remarquer que pour bâtir une ontologie à partir de textes, on utilise soit des ressources linguistiques externes, soit le corpus constitué des documents. Les outils supportant ces méthodes utilisent des techniques linguistiques pour retrouver les formes terminologiques dans l'analyse des textes. A Maedche et S. Staa [6] décrivent différents types d'approches distinguées en fonction du support sur lequel elles se basent : les plus courantes sont comme ci-dessus à partir de textes, de dictionnaires, d'autres à partir de bases de connaissances, ou encore de schémas semi-structurés et de schémas relationnels. Les travaux de [7] et [8] proposent une approche permettant de construire une ontologie minimale ; le processus consiste " *in extracting from texts specific types of information, rather than general-purpose relations. Accordingly, they produced remarkable efforts to conceptualize their competence domain through the definition of a core ontology*".
Comme déjà mentionné, nous nous intéressons plus particulièrement aux méthodes permettant de transformer un thésaurus en ontologie du domaine. Dans [9], l'approche présentée permet de transformer le thésaurus à facettes de l'art et de l'architecture AAT en ontologie pour indexer des images. Cette approche est entièrement manuelle. L'ontologie est formalisée en RDFS. Deux étapes d'identification de concepts et d'augmentation des concepts grâce à des propriétés permettent de définir cette ontologie. La méthode explicitée dans [10] repose sur trois étapes. Cette dernière a permis la transformation du thésaurus AGROVOC couvrant le domaine de l'agriculture, de la forêt,

de la nourriture et des domaines reliés tel que l'environnement. L'originalité se base sur une phase d'apprentissage afin d'extraire des relations supplémentaires augmentant ainsi la sémantique liée au thésaurus de base. Nos travaux actuels se rapprochent de ceux développés d'une part par [11], [12] et [13] qui s'appuient sur un thésaurus et un langage ontologique tel OWL pour améliorer l'interopérabilité entre outils et pour donner accès à ce dernier à une plus large communauté et d'autre part ceux de [14] qui simplifient l'opération de création d'ontologie à travers une approche permettant d'enrichir un thésaurus pour créer une ontologie à partir de sources de connaissances du domaine (vocabulaires, thésaurus, etc). Ces sources formalisées, contenant des termes représentant le domaine et (pour les thésaurus) des relations entre ces termes, apportent alors un plus sémantique indéniable à la représentation du domaine étudié.
En accord avec [11], l'une des étapes importantes pour transformer un thésaurus en ontologie est d'avoir une représentation des concepts et de leurs relations dans un format « traitable » par une machine. Nous avons choisi, dans un premier temps, de formaliser notre structure sémantique du domaine sur la base des Topics Map et sur OWL. D'une part, les TM sont le formalisme le plus adapté à des fins de navigation dans la carte de concepts[4] et dans leurs instances, ce qui nous a permis de concevoir un premier prototype. Nous avons ensuite travaillé sur une représentation OWL pour ses propriétés d'interopérabilité. Ce travail doit encore être approfondi. Le but n'est pas de représenter automatiquement le thésaurus en OWL mais de représenter le thésaurus dans un langage comme OWL. Ainsi, les travaux décrits dans [11] et [12] ont abordé ce thème de recherche lié à cette transformation. Plus récemment, [13] en transformant le thésaurus NCI en OWL DL ont rencontré des problèmes de représentation de connaissances dont nous pourrons tirer profit dans la construction de l'ontologie.
Nous allons aborder dans la partie suivante la démarche adoptée pour construire le thésaurus TERRIDOC puis nous nous intéresserons aux éléments de la méthodologie qui permettent de transformer le thésaurus en une ontologie.

4 Du fonds documentaire indexé à l'ontologie

Nous présentons à travers un exemple la méthodologie adoptée pour enrichir un premier vocabulaire de termes provenant du thésaurus

[4] Nous définissons une carte de concepts comme un triplet formé de : une liste de concepts, des relations entre ces concepts et des étiquettes (optionnelles) précisant ces relations.

RAMEAU afin de créer un thésaurus adapté, et les étapes à suivre pour transformer ce thésaurus en une ontologie.

4.1 Représentation sémantique de connaissances expertes

Nous nous appuyons dans notre démarche sur la base de notices descriptives correspondantes aux documents (figure 1) ainsi que sur le thésaurus RAMEAU. Dans notre phase d'extraction et de structuration des connaissances, l'exploitation des relations va nous permettre de construire le thésaurus TERRIDOC.
La première étape du traitement consiste à identifier et extraire automatiquement tous les termes (autorités matières RAMEAU) utilisés pour décrire le contenu du document dans les notices descriptives XML. Lors de la phase d'indexation, ces autorités sont sélectionnées par les bibliothécaires dans RAMEAU et utilisées dans les notices via la ou les balise(s) DEE (figure 1).

```
<DEE>Stations climatiques, thermales, etc. -- Barèges (Hautes-
Pyrénées) -- 18e siècle</DEE>
<DEE>Eaux minérales -- Pyrénées (France) -- 18e siècle</DEE>
<TITRE>Précis d'observation sur les eaux de Barèges et les eaux
minérales de Bigorre et du Béarn</TITRE>
<LEGENDE> Théophile de Bourdeu est à l'origine de la mode du
thermalisme pyrénéen</LEGENDE>
```

Figure. 1 – Extrait de notice descriptive 1

Chaque balise DEE correspond à une vedette-matière composée d'une ou plusieurs autorités séparées par l'élément « -- ». Chaque vedette-matière correspond à un thème estimé par l'expert comme important (l'autorité décrivant le thème est utilisée en tête de vedette) pour la description du contenu du document (*Stations climatiques, thermales, etc.* et *Eaux minérales* dans la figure 1). Nous obtenons en résultat de ce premier traitement un ensemble de termes. L'extraction automatique de cet ensemble de termes et leur mise en correspondance grâce au thésaurus RAMEAU dans un graphe conceptuel nous permet d'obtenir une première représentation sémantique du fonds documentaire. En exploitant le thésaurus RAMEAU, nous enrichissons automatiquement le vocabulaire obtenu ci-dessus avec : (i) les termes « génériques » et « employés pour » ; (ii) les relations qui leurs sont associées ; (iii) les relations entre termes associés s'il en existe.
Il faut noter que les relations hiérarchiques incluent la relation générique (genre-espèce), la relation partitive (tout-partie), la relation d'instance et les relations poly-hiérarchiques. Les travaux de D.H. Fischer [15] soulignent cette ambiguïté par le fait que la définition de ces relations « terme plus spécifique », « terme plus générique » est orientée par l'utilisation faite des thésaurus, c'est-à-dire l'aide au travail du documentaliste (indexation, recherche), et non par la formalisation de la

connaissance du domaine. Nous ne pourrons lever toutes les ambigüités liées aux relations « terme plus spécifique », « terme plus générique ». Si nous prenons l'exemple de la relation *Lieu de villégiature >GENERIQUE> Tourisme* , elle peut sembler incorrecte. Le but premier est d'avoir un outil de navigation à travers toute la structure sémantique.
Nos premières expérimentations ont été menées sur un corpus de 750 notices descriptives et leurs documents associés relatant du patrimoine culturel pyrénéen. Nous obtenons un ensemble de 1449 termes que nous enrichissons ensuite par les termes « employés pour », « associés » et « génériques » joints et par les relations correspondantes en nous appuyant sur RAMEAU. Le thésaurus obtenu offre une première structure synthétique représentant le travail des bibliothécaires. Seule, cette structure ne peut être exploitée par les experts pour observer et analyser l'ensemble des saisies. Nous en proposons donc une représentation sous forme de carte de concepts (figure 2) pour permettre aux experts d'appréhender de façon synthétique l'ensemble du travail d'indexation d'un fonds documentaire donné, réalisé par les différents bibliothécaires y ayant contribué.

Figure. 2 – Extrait du thésaurus TERRIDOC

4.2 Transformation du thésaurus en ontologie légère

Une contribution importante de notre travail concerne les étapes permettant de déceler puis de labéliser les ES ainsi que les relations associatives entre ces dernières et les concepts identifiés à partir du thésaurus RAMEAU.
Dans la volonté de définir une représentation d'un territoire en traitant la composante spatiale, la première étape de notre traitement consiste à identifier dans le thésaurus TERRIDOC les termes correspondants à des

ES qui vont devenir des instances, et les autres qui deviendront concepts. Nous utilisons pour cela la base Système d'Informations Géographiques de l'IGN, contenant la majorité des entités nommées spatiales françaises. Par exemple, les termes du thésaurus TERRIDOC «Bagnères-de-Bigorre (Hautes-Pyrénées)» et «Barèges (Hautes-Pyrénées) » sont identifiés comme entités spatiales, ce qui nous permet de créer un concept «entité spatiale» ainsi qu'une relation d'instance *instance_of* entre le concept entité spatiale et les deux instances «Bagnères-de-Bigorre (Hautes-Pyrénées)» et «Barèges (Hautes-Pyrénées)». Les autres termes du thésaurus sont ensuite définis comme concepts en leur ajoutant comme propriétés les définitions et/ou explications provenant du thésaurus RAMEAU. Dans l'extrait de l'ontologie (figure 4), «eaux minérales» et «stations climatiques, thermales, etc.» sont ainsi définis en tant que concepts. Cela nous permet de préciser les relations génériques avec les instances «Bagnères-de-Bigorre (Hautes-Pyrénées)» et «Barèges (Hautes-Pyrénées)» en relations d'instance *instance_of*. Cette première règle nous permet, en nous appuyant sur une ressource externe type SIG, de définir une ontologie légère offrant une première représentation sémantique d'un territoire. De ce fait, l'ontologie créée permet de faire les inférences élémentaires découlant de la taxonomie des concepts (p.ex. l'héritage des propriétés) sur ces concepts particuliers.

En plus du travail de description du contenu des documents, les notices descriptives renferment des informations riches pouvant décrire un territoire. Nous proposons en deuxième étape une chaîne de traitement linguistique (syntaxique et grammaticale) afin de capturer les ES ainsi que tous les termes les qualifiant. Afin de repérer des relations sémantiques [16], nous utilisons des patrons lexico-syntaxiques. Un patron lexico-syntaxique représente une expression régulière, formée de mots, de catégories grammaticales ou sémantiques, et de symboles. Il permet d'extraire des éléments de texte respectant l'expression. Dans notre cas, les patrons exploitent les étiquettes morpho-syntaxiques ou sémantiques attribuées par Linguastream (figure 3).

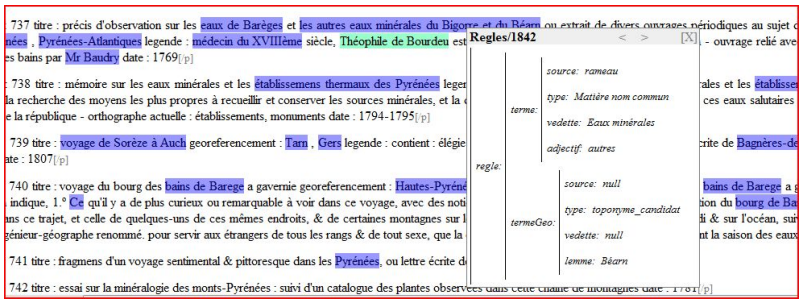


Figure 3. Extrait du traitement linguistique

La dernière phase de l'approche consiste à associer ces termes identifiés à l'ontologie par des relations de sens contenues dans les notices descriptives. Ainsi en reprenant les extraits de notices présentées figure 1, sont aussi retenus comme entités spatiales candidates les entités nommées «Bigorre» et «Béarn» que nous validons ensuite en tant qu'ES via l'appel au SIG. Un lien sémantique est alors créé entre le concept «Eaux minérales» et les instances de type spatial «Bigorre» et «Béarn» que l'on nomme *instance_of* (Figure 4).

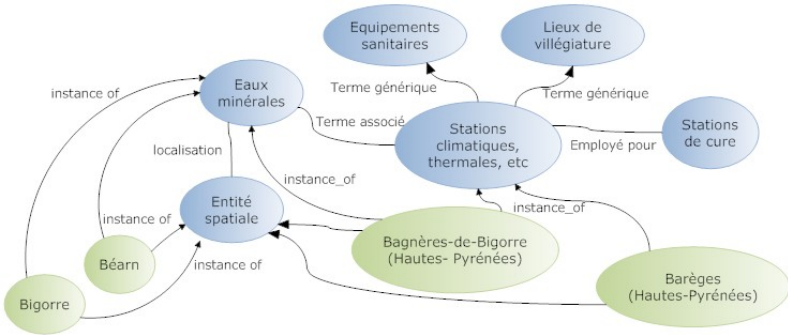


Figure 4. Extrait de l'ontologie générée

Nos travaux actuels cherchent à typer explicitement dans l'ontologie les relations classiques provenant du thésaurus TERRIDOC. Dans notre cas, un SIG peut nous permettre d'identifier, par calculs topologiques et géométriques sur les instances, les relations spatiales entre concepts. Nous cherchons aussi à caractériser l'ensemble des termes RAMEAU qui ne sont pas identifiés comme des instances de type spatial sous forme de concepts (possédant un nom, des caractéristiques propres sous forme d'attributs, etc.).

5 Conclusion

Comme nous l'avons expliqué, notre premier objectif est d'expliciter la sémantique informelle du thésaurus autour de concepts décrivant un territoire en se restreignant à l'aspect spatial dans le but de spécifier des relations ambigües entre termes présentes dans les thésaurus. Les traitements effectués sur le corpus de documents mis à disposition par la MIDR nous ont permis de modéliser la phase de création d'une ontologie enrichie du territoire reposant sur quatre étapes principales : (i) l'extraction d'informations du corpus via les notices XML associées aux documents que l'on organise sous forme d'un vocabulaire contrôlé, (ii) la définition d'un thésaurus (thésaurus TERRIDOC) caractérisant le

territoire issu du vocabulaire contrôlé et du thésaurus RAMEAU (thésaurus qui a assisté l'indexation manuelle par des documentalistes des documents du fonds documentaire de la MIDR), (iii) l'identification de toponymes du territoire et de relations les associant aux concepts déjà présents dans le thésaurus, (iv) la transformation du thésaurus en ontologie légère en l'enrichissant par les toponymes et relations spatiales identifiées.

Le processus mis en place est automatique. Notre ontologie est utilisée à des fins de recherche d'information et d'explicitation du travail des experts bibliothécaires. Aussi, nous travaillons actuellement à la spécification des relations que nous créons lors de la phase de transformation d'un thésaurus en ontologie. Nous envisageons d'appliquer notre méthodologie sur le thésaurus RVM Laval afin de conforter notre approche.

6 Références bibliographiques

[1] Alistair Miles, Brian Matthews, Dave Beckett, Dan Brickley, Michael Wilson and Nikki Rogers, SKOS Core: Simple Knowledge Organisation for the Web, 2005, http://www.w3.org/2004/02/skos/references, Dernier accès Web le 8 juillet 2009

[2] Lesbegueries J., C. Sallaberry, and M. Gaio, « Associating spatial patterns to text-units for summarizing geographic information ». 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval - GIR (Geographic Information Retrieval) Workshop, pp. 40-43, www.geo.unizh.ch/~rsp/gir06/accepted.html, ACM SIGIR 2006.

[3] Abolhassani, M., Fuhr, N., Govert; N. (2003). Information Extraction and Automatic Markup for XML documents, Intelligent Search on XML Data, LNCS Springer, p. 159–178.

[4] Gruber Thomas R.. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199-220, 1993.

[5] Bourigault D., N. Aussenac-Gilles, J. Charlet. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. Revue d'Intelligence Artificielle (RIA). Numéro spécial sur les Techniques Informatiques et Structuration de Terminologies. PIERREL J.M. et SLODZIAN M. (Ed.). Paris : Hermès. 18 (1), pp 87–110. 2004

[6] Maedche A., S. Staab, Ontology Learning for the Semantic Web, IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2), 2001

[7] Navigli R., P. Velardi. Ontology Enrichment Through Automatic Semantic Annotation of On-line Glossaries, Proc. of the 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2006), Podebrady, Czech Republic, October 2-6th, 2006, LNAI no. 4248, Springer, pp. 126-140

[8] Doerr Martin, J. Hunter, Carl Lagoze, *Towards a Core Ontology for Information Integration*, 2003, In Journal of Digital information, volume 4 issue 1, April 2003

[9] B. Wielinga, G. Schreiber, J. Wielemaker, and J. A. C. Sandberg. From thésaurus to ontology. Internation Conference on Knowledge Capture, Victoria, Canada, Octobre 2001

[10] Soergel D., B. Lauser, A. Liang, F. Fisseha, J. Keizer, S. Katz, Reengineering Thesauri for New Applications: the AGROVOC Example, Journal of Digital Information, Volume 4 Issue 4, 2004

[11] Jennifer Goldbeck, Gilberto Fragoso, Frank Hartel, James Hendler, Bijan Parsia, and Jim Oberthaler. The National Cancer Institute's Thesaurus and Ontology. Journal of Web Semantics, 1(1), Dec 2003. URL: ttp://www.mindswap.org/papers/WebSemantics-NCI.pdf

[12] L.F. Soualmia, C. Goldbreich, and S.J. Darmoni. Representing the mesh in owl: Towards a semi-automatic migration. In Proceedings of the First International Workshop on Formal Biomedical Knowledge Representa tion (KR-MED 2004), pages 81– 87, Whistler, Canada, June 2004

[13] Natalya F. Noy , Sherri de Coronado , Harold Solbrig , Gilberto Fragoso , Frank W. Hartel and Mark A. Musen Representing the NCI Thesaurus in OWL DL: Modeling tools help modeling Language, Applied Ontology 3 (2008) 173-190, DoI 10.3233/ A0-2008-0015, IOS Press

[14] Chrisment C., F. Genova, N. Hernandez et J. Mothe. «D'un thesaurus vers une ontologie de domaine pour l'exploration d'un corpus». *ametist,* Numéro 0AMETIST. http://ametist.inist.fr/document.php?id=152

[15] D. H. Fischer, From Thesauri towards Ontologies?, In Structures and Relations in Knowledge Organization : Proceedings of the 5th International ISKO Conference, W.M. Hadi, J. Maniez, S. Pollitt (Eds.),Würzburg: Ergon, pp. 18-30, 1998

[16] Auger, A., Barriere, C., 2008. Pattern based approaches to semantic relation extraction: a state-of-the-art. Terminology, John Benjamins, 14-1,1-19

Restructuration physique et logique de documents électroniques textuels

Jean-Luc BLOECHLE, Rolf INGOLD

Département d'Informatique, Université de Fribourg, CH-1700 Fribourg, Suisse.

Mots-clés : PDF, OCD, XML, structure physique, structure logique, modèle de document

Keywords : PDF, OCD, XML, physical structure, logical structure, document model

Résumé : La reconstruction des structures physiques et logiques de documents électroniques reste une problématique ouverte. Cet article présente une approche flexible et efficace permettant de régénérer de telles structures à partir de documents PDF. Une brève introduction présente tout d'abord le format PDF, ses atouts ainsi que ses défauts. Les principaux travaux dans le domaine de la restructuration de documents électroniques sont présentés. Un système complet de rétro-ingénierie du format PDF est ensuite exposé, celui-ci est basé sur une représentation intermédiaire appelée le document canonique, et permettant d'exprimer la structure physique tout en conservant l'apparence originale du document. L'étape finale de notre système d'analyse, la restructuration logique, est particulièrement mise en évidence. L'article conclut en exposant les travaux actuels et les éventuels améliorations futures.

Abstract : Physical and logical structure recovering from electronic documents is still an open issue. In this paper, we propose a flexible and efficient approach for recovering document structures from PDF files. After a brief introduction of the PDF format and its major features, we report about different existing works for PDF content extraction and analysis. To overcome the weaknesses of these systems, we propose a new analysis strategy, based on an intermediate representation, called canonical document, which enables representing physical structures in a canonical way. This paper then describes the PDF reverse engineering workflow and focuses on the document logical restructuring. Finally, the paper concludes with potential future improvements.

1 Introduction

Depuis sa publication en 1993, le format PDF de Adobe Systems est devenu le format standard pour l'échange et l'archivage de documents électroniques textuels et graphiques. En effet, le format PDF permet de restituer fidèlement l'apparence d'un document électronique quelconque aussi bien sur un écran que sur une imprimante. D'après Adobe Systems Incorporation, plus de 200 millions de documents PDF sont disponibles sur le web. Le format PDF peut être considéré comme un format universelle dans le sens où il est capable de reproduire toute information imprimable telle que du texte, des graphiques, des images, etc. Dans l'article "Why PDF is Everywhere" [1], McKinley met en évidence les points forts de ce format pour la gestion de documents et la recherche d'information. Le format PDF est d'ailleurs reconnu par les industries et gouvernements du monde entier. Dernièrement, un standard ISO a même été développé par l'organisation internationale pour la standardisation dans le but de spécifier un format PDF épuré nommé PDF/A et destiné à l'archivage à long terme.
Malgré toutes les qualités précitées, le format PDF n'est de loin pas parfait. En réalité, la spécification PDF a été définie afin de pouvoir reproduire tout document imprimable fidèlement et ceci au détriment de sa structure interne. Bien que les récentes spécifications du format PDF permettent d'incorporer des méta-données au contenu, la plupart des imprimantes PDF actuelles n'utilisent pas de telles possibilités. En conséquence, beaucoup de caractéristiques intéressantes liées aux structures du document sont perdues, alors qu'elles existaient au moment de l'édition. Cette perte d'information limite grandement la réutilisation de documents PDF, par exemple, la réédition ou le reformatage sont impossible, tandis que même des opérations aussi simple que copier/coller sont compromises.

Figure 1 : trois types de segmentation textuelle originale de documents PDF.

Il est intéressant de constater que la segmentation textuelle originale des documents PDF est totalement imprévisible. Aucune segmentation en mots ou unités lexicales n'est assurée par le format PDF, puisqu'il a pour unique but un rendu correct de telles entités, laissant leur représentation

interne au bon vouloir de l'imprimante PDF. La Figure 1 expose justement trois extraits de journaux au format PDF ayant des segmentations textuelles très diverses.
Au niveau logique, la séquence des blocs de texte n'est également pas assurée. De ce fait, la sélection ou l'exportation de texte avec Adobe Acrobat peut engendrer quelques surprises comme le présente la Figure 2.


Is online piracy unstoppable?

By Steve Lohr

The recording industry's long-running battle against online music piracy has come to resemble one of those "whack-a-mole" arcade games, in which the player hammers one rubber rodent's head with a mallet only to see another pop up nearby. Conk one, and up pops another, and so on. Three years ago, the music industry sued Napster, the first popular music file-sharing network on the Internet.

That sent Napster reeling, but other networks for trading copyrighted music — Kazaa, Grokster, Morpheus and others — sprang up. Last week, in the latest swing of the hammer, the Record-ing Industry Association of America filed 261 lawsuits against individual file sharers, which will surely make some of their estimated 60 million compatriots think twice — for now.

News Analysis

Earth Station Five, a company based in the West Bank, surfaced recently with claims of being at war with the industry association. It promises the latest in anonymous Internet file sharing. Its motto: "Resistance is futile."

Since Gutenberg's printing press, new technologies for creating, copying and distributing information have eroded the power of those in control of various media. In the last century, the pattern held true, for example, when recorded music became popular in the early 1900's, radio in the 1920's and cable television in recent years. But the heritage and design of the Internet present a particularly disruptive technology. Today's global network had its origins in the research culture of academia with its ethos of freely sharing information. And by design, the Internet turns every user in every living room into a mass distributor of just about anything that can be digitized.

See MUSIC, Page 7

Musicians have conflicting views on song-swapping. Page 9

Figure 2 : une sélection multicolonnes erronée ne respectant pas l'ordre de lecture.

2 Taxonomie des méthodes existantes pour l'analyse de PDF

Un nombre restreint de travaux et recherches ont été accomplis [2] afin d'exploiter le contenu des documents PDF, d'en extraire les structures physiques et logiques, et d'en dériver certaines annotations.
L'analyse de l'image du document bénéficie de méthodes qui ont mûri durant ces dernières décénies, de telles méthodes peuvent également être appliquées à des documents synthétique, sans bruits et imprimés en haute résolution [3], afin de retrouver le contenu et les structures originales de documents électroniques. Tandis que l'analyse direct du contenu électronique du document [4] profite de techniques partiellement dérivées de celles de l'analyse d'image. Ces méthodes récentes utilisent les primitives internes des document PDF [5]. Dans [6, 7], nous avons proposé de mélanger les deux méthodologies afin de pouvoir analyser tout type de PDF.
L'analyse du contenu électronique est à son tour composée de méthodes extensives et de restructuration. Les premières analysent le contenu du document afin de reconstituer les structures originales et y ajouter des annotations (tags PDF) sans réorganisation des primitives du document électronique. Ces techniques ont été appliquées avec des résultats intéressants dans plusieurs travaux [8, 9, 10]. L'objectif des techniques de restructuration est de représenter le document électronique en utilisant un

format différent du PDF, par exemple XML, pour permettre d'accéder facilement à l'information. Le cas le plus intéressant de restructuration est celui de la ré-ingénierie, qui vise à réorganiser le contenu du document en fonction des structures découvertes [11, 12, 13, 14, 15]. La conversion est un cas particulier de restructuration dans lequel aucune structure n'est extraite, le fichier PDF étant simplement transformé dans un format plus facile à manier [2].

3 Format canonique et restructuration physique

Le format canonique est un format développé au sein de notre groupe de recherche préservant fidèlement l'apparence d'un document électronique tout en y incorporant ses structures physiques. Le processus permettant de générer un tel document est le suivant : le contenu d'un fichier PDF est tout d'abord extrait par XED [7], puis la restructuration physique du document au format canonique est effectuée en utilisant une approche hybride. La restructuration physique a pour but de segmenter l'information textuelle en paragraphes homogènes composés de lignes elles-mêmes composées d'unités lexicales. L'algorithme de restructuration est divisé en trois phases :

- pré-traitement : normalisation, cristallisation, tri;
- phase ascendante : lexicalisation, linéarisation, fusion en blocs, fusion rétroactive, post-linéarisation;
- phase descendante : détection de changement d'interligne, détection de changement d'alignement.

Figure 3 : texte PDF brut à gauche et document canonique à droite

Toutes les étapes de l'algorithme utilisent des seuils dynamiques, relatifs à la taille de la police courante, permettant de fusionner ou segmenter le texte avec précision. La recherche des seuils a été faite empiriquement, tout d'abord par une estimation a priori de leurs valeurs, puis par un affinage minutieux sur un corpus éclectique de documents PDF. Quatre seuils ont été nécessaires au bon fonctionnement de l'algorithme: un seuil pour la fusion des caractères en mots, un seuil pour la fusion des mots en

lignes, un seuil pour la fusion des lignes en blocs de texte, et finalement un seuil plus général appelé seuil de précision (utile pour des tests d'alignement ou d'interligne par exemple). Une présentation détaillée de l'algorithme a déjà été présentée dans [2] et [16]. La Figure 3 ci-dessous présente un extrait de texte PDF brut à gauche, puis sa version segmentée à droite.
L'extraction de la structure physique a été appliquée sur trois documents différents, dont deux à structures complexes, les résultats obtenus sont exposés sur le Tableau 1.

Document title	number of text blocks	number of errors	correctness
French newspaper - Le Monde 2009/01/16	1827	35	98.08%
Swiss newspaper - La Liberté 2009/01/15	1410	16	98.87%
E-book - Alice's Adventures in Wonderland	1108	7	99.37%

Tableau 1 : résultats de l'extraction de la structure physique sur trois documents.

4 OCD, un formalisme XML optimisé pour le contenu physique

Le stockage permanent d'un document canonique au format OCD (Optimized Canonical Document) [16] permet à la fois de représenter la structure physique et de garantir la reproduction fidèle de ce document. Le format OCD est une description XML compacte et simple permettant le stockage permanent d'un document au format canonique sur un support physique. Son but n'est pas de concurrencer un quelconque autre format, mais bien de conserver un document structuré tout en préservant son aspect visuel d'origine, et cela d'une manière simple et synthétique. L'accès aux informations d'un tel format doit être facilité au maximum.

```
<block x="96.78" y="97.8">
 <line>
  <token tm="13.5" font-id="1" cs="0" ts="0">
43 48 41 50 54 45 52</token>
  <token ws=".3328"/>
  <token>49 49</token>
  <token>3a</token>
  <token/>
  <token>54 68 65</token>
  <token/>
  <token>50 6f 6f 6c</token>
  <token/>
  <token>6f 66</token>
  <token/>
  <token>54 65 61 72 73</token>
 </line>
</block>
```

Figure 4 : un extrait du format canonique représenté en OCD.

OCD supporte trois sortes de primitives graphiques : texte, image, et graphique vectoriel. Chaque primitive textuelle, graphique ou image y est décrite relativement à un état graphique de la page courante. Ainsi, un attribut est déclaré uniquement si celui-ci a changé de valeur relativement à l'état graphique qui lui-même mis à jour avec la nouvelle valeur de l'attribut. Les représentations des primitives utilisent des descriptions synthétiques. Les images sont compressées au formats JPG ou PNG puis insérées dans le document XML sous forme de flux hexadécimal. Les graphiques utilisent une description similaire à SVG, des coordonnées relatives sont employées à l'intérieur d'un même graphique. La représentation du texte bénéficie grandement du regroupement homogène des entités textuelles du format canonique permettant ainsi une description très réduite. Les primitives textuelles utilisent les largeurs de caractère de la fonte courante ainsi que des opérateurs d'espacement de caractère, de mot et d'interligne (cf. Figure 4). Le positionnement de chaque caractère est de ce fait respecté avec précision et cela avec un minimum d'espace disque. Finalement le fichier XML résultant est compressé en suivant le standard GZIP.
Ainsi, bien que OCD soit basé sur une représentation XML, sa taille est extrêmement réduite. Le Tableau 2 montre en effet que, par rapport au format PDF, notre format OCD permet de substantielles réductions de tailles de fichiers sur des documents textuels. Le tableau compare également notre format de fichiers aux formats XPS (le format de Microsoft) et XCD (ou XCDF, notre ancien format de stockage de documents canoniques [2]).

Document title	PDF	XPS	OCD	XCD
Aesop's Fables - 93 pages	243 KB	441 KB	91.9 KB	7'014 KB
Around the World in 80 Days - 339 pages	766 KB	1'912 KB	422 KB	36'045 KB
The Odyssey - 550 pages	1280 KB	3082 KB	850 KB	63'693 KB
The Last of the Mohicans - 698 pages	1'605 KB	3'944 KB	924 KB	80'896 KB
Ulysses - 1305 pages	2'953 KB	7'334 KB	1'743 KB	-

Tableau 2 : évaluation du format OCD par rapport à PDF, XPS, et XCD.

5 Dolores : un outil interactif pour la restructuration logique

A partir d'un document au format canonique, Dolores [17] (Document Logical Restructuring) permet de régénérer une structure logique par apprentissage interactif incrémental. L'utilisateur créé un modèle par interaction, apprentissage et correction. Il peut ensuite l'appliquer à d'autre documents d'une même classe et améliorer ce même modèle grâce

à l'apprentissage incrémental (cf. Figure 5). Trois phase principales peuvent être mise en évidence dans ce processus : l'extraction des caractéristiques, l'étiquetage logique et l'apprentissage.

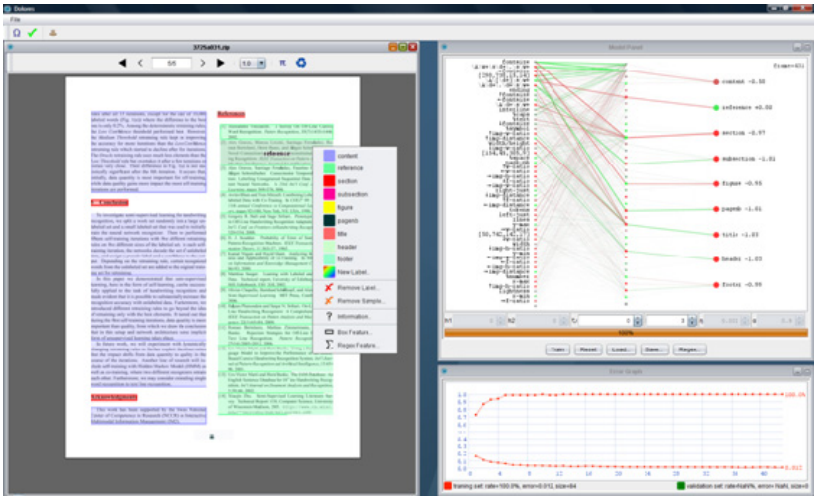


Figure 5 : Capture d'écran de Dolores, à gauche le document étiqueté, à droite le modèle.

6 Extraction des caractéristiques

L'extraction des caractéristiques est une tâche primordiale préalable à la phase d'apprentissage du système. Le choix des caractéristiques extraites, leur nombre, leur pertinence a un impact direct sur la création du modèle de document et donc sur les résultats de la classification. Dolores extrait un ensemble de caractéristiques de natures diverses sur chaque bloc textuel : géométriques, typographiques, topologiques.
Les caractéristiques extraites sur chaque bloc textuel sont les suivantes: coordonnée x/y, largeur, hauteur, rapport largeur/hauteur, taille de la fonte, interligne, luminosité de la fonte, écart type de la justification à gauche/droite, nombre de mots, nombre de lignes, pourcentage de majuscules, pourcentage de symboles, pourcentage de mots, pourcentage de nombres, pourcentage d'espaces, présence d'un caractère de ponctuation en fin de bloc, numéro de page, distance aux blocs textuels voisins (supérieur/inférieur/droite/gauche), tailles des fontes des blocs textuels voisins, rapports de la fonte courante aux tailles des fontes des blocs textuels voisins, rapport de la largeur du bloc courant aux blocs textuels voisins, distance aux images voisines, rapport de largeur du bloc courant aux images voisines.

Deux autres classes de caractéristiques sont également prises en compte : les régions et les expressions régulières. Concernant les régions, l'intersection des surfaces des blocs de texte (boîte englobante) d'une même classe est calculée, si celle-ci n'est pas nulle, la boîte englobante résultante est ajoutée comme caractéristique au modèle. La valeur de la caractéristique est le pourcentage de recouvrement de la surface d'intersection avec le bloc de texte courant. Concernant les expressions régulières, le principe est le même, une expression régulière est générée pour chaque échantillon (bloc de texte), l'expression régulière commune à chaque classe est recherchée, en cas de succès, celle-ci est ajoutée au caractéristiques du modèle.

6.1 L'étiquetage logique

La figure 5 montre l'interface de Dolores. L'étiquetage logique y est effectué d'une manière interactive. En effet, l'utilisateur peut ajouter ou supprimer des étiquettes lorsque bon lui semble. Le système d'apprentissage ajoute dans le modèle tout nouveau bloc de texte étiqueté. Une phase d'entraînement est ensuite instantanément effectuée, les blocs de texte sont alors étiquetés à la volée. L'action de l'utilisateur (l'étiquetage) est directement suivi de la mise à jour du modèle et reflété au travers de l'interface. L'utilisateur voit les erreurs d'étiquetage et corrige celles-ci de manière itérative. L'utilisateur peut étiqueter un bloc par l'intermédiaire du menu contextuel de la souris, ou alors directement en cliquant sur celui-ci si la classification actuelle est adéquate. De plus, dans le cas ou tous les blocs de texte d'une page sont correctement étiqueté, l'utilisateur peut insérer ceux-ci en vrac en allant dans le menu contextuel et en cliquant sur "étiqueter page" (ce menu ne peut apparaître que lorsque le pointeur de souris est à l'extérieur de tout bloc de texte et que l'utilisateur clique sur le bouton droit).
L'interface fournit des informations cruciales à l'utilisateur, lui permettant d'effectuer son étiquetage aisément et rapidement. Par exemple, la classe (l'étiquette logique) attribuée à chaque bloc de texte par le modèle est représentée par une surface rectangulaire colorée et semi-transparente (la couleur étant définie au préalable par l'utilisateur). Chaque bloc de texte contenu dans l'ensemble d'entraînement est encadré par un rectangle englobant dont la couleur correspond à celle de son étiquetage. Une barre horizontale est également affichée en-bas de chaque bloc de texte, son pourcentage de remplissage exprime le taux de confidence de l'étiquette attribuée par le modèle. Ainsi un taux de confidence bas indique qu'il est préférable de continuer à étiqueter la classe correspondante. Finalement, lorsque l'utilisateur passe sur un bloc de texte, celui-ci est mis en évidence par la superposition d'une surface rectangulaire grise semi-transparente, son étiquette logique s'affiche au centre de celui-ci, le code

couleur pouvant parfois s'avérer insuffisant (s'il y a beaucoup de classes par exemple).

6.2 Modèle et apprentissage

L'apprentissage est géré par un perceptron multicouches. Le modèle de document comprend à la fois l'ensemble des échantillons étiquetés (blocs de texte) ainsi que les données définissant le réseau de neurone. Une interface simple et conviviale implique que l'apprentissage soit totalement automatisé et instantané. De ce fait, la topologie du réseau est dynamique, elle s'adapte automatiquement au nombre d'entrées et de sorties. Le réseau contient une couche cachée. La couche d'entrée est totalement connectée à la couche cachée tandis que chaque neurone de la couche de sortie est connecté à quatre neurones de la couche cachée. Ceci assure à chaque neurone de sortie un nombre égale de neurones caché et évite que ceux-ci soit accaparés par un autre neurone de sortie dont la probabilité a priori est beaucoup plus élevée. Sans entrer dans les détails, l'algorithme d'entraînement du réseau est une rétro-propagation stochastique avec moment d'inertie. Le taux d'apprentissage diminue en fonction de l'erreur en sortie d'un neurone. Ces caractéristiques assurent un apprentissage convergeant et rapide, tout en minimisant le risque de stagner dans des minima locaux. Actuellement, l'apprentissage s'arrête lorsque le taux de reconnaissance est de 100% sur un minimum de 30 cycles consécutif (avec une borne temporel).
L'affichage du réseau neuronal met en évidence la force des pondérations ainsi que la pertinence de chaque caractéristique d'entrée par rapport à l'ensemble des classes ou alors pour une classe donnée (en pointant un neurone de sortie avec le curseur de la souris). Ceci permet à l'utilisateur d'appréhender d'un seul regard les caractéristiques discriminantes du réseau dans sa globalité ou pour chaque classe séparément. L'interface du réseau de neurone offre également la possibilité de désactiver un neurone d'entrée, afin de voir son impact sur le modèle. Un graphe d'erreur est affiché en dessous du réseau de neurones, il contient la courbe d'erreur ainsi que le taux de reconnaissance sur l'ensemble d'apprentissage et éventuellement sur un ensemble de validation/test. Enfin, il est possible de sauvegarder et d'ouvrir les modèles afin de les appliquer sur d'autres documents, ou éventuellement de les améliorer.

7 Conclusion

Cette article présente un système complet d'analyse de documents électroniques textuels. A partir d'un document PDF, ou tout autre document électronique textuel imprimable, le système extrait toutes les données textes, images et graphiques. Une restructuration physique est

ensuite effectuée sur le document, le résultat est alors sauvegardé au format OCD. L'étape de restructuration logique est assurée par Dolores, un outil interactif pour l'apprentissage incrémental de modèles de documents. Actuellement, seul les étiquettes logiques sont supportées par le modèle. La reconstruction de la hiérarchie fait partie des travaux futurs. Tandis que l'étude approfondie de la génération des modèles, ainsi que l'impact des divers paramètres d'apprentissage sur le taux de reconnaissance sont en cours d'évaluation. Le résultat de la restructuration logique d'un document peut finalement être conservé directement dans le format canonique au moyen de liens internes et sauvegardé sur disque grâce à un format étendant OCD nommé OCDL.
Le développement d'un processus complet permettant la réutilisation de contenus PDF est une gageure qui ne saurait être mise de côté, en effet, un tel processus permet de réactiver le cycle de vie des documents électroniques.

8 Références bibliographiques

McKinley, T. Why PDF is Everywhere. Inform, the journal of AIIM, 11(8), 1997.

Bloechle, J.-L., Rigamonti, M., Hadjar, K., Lalanne, D. and Ingold, R. XCDF: A canonical and structured document format. In 7th International Workshop, DAS'06, pages 141-152, Nelson, New Zealand, February 2006. Springer-Verlag.

Hadjar, K. and Ingold, R. Arabic Newspaper Page Segmentation. In Proceedings of the Seventh international Conference on Document Analysis and Recognition - Volume 2 (August 03 - 06, 2003). ICDAR. IEEE Computer Society, Washington, DC, 895.

Paknad, M.D. and Ayers, R.M., Method and apparatus for identifying words described in a portable electronic document, U.S. Patent 5,832,530, 1998.

Rigamonti, M., Bloechle, J.-L., Hadjar, K., Lalanne, D. and Ingold, R. Towards a Canonical and Structured Representation of PDF Documents through Reverse Engineering. ICDAR'05, 2005, pp. 1050-1054.

Hadjar, K., Rigamonti, M., Lalanne, D. and Ingold, R. Xed: a new tool for eXtracting hidden structures from Electronic Documents. DIAL'04, 2004, pp. 212-221.

Rigamonti, M., Hadjar, K., Lalanne, D. and Ingold, R. Xed: un outil pour l'extraction et l'analyse de documents PDF, CIFED'04, 2004, pp. 85-90.

Bagley, S.R., Brailsford, D.F. and Hardy, M.R.B. Creating reusable well-structured PDF as a sequence of component object graphic (COG) elements. DocEng'03, 2003, pp. 58-67.

Hardy, M.R., Brailford, D. and Thomas, P.L. Creating Structured PDF Files Using XML Templates, DocEng'04, 2004, pp. 99-108.

Lovegrove, W.S. and Brailsford, D.F. Document analysis of PDF files: methods, results and implications. Electronic Publishing, 1995, pp. 207-220.

Anjewierden, A. AIDAS: Incremental logical structure discovery in PDF document. ICDAR'01, 2001, pp. 374-377.

Chao, H. and Fan, J., Capturing the Layout of electronic Documents for Reuse in Variable Data. ICDAR'05, 2005, pp. 940-944.

Dejan, H. and Meunier, J.L.,A System for Converting PDF Documents into Structured XML Format. DAS'06, 2006, pp. 129-140.

Futrelle, R.P., Shap, M., Cieslick, C. and Grimes, A.E. Extraction, layout analysis and classification of diagrams in PDF documents. ICDAR'03, 2003, pp. 1007-1012.

Rahman, F. and Alam, H. Conversion of PDF documents into HTML: a case study of document image analysis. Asilomar CSS'03, 2003, pp. 87-91.

Bloechle, J.-L., Lalanne, D. and Ingold, R. OCD: An Optimized and Canonical Document Format. In 10th International Conference on Document Analysis and Recognition, ICDAR'09, Barcelona, Spain, July 2009, pp. 236-240.

Bloechle, J.-L., Pugin, C. and Ingold., R. Dolores: An Interactive and Class-Free Approach for Document Logical Restructuring. In 8th International Workshop, DAS'08, pages 644-652, Nara, Japan, September 2008.

Une approche de catégorisation structurelle de documents numériques pour une meilleure exploitation du patrimoine juridique décisionnel

Jin YAO (1), Jacques MADELAINE (1), Khaldoun ZREIK (2)

(1) DoDoLa - GREYC, Université de Caen, France
(2) CITU - Paragraphe, Université de Paris 8, France

Mots-clés : catégorisation de documents semi-structurés, extraction de connaissance, recherche d'information, patrimoine juridique décisionnel

Keywords: semi-structured document clustering, knowledge discovery, information retrieval, decision support for legal heritage

Résumé :

Le patrimoine de document juridique (loi, jurisprudence, brevet) s'est bien approprié l'univers de numérisation pour permettre une diffusion et une exploitation accrues des informations juridiques par des applications diverses. En conséquence, l'usage des bases documentaires juridiques partageables est devenu de plus en plus ouvert et fréquent favorisant ainsi un débit d'alimentation « semi-automatique » assez important. Constat 1 : par semi-automatique, on entend un processus de dépôt direct des documents dans des bases contrôlées par des SGBDs qui exigent une intervention humaine réduite surtout au niveau de l'indexation et de la classification. En effet, ce sont les modèles de documents (leurs structures logiques et physiques modélisées par le langage de balisage) qui assurent un rôle important dans les processus d'indexation et de gestion. Donc ces modèles incorporent indirectement connaissance et savoir-faire. Constat 2 : devant une telle masse de données « très souvent textuelles », il devient indispensable d'adopter aussi une approche pour gérer les documents électroniques juridiques en tant que supports de connaissance et de savoir faire. Ceci nous mène vers des problématiques de recherche d'information et d'extraction de connaissance. Ces deux constats nous conduisent à formuler une hypothèse de classification automatique qui tiendra compte de connaissance et de savoir-faire incorporés dans les structures des modèles de documents électroniques juridiques. Aussi on constate que ces connaissances ou savoir-faire ne sont pas toujours explicites dans les corps de documents. Cela nous dirige vers une approche de catégorisation pour extraire des catégories décisionnelles. Cet article présente une méthode de représentation de document semi-structuré

permettant d'analyser précisément les connaissances et le savoir-faire incorporé dans les contenus et les structures du document. Les expériences sur un corpus juridiques réel montrent que la prise en compte à la fois du contenu et de la structure conduit à une amélioration remarquable de qualité des catégories décisionnelles.

Abstract :

The legal document (law, case law, patent) uses commonly scanning facilities for dissemination and exploitation of legal information through various applications. Thus, the use of legal documentary databases has become more and more open and frequent, leading to a fairly important "semi-automatic" feeding mode. Observation 1: we intend to make a "semi-automatic" process to deposit directly documents in databases controlled by DBMS, including indexing and classification with a limited human intervention. In fact, it is the documents templates (the logical and physical structures modelled by the markup language) that take an important place in the process of indexing and management. Then the templates incorporate indirectly the knowledge and the expertise. Observation 2: in the presence of such a mass data (very often textual), it becomes essential to adopt an approach to manage the electronic legal documents as carriers of knowledge and expertise. This shifts the problem to domains of information retrieval and knowledge discovery. These two observations lead us to formulate an hypothesis for automatic classification that considers the knowledge and expertise incorporated in the structures of the legal electronic documents. This is motivated as we find that the knowledge or expertise are not always explicit in the document body. That pilots us to an approach of categorization to discover decision-making clusters. This article presents a representation method for semi-structured document who allows to analysis very precisely the knowledge and expertise incorporated in both contents and structures of document. The experiments upon a real legal corpus show that incorporation of content and structure produces a remarkable improvement of the quality of decision-making clusters.

1 Introduction

Dans le domaine du document juridique, chaque sous-domaine spécifique (brevets, jurisprudences, par exemple) respecte pratiquement la même structure de rédaction. Ceci peut expliquer l'usage réussi des langages de balisage comme XML (*eXtensive Markup Language*) pour la gestion et l'archivage de documents dans ce milieu. Ainsi nous travaillons sur un ensemble de documents juridiques en format XML, qui représente une base de données semi-structurées.
Partant du fait que les travaux en fouilles des données et la recherche d'information ont montré l'efficacité d'extraction d'information et de recherche d'information à partir des données fortement structurées (cas des bases de données relationnelles), nous supposons que les

informations incorporées dans la structure de documents semi-structurés peuvent aider à mieux catégoriser ces derniers afin d'améliorer la recherche d'information ou la découverte de nouvelles connaissances.
Un des objectifs indirects de cette étude est de présenter la structuration de document comme une démarche anticipative pour la gestion de patrimoine « numérique » dans le domaine de droit. Le document juridique dont la forme doit respecter des règles de rédaction strictes, nous semble fortement intéressant comme objet de recherche.
Nous proposons une méthode de catégorisation structurelle pouvant regrouper automatiquement et efficacement les documents similaires dans les mêmes classes sans aucune connaissance du domaine juridique a priori. Cette méthode d'apprentissage automatique, non-supervisé, peut être considérée comme faisant partie d'un processus de prétraitement de documents en vue de recherche ou d'extraction d'information.
Nos travaux de recherche ont montré que les caractéristiques du domaine juridique, fortement structuré, présentent un facteur favorisant l'extraction d'information à partir de la structure. L'extraction d'information dans notre projet est limitée à l'extraction de catégories décisionnelles pouvant aider le juriste à prendre des décisions en classant un cas d'étude ou de procès en cours.
Nous allons présenter la démarche du travail dans la première section. Dans la deuxième section, nous présentons le modèle de représentation que nous avons retenu. Puis nous détaillons des expériences effectuées sur la jurisprudence du Conseil Constitutionnel français. Avant la conclusion, nous analyserons les résultats des expérimentations.

2 Démarche

Les documents électroniques semi-structurés utilisent des balises XML ayant des propriétés structurelles. Cette opportunité a offert de nouveaux défis à l'apprentissage automatique, et particulièrement à la catégorisation. Plusieurs approches et méthodes ont été proposées à ce propos et peuvent être réparties en deux catégories :
Dans la première catégorie, les travaux ne considèrent que la structure du document. [i] adoptent une approche de traitement de signal pour catégoriser les documents. Les balises XML sont ainsi représentées comme une série temporelle. Et la similarité entre les documents est calculée en analysant des coefficients de transformation de Fourier. [ii] et [iii] proposent d'analyser directement la structure du document XML qui est représentée sous la forme d'un arbre de balises. La catégorisation par la structure du document permet de réduire la structure hétérogène d'un semble de documents. L'inconvénient principal de cette approche réside dans la complexité polynomiale des algorithmes utilisés.

La deuxième catégorie tient compte à la fois du contenu et de la structure d'un document XML. Dans [iv] [v] [vi], l'arbre du document XML est transformé en un sac de chemins, un sac de mots ou un sac mixte de chemins et de mots. Pour représenter l'ensemble de ces descripteurs linéaires, ils adoptent le modèle vectoriel proposé par Salton [vii]. [viii] ont étendu le modèle vectoriel en combinant le contenu, les éléments et les hyperliens dans le document XML. Ces travaux ont montré qu'une approche de catégorisation par l'information de contenu et l'information de structure donne une meilleure précision de regroupement si la structure de la collection en question est homogène.
Nous nous intéressons à découvrir la connaissance et le savoir-faire menés par le contenu et la structure du document. Nos expériences précédentes [ix], [x] montre que l'hétérogénéité de la structuration du document général affecte peu la qualité de la catégorisation thématique. Dans cet article, nous nous concentrons sur les documents juridiques à structuration homogène. Nous réalisons un processus heuristique pour comparer au fur et à mesure les différents descripteurs de document semi-structuré : d'abord, le descripteur de mot classique est utilisé ; ensuite, les descripteurs de structure seule sont examinés ; à la fin, le contenu et la structure hiérarchique du document sont pris en compte globalement. En comparant les résultats de trois séries de catégorisation, nous pouvons explorer le savoir-faire de la structure pour le prétraitement de patrimoine de documents juridiques.

3 Spécificités du document semi-structuré

Conserver le patrimoine exige de ne pas perdre l'information, donc on s'oriente vers une approche de traitement et de prétraitement qui concerne au maximum l'information encapsulée dans un document. Le document semi-structuré propose un modèle hiérarchique qui est généralement considéré comme un arbre. Les travaux existant ont montré que la complexité de catégorisation des arbres est élevée. Nous adoptons une méthode qui transforme une représentation arborescente du document en une représentation vectorielle sans pourtant perdre les informations hiérarchiques de l'arbre.
La figure 1 montre un exemple de document du Conseil Constitutionnel français structuré au format XML. On représente ce document en structure arborescente par des composants linéaires. Chaque composant représentant un type de l'information de contenu ou de l'information structurelle est un descripteur du document. Le modèle de chemins est choisi pour représenter l'information hiérarchique de la structure. Un chemin est une séquence ordonnée d'éléments qui représente une série consécutive de relation parent-enfant. Un chemin complet est une

séquence d'éléments qui commence à l'élément racine et se termine à un élément feuille (voir la figure 2). La longueur d'un chemin est le cardinal de l'ensemble d'éléments dans la séquence. En limitant la longueur d'un chemin complet, on peut créer différents types de sous-chemins. A partir de l'élément racine, après avoir compté *n* éléments, un chemin enraciné de longueur *n* est crée. À l'inverse, un chemin feuillu est crée à partir d'un élément feuille. En attachant le mot contenu dans un élément d'un chemin, on crée un chemin textuel qui comprend à la fois l'information de contenu et l'information de structure

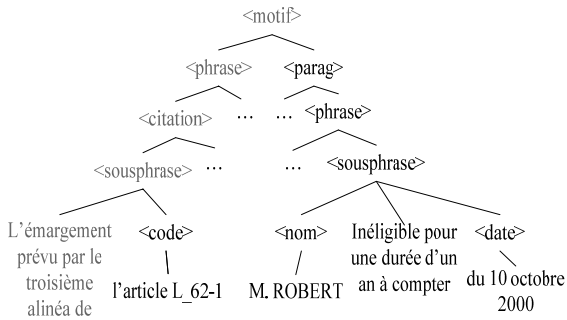


Figure 1. Un morceau d'un document du Conseil Constitutionnel français en XML

Descripteurs	Exemples
chemin complet	\motif\phrase\citation\sousphrase\
|chemin enraciné|=3	\motif\phrase\citation\
|chemin feuillu|=2	\ciataion\sousphrase\
|chemin enraciné&feuillu|=2	\motif\phrase\ \ciataion\sousphrase\
chemin textuel complet	\motif\phrase\citation\sousphrase\alinéa
|chemin textuel enraciné|=3	\motif\phrase\citation\alinéa
|chemin textuel feuillu|=2	\ciataion\sousphrase\alinéa
|chemin textuel enraciné & feuillu|=2	\motif\phrase\alinéa \ciataion\sousphrase\alinéa

Figure 2. Descripteurs structurels du chemin '\motif\phrase\citation\sousphrase\'

Un document peut être représenté par un ensemble de composants de même type (par exemple, les mots, les chemins complets, les chemins textuels enracinés), ou de types différents (par exmple, le mixte de chemin enraciné et de chemin feuillu). Le descripteur de l'approche structurel (le chemin ou le chemin textuel) peut être représenté, comme le descripteur de l'approche de contenu (le mot), dans un vecteur dont chaque dimension correspond à un descripteur. Donc on peut adopter

directement le modèle vectoriel de Salton. Selon l'approche statistique, le nombre d'occurrence peut être un facteur pour calculer l'importance d'un descripteur. Nous utilisons le coefficient TF-IDF pour mesurer son importance. La fréquence d'un descripteur *t* dans un document *d* est définie par l'équation suivante:

$$TF_{t,d} = \frac{n_{t,d}}{\sum_k n_{k,d}}$$

$n_{t,d}$ est le nombre d'occurrence d'un descripteur *t* dans un document *d* ;
$\sum_k n_{k,d}$ est le nombre d'occurrence de tous les descripteurs dans un document *d*.

$$IDF_t = \log(\frac{N}{df_t})$$

N est le nombre total de documents dans la collection ;
df_t est le nombre de documents contenant un descripteur *t*.

4 Expérimentations

4.1 Corpus du Conseil Constitutionnel français

Notre corpus est extrait de la base de documents du Conseil Constitutionnel français qui collecte toutes les publications du Conseil. 2204 documents au sujet de l'élection parlementaire entre 1958 et 2003 ont été sélectionnés. Chaque document décrit des jugements du Conseil sur le contentieux électoral en trois domaines: l'éligibilité de la candidature, le déroulement des opérations et le respect des règles de financement des campagnes. Parmi eux, le contrôle de financement de campagnes couvre une grande partie (53,9%) de la collection. Un document se compose une description des analyses des moyens invoqués, une indication des principes applicables et une réponse à la requête. Deux réponses sont majoritaire : l'inéligibilité de la candidature (49,6% de jugements) et le rejet de la saisie (47,2% de jugements). Donc, nous avons attribué manuellement à chaque document deux types d'étiquette de classes : un sur le sujet du contentieux (« financement » ou « autre »); un autre sur la décision rendue à répondre à la requête (« inéligibilité », « rejet » et « autre »). La structuration de tout le document respecte strictement une règle de rédaction. Autrement dit, les structures de l'ensemble de documents sont homogènes. La figure 1 montre un exemple de la structure du document en XML.

4.2 Prétraitement

Le prétraitement du document consiste à sélectionner les descripteurs pertinents pour la catégorisation. La catégorisation s'appuie sur la comparaison de similarité entre les documents. Plus les documents apportent des descripteurs communs, plus similaires ils sont. Cependant, les descripteurs non contributifs pour la comparaison doivent être éliminés. Par exemple, pour le descripteur de contenu, les mots non significatifs (« le », « de », etc.) sont enregistrés dans une liste (*stoplist*) et sont enlevés avec les chiffres. Les mots sont rendus à leurs formats canoniques en appliquant l'algorithme de *Porter Stemming* [xi] pour réduire le bruit. Les descripteurs couvrant seulement au-delà de 80% des documents dans la collection, et ceux qui se présentent dans quelques documents particuliers (en pourcentage inférieur à 0,5%), sont considérés peu contributifs pour la comparaison de similarité des documents et sont retirés. Avec l'algorithme de prétraitement, 11 types de descripteur sont crées. Chacun est modélisé par une matrice construite de la même façon. Ces matrices sont envoyées à un algorithme de catégorisation hiérarchique.

Algorithme 1: Algorithme de prétraitement

Input: Collection de documents XML : *C*,
Liste de mots vides : *Stoplist*,
Seuils de filtrage : *haut* = 80% et *bas* = 0,5%
Output: 11 matrices correspondent à 11 représentations

```
1 :  Index[mot] = BuildIndex(C)
2 :  Index[mot] = Supprimer(Index[mot], Stoplist)
3 :  for each mot m do
4 :    if m n'est pas un chiffre then
5 :      Index[mot].PorterStemming(m)
6 :    end if
7 :  end for
8 :  MotDescripteur [ ] = MotDescripteurCreation (C, Index, haut, bas)
9 :  for each chemin ch et longueur de chemin chl do
10 :   CheminDescripteur.ch[ ] = CheminCreation (C, ch, chl, haut, bas)
11 : end for
12 : for each chemin textuel cht et longueur de chemin textuel chtl do
13 :   CheminTextuelDescripteur.ct[ ] =
           CheminTextuelCreation(C, DescripteurMot, cht, chtl, haut, bas)
14 : end for
15 : for each descripteur d do
16 :   return MatriceCreation (d, C)
17 : end for
```

4.3 Méthode de catégorisation

Un algorithme de partition hiérarchique agglomératif proposé par l'outil CLUTO [xii] est utilisé. Cette méthode traite la catégorisation comme un

processus d'optimisation dont l'objectif est de maximiser une fonction de critères particuliers définies localement sur l'ensemble des solutions de catégorisation [xiii]. Une partition de K-parcours est obtenue via bi-sections répétées. Une bi-section consiste à une application récursive de la procédure d'optimisation de catégorisation de 2-parcours. Voici la fonction de critère utilisée

$$\sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} sim(u,v)} \quad \text{où} \quad sim(u,v) = \cos\theta = \frac{u \cdot v}{\|u\|\|v\|}$$

u et *v* sont deux vecteurs documentaires. Le processus d'optimisation doit maximiser cette fonction. La similarité entre deux vecteurs documentaires est mesurée par le cosinus.

Algorithme 2: Algorithme de catégorisation

Input: Matrice *M*,
Nombre de catégorie souhaitée : *k*,
Fonction de critère : *f*

Output: *k* catégoires : C[ ]

```
1 :  Ctemp = CategorisationInitiale(M)
2 :  i = 0
3 :  while i < k do
4 :    {C0, C1} = PartionEnDeuxCategories(Ctemp, f)
5 :    Ctemp = ChoisirUneCategorie(C0, C1)
6 :    if Ctemp == C0 then
7 :      C[i] = C1
8 :    else
9 :      C[i] = C0
10 :  end while
11 :  return C[ ]
```

4.4 Résultats

La catégorisation est évaluée quantitativement par l'entropie et la pureté ([xiv]). Deux évaluations sont proposées sur le corpus : l'évaluation de catégorisation thématique est une approche traditionnelle ciblée à la recherche d'information ; alors que l'évaluation sur la décision rendue s'adresse à extraire des catégories décisionnelles.
Pour l'évaluation thématique, deux séries sont lancées en différenciant le nombre de catégories. Pour le descripteur « mot », la qualité mesurée par deux coefficients augmente nettement : 26,9% pour l'entropie et 9,3% pour la pureté avec l'augmentation du nombre de catégories. La même tendance est trouvée également pour certains descripteurs. Un constat intéressant est que la qualité de catégorisation pour le descripteur « chemin feuillu » et les descripteurs « chemin textuel feuillu », « chemin textuel enraciné et feuillu », et « mixte de balise seule et mot » restent constant malgré une augmentation du nombre de catégories. L'approche

du chemin textuel permet une meilleure qualité que les deux autres approches quand le nombre de catégories est fixé à 2.

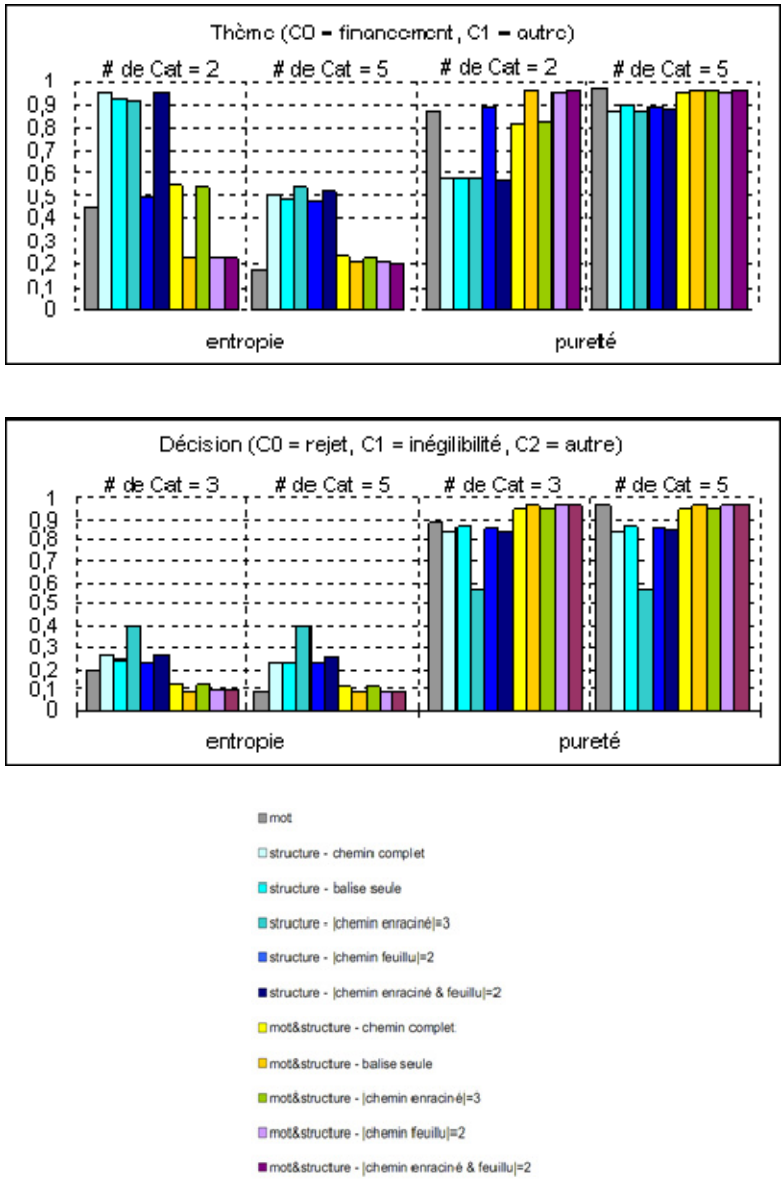


Figure 3. Résultats de l'évaluation thématique et de l'évaluation décisionnelle sur 11 descripteurs de trois approches

Au point de vue traditionnel pour une catégorisation thématique, deux documents proches partagent une partie de mots communs significatifs.

Le vocabulaire du document joue un rôle important dans ce cas. La structure du document n'apporte pas de vocabulaire approprié au thème du document. Pour cela, la qualité de l'approche de la structure seule reste limitée. Cependant, on observe que la structure offre une stabilité considérable. En combinant le mot et la structure, la qualité de catégorisation est nettement augmentée. La qualité brillante du descripteur « mixte de balise seule et mot » implique l'importance du vocabulaire de structure. Les productions de descripteur « chemin (textuel) feuillu » et de « chemin (textuel) enraciné et feuillu » montrent l'importance de l'information hiérarchique de la structure. Le descripteur « chemin (textuel) feuillu » prenant une sous-structure reposant sur les éléments feuilles est plus intéressante que le descripteur « chemin (textuel) enraciné » basé sur l'élément racine et ainsi que le descripteur « chemin (textuel) complet » reflétant la hiérarchie complète.
En ce qui concerne l'évaluation de la décision rendue, on constate que, à l'exception du descripteur « mot », le nombre de catégorie influence peu la qualité de catégorisation. Mis à part l'exception du descripteur « chemin (textuel) enraciné », tous les descripteurs produisent une qualité élevée : la valeur de l'entropie est inférieure à 0,3, et la valeur de la pureté est supérieure à 0,8. Parmi eux, l'approche du chemin textuel produit les meilleurs scores. L'approche du mot mène à une qualité élevée par rapport à l'approche de la structure seule. Même si cette dernière peut conduire à une qualité de catégorisation satisfait. Mais une combinaison du mot *et* de la structure offrent une qualité bonne et stable. Au contraire des résultats de l'évaluation thématique, tous les descripteurs de chemins textuels produisent de bons scores. Ces observations impliquent que le savoir-faire mené par la structuration du document est liée à la décision rendue.
En comparant deux évaluations effectuées, on constate que les résultats de catégorisation décisionnelle sont plus stables quand le nombre de catégories retenues augmente. Nous conclussions que les catégories retenues sont plutôt une partition des jugements qu'une partition thématique.

5 Conclusion

Dans cet article, nous proposons une méthode pour découvrir la connaissance et le savoir-faire du patrimoine de documents juridiques semi-structuré. Les résultats montrent que l'importance de l'information hiérarchique de la structure du document pour stabiliser la partition thématique de documents juridiques et pour l'extraction d'information décisionnelle par catégorie. En comparant avec le modèle classique « sac de mots », on remarque que la représentation tenue à la fois du contenu et

des sous-structures hiérarchiques du document améliore généralement ici la qualité de la tâche de prétraitement de documents juridiques. Et l'amélioration se trouve sous condition que la structuration de tous les documents soit homogène.
Malgré une approche structurelle testée sur un corpus homogène à la structure, notre méthode doit permettre de modéliser les documents à la structuration hétérogène qui est le cas pour la base documentaire hétérogène ou les documents en Web. Ceci doit être développé et testé dans nos futurs travaux.

6 Références bibliographiques

[i] Flesca S., Manco G., Masciari E., Pontieri L., Pugliese A. Detecting Structural Similarities between XML Documents. In *Proceedings of the International Workshop on the Web and Databases (WebDB)*. 2002

[ii] Nierman A., Jagadish H. V. Evaluating Structural Similarity in XML Documents. In *Proceedings of the Fifth International Workshop on the Web and Databases (WebDB 2002)*, Madison, Wisconsin, USA. 2002

[iii] Francesca F. D., Gordano G., Ortale R., Tagarelli A. Distance-based Clustering of XML Documents. In *L. De Raedt et T. Washio (Eds.), MGTS-2003 : Proceedings of the First International Workshop on Mining Graphs, Trees and Sequences*, pp. 75–78. 2003

[iv] Joshi S., Agrawal N., Krishnapuram R., Negi S. A bag of paths model for measuring structural similarity in Web documents. *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003

[v] Leung H., Chung F., Chan S.C.F., Luk R. XML Document Clustering Using Common XPath. *In WIRI'05 Proceedings of the 2005 International Workshop on Challenges*. 2005

[vi] Vercoustre A.M., Fegas M., Gul S., Lechevallier Y. A Flexible Structured-based Representation for XML Document Mining. *In: Workshop of the INitiative for the Evaluation of XML Retrieval (2005)*. page 443-457. 2005

[vii] Salton G. *Automatic Text Processing*. Addison-Wesley Publishing Company. 1988

[viii] Yang J., Chen X. A semi-structured document model for text mining. *J. Comput. Sci. Technol*. 17(5), 603–610. 2002

[ix] Yao J. et Zerida N. Rare patterns to improve path-based clustering of Wikipedia articles, In *XML data mining challenge INEX'07*, Dagstuhl, Germany, 2007

[x] Yao J. et Zreik K. La question de la structure dans la catégorisation de documents XML hétérogènes. In *Systèmes Intelligents*, Edited by Mustapha Bellafkih, Mohammed Ramdani, Khaldoun Zreik. ISBN 978-2-909285-53-3, Ed. Europia, Juin 2009

[xi] Porter M.F. An algorithm for suffix stripping. *Program*, 14(3) pp 130−137. 1980

[xii] Karypis G. CLUTO: A Software Package for Clustering High-Dimensional Data Sets. *University of Minnesota, Dept. of Computer Science*, Minneapolis, MN, Nov. 2003. Release

[xiii] Zhao Y. and Karypis G. Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, Vol. 10, No. 2, pp. 141 - 168. 2005

[xiv] Zhao Y. and Karypis G. Criterion functions for document clustering: Experiments and analysis. *Technical Report TR #01–40, Department of Computer Science, University of Minnesota*, Minneapolis, MN, 2001.

Design d'interfaces homme(s)-logiciel(s) : pouvoir transmettre et savoir transmettre

Laurence NOËL, Ghislaine AZÉMARD

Laboratoire Paragraphe - Université paris 8 – Saint Denis(1)

Mots-clés : design, interface, transmission, communication, formulation

Keywords: design, interface, transmission, communication, formulation

Résumé : L'évolution des moyens de communication a non seulement un impact sur la nature de ce qui peut être transmis mais elle implique aussi une redistribution des rôles au sein de la chaine de médiation. Dans cet article, nous analysons le rapport existant entre activité de conception et formulation de l'information. Nous décrivons le rôle joué par le designer dans le cadre de la conception d'un système d'interfaces et nous analysons la façon dont les technologies du web et du numérique étendent le champ des ressources que le designer analyse et peut exploiter.

Abstract : The evolution of our communication means has an impact on the nature of what can be transmitted, it also implies a redistribution of the roles within the mediation workflow. In this article, we analyze the relation between design activity and information transmission. We describe the part played by the designer in the contruction of a system of interfaces and web analyze the way web and digital technologies have extended the field of ressources a designer can analyze and exploit.

1 Introduction

Les développements liés aux technologies du numérique et de l'internet ont fait évoluer l'ensemble des ressources que nous pouvons utiliser pour partager nos savoirs, nos idées, nos expériences. Ces changements entraînent parallèlement des modifications au niveau des compétences et des savoirs requis pour pouvoir exploiter pleinement ces ressources lors du processus de médiation ainsi qu'au niveau des acteurs intervenant pendant ce processus. Si nous parlons de processus de médiation, c'est que la transmission d'information implique qu'il y ait création d'un artefact médiateur. L'interface correspond à la partie émergente des

complexes artificiels que les technologies du numérique et du web nous permettent de créer. Une interface est un espace qui est interprété de manière double : elle est constituée d'objets hypermédias perçus et manipulables par l'utilisateur, mais ces objets hypermédias correspondent eux-mêmes à des données qui ont été interprétées par une application logicielle. Dans cet article, nous proposons tout d'abord d'analyser le processus de conception de manière à pouvoir décrire la formation d'objets et la formulation d'un énoncé selon un même paradigme, la conception d'objets hypermédias relevant tout autant de l'un de que de l'autre. Nous analysons ensuite le rôle joué par le designer d'interfaces, son intervention étant selon nous liée au fait qu'un système d'interfaces est une construction complexe, constituée d'éléments qui entretiennent des relations d'interdépendance, et qu'il faut apprendre à exploiter.

2 L'activité de conception : de la formation d'objets à la formulation d'énoncés

Mots et choses, artefacts sémiotiques et artefacts usuels, signes et objets : ce qui est finalement le plus difficile à faire, ce n'est pas tant de trouver les points communs mais l'opposition pertinente qui permet de différencier les deux ensembles d'entités auxquels ces termes font référence puisque tout objet fait signe et que tout signe a une dimension matérielle, et que l'on utilise les artefacts sémiotiques, tout autant que l'on est amené à interpréter les artefacts usuels. Parallèlement et inversement, la formation d'objets et la formulation d'un énoncé restent des processus qui sont analysés selon des modèles différents, alors qu'au regard d'un certain niveau d'abstraction, ils relèvent d'une activité de conception similaire [1]. Dans cette partie, nous proposons d'analyser l'activité conceptuelle et de monter que c'est à travers l'analyse de ce processus commun que l'on peut établir une opposition pertinente entre nos deux ensembles de départ. Pour cela, nous proposons tout d'abord de revenir sur les notions de forme et de de fonction, qui sont pour nous, deux notions au coeur l'activité de conception, puis nous décrivons ce processus et le rôle que le designer, en tant que praticien professionnel, peut venir jouer au cours de ce processus.

2.1 Forme et fonction

Si la notion de forme est souvent employée dans les réflexions sur le design, c'est parce qu'à travers elle, on peut non seulement faire référence à l'aspect d'une entité, mais aussi à l'ensemble des traits caractéristiques qui nous permettent de la reconnaître. Selon Merleau-Ponty [2], « la forme des objets n'en est pas le contour géométrique : elle a un certain

rapport avec leur nature propre et parle à tous nos sens en même temps qu'à la vue. La forme d'un pli dans un tissu de lin ou de coton nous fait voir la souplesse ou la sécheresse de la fibre, la froideur ou la tiédeur du tissu. » En considérant la forme d'une entité (ou ses propriétés formelles), on s'intéresse à ce qui fait son unité, à ce qui rend son être concret (cet objet est de forme creuse, allongée, transparente, lisse), alors qu'en considérant ses fonctions (ou ses propriétés fonctionnelles), on se positionne sur l'axe des relations, du faire et on cette entité considère en sa qualité de médium, d'entité intermédiaire permettant de réaliser une action (je peux utiliser cet objet pour planter un clou). Examiner un élément au regard de ses fonctions, c'est donc analyser les rapports d'interaction que cet élément peut avoir avec son environnement et les rapports d'interaction qu'il peut nous permettre d'établir avec cet environnement[1].

Forme et fonction sont donc à considérer comme étant deux perspectives distinctes qui peuvent être communément adoptées pour caractériser une même réalité et qui entretiennent entre elles un rapport d'interdépendance : le réel est un tissu complexe d'interactions entre entités, et ce sont ces relations entre entités qui déterminent leur état, leur forme, de manière dynamique, tout comme la forme d'un élément à un instant *t* définit le champ potentiel des interactions qu'il peut nous permettre d'établir avec notre environnement.

2.2 L'activité de conception : traduction et transformation

Dans le cadre de la théorie de l'activité [4], la notion d'artefact médiateur permet d'englober aussi bien les « outils physiques » que les « outils psychologiques ». Selon cette approche, la conception d'artefacts est lié à un processus d'extériorisation des connaissances, mais cette expression ne nous paraît pas appropriée puisque c'est justement parce que nous ne pouvons pas extérioriser nos schémas d'activité cognitive ou intérioriser directement ceux d'autrui que nous devons utiliser et transformer des ressources qui sont elles communément perceptibles.

Pour caractériser l'activité de conception, nous préférons donc parler d'activité traductrice puisque l'entité est conçue au regard de l'ensemble des fonctions qu'elle doit pouvoir remplir et en prenant en compte un ensemble défini de ressources. Si l'activité de conception peut être considérée comme une activité de traduction du point de vue du sujet effecteur, elle implique par ailleurs un processus de transformation du point de vue des ressources utilisées pour traduire un schéma d'idées.

1La notion de fonction peut paraître proche du concept d'affordance de Gibson [3], elle renvoie aux « propriétés actionnables » de l'objet. Mais Gibson postule que ces affordances sont propres à l'objet : elles existeraient indépendamment du fait qu'elles soient perçues ou non, alors que selon notre perspective, les fonctions sont attribuées par le sujet interprétant et dépendent de la relation d'interaction entre l'interprétant-utilisateur et l'objet.

Cette transformation peut aller du simple ré-agencement d'un ensemble formé par des ressources, à l'hybridation ou bien encore à la fusion partielle ou globale de certaines ressources entre elles. Deux types de ressources sont alors à distinguer : les ressources transformables qui interviennent en tant que composant de la future entité et les ressources transformatrices qui permettent d'agir sur les ressources transformables. Pour créer une statue, un sculpteur va par exemple utiliser un outil de taille (ressource transformatrice) et du bronze (ressource transformable) pour donner forme à un objet qui est la traduction d'un schéma d'idées. De la même manière, lorsque nous discutons, nous utilisons notre appareil phonatoire (ressource transformatrice) pour agir sur l'air (ressource transformable) afin de produire un flux organisé de formes sonores dont l'interprétation dépend de la connaissance d'un code qui est lui-même une construction sociale (ressource transformable). Lorsque nous communiquons par écrit, le stylo (ressource transformatrice) nous permet d'agir sur un flux d'encre et un support papier (ressources transformables) pour produire un document écrit qui est le résultat global du processus de traduction/transformation et qui est interprété dans son ensemble tout autant qu'au regard des différents niveaux d'interprétation que ces composants offrent.

Les fonctions n'étant pas de l'ordre du tangible, la forme de l'entité finale est ce qui est généralement perçu comme étant le résultat du processus de conception, mais en spécifiant cette entité, on s'intéresse tout autant au contexte de la forme qu'à la forme [5, 6] au milieu associé qu'à l'objet technique [7]. Un objet peut être détourné dans son usage[2], et il peut avoir des propriétés autres que celles qui ont été initialement prévues, mais l'objectif, du point de vue du concepteur, est que l'artefact final respecte *au moins* l'ensemble des fonctions qui doivent pouvoir lui être attribuées : il est nécessaire qu'il puisse être interprété et utilisé de la façon initialement prévue, ce qui n'empêche pas que l'utilisateur puisse l'utiliser autrement. De la même manière qu'un phonème peut avoir plusieurs allophones, on peut considérer que plusieurs réalisations allomorphiques peuvent répondre aux objectifs fixés, les différences qui séparent chacune de ces réalisations étant alors considérées comme non pertinentes par rapport à un système déterminé de valeurs.

Si ces spécifications sont nombreuses, définir une entité finale qui les respecte peut représenter un défi en soi. On peut choisir d'utiliser une ressource parce qu'elle possède une certaine propriété mais il faut aussi considérer ses autres propriétés et s'assurer qu'elles ne sont pas conflictuelles avec les spécifications initiales : si on veut créer un objet brillant mais pas cher, on ne pourra pas utiliser l'or comme ressource

2 Perriault [8] s'est ainsi intéressé aux pratiques déviantes, qui correspondent à une volonté de la part de l'usager de détourner un instrument de son usage initial. Il distingue notamment les usages prescrits des usages effectifs et qualifie d'usages conformes ceux qui correspondent aux prescriptions du concepteur.

transformable puisque cette matière répond à l'un des objectifs de conception mais est en contradiction avec l'autre. Pour Lawson [9], c'est le développement des technologies et leur complexification croissante qui a conduit à faire émerger la profession de designer. Parler de complexe technologique, c'est souligner que les ressources artificielles dont nous disposons ne sont pas indépendantes les unes des autres mais qu'elles appartiennent à des ensembles techniques évolutifs, constitués de composants matériellement distincts mais fonctionnellement liés entre eux, ce qui accroît le nombre des rapports d'interdépendance que le concepteur doit analyser. Si, selon les dires de Simons [1], « quiconque imagine quelque disposition visant à changer une situation existante en une situation préférée est concepteur », tout le monde n'a cependant pas pour profession d'être designer et la question est alors de savoir ce qui permet de caractériser le savoir-faire de ce dernier et de délimiter son champ d'intervention.

2.3 Le savoir-faire du designer

Selon Krippendorff [10], le designer a en fait pour mission de définir une entité réalisable - ou un prototype - au regard des spécifications qui lui ont été fournies. Si l'idée initiale du projet n'est pas forcément la sienne, c'est cependant lui qui doit être capable de développer et d'enrichir cette idée, grâce au savoir qu'il possède concernant les ressources qui peuvent être utilisées et les processus de transformation qu'elles peuvent subir et en prenant en compte les contraintes et des demandes qui ont été formulées. C'est lui qui doit être en mesure de faire la part entre ce que le commanditaire souhaite et imagine, et ce qui est concrètement réalisable : créer un objet qui réponde à l'ensemble des exigences initiales d'un commanditaire n'est pas toujours possible puisque les ressources utilisées ont des propriétés multiples et qu'elles peuvent entretenir des relations conflictuelles. Le rôle du designer est alors de voir avec le commanditaire comment les spécifications fonctionnelles peuvent être hiérarchisées : il a alors pour mission d'analyser quelles fonctions doivent prioritairement pouvoir être attribuées à un objet de manière à ce que celui-ci puisse effectivement être réalisé en respectant une cetaine hiérarchie fonctionnelle. Le designer est donc celui qui définit ce qui peut être réellement créé et qui définit les limites du champ des possibles, tout en s'assurant que l'univers du réalisable a bien été complètement exploré et pris en compte.

Si la notion d'esthétisme est souvent associée au design dans les esprits, c'est qu'un designer porte effectivement une attention particulière à l'aspect extérieur des objets, non seulement parce que « la laideur se vend mal » [11], mais aussi parce que pour qu'un objet puisse être utilisé, il faut aussi que sa forme permette à l'utilisateur de comprendre comment cet objet peut être utilisé. Tout objet faisant l'objet d'un processus

d'interprétation, le designer est nécessairement amené à analyser la fonction informative que l'interprétant-utilisateur peut potentiellement attribuer à un objet. Dans le cas où l'entité à créer a pour fonction principale de permettre à une personne d'agir sur un autre objet du monde extérieur, la fonction informative sera donc prise en compte mais sera de traitée de manière secondaire, au regard de l'ensemble hiérarchisée des spécifications à respecter. Les artefacts informationnels sont alors ceux, parmi l'ensemble plus large des artefacts médiateurs, qui ont une fonction informative pour fonction principale, c'est à dire ceux qui sont conçus pour agir sur le champ cognitif d'autrui de manière à ce que les schémas cognitifs activés permettent d'évoquer un même référent entre les deux communicants.

3 Concevoir la partie émergente d'un complexe artificiel

Si la conception d'un système d'information peut nécessiter l'intervention d'un designer d'interface, c'est que l'interface est en elle-même une construction dynamique et correspond à la partie émergente d'un complexe artificiel. Elle n'est réalisée que si un ensemble de composants sont communément activés et peuvent être utilisés de manière combinée. Le terme d'interface utilisateur permet de faire référence à l'ensemble des artefacts créés pour permettre à l'homme de communiquer et d'interagir avec une application logicielle (clavier, écran, souris peuvent par exemple faire partie de cet ensemble). Nous nous intéressons ici plus particulièrement à la partie numérique de l'interfaçage, à sa partie programmable. Un iceberg possède une partie visible et une partie immergée mais il n'est finalement constitué que de glace et il est en de même pour les applications logicielles qui ont une interface utilisateur : elle sont entièrement constituées de données, mais une partie de ces données est invisible pour l'être humain tandis que l'autre partie vise justement à rendre l'application perceptible, appréhendable par ce dernier. Les technologies du numérique et du web étendent notre « pouvoir transmettre » : le rôle du designer est alors de savoir exploiter ce champ des possibles de manière à faciliter l'appropriation de l'information par l'utilisateur.

3.1 Extension des formes possibles de représentation

Si un être humain est capable de percevoir des couleurs, des textures, des sons et de multiples autres types de phénomènes, ses capacités corporelles ne lui permettent pas de les reproduire directement. Communiquer par l'intermédiaire d'objets hypermédias devient donc un

moyen de ré-établir un certain équilibre entre l'ensemble des formes que nous pouvons percevoir et l'ensemble des formes que nous pouvons produire pour transmettre une idée. L'activité perceptive relève d'un phénomène d'intégration de sensations plurielles, et cette activité perceptive laisse une trace multiple [12, 13] dans notre mémoire. En nous permettant de reproduire des formes du réel, et donc de nous exprimer à travers des formes qui vont permettre d'activer directement le schéma d'activité correspondant à l'ensemble des sensations perçues, et non à ceux qui correspondent aux concepts que nous leur avons associé suite à notre apprentissage du système linguistique, les technologies du numérique invitent donc à opérer à décentrement par rapport à ce modèle de l'expression linguistique qui place la notion de code en son coeur : les formes d'expression dont nous disposons s'étendent alors le long d'un continuum défini par le degré de motivation existant entre la forme construite pour représenter un référent et ce référent (l'apprentissage d'un code étant nécessaire lorsque la forme est immotivée).

Si l'expression linguistique garde un place importante dans la communication hypermédia, c'est qu'elle nous permet de tout décrire, que ce soit des événements, un raisonnement, ou bien encore une expérience. Celle-ci n'est cependant pas toujours le mode d'expression qui est le plus adapté pour transmettre un schéma d'idées : plutôt que de décrire un itinéraire, on peut par exemple proposer un mode de représentation cartographique qui permettra à l'interprétant de mieux appréhender les relation spatiales entre les objets. Décrire, raconter, montrer, faire écouter : chaque mode de représentation permet de faire découvrir un objet selon une certaine perspective. Formuler par l'hypermédia, c'est alors apprendre à utiliser différentes modalités d'expression dans leur complémentarité, et savoir choisir quel mode de représentation sera le plus adapté pour tel usage.

La question de la représentation ne se pose cependant pas seulement du point de vue des contenus, mais aussi du point de vue des contenants. Dans le cas d'un document papier, le support d'inscription est un bloc, et les possibilités qu'une personne a d'interagir avec le contenu sont définies par ce support. La forme de ce support est pré-définie, elle n'est que peu modifiable par l'énonciateur, et l'ensemble des formes d'inscription contenues par le support font donc partie d'un seul et même contenant. En utilisant les technologies du numérique, on a la possibilité de définir la forme des contenants tout autant que la forme des contenus, et il est d'ailleurs nécessaire de représenter les possibilités que l'utilisateur à d'interagir avec ce contenu. Le designer a alors pour rôle de spécifier une structure dynamique multi-componentielle, et pour chacun des composants de cette structure, la question du mode de représentation se pose selon des critères multiples : elle peut ainsi varier en fonction du type de média utilisé, du rapport entre la quantité de données à afficher et

l'espace qui est alloué pour afficher ces données au sein de la page, de l'usage qui est prévu pour ces données, etc.

3.2 Extension des opérations logiques que la machine peut effectuer à partir des données

Lorque le designer conçoit des interfaces, il est amené à considérer de manière conjointe deux types d'interprétants différents - l'interprétant logiciel et l'interprétant humain. Les codes et langages qui sont utilisés par l'un et par l'autre sont cependant différents : pour étendre l'espace d'intercompréhension entre les deux et donc améliorer les échanges qui peuvent être effectués au niveau de l'interface, une des possibilités est d'essayer de faire en sorte que l'interprétant logiciel soit en mesure de comprendre le “sens des données.” Parler du sens des données, c'est en fait ici faire référence à la signification que l'être humain attribue à une forme d'inscription qui est considérée au regard des relations qu'elle entretient avec d'autres formes au sein d'un système de représentation. Pour que l'interprétant logiciel puisse lui-même réaliser des opérations logiques en se basant sur le sens que l'homme attribue aux données, il faut donc que les relations sémantiques que ce dernier est capable d'établir entre ces données soient elles-mêmes formulées. C'est en se basant sur ces schémas de relation entre données que les opérations de traitement de l'interprétant logiciel peuvent être programmées pour se rapprocher du type d'opérations cognitives que l'être humain effectue à partir de ses propres connaissances : la machine devient alors capable non pas de comprendre les données, mais de suivre un fil de raisonnement qui se rapproche de celui suivi par l'être humain.

3.3 Extension de l'espace interactionnel

Les technologies du web amènent à prendre en compte non plus seulement le niveau individuel, mais aussi le niveau collectif. Plutôt que des interfaces homme-machine, ce sont des interfaces homme(s)-logiciel(s) que le designer web doit concevoir. Si le web 2.0 a été caractérisé de « web social », c'est en partie parce qu'il a été caractérisé par des applications, telles que les blogs et les plateformes de collaborative, qui ont été réalisées de manière à faciliter la participation des utilisateurs et la construction de ce web social. Ces applications ont été définies en intégrant, durant le processus de coneption, la vision des communautés sociales qu'elles permettent de tisser (Hendler et. Al, 2008). Lors de la conception des interfaces, il faut donc à la fois prendre en compte l'utilisateur en tant qu'individu mais aussi en tant que membre d'une communauté. Il faut penser non seulement à la façon dont cet

individu peut interagir avec le système, mais aussi aux actions qu'un groupe d'utilisateurs peut réaliser de manière collaborative. De manière parallèle, il faut penser la conception d'une ressource web, non pas de manière isolée, mais en prenant en compte le fait que celui-ci peut être connectée à d'autres applications et que les objets qu'elle contient peuvent être partagés par plusieurs applications.

4 Conclusion

Dans cet article, nous avons décrit le savoir-faire du designer comme correspondant à un savoir-traduire et à un savoir-transformer. Ce que le designer conçoit, c'est une forme fonctionnelle, c'est à dire une entité dont la formation dépend du champ d'action qu'on souhaite lui voir attribuer. Selon cette approche, l'échange d'informations entre deux personnes passe par la transformation de ressources communément perceptibles et c'est au regard des propriétés de l'ensemble formé par transformation de ces ressources que l'on élabore le contenu d'un message. Pour le designer, il est tout aussi important d'adopter une perspective communicationnelle que d'avoir une attitude technologique, parce que les interfaces qu'il doit concevoir sont interprétées de manière double – par des interprétants humains et des interprétants logiciels. Lors de son analyse, le designer doit donc prendre en compte le fait que les ressources dont il dispose ne sont pas indépendantes les unes des autres, mais qu'elles appartiennent à des ensembles techniques évolutifs, constitués de composants matériellement distincts mais fonctionnellement liés entre eux : il ne peut ré-assembler ces ressources, les combiner entre elles qu'en comprenant les rapports d'interactions existants au sein des complexes artificiels qu'elles contribuent à former. C'est en analysant et en prenant en compte l'évolution des ressources dont nous disposons pour communiquer que l'on peut espérer traduire un schéma d'idées avec une efficience informationnelle accrue et faciliter leur appropriation par l'interprétant-utilisateur.

5 Références bibliographiques

H. Simon. The Sciences of the Artificial, MIT Press, Cambridge. 1969

M. Merleau-Ponty, Phénoménologie de la perception, Gallimard, Paris. 1945

J. Gibson, The Theory of Affordances, Perceiving, Acting and Knowing, R. Shaw and J. Bransford (eds.), pp. 67-82. 1977

V. Kaptelinin, et B. Nardi, Acting with Technology: Activity Theory and Interaction Design, MIT Press, Cambridge. 2006

C. Alexander, Notes on the Synthesis of Form, HUP, Harvard, 1964.

B. Burdek, Design : History, Theory and Practise of Product Design, Birkhaüser, Berlin. 2005

G. Simondon. Du mode d'existence des objets techniques, Aubier, Paris. 1989

J. Perriault, La logique de l'usage. Essai sur les machines à communiquer, Flammarion, Paris, 1989

B. Lawson, How Designers Think: The Design Process Demystified, Elsevier, 2006

K. Krippendorff, The Semantic Turn: a New Foundation for Design. CRC, 2005

R. Loewy, La laideur se vend mal, Gallimard, Paris, 1990

D. Hintzman, Schema abstraction in a multiple-trace memory model, Psychological Review, Vol. 93, pp. 411-428. 1986

B. Whittlesea, Preservation of specific experiences in the representation of general knowledge, Journal of Experimental Psychology: Learning, Memory, and Cognition,Vol. 13, pp. 3-17. 1987

J. Hendler et. al, Web Science: An interdisciplinary approach to understanding the World Wide Web. In Communications of the ACM , Vol. 51, No 7, pp 60-69. 2008

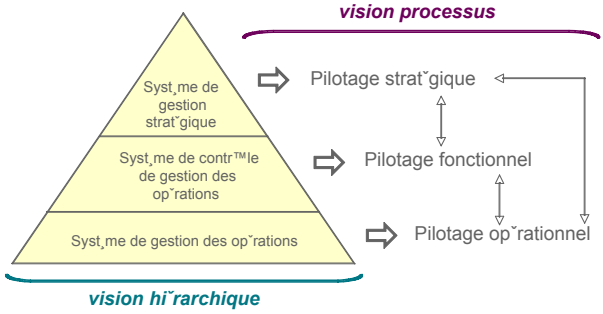
Analyse de la pratique de mise en document de l'information de pilotage en entreprise : le document au cœur de la structuration et de la restitution d'indicateurs

Samuel PARFOURU (1), Rodolphe BECK (2)

(1)EDF Recherche et Développement – Département Simulation et Traitement de l'Information pour l'Exploitation et la Production
(2)IBM Global Business Services – Secteur Energy et Utilities Consulting

1 Introduction

Le pilotage d'une entreprise est souvent associé à la vision hiérarchique et pyramidale de nos organisations. De nombreux travaux sur les systèmes d'information de pilotage montre une réalité plus complexe liée à l'évolution des mécanismes organisationnels de ces dernières années (cf. Figure 1). En effet, *« il s'agit désormais de raisonner en termes de processus car il devient difficile d'isoler clairement dans l'organisation qui (ou quoi) est à l'origine de la valeur »* [1].



La vision processus du pilotage de l'entreprise réinterroge nos organisations hiérarchiques

Ces processus tendent à définir des organisations transverses associées à une thématique, une fonction particulière. En outre, ce mode de fonctionnement se traduit par une multiplicité d'acteurs qui échangent de l'information en particulier via des tableaux de bord [1].
Le tableau de bord (cf. Figure 2) est un outil récurrent dans les entreprises et les organisations [2]. Il s'agit d'une vue synthétique qui doit permettre de rapidement évaluer une situation. Il doit proposer une visualisation efficace [3] qui en ce sens doit soutenir le raisonnement de celui qui l'utilise [2].
Le tableau de bord est par nature synthétique et ne peut être vu au singulier. Lorsque l'on aborde l'information de pilotage en entreprise, il est classique d'être confronté à une multiplicité de tableaux de bord complémentaires. Ceci conduit à les mettre en relation sous forme d'hypertextes [4].

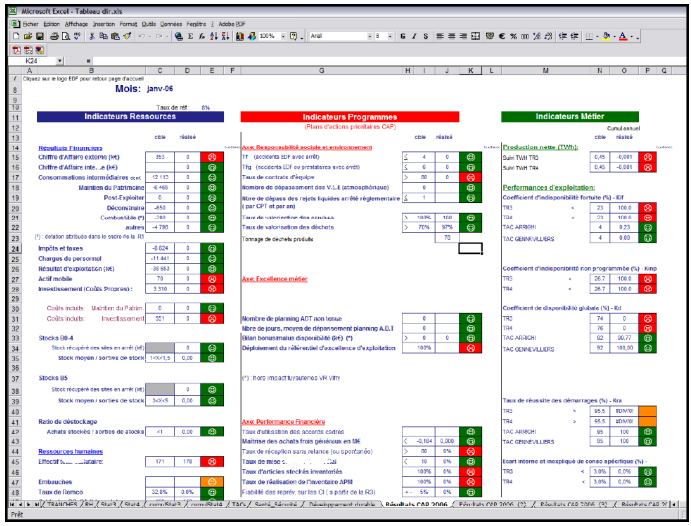


Figure 1. Un exemple de tableau de bord au sein d'Excel

2 Problématique : la construction des tableaux de bord

Au sein des organisations basées sur des processus, *« de plus en plus d'acteurs concourent ensemble au processus de création de valeur […] Ce qui signifie que le pilotage ne concerne plus des activités séparées mais des activités souvent imbriquées et interdépendantes […] ce qui a* ***un impact sur la construction et le pilotage des tableaux de bord*** *»* [1].

La construction des tableaux de bord est une activité mobilisant divers acteurs qui s'appuient sur un outillage plus ou moins complexe. En outre, nous admettons que la mise en place des organisations basées sur un mode processus ne s'est pas nécessairement accompagnée d'un déploiement d'outils adaptés. Ainsi, dans de nombreuses situations, une grande majorité des personnels se sont « auto outillés », en particulier en transformant l'informatique bureautique en réel *instrument* [5] de construction, de maintien et de restitution d'information décisionnelle.
L'utilisation des suites bureautiques se traduit par une construction empirique et peu encadrée des tableaux de bord. Chaque nouvelle situation (ex : nouvel indicateur) réinterroge le tableau de bord, sur le fond ou la forme, ou encore conduit à en produire une nouvelle instance.

3 Cadre théorique et démarche d'analyse

Notre étude porte sur *la construction des tableaux de bord.* Le cadre théorique que nous mobilisons gravite autour de la notion de document. Nous allons ainsi considérer les tableaux de bord comme une classe particulière de documents, notre analyse devant conduire à identifier le processus de mise en tableaux de bord (ie. mise en document) de l'information en considérant l'environnement informationnel dans lequel ils s'insèrent et en particulier les documents connexes.
Le document numérique est largement influencé par la création de formats numériques. Bien que le numérique implique une dématérialisation, ces formats se sont inspirés de la réalité et en particulier du « support final » (architexte [6] ou support matériel) visé. Ainsi, les documents que nous considérons dans notre étude correspondent à une « forme documentaire » de type : *contenu orienté présentation* [7]. Aussi, nous nous appuyons sur le travail interdisciplinaire synthétisé dans [8] qui identifie trois dimensions d'analyse pour le document que sont la *forme*, le *signe* et le *medium* comme prisme d'analyse.
Au-delà d'appréhender le document comme une instance stable et finalisée, nous mettons l'accent sur leur dynamique de construction. Nous considérons ainsi la notion de document pour l'action (DopA) [9] qui permet de redéfinir *« le concept de document en insistant sur la dimension collective de l'activité rédactionnelle il permet d'analyser les documents comme relevant de processus de communication pour partie différés ».*
Enfin, considérant le Tableau de bord comme un document, leur multiplicité et leur dynamique de construction nous mobilisons la notion de *dossier*, en introduisant la notion de « dossier décisionnel ». En effet, *« même si le document en lui-même est porteur d'information, ce n'est*

qu'en relation avec d'autres documents générés dans le cadre d'une activité qu'il prend véritablement tout son sens »[10].

4 Analyse d'un dossier décisionnel

Notre analyse s'est portée sur un dossier décisionnel relatif à un processus de pilotage réel désigné : Contrat Annuel de Performance (CAP). Nous avons ainsi étudié :

- La construction du CAP sous l'angle du processus social de mise en document (dimension forme et DopA),
- Les documents constituant le dossier décisionnel (document décrivant le CAP, tableaux de bord…) en s'intéressant à leur structure logique et physique en particulier (dimension forme et signe du contenu d'un dossier).

La première partie d'analyse permet de souligner que les organisations par processus introduisent nécessairement un « effet réseau » et une transversalité dans la construction. Ainsi, le CAP initialement défini à la tête de l'entreprise va être décliné par les différentes entités de l'organisation, chaque entité organisationnelle l'adaptant et l'enrichissant des éléments qu'elle souhaite considérer au-delà du socle référentiel qui lui est demandé ou « imposé » (cf. Figure 3).

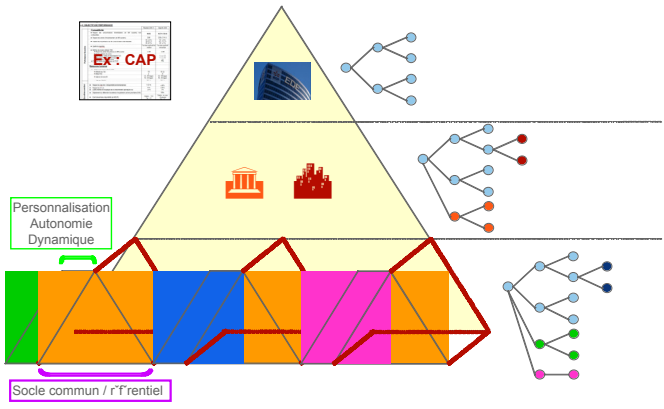
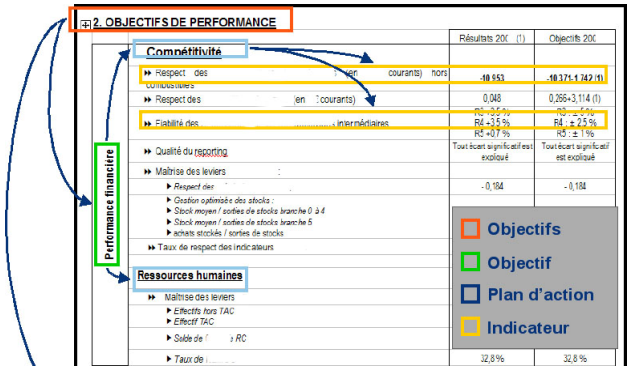


Figure 2. La structure du CAP comme résultant d'une construction collective

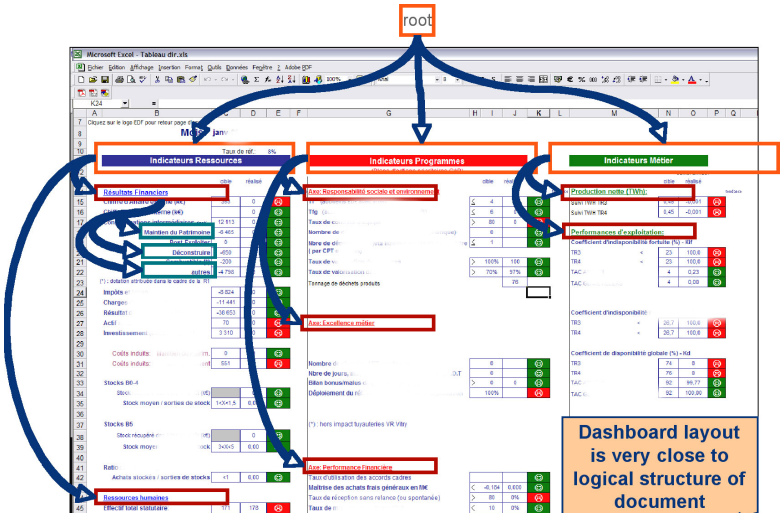
La seconde partie d'analyse souligne le caractère éminemment structuré des éléments (document, TdB) constituant le dossier décisionnel, laissant d'ailleurs apparaître des « micro formats » [11]. La structure logique (cf. Figure 4) du document décrivant le CAP ainsi que les tableaux (cf. Figure 5) - qui constituent une forme d'écriture support au traitement de

l'information [12][3][13] - permet de dissocier les « briques élémentaires » d'information contenues dans le document ainsi que leur imbrication et de là leurs dépendances.
L'analyse du tableau de bord sur le plan de son organisation logique mais également de la structure physique en particulier sur les interfaces de synthèse d'indicateurs (cf. Figure 6), permet d'aboutir au même constat avec une similarité forte entre les briques élémentaires identifiées dans le document CAP.

Structure logique du document CAP étudié



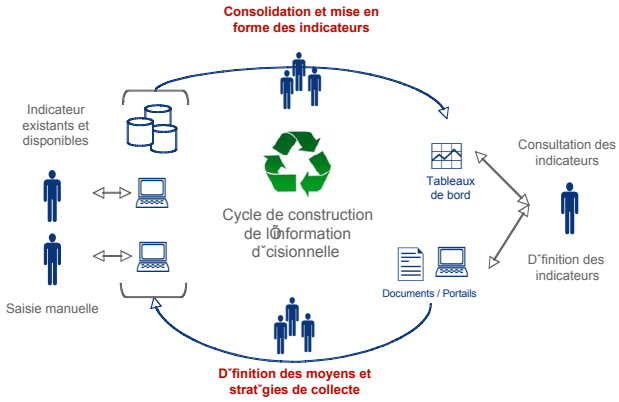
Un tableau extrait du CAP qui traduit une hiérarchie d'informations



La synthèse des indicateurs du CAP au sein d'un tableau de bord

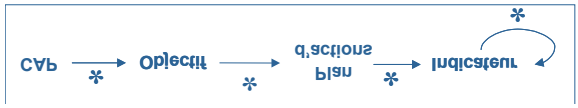
5 Proposition : Tableau de bord = *f*(document, utilisateurs)

Notre analyse nous conduit à considérer le tableau de bord comme une « ré-éditorialisation » du document le décrivant, dans laquelle on enregistre et fait évoluer des valeurs. Dans cette tâche, des *knowledge workers* [14], jouent un rôle important (cf. Figure 7) puisqu'il faut se prémunir de considérer le document comme un simple *lego* combinable à souhait [15].



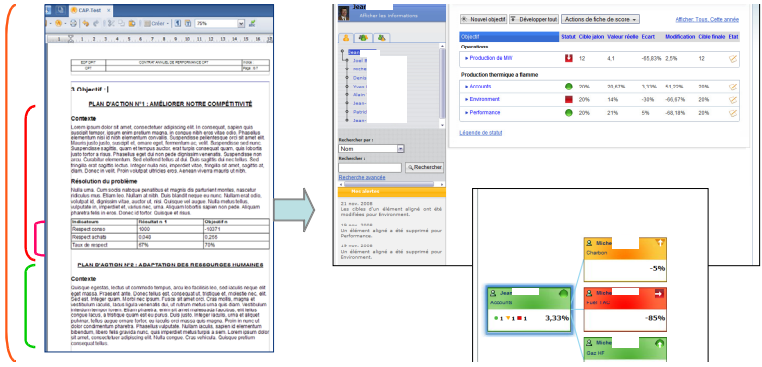
Cycle d'éditorialisation du CAP

Nous proposons ainsi une approche appréhendant le tableau de bord selon l'équation : tableau de bord = *f*(document, utilisateurs). La fonction *f* est alors conditionnée par le document, dont on peut extraire un modèle (cf. Figure 8), et l'utilisateur qu'il soit auteur, éditeur, lecteur voir « annotateur » (contributeur) sur le tableau de bord.



Meta modèle de CAP

Nous avons ainsi initié des développements en accord avec cette approche, en intégrant au sein d'une suite bureautique des mécanismes permettant d'identifier les briques élémentaires du document afin d'être en mesure de générer un tableau de bord au sein d'une plateforme robuste (cf. Figure 9).



Du document à la configuration du tableau de bord

6 Références bibliographiques

[1] C. Marsal et D. Travaillé, "Systèmes d'information de pilotage et tableaux de bord," *Encyclopédie de l'informatique et des systèmes d'information*, Vuibert, 2006, pp. 1351-1362.

[2] A. Fernandez, *Les nouveaux tableaux de bord des managers - Le projet décisionnel dans sa totalité*, Eyrolles, 2008.

[3] J. Bertin, *Sémiologie graphique - Les diagrammes - Les réseaux - Les cartes*, Editions de l'Ecole des Hautes Etudes en Sciences, 1973.

[4] S. Parfouru et R. Beck, "De l'écriture d'un document à la génération d'un hypertexte décisionnel - Analyse et instrumentation du cycle de vie d'un tableau de bord de pilotage dans le secteur de la production d'énergie."

[5] P. Béguin et P. Rabardel, “Concevoir pour les activités instrumentées,” *Interaction Homme-système, perspectives et recherches psycho ergonomiques, Revue IA*, 2000, pp. 35-54.
[6] Y. Jeanneret et C. Tardy, *Métamorphoses médiatiques, pratiques d'écriture et médiation des savoirs*, 2007.
[7] V. Lux-Pogodalla et J. Vion-Dury, “Réflexions sur la modélisation des documents,” *Information-Interaction-Intelligence*, vol. 4, 2004.
[8] R.T. Pédauque, *Le document à la lumière du numérique*, Caen: C&F Editions, 2006.
[9] M. Zacklad, “Processus de documentation dans les Documents pour l'Action (DopA) : statut des annotations et technologies de la coopération associées,” Montréal: 2004.
[10] S. Mas et LouiseGagnon-Aragon, “Pour un approfondissement de la "notion" de dossier dans la gestion de l'information organique et consignée d'une organisation,” *Archives*, vol. 35, 2003, pp. 29-48.
[11] Francesc Campoy Flores, Vincent Quint, et Irème Vatton, “Templates, Microformats and Structured Editing,” 2006.
[12] J. Goody, *La raison Graphique - la domestication de la pensée suavage*, Les Editions de Minuit, 1979.
[13] E.R. Tufte, *Visual and Statistical Thinking: Displays of Evidence for Makink Decisions*, Graphics Press, 1997.
[14] I. Alberts et S. Bertrand-Gastaldy, “Pratiques textuelles et genre en contexte de travail au gouvernement,” Fribourg, Suisse: ADBS, 2006, pp. 227-247.
[15] D. Cotte et M. Després-Lonnet, “Le document numérique comme " lego " ou La dialectique peut-elle casser des briques ?,” *Information-Interaction-Intelligence*, vol. 4, 2004.

Intégration des Langages pour la Gestion de Documents

Une Nouvelle Etape dans l'Evolution de XML

Catherine PUGIN, Rolf INGOLD

Département d'Informatique, Université de Fribourg

Mots-clés : XML, modélisation, transformation, intégration

Keywords: XML, modeling, transformation, integration

Résumé : Cet article présente une proposition d'intégration de trois langages principaux pour le traitement des documents semi-structurés : langages de balisage, de modélisation et de transformation. XSD (XML Schema) et XSLT sont les langages de modélisation et de transformation les plus connus. Ils enrichissent le langage de balisage XML car ce sont des applications XML pour XML lui-même. Toutefois, le traitement des documents manque de rigueur car ces langages manquent d'une intégration forte. Nous proposons donc trois nouveaux langages développés sur des fondements solides et des concepts clarifiés : YML (balisage), DML (modélisation) et DGL (transformation). Le développement de ces langages est basé sur l'intégration qui est le point de départ de ce projet. Cet article définit ce que signifie « intégration » et quelles sont les propriétés d'un système intégré. Finalement, nous introduisons une application concrète qui illustre quelques-uns des concepts présentés.

Abstract: This paper presents a study on the integration of three core languages dedicated to the processing of semi-structured documents. As core languages, we define markup, modeling and transformation languages. XSD and XSLT are well-known languages that enhance the XML language since they are XML applications for XML. However, the treatment of documents is not as rigorous as it could be because these languages are not well integrated. Therefore, we introduce three new languages built upon solid foundations: YML (markup), DML (modeling) and DGL (transformation). We apprehend them from the point of view of integration. The paper also defines precisely what is meant by integration in this proposal and what the requirements of an integrated system are. Finally, some of the concepts of integration are illustrated by a typical application.

1 Introduction

Publié pour la première fois en 1998 par le W3C, le langage de balisage XML [1], conçu initialement pour la représentation de documents semi-structurés, a rapidement conquis le monde scientifique et celui de la gestion. Il est aujourd'hui utilisé dans des contextes divers et par des équipes académiques et industrielles variées. Son succès est aussi dû au développement de nombreuses applications XML qui facilitent et uniformisent le traitement des documents, en particulier XSD [6] et XSLT [2] qui, en tant qu'applications XML pour le traitement du XML, enrichissent le langage.

Il a fallu une période de 40 ans aux langages de programmation traditionnels pour atteindre la maturité communément admise du paradigme orienté-objet. XML, bien que riche de ses 10 ans d'expérience et de ses cinq éditions, verra immanquablement sa spécification évoluer vers plus de maturité. Ainsi, XML n'est, à ce jour, qu'un sous-ensemble de SGML dépourvu de ses propriétés les plus obscures [5]. Les langages de modélisation et de transformation, qui, de notre point de vue, forment avec XML les langages centraux pour la gestion des documents semi-structurés, souffrent aussi d'un manque de maturité. Solutions pragmatiques et manque d'interaction entre les différents langages - dû à des intérêts commerciaux évidents - sont souvent justifiés par une rétrocompatibilité indispensable au vu de la large diffusion des documents au format XML.

Nous présentons une expérience qui vise à faire évoluer ces langages vers une nouvelle étape de maturité. La clé de voûte de ce travail est l'intégration des langages de balisage, de modélisation et de transformation. Nous souhaitons développer un système intégré qui soit auto-suffisant et qui permette de traiter des applications complexes de manière simple et naturelle, et ceci, grâce au concept d'intégration.

Les travaux existants dans le domaine de la modélisation et de la transformation de documents sont riches mais, même si certains traitent un certain type d'intégration entre les langages, aucun, à notre connaissance, n'est mené avec l'intégration comme véritable point de mire dès le début de la recherche.

Le système que nous proposons est basé sur trois langages : YML pour le balisage, DML pour la modélisation et DGL pour la génération et la transformation des documents. DML et DGL sont des applications YML et leurs instances respectives des documents YML.

Par intégration, nous entendons les propriétés suivantes pour le système : (1) chaque langage est conçu pour remplir idéalement sa tâche propre et est basé sur des concepts clairs, solides et uniquement dédiés à sa fonction ; (2) les langages sont développés en parallèle et poursuivent le même but, c'est-à-dire une gestion plus simple et plus naturelle des

documents ; (3) ils interagissent les uns avec les autres ; (4) ils bénéficient les uns des autres ; (5) finalement, ils sont soutenus par des spécifications formelles complètes qui assurent une implémentation rigoureuse et dépourvue d'erreurs.
Cet article est organisé de la manière suivante : la section 1 présente un bref état de l'art des langages de modélisation et de transformation pour les documents semi-structurés. La section 2 définit la notion d'intégration. YML et DML sont brièvement introduits dans la section 3 tandis que la section 4 développe plus spécifiquement le langage de transformation DGL. La section 5 s'étend sur l'intégration de ces langages à l'aide d'un exemple concret d'application intégrée. Finalement, la dernière section conclut l'article tout en présentant les perspectives futures de ce travail.

2 Bref état de l'art

Nous présentons dans cette partie les langages les plus significatifs qui permettent de modéliser ou de transformer des documents semi-structurés. Nous constatons ainsi que si l'intégration des concepts est introduite par certains travaux, elle n'est jamais considérée comme point de départ de la recherche.
DTD [1], XSD [6] et Relax NG [4] sont les langages de modélisation les plus utilisés et les plus populaires. DTD et XSD sont des propositions du consortium W3 tandis que Relax NG a été proposé par le consortium OASIS comme une alternative aux deux autres langages. Selon la taxonomie de Murata [10], l'expressivité de ces trois langages est différente : DTD décrit des grammaires locales (deux éléments partageant le même nom ont obligatoirement le même contenu), XSD des grammaires de type simple (deux éléments frères partageant le même nom ont obligatoirement le même contenu), tandis que Relax NG décrit des grammaires d'arbre régulières et est par conséquent le plus expressif des trois.
En plus de ces différences d'expressivité, la conception de chaque langage est dépendante de certains aspects particuliers. Ainsi, DTD est un héritage direct de la famille de langage SGML. Tandis que XML est une simplification de SGML, DTD est une simplification du langage de modélisation dédié à SGML (SGML DTD). DTD possède toujours une syntaxe non-XML qui empêche le langage d'être lui-même modélisé. De plus, il ne considère pas les espaces de nommage qui furent introduits après sa propre spécification. XSD est proposé en 2001 pour répondre aux critiques adressés à DTD. La syntaxe XML, la gestion des espaces de nommage et l'introduction de nombreux types de données prédéfinis en font une bonne alternative à DTD. Malheureusement, le langage, au vu de

sa lourde spécification, est complexe et difficile à appréhender pour des utilisateurs peu expérimentés. Ainsi, Relax NG se profile comme un bon compromis. Il est plus riche que DTD tout en restant plus simple que XSD. Facile à apprendre et à utiliser de manière efficace, il séduit largement, si bien que certains groupes d'intérêt du W3C l'utilisent aujourd'hui pour leurs travaux.
En 1999, le W3C propose le premier langage dédié à la transformation de documents XML : XSLT est de ce fait le langage le plus connu et le plus utilisé pour le traitement de données semi-structurées. Il est indissociable à XPath, un langage pour localiser et sélectionner les différents nœuds d'un document. Toutefois, la version 1.0 du langage souffre d'une spécification peu rigoureuse. Il n'existe ainsi aucune formalisation sémantique du langage et aucun modèle XSD complet n'y est associé, ce qui rend difficile la validation les transformations XSLT, qui sont elles-mêmes des documents XML. Le manque de spécifications claires avait jusqu'à récemment comme conséquence que les différentes implémentations de XSLT ne rendaient pas exactement les mêmes résultats. La deuxième version de XSLT [8], publiée en 2007, inclut une certaine notion d'intégration (« *schema-awareness* ») qui permet, d'une part, de valider les documents d'entrée et de sortie avant l'exécution de la transformation et, d'autre part, de définir des types pour des fonctions XSLT. Toutefois, cette intégration de XSD entraine de nombreuses constructions relativement obscures. De plus, ces concepts ne doivent pas obligatoirement être implémentés et restent ainsi peu usités.
La plus grande faiblesse communément admise de XSLT 1.0 est l'impossibilité d'effectuer un typage statique des transformations. Grâce à celui-ci, il est possible de vérifier avant l'exécution de la transformation que le document produit est valide par rapport au modèle attendu en sortie. Pour répondre à ce manque, différents projets ont vu le jour. Tozawa [15] intègre le typage statique dans un sous-ensemble de XSLT. XDuce [7] permet d'exprimer des contraintes structurelles sur les documents grâce à l'utilisation d'expressions régulières. Il a largement inspiré d'autres projets. Finalement, la librairie XACT [9] pour Java utilise XSD comme formalisme de typage, ce qui consiste également en une certaine forme d'intégration.

3 Propriétés d'intégration

Avant d'introduire les trois langages que nous proposons (YML pour le balisage, DML pour la modélisation et DGL pour la transformation, tous trois dans une syntaxe XML), nous définissons les propriétés du système intégré. Tout d'abord, pour favoriser l'intégration, DML et DGL sont des applications YML de telle sorte que des modèles DML puissent être

définis pour chaque langage et que les instances DML et DGL puissent être validées à l'aide du même outil que les instances YML en général. De la même manière, les instances DML et DGL peuvent être traités par des transformations DGL. De plus, celles-ci produisent tout type d'instances. Ainsi le système peut être qualifié d'auto-suffisant : toutes les instances dans le système peuvent être validées et peuvent être considérées comme le résultat d'une transformation. Aucun autre outil ou langage n'est donc nécessaire pour que le système intégré soit fonctionnel.

L'intégration des langages va encore plus loin et permet d'étendre l'expressivité des langages sans ajouter de nouvelles constructions mais en combinant les langages. Ainsi, si le processus de sélection des nœuds imaginé dans la version de base de DGL permet de sélectionner les nœuds seulement par rapport à leur genre (nœuds textes, éléments, etc.), l'intégration de constructions du langage DML dans DGL permet de sélectionner de manière naturelle les nœuds par rapport à leur type (structure interne). Parallèlement, l'inclusion de constructions du langage DGL dans DML permet de définir des structures dynamiques, c'est-à-dire des structures qui s'adaptent à un contexte donné.

Pour réaliser ce type d'intégration, il est indispensable que les concepts sur lesquels repose chaque langage soient des concepts simples et clairement définis. Pour ce faire, une importance toute particulière est accordée à la définition d'une sémantique rigoureuse. Cette sémantique, définie à l'aide de la sémantique dénotationnelle [14], ne doit pas être considérée comme une valeur ajoutée du projet, qui serait introduite après la spécification des langages. En effet, le développement de chaque langage repose sur la sémantique qui en est le fondement. Durant le développement, chaque concept difficile à formaliser est assimilé à un concept trop compliqué et est alors repensé et simplifié.

De plus, les relations entre les différents langages et les instances de chaque langage doivent être étudiées pour éviter des chevauchements de concepts comme cela arrive parfois dans le monde XML, où le problème des espaces blancs est géré par trois constructions différentes dans chacun des langages : l'attribut *xml:space* dans XML, l'élément *xsl:strip-space* dans XSL et l'élément *whiteSpace* dans XSD. Ainsi, les modèles DML sont organisés de manière hiérarchique avec, au sommet, un modèle universel qui décrit toute instance YML. Les autres modèles héritent de celui-ci. Au bas de la pyramide, chaque modèle permet de valider une seule instance YML. Des mécanismes de dérivation par restriction sont mis en place pour décrire les relations entre les modèles. Comme les modèles sont basés sur des expressions régulières, il suffit de tester leur inclusion pour déterminer si un modèle est plus restreint qu'un autre. De leur côté, les transformations DGL peuvent être typées statiquement. Avec la transformation DGL et le modèle d'entrée, un modèle canonique

est calculé. Si ce modèle canonique est égal ou inclus dans le modèle de sortie, alors la transformation est assurée de produire des documents valides, pour autant que les documents d'entrée soient valides par rapport au modèle d'entrée.
Finalement, il est important que chaque langage possède une syntaxe commune. Ainsi la syntaxe choisie est la syntaxe XML qui, grâce à un mécanisme d'espaces de nommage légèrement modifié, permet de concrétiser l'intégration des langages.
L'avantage de ce système est la possibilité de définir de manière simple et naturelle, des transformations génériques. Une transformation générique est une transformation qui peut être appliquée indépendamment à différentes entrées pour produire une transformation spécialisée qui permet de traiter un certain type de documents. L'utilisateur évite la définition d'un nombre conséquent de transformations, puisque celle-ci est automatisée.

4 Documents et modèles

Les langages de balisage (YML) et de modélisation (DML) ont déjà fait l'objet de diverses publications et sont de ce fait seulement brièvement introduit dans cette partie.
D'une part, YML, dont une première version, fortement remaniée, est présentée dans [11], est un langage de balisage, sous-ensemble de XML, développé après une étude critique minutieuse de XML et conçu dans le but de se départir de toutes les lourdeurs de l'héritage XML : gestion des espaces de nommage, traitement des espaces blancs, unicité de l'élément racine, etc. Après avoir songé à le détacher totalement du monde XML et avoir défini une nouvelle syntaxe pour ce langage, nous avons décidé de nous rapprocher du monde XML afin de profiter des nombreux outils déjà existants. Ainsi un document YML est un document XML soumis à aux restrictions syntaxiques suivantes : (1) la représentation du contenu textuel est différente. Des crochets carrés sont ajoutées avant et après le texte ainsi qu'un croisillon optionnel qui indique un retour à la ligne (*[texte], [texte]#*). Cette représentation permet de distinguer au niveau des instances directement quels espaces blancs sont significatifs (entre les délimiteurs) et lesquels ne le sont pas (à l'extérieur). (2) La déclaration XML est conservée tandis qu'un élément racine *yml* est rendu obligatoire pour se conformer aux règles de XML. Le document YML lui-même est composé des attributs de cet élément ainsi que des nœuds qui constituent son contenu. (3) La déclaration YML est composée de deux attributs inclus dans l'élément *yml*. Il indique la version et l'encodage du document.

D'autre part, DML [13] est une application YML et un langage de modélisation basé sur des grammaires d'expressions régulières. Il peut être analysé de manière totalement déterministe. Un méta-modèle DML normatif permet de valider n'importe quel modèle DML. Le langage est donc complètement auto-descriptif. Les modèles DML sont appréhendés dans une hiérarchie avec, au sommet, un modèle universel qui valide tout document YML. Un mécanisme d'héritage entre les modèles permet, entre autres, de simplifier des opérations de typage. Puisque les contenus sont décrits par des expressions régulières, il suffit de tester l'inclusion de celles-ci pour déterminer les relations d'héritage entre deux modèles.
Dans son approche, DML est proche de Relax NG. L'idée principale est de trouver un compromis entre DTD et XSD. Toutefois DML diffère sur les points suivants : Relax NG ne propose aucun mécanisme d'héritage qui compliquerait le langage. Dans son souci principal d'intégration, cette vision d'ensemble des modèles est importante à maintenir pour DML. Par ailleurs, si Relax NG traite les éléments et les attributs de manière équivalente, DML insiste sur le fait que les attributs représentent des métadonnées sur les éléments et de ce fait ne peuvent être assimilés à des éléments. Deux éléments partageant le même nom mais avec des attributs différents sont considérés comme différents. Finalement, la relation entre documents et modèles est sensiblement différente. Pour Relax NG, la validation est simplement une opération qui peut être appliquée à un document. Pour DML, la validité est une propriété intrinsèque du document, qui doit toujours être valide, au moins par rapport au modèle universel.

5 Génération et transformation de documents

Le langage DGL est une application pour la génération et la transformation de documents YML. Il permet de donc de générer de nouveaux documents YML sans qu'aucun document d'entrée ne soit défini ou de manière plus traditionnelle de transformer un document en un autre. Un programme DGL peut prendre de zéro à plusieurs documents en entrée. Si plusieurs documents sont nécessaires à la transformation, des identifiants sont associés à chaque document et conservés en mémoire. Le modèle de traitement de DGL [12] est proche de XSLT. Un programme DGL est composé de *templates* et de *patterns*. Un template principal initialise la transformation puis les patterns sont constitués de nouveaux templates réunis sous forme de règles. Trois opérations simples forment les templates : un nœud peut être copié, créé ou alors un pattern peut être appliqué sur une liste de nœuds sélectionnés.
Le processus de sélection des nœuds est fortement simplifié par rapport à XPath [3] mais est suffisant pour cette expérience. Il s'effectue étape par

étape à partir d'une première liste de nœuds. Celle-ci contient soit la racine qui est un composant abstrait représentant le document, soit le nœud courant, soit des nœuds définis par un mécanisme de variables qui permet de tenir compte de certains axes. A chaque étape, les nœuds sont sélectionnés selon leur genre jusqu'à la dernière étape. Entre autres spécificités, il est possible d'accéder non seulement aux nœuds mais également à leur contenu, par exemple à la chaîne de caractères qui constitue le nom d'un élément. Ceci permet de copier, par exemple, le nom d'un élément comme valeur d'un nœud texte.
DGL peut être typé de manière statique, comme expliqué précédemment, et de manière dynamique dans les cas où le typage statique est insuffisant. En associant des structures DML aux templates de la transformation, il devient possible de valider les parties d'un document par rapport aux parties de différents modèles.
De plus, grâce à l'intégration de DML, le mécanisme de sélection des nœuds peut être grandement enrichi. En effet, en associant, à chaque étape de la sélection, des structures DML décrivant le contenu d'un élément ou des types DML définissant les expressions régulières des nœuds textes, il devient possible de sélectionner les nœuds non seulement par rapport à leur genre mais également par rapport à leur contenu ou type. Ainsi le mécanisme de sélection devient très riche et permet de définir des transformations de plus en plus précises. Les structures DML associées à la sélection peuvent de plus être dynamiques, c'est-à-dire évolutives au fil de la transformation, suivant le contexte et les nœuds de l'input rencontrés.
Il faut encore souligner que le langage DGL est totalement et formellement spécifié, grâce la sémantique dénotationnelle. Ceci rend son implémentation très rigoureuse et permet de comprendre et d'appréhender le langage de manière plus saine.
Grâce à la simplicité de la version initiale de DGL ainsi qu'à l'intégration naturelle de DGL et de DML, il est possible de définir des transformations génériques qui permettent d'exécuter de manière automatique des tâches répétitives telles que celle décrite dans la prochaine partie.

6 Traduction de documents : une application intégrée

L'application que nous présentons est, parmi d'autres, un bon exemple de la puissance des transformations génériques réalisables grâce au trio YML-DML-DGL. Si, dans cet exemple, ce sont plutôt les concepts généraux de l'intégration qui sont mis en avant, d'autres applications comme la génération automatique de documents à partir de n'importe

quel modèle mettent en avant l'intégration concrète de DML et DGL. Elles ne sont toutefois pas présentées ici.
Cette application simple et automatique concerne donc la traduction de documents : si plusieurs modèles, qui possèdent d'une part une structure similaire et d'autre part des balises dans différentes langues sont disponibles dans le système, alors il est possible d'écrire une transformation DGL unique et générique qui prend en entrée deux modèles, le premier rédigé la langue source et le second dans la langue cible, et produit la transformation spécifique qui pour chaque document YML de la langue source produit le document traduit. La figure 1 illustre cette application pour un exemple simple de carnet d'adresses. Le modèle existe en anglais (source) et en français (cible). La traduction générique (*translation.dgl*) est écrite par l'utilisateur et la traduction spécialisée (*english-french.dgl*) est automatiquement produite par la traduction générique.

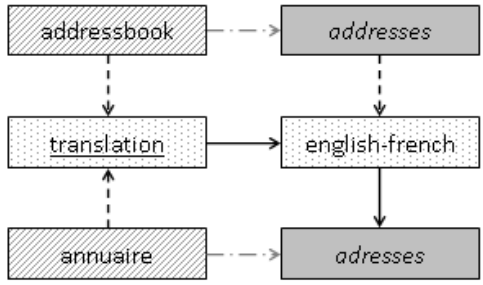


Figure 1. Traduction générique de documents

De manière plus concrète, la transformation générique produit pour chaque élément du modèle source une règle qui sélectionne ces éléments et crée de nouveaux éléments dont le nom est issu du modèle cible. Puisque la structure générale des modèles est la même, l'appariement entre les éléments est évident. L'information en tant que telle est contenue la plupart du temps, et c'est le cas dans cet exemple, dans les nœuds textes. Ceux-ci sont dès lors simplement copiés à des positions et niveaux équivalents. De plus, comme la transformation est générique, les modèles source et cible peuvent être interchangés sans que la transformation ne doive être modifiée. La transformation spécialisée ainsi générée peut être appliqué à n'importe quel document valide par rapport au modèle source pour être traduit.
Cette application montre la robustesse et les potentialités de l'intégration des langages. Appréhender ces langages sous la contrainte d'une intégration forte permet donc d'envisager une évolution positive de la gestion des documents semi-structurés.

7 Conclusion et perspectives

Dans cet article, nous avons présenté une étude qui vise à intégrer les trois langages principaux pour le traitement des documents semi-structurés. Afin d'assurer des fondations solides à ce projet, nous proposons trois nouveaux langages afin d'éviter les potentielles faiblesses des langages existants et de garantir que ces langages se focalisent avant tout sur une future intégration. YML pour le balisage, DML pour la modélisation et DGL pour la génération et la transformation de documents visent ce but. Chacun de ces trois langages a été développé de manière à s'intégrer avec les autres pour proposer un traitement rigoureux des documents semi-structurés.
Le système a été complètement développé en Java. Sur la base de la sémantique formelle complète développée pour soutenir les concepts de DGL mais également de certaines parties du DML telles que le mécanisme d'héritage, l'implémentation des outils s'est révélée rigoureuse et sûre.
Nous avons brièvement introduit une application intégrée qui montre comment les instances des différents langages sont liées les unes aux autres dans le langage. D'autres applications existent qui démontrent les potentialités de l'intégration notamment en matière de généricité. Parmi elles, une application de génération automatique de documents permet de produire soit le document minimal correspondant à tout modèle, soit une transformation capable de générer ce document minimal.
L'intégration peut encore être renforcée afin d'améliorer le traitement des documents semi-structurés, qui, au vu de l'évolution du Web et de la transmission de l'information, reste un défi majeur du monde informatique. Ainsi, la validation des documents YML pourrait être traitée par des transformations DGL qui produiraient un document YML certifiant la validité d'un autre document. Une telle transformation existe déjà, elle permet de produire pour un certain nombre de modèles répondant à des restrictions encore importantes des transformations validant les documents associés. L'intégration de langages si proches dans la gestion de documents est ainsi, et au vu des premiers exemples produits et analysée, tout à fait pertinente.

8 Références bibliographiques

[1] T. Bray, J. Paoli, C.M. Sperberg-Mcquenn, E. Maler et F. Yergeau. Extensible Markup Language (XML) 1.0 (Fifth Edition). W3C, 2008 http://www.w3.org/TR/2008/REC-xml-20081126

[2] J. Clark. XSL Transformation Version 1.0. W3C, 1999 http://www.w3.org/TR/XSLT

[3] J. Clark et S. DeRose. XML Path Language (XPath) Version 1.0. W3C, 1999
http://www.w3.org/TR/XPath

[4] J. Clark et M. Murata. RelaxNG Specification, 2001.
http://www.relaxng.org/spec-20011203.html

[5] D. Connolly. The Evolution of Web Documents. xml.com, 1997
http://www.xml.com/pub/a/w3j/s3.connolly.html

[6] D.C. Fallside, P. Walmsley. XML Schema Part 0: Primer. W3C, 2004
http://www.w3.org/TR/xml-schema-0

[7] H. Hosoya, B.C Pierce. XDuce: A statically typed XML processing language. ACM Transactions Internet Technologies, 2003.

[8] M. Kay. XSL Transformation Version 2.0. W3C, 2007.
http://www.w3.org/TR/xslt20

[9] C. Kirkegaard, A. Moeller. Type Checking with XML Schema in XACT. Proceedings of the Workshop on Programming Language Technologies for XML (PLAN-X), 2006.

[10] M. Murata et al. Taxonomy of XML schema languages using formal language theory. ACM Transactions Internet Technologies, 2005

[11] C. Pugin et R. Ingold. YML : une version épurée de XML pour faciliter une spécification rigoureuse des modèles de documents et des transformations. Actes du Colloque International sur le Document Electronique 9 (CIDE9), Fribourg (Suisse) 2006

[12] C. Pugin et R. Ingold. Combination of Schema and Transformation Language Described by a Complete Formal Semantics. Proceedings of ACM Symposium on Document Engineering (DocEng07), Winnipeg (Canada) 2007

[13] C. Pugin et R. Ingold. Instances, modèles et transformations: tout en un. Actes du Colloque International sur le Document Electronique 11 (CIDE11), Rouen (France) 2008

[14] R.D. Tennent. The Denotational Semantics of Programming Languages. Communication of the ACM, vol.19, n.8, 1976.

[15] A. Tozawa. Towards static type checking for XSLT. Proceedings of ACM Symposium on Document Engineering (DocEng01), Atlanta, GA, 2001.