**Document numérique entre permanence et mutations**
Ouvrage collectif dirigé par Madjid Ihadjadene, Manuel Zacklad et Khaldoun Zreik
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 978-2-909285-67-2

Illustration couverture : Europia
Conception ouvrage : Europia

# Document numérique
## entre permanence et mutations

Actes du 13e Colloque international sur le Document Electronique
(CiDE.13) 16 – 17 Décmebre 2010, INHA, Paris

Sous la direction de :
Madjid Ihadjadene, Manuel Zacklad et Khaldoun Zreik

europia

# Table des matières

# Préface

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 par sa conception et ses usages et de diffuser les derniers résultats de la recherche académique et industrielle. Les caractéristiques apparemment très diverses des documents électroniques qui circulent sur le web et dans les institutions témoignent de la diversité des activités intellectuelles dont ils sont à la fois la résultante et le support. Mais derrière cette diversité apparente, **CiDE 13** vise à dégager un certain nombre d'invariant qui caractérisent les supports documentaires, les formes d'expression, les pratiques de communication… Par ailleurs, face aux évolutions très rapides des technologies de diffusion des documents semi-structurés, d'indexation et de recherche d'information, CIDE 13 se propose de revisiter les concepts utilisés pour définir ces outils qui constituent aujourd'hui l'épine dorsale du système d'information des organisations.

CiDE.12 ( Montréal, Canada) 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.

CiDE' 11 (Rouen, France).s'est attaché à **la notion de l'usage de documents** qui introduit un point de vue différent de celui de production qui n'est symétrique qu'a première vue car «à quoi serviraient tous les savoirs parcellaires sinon a être confrontés pour former une configuration répondant a nos attentes, a nos besoins et a nos interrogations cognitives ? » (E.Morin).

Pour sa dixième édition le thème retenu pour CIDE 10 (Nancy, France) était le document numérique dans le monde de la science et de la recherche. En effet, le monde scientifique est particulièrement touché par cette évolution ou les réflexions sur le rôle fondamental de l'information scientifique dans la nouvelle infrastructure de la recherche a l'ère du de l'information électronique sont de plus en plus actives. Le document numérique est omniprésent dans tous les articles ou rapports traitant des nouvelles pratiques scientifiques repérés par des mots-clés tels que "Cyberinfrastructure, e-Infrastructure, **e-Science** ou e-Recherche".

CiDE.9(Fribourg, Suisse) a abordé le document numérique en tant que vecteur de communication à l'ère du déploiement des techniques de gestion d'information basées sur l'Internet. Il vise à débattre de sa production et son usage, liés à des **situations organisationnelles,** contraints par des pratiques professionnelles et ancrés dans des habitudes sociales. Si les environnements technologiques, de plus en plus sophistiqués et élaborés, ouvrent de nombreuses perspectives il n'en subsiste pas moins qu'une utilisation judicieuse, en accord avec l'évolution des pratiques des utilisateurs et leur acceptation des technologies, constitue actuellement un défi d'envergure.

CiDE.8(Beyrouth, Liban) avait comme objectif de resserrer les liens entre l'ingénierie documentaire et l'**ingénierie linguistique** tout en considérant les différentes dimensions des documents électroniques à savoir : cognitive, structurelle et technologique. CIDE.8 a donné un intérêt particulier aux thèmes relatifs a la question du multilinguisme dans la conception et la perception du document.

CiDE'7(La Rochelle, France) avait porté sur les **aspects sémantiques** du document. La mise en avant du « sens » a en effet longtemps été regardée avec beaucoup de scepticisme au profit de traitements dits « de surface », s'attachant à « la forme » par opposition au « contenu ». Cette perception est en train de changer. Des progrès significatifs ont été réalisés au cours des dernières années, d'abord sur le document textuel (extraction d'informations, question answering, résumé automatique...), puis relayés de plus en plus dans les autres médias (extraction d'information et indexation de documents sonores et vidéo par le contenu, résumé d'oeuvres...). Par ailleurs, les travaux déployés autour du thème du « web sémantique » visent à décrire le contenu des documents ou ressources de toutes sortes de manière à les rendre accessibles et interopérables

CiDE.6(Caen, France ) (2003) concerne les méthodes pour traiter de la plurimodalité et du **multimédia.** Ces méthodes trouvent leurs sources dans l'ingénierie du document et certains domaines des sciences humaines et de la psychologie.

CiDE.5 (Hammamet, Tunisie ) concerne l'exploitation du document électronique dans le cadre de l'**éducation et la formation**. Un intérêt particulier a été octroyé aux travaux de recherche visant l'exploitation des documents mobiles dans la mise en oeuvre de systèmes éducatifs hybrides tirant parti des technologies émergentes dans le domaine des périphériques "sans fil".

CiDE.4 (Toulouse, France) concerne les méthodes, les démarches et les techniques cognitives. En effet, si l'on retient que les sciences de la cognition ont pour objet l'étude des relations de l'esprit humain avec le monde et les autres esprits (dont lui-même), il apparaît que le document électronique - lieu, support et/ou objet-même de communication - s'offre comme un terrain privilégié d'investigation, par les **sciences cognitives,** des différents processus dont ces documents sont l'objet, et des spécificités ouvertes par la modalité numérique.

CiDE.3 (Lyon, France) a plus particulièrement insisté sur la conception et l'utilisation de systèmes d'information documentaire ainsi que sur les aspects dynamiques du document. Un accent particulier est donc mis sur les **bibliothèques numériques.** Les sujets développés ont concerné notamment le cycle de vie du document et ses aspects ergonomiques et perceptifs.?

CiDE.2 (Damas, Syrie ) avait fait porter la réflexion porte en priorité sur l'**aspect dynamique du document**. Le document devenant en même temps entité et processus il acquiert, tout a la fois un état achevé mais aussi perpétuellement provisoire, un statut d'entrée autant que de sortie. Il convient désormais de percevoir qu'au travers cette propriété émergente sont a reconsidérer les acquis méthodologiques, techniques, d'usages et de comportements.

La première édition CiDE (Rabat, Maroc ) avait porté essentiellement sur les problèmes posés par l'utilisation du document électronique comme outil spécifique dans **le processus de communication écrite** (que le document électronique soit enjeu ou simple support de cette communication). La réflexion s'est organisée autour de : la nature hétérogène des objets contenus dans le document, la richesse des éléments structurels et leurs contributions a la description du document et enfin la nature du support physique d'inscription

*Nous tenons à remercier tous les membres du comité de programme pour leur collaboration qui contribue à la qualité de ce colloque international ainsi que les membres du comité d'organisation. Nous remercions également également l'INHA de nous accueillir dans ses locaux à Paris.*

**M.Ihadjadene, M.Zacklad, K.Zreik**
**Co-présidents de CiDE.13**

**Comité du Programme**

Ghislaine Azemard, Paragraphe - Université de Paris 8, france
Bruno Bachimont Université de Technologie de Compiègne, France
Thierry Baccino, Lutin Université de Nice-Sophia Antipolis, France
Laurence Balicco, Gresec Université Grenoble 3, France
Abdel Belaid, Loria Université de Nancy 2, France
Evelyne Broudoux, Dicen Université de Versailles-Saint-Quentin, France
Anne Laure Brisac, INHA Paris, France
Jean Caelen, CLIPS, Grenoble, France
Ghislaine Chartron, Dicen Cnam, France
Stéphane Chaudiron, Geriico, Université de Lille 3, France
Viviane Couzinet, Lerras Université Toulouse 3, France
Jacques Ducloy, DRRT-Lorraine, France
Bernadette Dufrenne, HAR Université de Paris 10, France
Laurence Favier, Université de Bourgogne, France
Mauro Gaio, Liuppa Université de Pau, France
Joël Gardes, Orange Labs, France
Brigitte Guyot, Dicen Cnam, France
Patrick Gallinari, LIP6, Université de Paris 6, France
Mohamed Hassoun, Enssib-Lyon, France
Maryvonne Holzem , Lidifra Université de Rouen, France
Geneviève Lallich-Boidin, Elico Université de Lyon1, France
Omar Larouk, Enssib Lyon, France
Jacques Labiche, Litis Université de Rouen, France
Sylvie Leleu-Merviel, Université de Valenciennes et du Hainaut Cambrésis, France
Jacques Madelaine, Greyc Université de Caen, France
Yves Marcoux Ebsi Montreal, Canada
Ghassan Mourad, LaLICC Université Libanaise, Liban
Giovanni De Paoli, Université de Montréal, Canada
Samuel Parfouru EDF R&D Paris, France
Martine Poulain, INHA Paris, France
Jean-Pierre Raysz, Jouve-R& D, France
Jean Revez, Université du Québec, Canada
Imad Saleh, Paragraphe - Université Paris 8, France
Emmanuel Souchier, Gripic-Celsa France
Lynda Tamine-Lechani, Irit,, Toulouse, France
Said Tazi, Laas-CNRS, Université Toulouse 1, France
Eric Trupin, Litis Université de Rouen, France
Lise Vieira, Mica-Gresic Université de Bordeaux 3, France

**Comité d'organisation**

B. Dufrene (Université Paris 10)-Présidente
M. Flicoteaux (Université de Paris 10)
J. Gardes (Orange)
S. Ranjavely (Paragraphe, Université Paris 8)
C. Payeur (DICEN-Cnam)

**Présidents du colloque**

Madjid Ihadjadene, (Paragraphe, Université Paris 8)
Manuel Zacklad (DICEN-Cnam)
Khaldoun Zreik (Paragraphe - Université Paris 8)

# Construire ensemble des mémoires numériques durables : l'archivage numérique pérenne

**Benoît Habert**
*benoit.habert@ens-lyon.fr*
*ICAR – Ecole Normale Supérieure de Lyon et EDF R&D*

**Résumé.** Le numérique natif et la numérisation de données analogiques sont désormais conçus comme la manière privilégiée de transmettre l'existant. L'obsession mémorielle actuelle contribue à cette tendance. Les fragilités intrinsèques du numérique contraignent en fait à s'interroger sur les conditions techniques mais aussi et surtout sociales et humaines qui permettent de produire du « numérique durable », c'est-à-dire tel qu'il permette un accès maintenu, à long terme, des communautés aux connaissances et aux savoirs qui leur sont précieux. Les dimensions de cette pérennisation (le modèle utilisé pour ces opérations, sa mise en œuvre possible) sont abordées à partir d'un projet pilote d'archivage numérique de données orales, dans le cadre d'une Très Grande Infrastructure de Recherche en sciences humaines et sociales.
**Mots-clés.** archivage numérique pérenne, très grandes infrastructures de recherche, OAIS (Open Archival Information System), transmission des connaissances et des données

## 1 Introduction

A l'URL <http://www.futura-sciences.com/fr/news/t/technologie-1/d/au-clair-de-la-lune-ecoutez-le-plus-vieil-enregistrement-sonore-du-monde_15096/>, si vous cliquez, vous entendez le plus vieil enregistrement sonore du monde. Il dure dix secondes et date très précisément du 9 avril 1860 (soit 17 ans avant le phonogramme d'Edison et 28 ans avant le premier enregistrement précédemment connu, celui d'un oratorio de Haendel sur un rouleau de cire). C'est probablement une femme qui chante un air simple : « Au clair de la lune, Pierrot répondit… » Edouard-Léon Scott de Martinville a « fixé » ce fragment grâce à son invention : le phonautographe. Cet appareil (Figure 1) inscrit au moyen d'un stylet relié à un résonateur une ligne blanche ondulant et analogue au son sur une bande papier recouverte de noir de fumée. David Giovannonni et Patrick Feaster, de l'association américaine First Sounds, qui se consacre à la reconstitution des plus enregistrements connus, ont retrouvé les dépôts pour brevet d'E.-L. Scott de Martinville à l'INPI (Institut national de la propriété industrielle) et à l'Académie des sciences celui que vous pouvez écouter. L'inventeur avait mis au point un dispositif destiné à noter visuellement le son, pas à le rejouer, contrairement à T. Edison. D. Giovannonni et P. Feaster ont donc fait appel aux techniques mises au point et adaptées par Vitaliy Fadeyev et Carl Haber (Lawrence Berkeley

National Laboratory). Ces techniques permettent de reconstituer à l'aide d'un laser la forme du sillon, analogue au signal sonore initial. Cette archéologie numérique fait que cette voix demeure et ne reste pas captive pour toujours d'une technique disparue. Il n'est pas sûr que le numérique que nous créons actuellement ait cette chance si nous ne mettons pas en place son archivage pérenne[1].

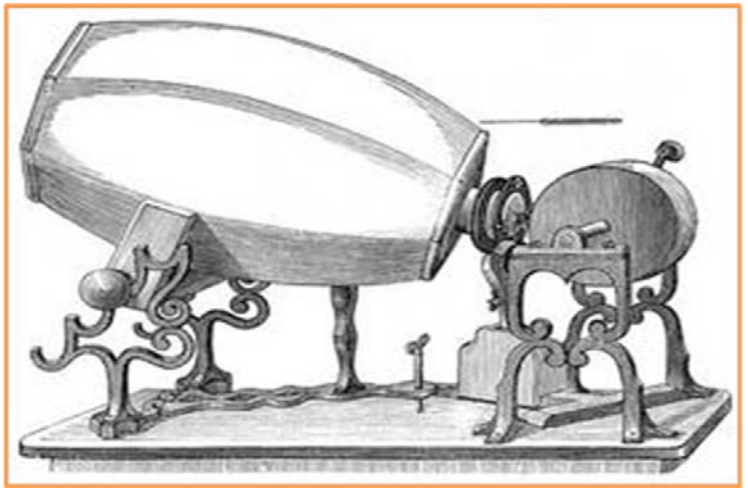

**Figure1.** Phonautographe d'E.-L. Scott de Martinville – modèle 1859 In Franz Josef Pisko Die neuere Appparate der Akustik, Vienne, 1865

La partie 1 oppose la place grandissante du numérique pour la transmission de l'existant à sa fragilité largement méconnue. La partie 2 présente la norme OAIS, modèle de référence pour la pérennisation du numérique. La partie 3 introduit un projet pilote de pérennisation en sciences humaines et sociales, dans le cadre d'une très grande infrastructure de recherche. La partie 4 examine la mise en œuvre du modèle OAIS dans ce contexte. La partie 5 insiste sur les enjeux de l'archivage numérique en termes de représentations partagées. La partie 6 souline les conditions de pérennisation d'un tel projet. La partie 7 conclut sur les changements de mode de travail qui résultent de l'archivage numérique pérenne et sur l'importance de réapprendre à oublier pour mieux pérenniser.

## 2 Mauvais souvenirs numériques

La première décennie du XXIe siècle prolonge la « compulsion mémorielle » qui s'installe dans le dernier quart du siècle précédent. R. Robin [1] et E. Hoog [2], parmi d'autres, analysent ce phénomène. Les signes en sont nombreux. Citons-en quelques-uns. « Selon les estimations du cabinet IDC, 420 milliards de photos ont été prises dans le monde en 2007, soit près de 50 millions par heure » [2]. P. Ricœur [3] cite P. Nora qui parle dans Les lieux de mémoire d'un « moment-mémoire » pour définir notre époque. Le succès de ce dernier ouvrage de 3 volumes et près de 5 000 pages est lui-même un symptôme. Paru initialement entre 1984 et 1992 en édition reliée, il est désormais disponible en « livre de poche ». Tout se passe comme si le peuple français, « recru d'histoire »

[1] *Quelle est l'espérance de vie de l'URL mentionnée supra ?*

(De Gaulle), se projetait malaisément dans le futur et se réfugiait dans le passé, sa conservation, voire son embaumement. Sans doute s'agit-il d'ailleurs d'une tendance –la recomposition des racines – partagée aussi bien par les « vieilles nations », en y incluant les Etats-Unis, que par les plus « récentes », comme celles issues de la réorganisation de l'ex-Europe de l'Est.
La généralisation du numérique ouvre des horizons nouveaux à cette obstination mémorielle. Le numérique dissocie en effet l'information représentée d'un support spécifique : une image, des sons, un texte, un programme se représentent uniformément par des suites de 0 et de 1. Celles-ci peuvent être stockées indifféremment sur des disquettes, des disques durs, des mémoires flash, des CD-Rom ou des DVD, etc., dont la taille ne cesse d'augmenter. La numérisation de données analogiques répond alors à ce désir de ne pas perdre ses « souvenirs », à titre individuel mais surtout collectif. Les institutions patrimoniales nationales, comme la Bibiothèque Nationale de France (BNF) ou la DAF (Direction des archives de France), sont engagées fortement et depuis de nombreuses années dans ce mouvement. Les pouvoirs régionaux soutiennent également des actions nombreuses et importantes en ce sens. Le numérique « natif » a cru dans des proportions incomparablement plus fortes. Chacun(e) de nous dans sa vie professionnelle comme personnelle vit et produit du numérique et dépend crucialement des mémoires offertes par son/ses ordinateur(s), ses sauvegardes, comme les « pertes » accidentelles ou les vols le rappellent cruellement. Pour reprendre les mots d'A. Ernaux [4] : « La recherche du temps perdu passait par le web. [...] On était dans un présent infini. On n'arrêtait pas de vouloir le 'sauvegarder' en une frénésie de photos et de films visibles sur le champ. Des centaines d'images dispersées aux quatre coins des amitiés, dans un nouvel usage social, transférées et archivées dans des dossiers – qu'on ouvrait rarement – sur l'ordinateur. Ce qui comptait, c'était la prise, l'existence captée et doublée, enregistrée à mesure qu'on la vivait, des cerisiers en fleur, une chambre d'hôtel à Strasbourg, un bébé juste né. Lieux, rencontres, scènes, objets, c'était la conservation totale de la vie. Avec le numérique, on épuisait la réalité ».

**Figure 2.** *Dégradations numériques minimes*

Cet emballement à produire du numérique, par conversion ou nativement, s'accompagne curieusement d'une inconscience certaine quant à la durée de vie effective des données numériques résultantes. La peur de perdre « leur passé » travaille les individus et les peuples : ils risquent pourtant de se trouver confrontés très rapidement à de mauvais souvenirs numériques. Le numérique conjugue en effet les fragilités. Les supports physiques du numérique vieillissent mal. « Une étude réalisée par Google sur son propre parc informatique a ainsi révélé que 8% des disques vieux de deux à trois ans devaient être remplacés en raison de leur défaillance » [2]. D'une recopie d'un support à l'autre, un bit peut changer. Les formats évoluent rapidement et « piègent » l'information qui est représentée par leur intermédiaire. Les logiciels qui donnent accès à un moment donné à l'information numérisée meurent d'eux-mêmes. Enfin, l'information peut se retrouver inaccessible, parce qu'on ne sait plus qu'elle existe ou parce que les

termes qui permettent d'y accéder (métadonnées) ont changé. Cette fragilité numérique se rapproche parfois des dégradations minimes observées pour l'analogique. C'est le cas lorsque les différences de ressources (polices, par exemple) d'un ordinateur à l'autre déforment le rendu. Figure 2, la flèche de réécriture (→) s'est ainsi métamorphosée inopinément en un D orné. Plus fréquemment la dégradation est « catastrophique ». Alors qu'un manuscrit endommagé, un papier jauni, cassant, une photographie ancienne et un peu effacée restent au moins partiellement lisibles, un fichier numérique devient souvent totalement inutilisable. Au total, sans pessimisme exagéré, on peut estimer que l'essentiel du numérique actuel (natif ou non) a une espérance de vie limitée, de 5 à 10 ans maximum. Nous produisons aujourd'hui sous forme numérique des données et des connaissances qui se révèleront inutilisables à long terme (une génération et plus) si nous ne prenons pas les mesures appropriées.

## 3 Comprendre le problème posé par la pérennisation du numérique : le modèle de référence OAIS

Le domaine spatial a été l'un des premiers à utiliser les technologies numériques de façon massive dès la fin des années 1960 : l'information transmise au sol par les sondes scientifiques spatiales était nécessairement de nature électromagnétique ; le volume d'information imposait des traitements automatisés. Les informations recueillies étaient le plus souvent uniques et irremplaçables. Lorsqu'on observe une comète qui passe au voisinage de la terre ou une éruption solaire, lorsqu'on établit une cartographie précise des forêts sur la terre à une date donnée, il ne sera pas possible de reconstituer ces informations si l'on a perdu les données correspondantes. Après qu'aient été accumulé des observations pendant plus de 20 ans, après qu'aient été subies les premières mutations technologiques du numérique, la question de la pérennisation de ce patrimoine d'observations scientifiques s'est posée avec acuité dès le début des années 1990. Les agences spatiales, la NASA aux Etats-Unis et le CNES en France en particulier, avaient commencé à apporter des réponses pragmatiques aux questions qui se posaient mais la nécessité d'une réflexion normative de fond sur le sujet s'est vite imposée.
Le CCSDS (Comité consultatif pour les systèmes de données spatiales) [5] est un organisme de standardisation commun aux agences spatiales. Il représente également le sous-comité de l'ISO (Organisation internationale de normalisation) dédié aux véhicules spatiaux. Les ingénieurs du CCSDS ont été sollicités en 1995 pour élaborer une norme en matière d'archivage long terme des observations spatiales. Ils ont alors eu l'intelligence de répondre de manière prospective sur deux points essentiels :

1 La question de l'archivage à long terme d'informations sous forme numérique n'est en rien une question spécifique au domaine spatial et il convient d'associer à une telle réflexion, les représentants d'autres secteurs d'activité qui sont ou seront confrontés aux mêmes besoins de conservation numérique,
2 Dans un contexte d'instabilité et d'obsolescence constante et rapide des technologies numériques, il est préférable de proposer un modèle de référence abstrait permettant de comprendre les spécificités particulières du numérique et de définir tous les concepts nécessaire à la compréhension et à la résolution du problème posé. Un tel modèle pourra être durablement stable alors que toute norme de mise en œuvre sera dépendante des technologies du moment et par conséquent éphémère.

C'est ainsi qu'est né le « Reference Model for an Open Archival information System », appelé plus simplement Modèle OAIS, [6] standardisé par le CCSDS en 2002 puis normalisé par l'ISO en 2003 (ISO 14721). Nous en retiendrons ici les points essentiels.
L'archive (au singulier) est définie dans ce modèle comme une organisation chargée de conserver l'information pour permettre à une communauté d'utilisateurs cible d'y accéder et de l'utiliser. On y retrouve l'idée de la pérennisation de l'information mais aussi, très explicitement, la nécessité de faire en sorte que dans le futur, cette information puisse être retrouvée, récupérée, comprise, interprétée convenablement par des utilisateurs qui n'ont pas participé à sa création et qui utilisent d'autres moyens numériques, ordinateurs, systèmes d'exploitation, logiciels que ceux sont qui ont été utilisés pour la création de cette information. Ceci nous amène tout de suite à une réflexion qui dépasse les questions purement technologiques : si nous considérons qu'un document est un ensemble d'informations enregistrées, considéré comme une unité, qu'est-ce qui permettra que ce document soit compréhensible dans le futur par ces utilisateurs ?
Les deux piliers du modèle OAIS sont constitués par un modèle d'information et un modèle fonctionnel.
Le modèle d'information apporte certainement la contribution la plus cruciale. Nous voulons pérenniser des informations numériques représentées en pratique par des données constituées de séquences de 0 et de 1 (les bits). Le modèle identifie et conceptualise les catégories d'informations supplémentaires qu'il faudra obligatoirement conserver en complément des données pour atteindre l'objectif fixé. Au cœur de ces informations complémentaires se trouve l'information de représentation : elle concentre tout ce qu'il est nécessaire de savoir, aujourd'hui et dans le futur proche ou lointain, pour pouvoir passer des séquences de bits à une information intelligible. Vaste question, d'autant que cette information de représentation étant elle-même numérique, le problème devient récursif. A côté, se greffent également l'information de provenance (savoir d'où vient le document, être sûr de son authenticité), l'information d'identification (qui n'est pas un concept nouveau et qui existe depuis longtemps avec les ISBN, ISSN… mais en numérique, comment créer un identifiant à la fois unique et pérenne ?). L'information d'intégrité nous assure que le document n'a pas été modifié et l'information de contexte replace ce document dans le cadre général d'un ensemble de documents organisés, hiérarchisés, reliés entre eux. En matière de terminologie, la distinction est donc faite dans le modèle OAIS entre l'information, définie ici comme une connaissance susceptible d'être échangée, et la donnée qui n'est que la représentation formalisée de cette information – représentation adaptée à la communication, l'interprétation ou le traitement. La donnée est porteuse d'information. Une difficulté majeure sur le long terme sera donc de toujours être en mesure de passer de la donnée à l'information contenue dans cette donnée puisque c'est la préservation de l'information qui nous intéresse.
La Figure 3 donne un aperçu concret des « couches » qui s'empilent pour tout objet numérique, si simple soit-il. En a, un « train de bits », de 0 et de 1, très exactement 120. En b, le découpage en 15 octets (suite de 8 bits). En c, chaque octet est interprété comme représentant un nombre, lui-même un numéro d'ordre dans un jeu de caractères (ici l'ISO-Latin1). En d, chacun de ces numéros d'ordre est interprété comme un chiffre. On obtient en e un nombre en concaténant ces chiffres. Ce nombre serait susceptible des opérations habituelles de son type : sommation, multiplication, etc. Alternativement, et c'est une autre interprétation, en f, en ajoutant des espaces « pour l'œil » (elles ne figurent à aucun des niveaux du

millefeuille numérique), on obtient un identifiant du type de celui qui figure sur la carte Vitale de chacun(e) de nous. Il s'agit très précisément d'un numéro d'inscription au répertoire (NIR), l'identifiant unique des individus inscrits au répertoire national d'identification des personnes physiques (RNIPP) géré par l'INSEE depuis 1946 <http://xml.insee.fr/schema/nir.html>. Le NIR est un numéro à treize caractères dont la composition est précisée dans l'article 4 du décret n° 82-103 du 22 janvier 1982 : « Le numéro attribué à chaque personne inscrite au répertoire comporte 13 chiffres. Ce numéro indique successivement et exclusivement le sexe (1 chiffre), l'année de naissance (2 chiffres), le mois de naissance (2 chiffres), et le lieu de naissance (5 chiffres ou caractères) de la personne concernée. Les trois chiffres suivants permettent de distinguer les personnes nées au même lieu, à la même période. » En f, se trouve donc un identifiant numérique de 15 chiffres composé du NIR (13 chiffres) et de sa clé (2 chiffres). Il permet de réduire les erreurs de saisie ou de transmission des identifiants NIR et est couramment utilisé dans le domaine médical, notamment dans les déclarations de sécurité sociale. La clé permet de vérifier que les 13 chiffres qui précèdent ont été transmis sans erreur. On calcule cette clé en soustrayant de 97 le reste de la division entière par 97 du nombre formé par les 13 chiffres précédents. Dans l'exemple présent : 97 – (1580875656192 modulo 97) = 97 – 23 = 65 → le NIR a bien été transmis. Supposons que l'identifiant transmis soit 158087564619265, c'est-à-dire la très minime erreur de recopie d'un bit sur 120 au 9e octet. 97 – (1580875646192 modulo 97) = 97 – 14 = 83 → ce qui a été transmis n'est pas un NIR.

| | | |
|---|---|---|
| a | « Train de bits » | 1,10E+117 |
| b | Octets | 00110001 00110101 00111000 00110000<br>00111000 00110111 00110101 00110110<br>00110101 00110110 . 00110001 00111001<br>00110010 …. 00110110 . 00110101 |
| c | Numéros d'ordre dans un jeu de caractères | 49 53 56 48 56 55 53 54 53 54 49 57 50 54 53 |
| d | Chiffres | 1 5 8 0 8 7 5 6 5 6 1 9 2 6 5 |
| e | Nombre | 158087565619265 |
| f | Identifiant | 1 58 08 75 656 192 65 |

**Figure 3.** *Le « millefeuille » du numérique*

La donnée est le train de 120 bits en a de la Figure 3. Les couches successives b, c, d, f sont autant d'informations de représentation nécessaires à la sauvegarde de l'information correspondant à cette donnée. Elles ne suffisent d'ailleurs pas à permettre de se servir de cette information : il manque le contexte (NIR et sa structure ; règle d'utilisation de la clé de contrôle, etc.).
Archiver, dans ce cadre, consiste tout d'abord à « certifier » les données qui sont livrées (les trains de bits) en s'assurant qu'elles répondent effectivement au format qu'elles revendiquent. Dans le cas du NIR, c'est par exemple vérifier que le chiffre en première position est soit 1 ou 2 (codage du sexe), que celui en 3e et 4e position correspondent aux codages : 01, 02, 03, 04, 05, 06, 06, 08, 09, 10, 11, 12, c'est-à-dire à une représentation conventionnelle des mois de l'année, etc. Une archive accepte donc un nombre déterminé de formats d'archivage pour les données et opère au versement cette certification. Par exemple, pour reprendre

l'exemple de la Figure 2, un fichier PDF peut ne pas contenir toutes les ressources nécessaires pour être utilisé sur toutes les plateformes. Son caractère « auto-porteur » va donc être examiné. L'archive vérifie également la bonne formation des métadonnées correspondantes. Comme on l'a vu dans l'exemple de la Figure 3, le contexte d'utilisation de la donnée doit être fourni, dans toute sa complexité éventuelle, et dans un format lui aussi standardisé (par exemple à partir du Dublin Core [7]). Concrètement, l'archive reçoit d'un service versant (Figure 4) un paquet de versement (Submission Information Packet – SIP), associant données et métadonnées. Ce paquet correspond à une « grammaire » XML qui fait l'objet d'un accord entre archive et service versant. Le service versant a pour fonction de faire le truchement entre les données et métadonnées parfois encore « désordonnées » des producteurs et les spécifications de l'archive. L'archive met en place des chaînes de traitement qui valident ou invalident le paquet de versement. En cas d'invalidation, le service versant est notifié et doit « revoir sa copie ». En cas de validation, le paquet est archivé avec un numéro unique (comme AIP – Archival Information Packet).

La tâche centrale de l'archive – la certification – est complétée par d'autres volets. En premier lieu, l'archive et le service versant doivent s'accorder sur ce qui a été versé. Pour cela, les deux calculent avec le même algorithme une empreinte numérique du paquet de versement. L'archive envoie cette empreinte au service versant qui peut savoir alors si ce qui a été reçu et traité est bien ce qui a été envoyé. Cette propriété est particulièrement cruciale pour pouvoir remplir une des fonctions de l'archive, la réversibilité, c'est-à-dire la capacité à retourner à l'identique au service versant son paquet (par exemple, en cas de cessation de l'archive ou de changement d'archive pour le service versant). En second lieu, l'archive doit préserver les paquets archivés des destructions ou dégradations de support. Elle assure leur réplication sur un site distant. Elle vérifie régulièrement la stabilité de chaque paquet : elle compare l'empreinte numérique de départ du paquet et celle qu'on calcule – elles doivent être identiques, dans le cas contraire, elle restaure le paquet à partir de l'autre version. En troisième lieu, l'archive assure la migration des données et des métadonnées en fonction des évolutions des formats (ce qui suppose une activité de veille sur les formats). On notera ici un paradoxe au cœur même de l'archivage du numérique. La migration est nécessaire pour l'archive ne meure pas. Elle ne garantit pas pour autant la préservation exacte de l'information. Enfin, l'archive donne accès à l'information en diffusant des versions des paquets archivés (DIP – Dissemination Information Package).

Soulignons quelques distinctions de sens. Stocker n'est pas archiver, pas plus que sauvegarder, ou répliquer : c'est effectuer une « copie » (éventuellement multiple) de la donnée telle quelle, sans vérifier sa bonne conformité et sans lui associer obligatoirement des métadonnées (autres que celles du nom du fichier et de son type éventuel). Dans la pratique, hors le monde de l'archivage numérique pérenne, les mots s'emploient sans grande précaution, et l'on parle souvent d'archivage là où il ne s'agit guère que de sauvegarde ou de réplication.

Le modèle fonctionnel (Figure 4) vise à délimiter les responsabilités entre l'archive et les intervenants externes que sont les producteurs d'information, le service versant, les utilisateurs et ce qui est appelé le management et qui correspond ici à l'entité qui définit le mandat de cette archive et qui, souvent, lui fournit les ressources nécessaires à son fonctionnement. Ce modèle fonctionnel identifie également les flux de données et les fonctions essentielles qui sont prises en charge par l'archive. Les entités fonctionnelles « Entrées », « Stockage », « Gestion des données », « Accès » parlent d'elles mêmes. L'entité

« Administration » assure la supervision, la coordination continue du fonctionnement des autres entités de l'Archive. C'est elle qui prend les décisions internes. L'entité « Planification de la préservation » assure les fonctions relatives à la surveillance de l'environnement de l'archive et à la production de recommandations visant à ce que les informations archivées restent accessibles et compréhensibles sur le long terme dans un environnement en évolution permanente.

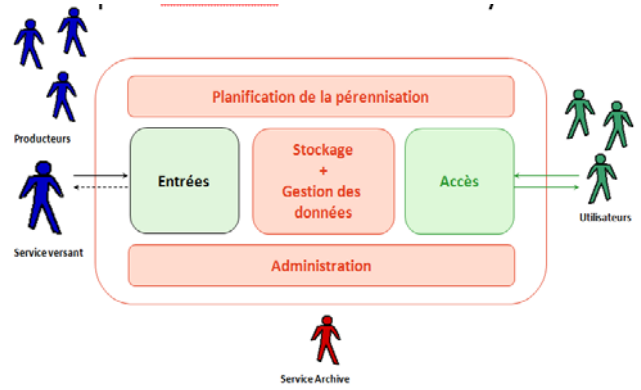


**Figure 4.** Le modèle OAIS

On voit la place centrale des formats. La Figure 5 matérialise la transformation entre les formats utilisés par le producteur, ceux acceptés par l'archive en entrée (Figure 6 – ceux de la plateforme d'archivage du CINES [8] – cf. infra), et ceux fournis à l'utilisateur. Un format acceptable est un format largement utilisé, mais surtout dont les spécifications sont publiées, de telle manière qu'on puisse valider son respect par une donnée qui s'en réclame. Les formats normalisés sont donc privilégiés, ainsi que les formats non propriétaires. Ce n'est pas exemple qu'assez récemment que le format Word a été publié et normalisé. La complexité actuelle de ce format n'en fait pas un candidat naturel à l'archivage. L'archivage peut donc contraindre à migrer des documents Word vers du PDF/A. Symétriquement, le format d'archivage n'est pas forcément celui dans lequel on donnera accès aux données. Par exemple, une image peut être archivée dans un format « lourd » comme TIFF tout en étant rendu accessible en PNG ou en JPEG pour des raisons de facilité de transfert, et assortie d'une vignette « dégradée » permettant le feuilletage commode des données archivées.

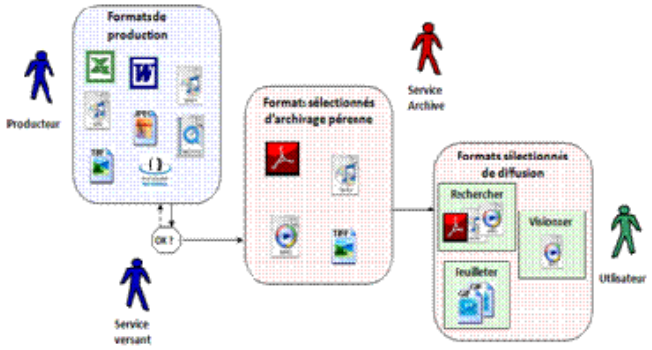


**Figure 5.** Les formats : du producteur à l'utilisateur via l'archive

| Type | Format |
|---|---|
| Texte | HTML, PDF, TXT, XML |
| Image | GIF, JPEG, TIFF, PNG, SVG |
| Audio | WAV, AIFF, AAC, VORBIS |
| Vidéo | MJPEG2000, MPEG4, THEORA |

**Figure 6.** *Les formats acceptés au CINES*

Dès la fin des années 1990, avant même sa normalisation définitive, le modèle OAIS s'est imposé comme une base conceptuelle incontournable pour la pérennisation du patrimoine numérique. Il a donné lieu à un nombre important d'implémentations. Le service d'archivage numérique mis en place à la Bibliothèque nationale de France autour du système [9] (Système de préservation et d'archivage réparti) a poussé très loin la recherche d'une totale conformité au modèle.

## 4 Un projet pilote de pérennisation des données orales en SHS (2008-2010)

Les Très Grandes Infrastructures de Recherche (TGIR) ou Très Grands Equipements (TGE) répondent à la nécessité de mettre à la disposition de la communauté scientifique des moyens dépassant la programmation à relativement court terme (3 à 4 ans maximum) et les budgets correspondants. Ce sont donc des équipements « lourds », dont les financements dépassent ceux des laboratoires et des projets et qui sont conçus pour le moyen terme et un usage « transverse ». Ils sont pilotés par le Ministère de l'Enseignement et de la Recherche [10] et pour certains en charge du CNRS [11]. Les TGIR ont pris naissance, ce n'est pas surprenant, en physique. Ce modèle d'outillage scientifique s'est progressivement étendu à d'autres disciplines. La notion d'« instrument » s'en est élargie d'autant. C'est ainsi qu'ont émergé en astronomie les observatoires virtuels [12]. Il s'agit d'un réseau d'observatoires qui améliorent la mutualisation de leurs observations et qui combinent les observations provenant de plusieurs télescopes comme s'il n'existait qu'un seul instrument. La nécessité d'« équiper » de manière similaire les sciences humaines et sociales (SHS) a émergé progressivement. Au niveau européen, la feuille de route élaborée en 2006 par ESFRI [13] (European Strategy Forum on Research Infrastructures) et actualisée en 2008, planifie la stratégie européenne de financement de la construction des futures TGIR. Elle comprend plusieurs TGIR destinées aux SHS, dans des degrés variés de « construction » [14]. Le CNRS a créé début 2007 le TGE Adonis [15], consacré à l'accès unifié aux données et documents numériques en SHS. Pour remédier à la situation esquissée dans la première section, un objectif privilégié par le TGE Adonis a été la pérennisation des données numériques, avec deux visées : la sûreté et la fiabilité ; la réduction – à terme – des coûts par la mise en commun des moyens et des ressources humaines compétentes.

Le TGE Adonis a lancé en novembre 2007 une étude sur les besoins et les offres en archivage numérique pour les SHS. Pour éviter une expertise mêlant juges et parties, cette étude a été confiée à O. Bärring, du CERN (Genève). Les conclusions de l'étude ont été prêtes début février [16]. O. Bärring conseillait de ne

pas créer d'entité nouvelle dédiée à la pérennisation numérique, mais bien plutôt de s'appuyer sur des centres de calcul « lourds » déjà existants, bien munis en compétences et en équipements. L'idée était de faire des économies d'échelle : l'archivage SHS constituerait une proportion relativement minime des activités de tels centres et ne nécessiterait dès lors pas d'investissements initiaux importants en matériel (serveurs, réseau) coûteux. Cette solution permettrait aussi de bénéficier de compétences en matière de stockage et d'échanges massifs de données et de dégager des forces spécifiques pour le volet Archivage numérique à proprement parler.

L'hétérogénéité de départ des SHS conjuguée à un contexte de restructurations fortes interdisaient l'idée d'une solution globale d'archivage numérique. L'instance décisionnelle du TGE Adonis, son Comité de pilotage, a opté en mars 2008 pour un projet pilote d'archivage limité à un secteur particulier : les données orales. Par données orales, on entend des enregistrements de parole (des conversations, des monologues) qui servent à la recherche en linguistique (apprentissage du langage, fonctionnement de l'oral, etc.) ou en ingénierie linguistique (données d'entraînement de systèmes de reconnaissance de la parole). Il s'agit donc en général d'enregistrements sonores, mais les enregistrements vidéo se font de plus en plus fréquents : ils fournissent des renseignements supplémentaires sur les interactions entre les interlocuteurs, sur la complémentarité entre ce qui se dit par la voix, par le geste et par le corps. À cela peuvent s'ajouter des mesures de paramètres physiologiques associés à la production de parole : électroglottographie, articulographie, palatographie etc. Ces données ne sont pas « nues », elles sont assorties d'annotations ou d'enrichissements : transcriptions textuelles de l'échange ou de la prise de parole, découpage en « sons » (phonèmes) et en syllabes, etc. Enfin des métadonnées riches permettent d'identifier ces données et leurs annotations : langue concernée, période et région d'enregistrement (par exemple, pour pouvoir étudier l'oral selon les régions de France ou de la francophonie), informations sur les locuteurs (âge, sexe, catégorie socio-professionnelle).

Plusieurs raisons ont motivé la volonté du Comité de pilotage du TGE Adonis de partir de ce type de données pour montrer la faisabilité d'un dispositif d'archivage numérique en SHS. En premier lieu, plusieurs sous-communautés scientifiques produisent et/ou utilisent des données orales : psycholinguistes, syntacticiens, phonéticiens et phonologues, etc. Une même annotation ou un même enregistrement peuvent donc être utilisés et « enrichis » selon des angles différents. La coexistence de ces intérêts multiples et de données potentiellement partageables a conduit à des confrontations sur les manières d'annoter ces données, mais aussi de les recueillir, en vérifiant que l'on préserve à la fois les droits des personnes qui sont enregistrées et les nécessités de la recherche. Ces mises en regard ont conduit tout particulièrement à un Guide des bonnes pratiques [17]. L'usage des données orales va d'ailleurs au-delà de la communauté scientifique. Les langues de France font partie du patrimoine. C'est à ce titre que la DGLFLF [18] (la Délégation Générale à la Langue Française et aux Langues de France) a appuyé le travail de convergences qu'est le Guide des bonnes pratiques et qu'elle soutient depuis 5 ans le recueil et la mise à la disposition du grand public de données orales enrichies, via un portail dédié. En second lieu, le champ des données orales, comparativement à d'autres, était relativement structuré institutionnellement. La section 34 du CNRS – consacrée aux sciences du langage – a en effet créé deux fédérations de linguistique, l'une consacrée aux universaux de langage et à la typologique des langues, TUL, et l'autre, ILF, rassemblant les

laboratoires travaillant en linguistique du français. Les deux fédérations animent un réseau d'équipes et de laboratoires producteurs et utilisateurs de corpus oraux. Elles travaillent avec la DGLFLF à la constitution de nouveaux corpus oraux mais aussi à la sauvegarde sous forme numérique de corpus analogiques. Elles ont contribué au Guide de bonnes pratiques. Par ailleurs, le CNRS a créé en 2006 des centres de ressources numériques [19] conçus pour faciliter l'utilisation mutualisée de données numériques. Ils ont été organisés par type de données : textuelles, géographiques, images et… orales. Un CRDO [20] (Centre de ressources pour la description de l'oral) a donc été mis sur pied, mais avec deux têtes, l'une au Laboratoire Parole et Langage d'Aix, l'autre alors au LACITO en région parisienne. Ces deux têtes ont agi, sinon de concert, du moins dans des directions finalement convergentes en mettant en ligne des données, en aidant au recueil et à la production de corpus. En troisième lieu, les données orales constituent un bon « banc d'essai ». Les données ne sont pas trop volumineuses ni trop complexes, par rapport par exemple aux données issues de la simulation en 3 dimensions. Dans le même temps, elles constituent déjà un défi sérieux. Au départ du projet, le volume à pérenniser avoisinait déjà les 2 téra-octets, soit deux mille milliards d'octets. Il mélangeait du son, de la vidéo et du texte. Les problèmes juridiques de la mise en ligne de ces données, s'ils avaient été défrichés par le Guide des bonnes pratiques, n'étaient pas pour autant entièrement résolus.
Volume et complexité « raisonnables » des données à pérenniser, structuration du champ, force des communautés scientifiques et d'usage, ces trois caractéristiques faisaient des données orales un point de départ propice pour la démarche expérimentale envisagée. Le Comité de pilotage a donc donné le feu vert au projet pilote en mars 2008. Il a fixé l'objectif d'une évaluation scientifique et technique à Pâques 2009. Il a donné mission de prendre contact avec la Direction des Archives de France (DAF). La DAF est en effet légalement en charge des données produites par les chercheurs dans le cadre de leurs fonctions : c'est donc cet organisme qui peut éventuellement donner délégation à un archiveur sur ce point. En reprenant une des conclusions du rapport Bärring , le Comité de Pilotage a en outre souhaité une solution assise sur deux importants centres de calcul, le CINES [21] à Montpellier et le CC-IN2P3 [22], à Villeurbanne.
La mise en place effective du projet pilote a débuté pendant l'été 2008. Alors directeur adjoint du TGE Adonis, j'ai assuré le suivi du projet au sein d'Adonis. Claude Huc, précédemment en charge de l'archivage numérique au CNES et responsable alors du groupe PIN – Préservation de l'informatique numérique [23], a accepté d'assurer la coordination technique et fonctionnelle du projet en tant que consultant. Les premières réunions ont été facilitées par les contacts pris à l'occasion de l'Université d'été organisée par le TGE Adonis en septembre 2008. Une évaluation à destination du Comité de Pilotage du TGE Adonis a été effectuée par Yves Marcoux, Professeur à l'Université de Montréal, en juin 2009. En octobre 2009, la DAF a examiné l'état du projet pilote. Le projet pilote s'est arrêté en juin 2010 : la mise en production de l'archivage des données orales en SHS a commencé alors.
La section suivante détaille les adaptations du modèle OAIS faites dans le cadre du projet pilote.

## 5 Du modèle OAIS à sa mise en œuvre

La mise en œuvre du modèle OAIS pour l'infrastructure mutualisée du TGE ADONIS s'appuie sur trois acteurs : le CRDO, porteur d'une expertise sur les

données orales et deux centres informatiques majeurs, ce qui évite d'être contraint à des investissements initiaux importants. Une répartition des fonctions – et donc des responsabilités correspondantes – entre les acteurs est présentée Figure 7.

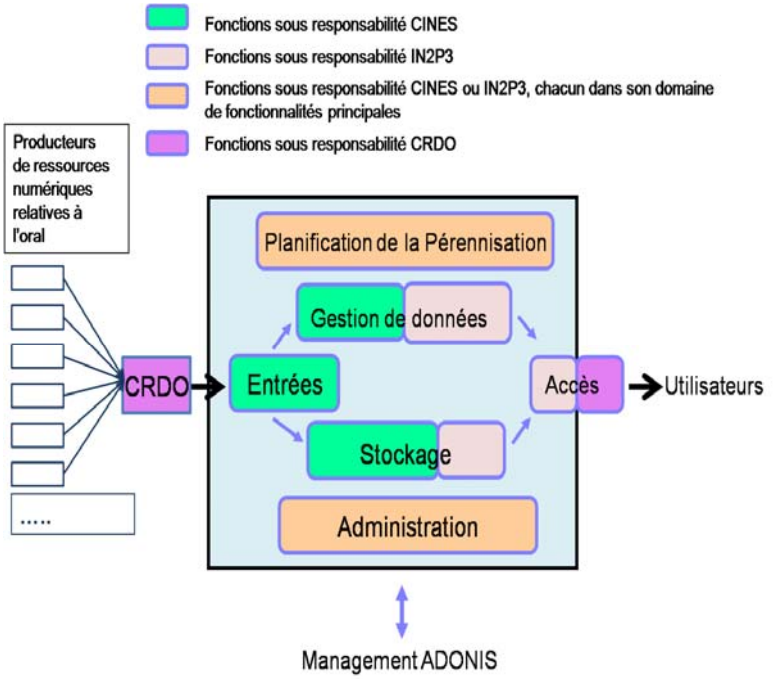


**Figure 7.** La mise en œuvre du modèle OAIS dans le projet pilote

Le Centre Informatique de l'Enseignement Supérieur – CINES [21] est depuis plusieurs années en charge de l'archivage des thèses numériques des universités françaises, de celui des revues SHS numérisées hors embargo – Persée [24] et, plus récemment, de l'archive ouverte de pré-publications en ligne HAL [25]. Il dispose d'une compétence établie en matière de préservation de l'information numérique. Cette compétence, reconnue par la tutelle en 2007, couvre un large spectre d'activités : standardisation de la structure des « paquets de versement » transmis par les producteurs, définition des métadonnées de préservation indispensables, identification des formats de données recevables pour une préservation à long terme, contrôle et validation en profondeur de la conformité de toutes les données versées par rapport à ces formats, gestion multi-site du stockage, surveillance et renouvellement des supports physiques d'enregistrement, usage des empreintes numériques pour garantir l'intégrité des données, capacité de réversibilité consistant à restituer à l'identique, les données et métadonnées au producteur, etc. Cet ensemble opérationnel constitue la Plate-forme d'archivage du CINES – PAC [8]. Par contre, le CINES dispose d'une expérience moindre en matière de moyens de recherche, d'accès et de récupération des données. Cette fonction fondamentale d'accès à l'information est assumée par l'ABES (Agence bibliographique de l'enseignement supérieur) pour les thèses, par l'université Lyon 2 pour les revues du projet PERSEE [24], par HAL [25] pour les pré-publications en ligne ou par le CC-IN2P3 [22] pour l'infrastructure pilote d'archivage mise en place par le TGE ADONIS.

Le Centre de Calcul de l'Institut national de physique nucléaire et de physique des particules – CC-IN2P3 [22], lui-même une Très Grande Infrastructure de Recherche, est depuis longtemps en prise directe avec une large communauté d'utilisateurs (en France, 25 laboratoires et 2 500 personnes) dont il assure la gestion. Il a acquis des compétences solides en matière de virtualisation du calcul et du stockage de données. Cette virtualisation permet de s'abstraire des solutions ou technologies utilisées face à un problème donné, elle conduit les développeurs à réfléchir de façon indépendante des outils. En outre, il dispose d'une infrastructure de stockage phénoménale (en 2009 5 péta-octets – Po – de données en ligne ; 4 Po de disques ; 30 po sur cartouches) pour les besoins de la physique subatomique. Pour l'archivage Adonis, le CC-IN2P3 développe et opère tout ce qui peut être générique dans la fonction de gestion de données et d'accès : stockage en ligne et organisation des données au sein d'un modèle sous le logiciel Open source Fedora Commons [26], gestion des droits, transformations en vue de fournir aux utilisateurs des données au format souhaité, etc. A terme, le CC-IN2P3 pourrait offrir aussi des mécanismes de recherche transverses couvrant tous les domaines des SHS.

Le CINES et le CC-IN2P3 ne disposent pas de compétences métier particulières, tant pour ce qui concerne les données orales que pour les autres domaines des SHS. C'est ce qui confère une importance majeure au CRDO, dans une fonction de médiateur entre la communauté des corpus oraux et l'infrastructure d'archivage. Le CRDO est né du besoin – identifié par le CNRS – de concentrer une compétence technique en matière de corpus oraux, compétence souvent absente des laboratoires et des petites structures de recherche. Il concentre en pratique un savoir-faire technique sur les corpus oraux, sur les formats de données et les métadonnées associées ainsi que sur la manière d'agencer les données afin de produire des entités archivables. Il permet donc une collecte intelligente des corpus et leur mise en paquets de versement (SIP) conformes pour archivage après une validation technique. Emanation de la communauté des corpus oraux, le CRDO garde également la maîtrise de l'interface utilisateur pour la recherche et la récupération de données archivées. Pouvoir prendre en compte, dans cette interface, les spécificités et la terminologie habituelle de cette communauté est en effet essentiel. C'est pourquoi, Figure 7, le CRDO intervient aussi bien à l'entrée qu'à la sortie du dispositif d'archivage.

## 6 Construire des représentations partagées pour transmettre le numérique

L'archivage numérique pérenne n'est pas ni en première analyse ni sur le fond un problème technique, au sens étroit. La place réelle des infrastructures techniques (bande passante, système d'authentification, serveurs et dispositifs de stockage, etc.), celle des normes comme OAIS [6] et des standards de métadonnées comme Dublin Core [7] n'est en fait que l'aboutissement d'une démarche longue et délicate de co-construction d'une solution assumée par chacune des parties prenantes. Il ne s'agit donc pas simplement d'« instancier » le modèle OAIS, mais de trouver ensemble la manière de lui donner sens dans un contexte particulier. L'inscription dans la durée d'une infrastructure de pérennisation de l'information numérique suppose, globalement la construction de représentations partagées entre les différentes parties prenantes. Il s'agit de s'accorder sur l'organisation des données et des métadonnées ainsi que sur le processus même d'archivage :

responsabilités des différents acteurs ; ensemble des processus à mettre en place, avec leurs variantes ; indicateurs de bon fonctionnement, etc.
Pour que l'infrastructure fonctionne et apporte effectivement, de façon fiable, les services qui ont été définis, elle doit par conséquent résulter d'une étape de développement et de validation maîtrisée et conduite avec méthode. Dans le cadre du projet pilote, s'agissant d'une infrastructure multipartenaires, il était crucial de constituer une équipe cohérente, motivée, au sein de laquelle chacun dispose d'une vue claire de l'ensemble du projet et des fonctions et responsabilités qui lui sont confiées. Tout cela est passé par une définition précise et tenue à jour des tâches et du calendrier, par des réunions « physiques » toutes les six semaines et des téléconférences bimensuelles, toujours consignées par un compte-rendu, par la mise en place de moyens de communication et de collaboration (liste de diffusion, gestionnaire de projet, wiki [27]), et par une coordination fonctionnelle et technique rapprochée et neutre. Le coordinateur du projet avait donc la responsabilité de trouver des solutions aux désaccords éventuels entre partenaires, à ce titre, il doit être indépendant de ces partenaires. Dans ce cas comme de manière générale, la solidité ultime de la solution proposée repose sur celle de l'accord ainsi construit, sur la capacité des acteurs à s'entendre globalement comme dans le détail ainsi qu'à s'accorder sur des bonnes pratiques. La lenteur est de mise, sous peine de mécomptes.
On peut faire l'hypothèse que l'émergence d'une norme ou d'un standard dans un domaine correspond à une réification des conventions de la communauté sous-jacente. Cette émergence serait pour l'immatériel, pour les processus et les données numériques, l'équivalent de ce que Latour entend par « boîte noire » pour les instruments. La communauté scientifique a débattu sur ce qui était à mesurer, sur les erreurs de mesure acceptables, etc. Elle stabilise un consensus, éventuellement réexaminable, en un instrument, devenu « boîte noire » : le débat qui lui a donné naissance est clos, provisoirement au moins. On retrouve une situation proche de celle décrite par A. Desrosières [28] pour l'émergence des concepts et indicateurs statistiques : « La constitution d'un espace rendant possible le débat contradictoire sur les options de la cité suppose l'existence d'un minimum d'éléments de référence communs aux différents acteurs : langage pour mettre en forme les choses, pour dire les fins et les moyens de l'action, pour en discuter les résultats. Ce langage ne préexiste pas au débat. Il est négocié, stabilisé et inscrit, puis déformé et défait peu à peu, au fil des interactions propres à un espace et une période historique donnés. Ce n'est pas non plus un pur système de signes reflétant des choses existant en dehors de lui [...]. » Que l'on pense au taux de chômage par exemple. A. Desrosières souligne les liens complexes entre la stabilisation des catégories statistiques et l'évolution des appareils d'Etat : le besoin d'agir sur la réalité implique de s'accorder sur des notions et leur usage ; inversement disposer de notions et de « jurisprudences » sur leur emploi fournit des leviers sur le réel tout en transformant ce réel. C'est tout l'enjeu par exemple du débat récent sur les statistiques d'origine « ethnique ». Peut-on agir sur l'équité de l'accès aux emplois des Français de différentes origines sans disposer de « mesures » de ces origines et de leur répartition ? Mais inversement, de telles mesures constituent des « manières de voir » dont la maîtrise est problématique. A. Desrosières [29] conclut : « L'information statistique est un de ces langages, parmi d'autres, grâce auquel les acteurs sociaux peuvent se coordonner. ». Les normes et standards dans le domaine de l'archivage numérique comme ailleurs correspondent à une telle coordination. Mais pour qu'ils puissent jouer ce rôle, il faut que les acteurs d'un système de pérennisation donné les adaptent à leur

« monde ». Le recours à des normes ou à des standards ne relève pas dans ce cas d'une simple « mise en musique », mais d'une démarche active mais lente d'appropriation. A. Desrosières (ibid.) ajoute : « ... ces espaces de formes durablement solidifiées [...] doivent à la fois être indiscutées pour que la vie suive son cours, et néanmoins discutables pour que la vie puisse changer de cours. » et invite à : « ... penser en même temps ces objets comme construits et réels, conventionnels et solides. ».
Le développement de l'archivage numérique pérenne depuis une quinzaine d'années dispose maintenant d'un corps de doctrine et d'expériences variées, y compris en France [30]. Ce développement peut être aussi considéré comme la conventionnalisation progressive d'un tel langage. Mais nous n'en sommes qu'aux premiers stades d'une telle évolution. Comme un instrument de mesure, le dispositif d'archivage, s'il réussit, devrait devenir une « boîte noire » pour les producteurs comme pour les utilisateurs de données, avec une vision et des protocoles clairs sur la manière d'y intégrer des données et d'y avoir accès. Nous n'en sommes pas là, loin s'en faut.

## 7 Pérenniser la pérennisation…

Les atouts du dispositif mis en place pour le projet pilote sont sa solidité, sa généricité et sa souplesse : solidité parce que le CINES et le CC-IN2P3 sont des structures stables, pérennes et expérimentées. Généricité parce que les moyens matériels et logiciels mis en œuvre pour l'archivage des données orales peuvent être utilisés sans changement pour les autres domaines des SHS quels qu'ils soient. Il s'ensuit une économie d'échelle considérable, et une réduction de la déperdition d'énergie au niveau des équipes de recherche, pour des activités qui ne relèvent pas des métiers des SHS. Souplesse parce que chaque domaine des SHS peut continuer à utiliser les métadonnées standards propres au domaine et disposer d'une interface d'accès aux données conforme à ses besoins. Cette souplesse est mise en évidence par le fait que les deux pôles du CRDO ont pour l'instant retenu deux localisations différentes pour l'implantation de leur application d'interface : l'une sera installée immédiatement sur le site du CC-IN2P3, l'autre pour quelque temps dans un laboratoire distant. Il en résulte, pour les autres domaines des SHS, que les autres centres de ressources numériques concernés (pour le texte, pour les images, etc.) pourront choisir la configuration la mieux adaptée à leur situation. Ces atouts du projet d'archivage ne doivent pas cacher les conditions à réunir pour le pérenniser.
Un premier volet concerne le dispositif humain et technique que préfigure le projet pilote. Il convient de fiabiliser les processus mis en œuvre : les logiciels (iRods [31] pour l'abstraction des systèmes de fichiers, des transferts et des actions déclenchées automatiquement par exemple pour préparer les données à diffuser ; Fedora Commons [26] pour les vues abstraites sur les données archivées) testés, installés, « réglés » pour un bon fonctionnement, incorporent, une fois opérationnels, toute une expérience à mémoriser (à archiver, donc…). Les choix techniques qui ont présidé à la sélection de tel ou tel composant doivent être documentés. Il en va de même des choix organisationnels. On l'a vu supra, il paraissait naturel de ne pas considérer directement les producteurs de données orales comme autant de « services versants » pour l'archive et de confier ce rôle à l'intermédiaire métier qu'est le CRDO. Maintenir la compréhension de ce positionnement compte pour le maintien et l'extension du projet pilote d'archivage : la situation est-elle comparable pour d'autres types de documents ?

Pour les reconstitutions archéologiques en 3 dimensions, par exemple, faut-il parier également sur un truchement du même ordre ? Par ailleurs, une grande rigueur technique s'impose en matière de gestion des évolutions de l'infrastructure et de surveillance de la continuité du service. Pour cela, on doit s'appuyer sur des ressources stables et pérennes. C'est là un point critique qui concerne tous les partenaires. La faiblesse principale vient de l'absence de statut et de garantie de ressources pour le CRDO de la part de sa tutelle, fragilité qui caractérise l'ensemble des dispositifs du même type mis en place par le CNRS pour d'autres types de documents. La maîtrise de l'infrastructure mutualisée, la capacité à assurer une maintenance rapide et efficace, à gérer l'enrichissement de l'archive, impliquent des personnels permanents et parfaitement formés à ces tâches. On le sait, l'obtention de postes permanents pour des tâches nouvelles se heurte à l'orientation inverse des politiques publiques. Enfin, doit être progressivement dégagé un modèle financier viable, qui s'appuie sur l'évolution escomptable des volumes et des services nécessaires pour déterminer les coûts à consentir. De multiples études ont été déjà réalisées sur la question des coûts de l'archivage numérique. C'est le cas du projet LIFE (Life Cycle Information for E-literature) ou encore de la « Blue Ribbon Task Force » de la National Science Foundation. Malheureusement, ces projets n'abordent pas ou peu la phase cruciale et décisive de mise en place. Il faut pourtant le savoir, la mise en place d'un dispositif d'archivage ne sera pas immédiatement génératrice d'économies : les laboratoires et les individus conservent leurs équipements et leurs procédures de sauvegarde. Les économies d'échelle n'émergeront que progressivement, quand des transferts de charge pourront être effectuées. Ces transferts supposent une pédagogie de la mutualisation, qui là encore, construise une représentation partagée de la solution. Une première étape est sans doute de faire émerger les « coûts cachés » que représentent les sauvegardes désordonnées actuelles en matériel mais plus encore en temps de travail. La seconde revient à rapporter les coûts du dispositif mutualisé aux volumes sécurisés, à la fiabilité obtenue et au service fourni. Ce qui apparaît comme une dépense importante parce que globalisée l'est moins une fois rapporté aux dépenses invisibles actuelles et à l'apport pour la recherche.

En second lieu, cette infrastructure doit être reconnue, adoptée, soutenue par la communauté scientifique. Le choix d'une totale transparence a été retenu dès le début du projet. Cette transparence doit être aussi relayée par des actions de communication, de concertation, d'écoute auprès de la communauté. Cette visibilité du projet est essentielle et doit préparer la phase opérationnelle au cours de laquelle le pilotage sera scientifique. La finalité est scientifique et c'est l'intérêt de la communauté scientifique qui doit prévaloir à toutes les décisions essentielles, même si, en arrière plan, l'infrastructure s'appuie sur des compétences et des moyens techniques. L'identification de ce qui sera utile et précieux pour le futur, la définition des priorités dans les actions, la mise sur pied de coopérations, la participation à des projets internationaux doivent être guidées par les chercheurs.

En troisième lieu, les données archivées sont produites, pour la plupart, dans le cadre des activités d'équipes de recherches appartenant à des organismes publics. Leur conservation relève donc, en première analyse, de la responsabilité des Archives nationales. Leur prise en charge par l'infrastructure Adonis requiert un accord et une collaboration avec ces dernières afin de faire coïncider les impératifs de conservation du patrimoine et les exigences d'une base d'information scientifique vivante.

Un dernier volet de la stabilisation du dispositif d'archivage concerne le contexte de réorganisation de l'Etat et de la recherche. Mais ce contexte est générateur

avant tout d'interrogations et d'inquiétudes. La première est celle de l'articulation inexistante actuellement entre ce dispositif à visée nationale, et les démarches qu'entament ou que ne manqueront pas d'entamer d'autres acteurs de la recherche : les nouveaux pôles universitaires (PRES et autres campus), les régions. La convergence des approches n'est pas assurée. Il n'existe d'ailleurs pas de « juge de paix » qui pourrait y œuvrer, et les forces centrifuges semblent dominer. La seconde inquiétude porte sur la volonté comme sur la capacité actuelle de l'Etat à une action cohérente et suivie en matière de Très Grandes Infrastructures et de Recherche en SHS. La lourdeur des investissements et les partenariats internationaux sécurisent pour une part les TGIR « historiques », en physique par exemple. Pour l'heure, malgré des effets d'annonce, les TGIR en activité en SHS occupent une portion congrue (1,5% du budget total annoncé des TGIR fin 2008…) et se résument sauf erreur… au TGE Adonis, lui-même menacé par les évolutions internes du CNRS, par la place mouvante et incertaine des SHS en son sein ainsi que par les désaccords entre ministère et CNRS tant sur la politique même d'équipement des SHS que sur les répartitions des rôles pour y parvenir.

## 8 Conclusion

La mise en place de l'archivage des données de la recherche en SHS peut contribuer à l'introduction de nouvelles manières de travailler. En concordance avec la mise en place du fonctionnement par projets « lourds » (à l'échelle des SHS…), l'archivage contribue à une vision « instrumentée » de la recherche qui dépasse le modèle de l'artisan ou de l'atelier d'artiste. C'est un tournant « industriel » qui s'esquisse peut-être. La préoccupation de la pérennisation, ici agissant a posteriori, sur des données déjà constituées, doit maintenant être intégrée dès l'amont : les projets ANR gagneraient à prévoir d'emblée les conditions de la pérennisation et du partage des données numériques qu'ils produisent. Cette évolution implique que la constitution de données « durables » soit considérée comme une dimension intrinsèque de l'activité de recherche, occupante une place spécifique dans l'évaluation des chercheurs et des laboratoires, sans quoi elle conservera une position ancillaire, gênant par là-même la perception des enjeux intellectuels sous-jacents. La stabilisation des données et de leur accès peut contribuer enfin à maintenir des liens étroits, directs (« cliquables ») entre les publications désormais souvent numériques et les données sur lesquelles elles s'appuient, à l'instar de ce qui a été exposé supra pour les observatoires virtuels en astronomie. Il en résulte de nouvelles manières de croiser les approches ou de vérifier les hypothèses et les conclusions d'une recherche.

Nous risquons aujourd'hui d'être submergés un « tout numérique » qui ne réussit pas à déterminer vraiment ce qui est précieux, ce qui doit rester vivant et qui accumule les méga, les téra, les pétaoctets au motif que « ça peut toujours servir » : « … aujourd'hui, la technique incite à tout garder, quelle que soit la nature de l'objet concerné. […] on ne conserve plus parce que c'est important, mais c'est parce que l'on conserve que c'est important. Ou plutôt qu'on lui permet de le devenir » [2]. C'est La mémoire saturée [1], où le présent et les futurs possibles n'éclairent plus ce qui du passé est à conserver et ce dont il faut faire table rase. La taille de la Toile et les volumes nécessaires à son indexation en sont l'écho : « Les différents datacenters de Google représenteraient à eux seuls une capacité de 200 pétaoctets de stockage » [2], soit 200 000 téraoctets ou 200 millions de giga-octets, ce qui ramenés à notre échelle, à quelque chose de plus familier, correspond au moins à une centaine de milliards de livres de taille moyenne (en texte seul, hors

mise en page et images). « Aujourd'hui, la mémoire et son culte font office d'"agents de liaison" entre un passé fantasmé, un présent inquiétant et un futur indéfinissable » [2]. P. Ricoeur [3], sous le titre « la mémoire empêchée » reprend à Freud [32] l'opposition entre la répétition du passé et sa remémoration, c'est-à-dire son élaboration, voire sa perlaboration, avec ce que cela suppose de choix et d'oubli. La répétition conduit à la pétrification du passé, la remémoration à sa réorganisation en fonction du présent et des futurs envisageables. L'hystérie mémorielle actuelle se conjugue aux nouvelles formes de momification que permet le numérique pour rendre difficiles histoire, oubli et élimination, alors que, pour reprendre M. Augé : « L'oubli est nécessaire à l'individu comme à la société » [33] et qu'« [il] est la force vive de la mémoire » (ibid.). Sans doute faut-il, dans le domaine de la recherche comme sur le plan personnel (ré)apprendre à oublier, pour savoir reconnaître et préserver ce qui est réellement précieux, ce qui doit rester vivant, puisqu'aussi bien « la mémoire est l'organisation collective d'un oubli sélectif » (Rony Brauman [34]).

**Remerciements**
Je remercie Claude Huc pour m'avoir permis de reprendre une partie du matériel d'un commun article sur cette thématique [35]. Je remercie également pour leurs éclairants commentaires et suggestions tant mes collègues du projet pilote Adonis sur les données orales (Bernard Bel – LPL ; Pascal Dugénie – CINES ; Michel Jacobson – DAF ; Thomas Kachelhoffer – CC-IN2P3 ; Nicolas Larrousse – CINES) que ceux du projet archivage numérique d'EDF R&D (Ariane Bonneau ; Thierry Chauvier – à qui je dois les figures 4 et 5 ; Martine Le Corroller).

## 9 Bibliographie

1. Robin, R. (2003) *La mémoire saturée*. Paris : Stock.
2. Hoog, E. (2009) *Mémoire année zéro*. Paris : Seuil.
3. Ricoeur, P. (2000) *La mémoire, l'histoire, l'oubli*. Paris : Seuil.
4. A. Ernaux *Les années* Gallimard 2008
5. CCSDS – Consultative Committee for Space Data Systems. http://public.ccsds.org/
6. OAIS – CCSDS, 650.0-B-1, Reference Model for an Open Archival Information System (OAIS) ,ISO 14721, janvier 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf
7. DUBLIN CORE – Jeu de métadonnées faisant l'objet d'un large consensus http://dublincore.org/
8. PAC – Plate-forme d'Archivage au CINES http://www.cines.fr/-l-application-PAC-.html
9. SPAR – Système de préservation et d'archivage réparti de la Bibliothèque nationale de France (BnF). http://www.bnf.fr/pages/infopro/numerisation/num_spar.htm
10. Site du Ministère de l'Enseignement Supérieur et de la Recherche (MESR) sur les TGIR http://www.roadmaptgi.fr/

11. Site du CNRS sur les TGIR http://www.cnrs.fr/fr/recherche/ups3019/feuilles-route-infrastructures.htm

12. OVS – Observatoire virtuel en astronomie – participation du CDS de Strasbourg http://cdsweb.u-strasbg.fr/CDS-f.gml

13. ESFRI – Coordination européenne des TGIR http://cordis.europa.eu/esfri/

14. Principales TGIR européennes en SHS : DARIAH http://www.dariah.eu/ ; CESSDA http://www.cessda.org/ ; CLARIN http://www.clarin.eu/

15. TGE Adonis – http://www.tge-adonis.fr/

16. Conclusions d'O. Bärring sur la mutualisation de l'hébergement et de l'archivage http://www.tge-adonis.fr/?Le-point-de-vue-d-Olof-Barring-du

17. Baude, O. (ed) (2006) *Corpus oraux – Guide des bonnes pratiques* 2006. Paris : Presses universitaires d'Orléans & CNRS Éditions. Egalement en ligne : http://hal.archives-ouvertes.fr/hal-00357706/fr/

18. DGLFLF – Délégation Générale à la Langue Française et aux Langues de France. http://www.dglf.culture.gouv.fr/ et ressources mises en place http://www.corpusdelaparole.culture.fr/

19. CRN – Centres de resources numériques SSH – pour les manuscrits : http://www.cn-telma.fr/ , les textes : http://www.cnrtl.fr/, les images : http://www.cn2sv.cnrs.fr/, et les données géographiques : http://www.m2isa.fr/ - pour l'oral, cf. [20].

20. CRDO – Centre de ressources pour la description de l'oral, dans son versant parisien : http://crdo.risc.cnrs.fr/ et aixois : http://crdo.fr/

21. CINES – Centre Informatique National de l'Enseignement Supérieur http://www.cines.fr/ et particulièrement http://www.cines.fr/-D-I-S-T-.html pour l'archivage pérenne

22. CC-IN2P3 – Centre de Calcul de l'Institut national de physique nucléaire et de physique des particules http://cc.in2p3.fr/

23. PIN] Groupe de travail sur la préservation de l'information numérique http://www-pin.aristote.asso.fr/

24. Persée – http://www.persee.fr/

25. HAL – http://hal.archives-ouvertes.fr/

26. FEDORA –. http://www.fedora-commons.org/

27. WIKI du projet pilote d'archivage des données orales au TGE Adonis – http://www.tge-adonis.fr/wiki/index.php/Accueil_Projet_pilote

28. Desrosières, A. (2000) *La politique des grands nombres - Histoire de la raison statistique.* Paris : La Découverte.

29. Desrosières, A. (2008) *L'argument statistique, vol. I: Pour une sociologie historique de la quantification*. Paris: Presses de l'École des Mines.

30. Banat-Berger, F. ; Duplouy, L. ; Huc, C. (2009) *L'archivage numérique à long terme – les débuts de la maturité ?* Paris : La Documentation Française.

31. iRODS – https://www.irods.org/

32. Freud, S. « Remémoration, répétition, perlaboration », in *La technique psychanalytique,* PUF, 1970

33. Augé, M. (1998) *Les formes de l'oubli*. Paris : Payot.

34. Brauman, R. & Finkielkraut, A. (2006) *La discorde : Israël-Palestine, les Juifs, la France* – Conversations avec Élizabeth Levy. Paris: Fayard.

35. B. Habert et C. Huc, Building together digital archives for research in social sciences and humanities, *Social Science Information*, vol. 49, n°3, septembre 2010, p. 415-443.

# Approche semio-rhétorique des couplages texte-mouvement dans le discours numérique

**Alexandra Saemmer**

alexandra.saemmer@univ-paris8.f
Laboratoire Paragraphe,
Université Paris 8

**Résumé** : Le but du travail de recherche présenté ici est de caractériser avec précision les particularités sémio-rhétoriques du discours numérique. Il s'agit plus particulièrement d'étudier les « figures d'animation média ». Seront d'abord identifiées les unités sémiotiques impliquées dans le couplage entre mouvement et texte. Seront ensuite analysés les mécanismes de construction de sens dans ces ensembles "pluricodes". Le modèle théorique est illustré par des exemples prélevés dans un corpus de bannières publicitaires en ligne et de créations poétiques sur support numérique.

**Mots-clés :** figure, signe, rhétorique, sémiotique, discours numérique, bannières publicitaires, conventions.

## 1 Introduction : Des unités sémiotiques du mouvement vers la figure

Après un temps où la critique circonscrivait les caractéristiques des écrits numériques par des discours centrés sur la notion générale de l'« hypertexte » (entre autres Landow 1992, Vandendorpe 1999) et le rapport numérique-papier (entre autres Eco 2005, Baccino/Colombi 2001), où elle fondait ses analyses sur des approches socio-sémiotiques (Le Marec 2001, Landowski 1989), se concentrait sur des aspects ergonomiques (Bastien/Scarpin 1993) ou essayait d'ériger des listes de recommandations à l'égard des web-designers (Nielsen/Loranger 2008), le but du travail de recherche présenté dans cet article est de caractériser avec précision les particularités sémio-rhétoriques du discours numérique[1]. Cette recherche a trouvé son point de départ dans des observations réalisées par plusieurs chercheurs depuis des années sur les œuvres d'art et de littérature numérique ; des approches plus

1 Pour une approche sémiotique des bannières publicitaires convoquant des figures de style de la rhétorique classique, voir par exemple Nicole Pignier. Le recours à des figures rhétoriques propres au texte s'imposait également à nous dans un premier temps (Bouchardon, Clément, Saemmer 2007). Elle nous paraît aujourd'hui dangereuse compte tenu du caractère pluricode du texte numérique animé et manipulable.

récentes des interfaces Web et les bannières publicitaires, ont également été intégrées dans ce travail en cours d'élaboration par des chercheurs du laboratoire Paragraphe : Philippe Bootz, Serge Bochardon, Jean Clément et l'auteur de cet article[2].
Nous avons opté pour une approche qui s'inspire du structuralisme tout en plaçant la notion de point de vue au cœur du modèle. Nous tentons ainsi d'élaborer une catégorisation des « signes » et « figures » du discours numérique sans qu'il s'agisse pour autant de fournir un système de « briques » immuable, dans lequel il suffirait de puiser pour exprimer par exemple automatiquement le dynamisme, l'urgence ou la tristesse. L'activation d'un « trait signifiant » reste hautement déterminée par le média (le texte ou l'image) auquel il est appliqué, par l'isotopie et le contexte de lecture.

Notre modèle prend en compte trois caractéristiques fondamentales des productions numériques, qui sont

- l'interactivité : que faisons-nous lorsque nous interagissons ? Nos gestes sont-ils signifiants ?
- La programmation : l'écrit numérique est généralement programmé. Quelle est l'influence de ce fait sur sa signification ?
- L'animation : comment le mouvement agit-il sur la signification d'un texte ?

Dans l'une des contributions de l'auteur du présent article à cette recherche, sont étudiées plus particulièrement des « figures d'animation média » identifiées à partir d'un corpus d'une centaine de sites web commerciaux et de bannières publicitaires en ligne et d'une vingtaine de créations littéraires[3]. Dans la présentation des résultats proposée ici, il s'agira d'abord de circonscrire les « unités sémiotiques » qui sont impliquées dans un couplage entre un mouvement et un média. Seront ensuite analysés les mécanismes de construction de sens dans ces ensembles pluricodes[4]. Pour chaque partie sera d'abord présenté un modèle archétypal construit à partir des observations faites dans le corpus ; ce modèle sera illustré par des exemples prélevés directement dans le corpus.

## 2 Les unités sémiotiques du mouvement

Tout visiteur de sites web est régulièrement confronté à ces annonces commerciales de toutes couleurs, animées et parfois manipulables[5], que l'on nomme « bannières » dans le jargon du e-marketing. Afin de montrer comment dans le discours numérique le sens se construit, observons l'un des grands classiques de la web-publicité (reconstruit ici sous forme de modèle dépouillé d'un certain nombre d'éléments contextuels) : une bannière dans laquelle, sur un fond plus ou moins multicolore, clignote à rythme rapide (environ 3 pulsations par seconde) le mot « soldes ».
Exemple archétypal d'un mouvement insistant

2 Un livre intitulé Signes et figures du discours numérique est en cours de finalisation.
3 Le corpus en cours d'élaboration est consultable à l'adresse <[http://www.alexandrasaemmer.fr/corpus/]>.
4 Le texte à l'écran est pluricode dans le sens d'une superposition de systèmes sémiotiques. Dans cet article sera principalement étudiée la superposition texte et mouvement.
5 Les ensembles pluricodes fondés sur une mise en relation entre contenu média et manipulation seront également étudiés dans le livre Signes et figures du discours numérique en cours d'élaboration.


Le mot « soldes » clignote à rythme rapide sur cette bannière qui pourrait se trouver sur un site commercial.
Modèle archétypal construit à partir d'exemples du corpus, consultable à l'adresse *http://www.alexandrasaemmer.fr/corpus/obsessionnel/soldes1.html*

Commençons par décomposer cet ensemble pluricode en dégageant deux systèmes signifiants : D'un côté le mot « soldes », avec ses significations que nous pouvons relever dans une encyclopédie : « Les soldes consistent à vendre avec une forte réduction sur le prix […] Elles ne peuvent être réalisés qu'au cours de deux périodes par année civile […] Leur durée maximale autorisée est de six semaines. »[6] D'un autre côté le mouvement du « clignotement ». Le mot « soldes » rentre dans la catégorie des « signes au sens strict », c'est-à-dire des signes arbitraires et non correspondants (le mot allemand pour « soldes » est « Schlussverkauf », et ne présente aucune similitude sonore avec le mot français). Un mouvement comme le clignotement fait partie des « signes motivés par ressemblance créés par des découpages non correspondants » (Klinkenberg 193), donc des icônes au sens que Charles Sanders Pierce a donné à ce terme. Le clignotement « rappelle », par sa matérialité même, des phénomènes du monde physique : l'apparition et la disparition rapides de signaux lumineux sur les bords d'autoroutes, l'animation des enseignes lumineuses dans les grandes villes, le clignotement des boutons sur des appareils électriques, ou alors notre cœur qui bat à un rythme rapide… Contrairement au « signe au sens strict », quelque chose du monde physique survit dans l'icône, et est reconnu comme tel.
Laissons pour l'instant de côté ce que nous associons éventuellement au clignotement comme significations symboliques : le danger, le stress… Laissons aussi de côté le média qui est mis en mouvement par clignotement : le mot « soldes » avec ses significations. Concentrons-nous sur le mouvement : Dans cet exemple publicitaire, un élément textuel apparaît et disparaît périodiquement et à rythme rapide. Est-ce l'apparition et la disparition qui sont déterminantes dans la construction de la signification de ce mouvement ?
Dans certains exemples du corpus, il arrive qu'un élément apparaît et disparaît très lentement, ou qu'il apparaît sans pour autant disparaître. Qu'est-ce que ces déclinaisons de l'apparition / disparition ont cependant en commun ? Certes, d'un point de vue technique, elles se « fabriquent » à peu près de la même manière ; il est pourtant probable que le lecteur n'y associe pas les mêmes expériences dans le monde physique, et qu'il ne fait pas signifier ces mouvements de la même façon. Une apparition-disparition lente par exemple ne fait-elle pas plutôt penser à l'apaisant va-et-vient des vagues de la mer qu'au clignotement d'un bouton d'alerte sur un appareil électrique ?
Si nous comparons en revanche une bannière où le mot « soldes » clignote rapidement et de manière réitérée, une bannière ou ce mot change de couleur au même rythme, et une bannière ou le mot change rapidement et périodiquement de taille, nous pouvons énoncer l'hypothèse que les trois animations seront perçues de manière assez semblable par le lecteur, malgré les différences évidentes entre les caractéristiques visuelles mobilisées ; en tout cas, les analyses faites à partir du

6 <[http://fr.wikipedia.org/wiki/Soldes]>.

corpus permettent d'affirmer que les créateurs de bannières utilisent visiblement ces trois animations dans des buts semblables.


Figure 2 : Le mot « soldes » change de couleur ou de taille à rythme rapide et de façon réitérée sur ces deux bannières qui pourraient se trouver sur un site commercial. Modèles archétypaux construits à partir d'exemples du corpus, consultables à l'adresse *http://www.alexandrasaemmer.fr/corpus/obsessionnel/soldes2.html, http://www.alexandrasaemmer.fr/corpus/onsessionnel/soldes3.html.*
Exemples du corpus consultables en capture vidéo :
http://www.alexandrasaemmer.fr/corpus/obsessionnel/

Alors qu'une catégorisation du mouvement selon des comportements purement visuels s'imposait de prime abord à cause d'une ressemblance forte entre les processus de fabrication (en effet, les outils de créations comme les générateurs de bannières[7] et les logiciels d'intégration multimédia classent les « effets » disponibles par comportements visuels, sans poser la question de leur signification), il semble donc possible d'établir des catégories d'unités sémiotiques du mouvement qui dépassent la simple énumération de caractéristiques visuelles, et regroupent les exemples du corpus autour d'unités qui, tout en étant basées sur des processus de fabrication différents, sont perçues de façon semblable par le lecteur.
Ce phénomène, dont la compréhension et l'exploitation peut avoir notamment des répercussions considérables pour la pratique du web-design, s'explique par le caractère iconique du mouvement. L'icône est, dans l'animation, un signe motivé par une ressemblance avec les choses. Cette relation est gérée par le « type » dont Jean-Marie Klinkenberg donne la définition suivante : « Il a été constitué par des processus d'intégration et de stabilisation d'expériences antérieures (…) J'ai déjà vu des chats et je sais qu'ils ont des moustaches, qu'ils griffent, et qu'ils miaulent… Bref, j'en connais un bout sur les chats, et ce bout fait partie du type » (Klinkenberg 385). De même, dans beaucoup de contextes culturels, nous connaissons un bout sur le clignotement : les enseignes lumineuses dans la rue, les boutons sur des machines, etc. Le lecteur perçoit donc le clignotement et, grâce au type, l'identifie comme tel. D'autres formes de mouvement avec des caractéristiques visuelles différentes font appel à ce même type, même si elles sont différentes au niveau visuel : le changement de couleur rapide et réitéré avec contrastes forts, le changement de taille rapide et répété. Pour la reconnaissance du type, le rythme et la réitération font donc partie des caractéristiques hautement déterminantes, alors que ce n'est pas le cas pour l'apparition et la disparition seules.
Ces mêmes caractéristiques liées au rythme et à la réitération jouent un rôle important dans la catégorisation des « unités sémiotiques temporelles » entreprise par les chercheurs du MIM (Laboratoire de recherche et création musicales et

7 Exemples de générateurs de bannières : <http://www.3dtextmaker.com/cgi-bin/3dtext.pl>, <http://www.generateur.net/banniere_flash/>

multimédia). 16 UST ont été identifiées jusqu'alors à partir d'un corpus musical[8]. Elles ont été validées par des tests auprès d'auditeurs, qui semblent en effet pouvoir communément reconnaître ces unités grâce à un certain nombre de caractéristiques ; souvent, celles-ci touchent au rythme et à la réitération. Les chercheurs du MIM ont décidé d'attribuer aux UST des noms résumant la présence d'un certain nombre de ces caractéristiques, et de leur adjoindre une description sémantique. À l'UST « sur l'erre » correspond ainsi, dans la description sémantique, l'« image d'un bateau » « qui, ayant affalé ses voiles ou coupé son moteur, continue à avancer sur terre grâce à sa vitesse acquise, ralenti lentement par la résistance de l'eau » ; elle est représentée par un extrait musical où le niveau sonore baisse progressivement jusqu'à devenir inaudible[9]. L'UST « trajectoire inexorable » est décrite par une prévisibilité de la non-fin : « qui ne finit pas d'avancer », par exemple ; l'un des extraits sonores sur le site du MIM est ainsi caractérisé par un son qui ne finit pas de monter[10]. L'UST « obsessionnel » est décrite par son « caractère insistant » ; elle est illustrée par un extrait musical composé d'une note de piano répétée plusieurs fois, de façon saccadée et rapide[11].
Lorsqu'on compare ce son « obsessionnel » avec le clignotement, le changement de taille ou le changement de couleur rapides du mot « soldes » dans les bannières, l'hypothèse s'impose que les UST constituent des unités abstraites, implémentables dans différents médias : du son, du texte, de l'image. Dans le cas de « l'obsessionnel » sonore comme de son correspondant visuel, ce sont les caractéristiques « par répétition », « mécanique », « à réitération régulière », « à énergie cinétique constante » qui sont déterminantes pour la reconnaissance du type[12]. L'UST sonore « sur l'erre », caractérisée par une baisse progressive du volume sonore, peut trouver son correspondant visuel dans la disparition progressive d'un élément média, mais aussi dans un mouvement de floutage, ou un changement de position d'un élément avec ralentissement progressif par perte d'énergie[13].

### 2.1 Exemples du corpus : Trajectoires

Regardons de près deux exemples publicitaires du corpus s'apparentant à l'UST sonore « trajectoire » (caractérisé par un son « qui n'en finit pas d'avancer »). Dans le domaine visuel, ce mouvement est soit continu et « infini » comme dans le domaine sonore (« trajectoire inexorable ») ; soit il atteint un niveau maximum et s'arrête (« trajectoire à but défini »)[14].

8 Voir présentation des UST sur le site du MIM, <[http://www.labo-mim.org/, onglet « recherche », « UST »]>.
9 Pour écouter l'extrait : <[http://www.labo-mim.org/site/index.php?2008/08/22/45-sur-l-erre]>.
10 Pour écouter l'extrait : <[http://www.labo-mim.org/site/index.php?2008/08/22/28-trajectoire-inexorable]>.
11 Pour écouter l'extrait : <[http://www.labo-mim.org/site/index.php?2008/08/22/36-obsessionnel]>.
12 Exemples de l'unité sémiotique du mouvement « obsessionnel » consultables sous forme de capture vidéo : [http://www.alexandrasaemmer.fr/corpus/obsessionnel/>.
13 Exemples de l'unité sémiotique du mouvement « sur l'erre » consultables sous forme de capture vidéo : [http://www.alexandrasaemmer.fr/corpus/surlerre/]>.
14 Autres exemples de l'unité sémiotique du mouvement « trajectoire » consultables sous forme de capture vidéo : [http://www.alexandrasaemmer.fr/corpus/trajectoire/>

Sur le site web des parfums de la marque Kenzo[15], des éléments textuels donnent l'impression de s'approcher progressivement du champ de vision du lecteur par changement de taille. Puis ils grandissent tellement qu'ils débordent du cadre de l'espace d'affichage. Le lecteur a l'impression que ce mouvement d'agrandissement se poursuivra indéfiniment hors-cadre ; entre-temps, les mêmes éléments textuels réapparaissent en très petite taille pour s'agrandir à nouveau. Cette animation s'apparente donc bien à l'UST « trajectoire inexorable ».


Figure 3 : Les éléments textuels changent progressivement de taille, puis disparaissent vers le hors-champ tout en réapparaissant en taille réduite. Exemple du corpus consultable sur www.kenzo.fr, onglet parfum.

Dans une bannière publicitaire pour le Club Med, un élément textuel juxtaposé à une image, d'abord flou, devient progressivement lisible, décrivant ainsi une « trajectoire à but défini ».


Figure 4 : L'élément textuel, d'abord flou, devient progressivement lisible. Exemple du corpus. Capture vidéo consultable à l'adresse *http://www.alexandrasaemmer.fr/corpus/trajectoire/loupe.mov*

## 3 La question de la signification

En passant en revue les 16 UST, l'on constate que les chercheurs du MIM ont adopté des termes souvent très suggestifs pour les catégoriser ; par ailleurs ont-ils procédé non seulement à une description morphologique, mais à des descriptions sémantiques parfois très imagées. En effet, une question s'impose depuis le début de cet article : qu'est-ce qu'un mouvement *signifie* ? Pour plus de précision, il faudra immédiatement nuancer la question en la reformulant ainsi : qu'est-ce qu'un mouvement signifie *pour un lecteur* et *en fonction du média impliqué* ? Car la construction de la signification ne peut évidemment pas être séparée du contexte de réception, ni des habitudes et connaissances du lecteur. François Rastier fait ainsi très justement remarquer que « le sens n'est ni dans l'objet, ni dans le sujet, mais 'dans' leur couplage, au sein d'une pratique sociale » (Rastier 125).

15 <[http://www.kenzoparfums.com/FR/kenzo.html]>.

Via des expériences physiques communes des déclinaisons visuelles de l'obsessionnel dans le monde physique, un lien de ressemblance se crée entre le clignotement et la chose. Cette expérience commune donne lieu à des interprétations diverses au niveau symbolique (le « symbole » est défini par Jean-Marie Klinkenberg comme un signe arbitraire créé par des découpages correspondants (194)) : grâce à sa connaissance des signalisations clignotantes sur les routes, le lecteur peut ainsi associer au clignotement des significations comme « attention » ou « danger » sans que cette signification soit directement motivée par les caractéristiques physiques ; le fait de sentir son cœur battre au rythme de l'obsessionnel, peut également lui faire créer un lien avec « l'émotion forte » ; le fait d'avoir vu clignoter le mot « soldes » sur des sites commerciaux lui suggère d'attribuer la signification « urgence » à ce mot. Seront appelés « traits signifiants possibles » les caractéristiques que le lecteur associe à un mouvement parce qu'elles ont été sélectionnées au long des expériences répétées dans une culture. La récurrence des mêmes combinaisons texte-mouvement dans les mêmes contextes crée une attente chez le lecteur, qui peut être satisfaite ou troublée dans une situation de lecture précise.

## 4 Couplages média conventionnels

Une unité sémiotique du mouvement est donc potentiellement porteuse d'un certain nombre de traits signifiants. Ceux-ci sont ensuite actualisés en fonction du média sur lequel le mouvement est appliqué, ainsi qu'en fonction du contexte de lecture. Par le terme « figure d'animation média » sera désignée une relation entre des médias et un mouvement, dans laquelle la sémiose est basée sur des processus d'intersection de traits signifiants associés au mouvement, au média et au contexte. Le mouvement apporte une confirmation, une précision ou un brouillage des traits signifiants du media et modifie le poids de ceux-ci dans la construction du sens.

Plus le champ d'intersection entre les traits signifiants du mouvement et les traits signifiants du média est étendu, et plus l'union des traits répond aux attentes du lecteur en fonction du contexte, plus la sémiose relève du « couplage conventionnel ». Le terme « convention » est utilisé ici non pas dans le sens d'une « norme » immuable constituée à l'intérieur de l'objet, mais comme un « degré zéro » dans le sens de « ce que le lecteur attend dans cette position » (Groupe μ 42).

Revenons au mot « soldes » qui clignote à rythme rapide sur un site commercial. Ce mouvement qui s'apparente à l'UST « obsessionnel » est porteur de traits signifiants possibles : l'urgence, le danger, le stress, l'excitation. Le contenu média sur lequel l'animation est appliquée, mobilise également un certain nombre de traits signifiants : « événement commercial », « réduction de prix », « période limitée dans le temps », « urgence d'en profiter ». Par une intersection de tous ces traits se forme un ensemble signifiant qui pourrait être paraphrasé ainsi : « la période des soldes est un événement important, limité dans le temps, il y a urgence d'en profiter ». Les facteurs d'iconicité comme le battement de cœur rapide soutiennent cet ensemble signifiant. Certains traits comme « danger » potentiellement associés au mouvement, ne rentrent pas dans le champ d'intersection ; ils ne semblent pourtant pas troubler la sémiose.

Bien sûr, lors de la sémiose, les habitudes du lecteur ainsi que d'autres facteurs contextuels jouent un rôle important : par exemple, les déclinaisons visuelles de l'obsessionnel créent chez certains malades d'épilepsie un risque de crise, qui est susceptible de réactiver le trait signifiant « danger ». On peut néanmoins supposer que l'ensemble mouvement-média répond globalement aux attentes du lecteur : les

traits signifiants partagés par le mouvement et le média sont si nombreux que le mouvement joue principalement un rôle de confirmation ou d'accentuation du sens mobilisé par le média. Les couplages conventionnels mettent ainsi de l'emphase sur les contenus, précisent leur signification et captent le regard du lecteur sans troubler ses attentes.

### 4.1 Exemple du corpus : Trajectoire à but défini

Reprenons à présent la bannière publicitaire pour le Club Med (voir captures d'écran plus haut). Le mouvement de « défloutage » s'apparente à l'UST « trajectoire à but défini », et mobilise des traits signifiants comme « matérialisation rassurante » ou « révélation ». Il est appliqué à un texte présentant une offre commerciale limitée dans le temps, qui insiste sur l'incitation « partez » : « Partez entre le 10 janvier et le 26 avril 09 ». Des traits signifiants comme « invitation au voyage », « prendre des vacances », et aussi « limitation de l'offre dans le temps » sont donc potentiellement activés chez le lecteur. L'isotopie (images entourant le texte) est constituée de scènes de vacances, auxquelles nous pouvons associer les traits « plaisir », « loisir », « sport », « agrément ». L'agrément et le plaisir sont mobilisés d'une part par la matérialisation rassurante associée à la « trajectoire à but défini », et d'autre part par l'isotopie suggérant également un certain nombre de plaisirs : tout comme le texte se stabilise progressivement et devient visible, les plaisirs de vacances pourraient devenir réalité pour le lecteur. En revanche, l'offre est limitée dans le temps. Le trait signifiant « urgence » pourrait troubler la sémiose. Le défloutage est pourtant immédiatement suivi d'un refloutage mobilisant les traits signifiants de l'UST « sur l'erre » (un mouvement de disparition avec des traits signifiants comme « perte ») ; la frustration ou l'énigme s'installent à nouveau. Pour le lecteur, il y a donc urgence de profiter de cette offre commerciale avant qu'elle ne redevienne un mirage.

## 5 Cas particulier du couplage conventionnel : le ciné-gramme

Le ciné-gramme constitue un cas particulier du couplage conventionnel : les traits signifiants mobilisés par le média et le mouvement se trouvent en intersection quasiment totale de sorte qu'une impression de synonymie se crée entre la signification du mouvement et celle du média. Le jeu avec une quasi-synonymie (mais qui n'est jamais entière), fait que le ciné-gramme peut dégager la même fascination que le calligramme dans l'univers papier.

### 5.1 Exemple archétypal : Obsessionnel

Dans un exemple archétypal reconstruit à partir de plusieurs exemples du corpus, observons le mot « cœur » qui clignote à rythme régulier et rapide sur un site de rencontres :


Figure 5 : Le mot « cœur » clignote à rythme rapide et de façon réitérée sur cette bannière qui pourrait se trouver sur un site de rencontres. Modèle archétypal construit à partir d'exemples du corpus, consultable à l'adresse *http://www.alexandrasaemmer.fr/corpus/obsessionnel/coeur-rencontres.swf*.

Les traits signifiants possibles associés à l'obsessionnel sont « urgence », « excitation ». Les traits signifiants potentiellement mobilisés par le mot cœur sont « organicité », « émotions », « organe vivant ». Les traits signifiants activés par l'isotopie (le site de rencontres) s'organisent autour des notions « émotions », « rencontre ». Le champ d'intersection des traits communs au média, au mouvement et à l'isotopie, est donc très étendu.
Bien sûr, le cœur ne clignote pas réellement dans la poitrine des êtres vivants ; grâce au type auquel le clignotement fait référence, le mot « cœur » clignotant semble néanmoins plus proche du référent « cœur » que le mot statique. Cette dimension littéralement « organique » est aussi présente lorsque le mot « soldes » clignote sur l'écran ; mais le *référent* du mot « soldes », bien que celles-ci soient effectivement limitées dans le temps et qu'il faille donc en profiter d'urgence, n'est pas caractérisé par un clignotement, alors que le mouvement de systole et de diastole du référent « cœur » semble pouvoir être imité par une unité sémiotique du mouvement qui s'apparente à l'UST « obsessionnel ». En même temps, un recouvrement complet des traits signifiants associés au mouvement et au média est évidemment impossible. Les plus beaux ciné-grammes jouent avec ce rêve d'identité entre deux systèmes sémiotiques en repoussant les frontières du couplage conventionnel : par exemple pourraient-ils, « en un seul mot », nous raconter l'histoire d'un cœur qui s'emballe lors d'une rencontre furtive, ou qui s'arrête de battre.

### 5.2 Exemple du corpus : Trajectoire inexorable

Dans une publicité pour les parfums Kenzo prélevée dans le corpus (voir captures d'écran plus haut), les noms des parfums sont animés en apparaissant lentement sur fond blanc et en grossissant progressivement comme s'ils s'approchaient du lecteur avant de disparaître vers le hors champ. Jamais les éléments textuels ne perdent en intensité lumineuse, suggérant ainsi au lecteur que leur mouvement continue indéfiniment, sans jamais s'épuiser. Ce mouvement s'apparentant à l'UST « trajectoire inexorable » mobilise des traits signifiants comme « matérialisation rassurante », « cohérence » et « persistance », qui sont couplés aux traits signifiants potentiellement mobilisés par un parfum : « odeur agréable », « persistance ». Bien évidemment, tout parfum n'est pas caractérisé par une grande persistance ; on dirait même que c'est plutôt l'effacement progressif qui est caractéristique du parfum. Il s'agirait donc dans cette publicité de faire rêver le lecteur à un parfum *idéalement* persistant, suffisamment fort et matériel pour s'imprégner dans la mémoire. La mise en relation entre chaque parfum de la gamme Kenzo et le mouvement de la trajectoire inexorable relève du couplage conventionnel. Le mouvement de l'ensemble des parfums rentre plus spécifiquement dans la catégorie du ciné-gramme : l'envahissement continu de l'espace par les noms de parfum suggère un mouvement centrifuge continu qui fait penser à l'action d'un diffuseur d'odeurs.

## 6 Couplages média non conventionnels

L'animation textuelle n'est pourtant pas réductible à l'exploration de tels effets « mimétiques ». Dans un couplage d'animation non conventionnel, la relation entre les médias, le mouvement et le contexte est certes également fondée sur une intersection de traits signifiants. Le champ d'intersection entre les traits associés au mouvement et les traits associés au média est pourtant plutôt réduit. Par ailleurs, les traits exclus du champ d'intersection entre le mouvement et le média continuent à jouer un rôle important dans la sémiose en fonction du contexte. Un différentiel se crée ainsi entre les attentes du lecteur et l'état réalisé dans l'animation. Pour que ce

différentiel ne soit pas considéré par le lecteur comme une erreur de sens, une médiation s'effectue éventuellement avec l'isotopie de sorte que le lecteur arrive à faire signifier au moins partiellement l'ensemble pluricode. Les couplages non conventionnels s'apparentent ainsi aux figures appelées « tropes » dans le domaine linguistique.

### 6.1 Exemple archétypal : Obsessionnel

Reprenons l'exemple archétypal de la bannière sur un site de rencontres : Cette fois-ci, ce n'est plus le mot cœur, c'est le mot « cerveau » qui est affecté d'un clignotement.

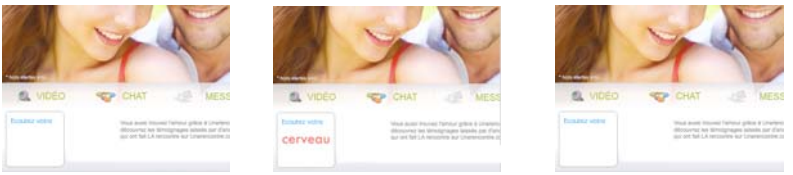


Figure 6 : Le mot « cerveau » clignote à rythme rapide et de façon réitérée sur cette bannière qui pourrait se trouver sur un site de rencontres. Modèle archétypal construit à partir d'exemples du corpus, consultable à l'adresse *http://www.alexandrasaemmer.fr/corpus/obsessionnel/cerveau-rencontres.swf*

Il est probable que dans le contexte d'un site de rencontres, cette animation déconcerte d'abord le lecteur. Des traits signifiants comme « urgence », « danger », « stress » et « vitalité », mais également « montée d'émotions » sont potentiellement mobilisés par le clignotement. L'isotopie (le site de rencontres) mobilise des traits signifiants comme « montée d'émotions », « rencontre amoureuse ». Nous avons vu que le lecteur ne s'étonnerait sans doute guère de voir clignoter le mot « cœur » dans ce contexte. L'organe « cerveau » en revanche, n'est pas caractérisé par un battement physique rapide ; il mobilise par ailleurs des traits signifiants comme « raison », « intellect » et « maîtrise » qui restent en dehors du champ d'intersection entre les traits associés au mouvement et au média.
Si le lecteur se détourne avec le constat que cette animation « ne fait pas sens » parce que le cerveau ne clignote pas dans le monde physique, le processus de production de la figure s'arrête. Compte tenu de l'isotopie, le principe général de coopération pourrait néanmoins être au moins partiellement sauvegardé. Dans ce cas, un contenu compatible avec le reste du contexte est superposé à l'élément « cerveau » identifié comme impertinent[16]. Ce contenu compatible peut être formulé comme « organe humain vivant réagissant à la perception ». Un « lien dialectique » s'établit entre cet élément impertinent et le contenu compatible avec l'isotopie ; il peut être formulé ainsi : « le cerveau est certes le siège de l'intellect, de la raison et de la pensée, mais il est aussi un organe humain vivant qui réagit, comme le cœur, avec émotion à une perception ». Le lecteur sait bien qu'en dépit de l'imaginaire communément mobilisé par le mot « cœur », c'est le cerveau qui est le siège des émotions.

16 Cette description du processus de médiation dans la figure s'inspire du schéma de « production de la figure » proposé par Jean-Marie Klinkenberg (344 ss).

Il apparaît pourtant clairement à quel point la réalisation de cette médiation est conditionnée par l'isotopie : il suffirait de placer cette animation dans le contexte d'un site pédagogique vantant les mérites du *brainstorming* pour que l'animation ne paraisse plus du tout impertinente. En revanche, le fait de trouver une bannière clignotante avec le mot « soldes » sur un site de rencontres, serait sans doute considéré par le lecteur comme foncièrement impertinent.
Le couplage non conventionnel comporte toujours un risque d'incompréhension, à cause du différentiel qui se crée entre les attentes du lecteur et l'état réalisé dans l'ensemble pluricode, et qui ne sera jamais entièrement résorbé par la médiation décrite. Même si certains créateurs de publicité expérimentent ainsi avec la surprise et le trouble provoqués par cette figure afin de créer des « sensations nouvelles » et afin de doter une marque d'une aura anti-conventionnelle, le terrain d'application le plus riche du couplage non conventionnel est le domaine de la littérature et des arts numériques.

### 6.2 Exemple du corpus : Trajectoire à but défini

Le poème numérique *The Last Performance*[17] de Judd Morrissey est un projet d'écriture collaborative à contraintes. Pendant deux ans, des contributeurs ont été invités à mettre à la disposition du projet un certain nombre de textes ou mots-clé autour de questions de réflexion, dont : « Collaboration as Architecture : Double Building », « Concerning lasts made » ou « Consider the style of old words in new times ». Certains de ces textes et mots-clé ont été montés à leur tour sous forme de structure « dansante » (*The dance*) changeant indéfiniment de configuration. Toutes ces structures ont été calquées sur les proportions du « dôme » de la *Dzamija* à Zagreb – un bâtiment qui, au cours de l'histoire, a changé plusieurs fois de fonction : d'abord musée, il a été transformé en mosquée, puis reconverti en musée.

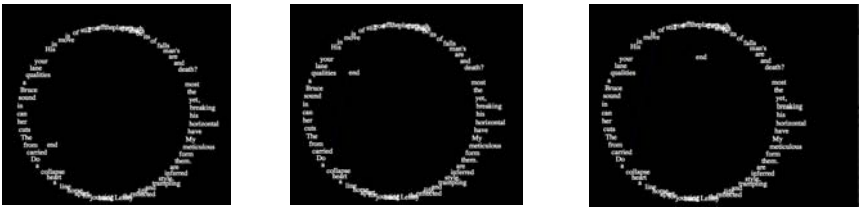

Figure 7 : Dans ce poème numérique intitulé *The Dance*, le mot « end » se différencie des autres mots en décrivant plusieurs trajectoires à but défini. L'œuvre est consultable à l'adresse http://thelastperformance.org/title.php.

Lorsque le lecteur lance l'animation *The Dance*, parmi tous les mots qui se configurent sous différentes formes (cercle, demi-lune, spirale), un seul mot s'échappe et décrit des mouvements de va-et-vient en bas de l'écran. Lorsque tous les autres mots se reconfigurent en cercle, le mot « end » continue à se différencier : il traverse plusieurs fois l'espace noir au centre du cercle formé par les autres mots. Ces « trajectoires à but défini » mobilisent potentiellement des traits signifiants comme « cohérence », « avoir un but », « transformation sans perte de matière ». Le mot « end » en revanche, active plutôt des traits signifiants comme « fin de l'histoire », « fin de la vie », point final ». Le fait de voir le mot « end » constamment

17 <[http://thelastperformance.org/title.php]>.

décrire des trajectoires sur l'écran, va donc certainement à l'encontre des attentes du lecteur, qui y associerait plutôt un mouvement de disparition progressive ou alors une immobilité complète. S'agit-il d'une incohérence de sens ?
Il nous semble que cette figure d'animation média peut rester partiellement interprétable grâce à un processus de médiation entre l'élément considéré comme déconcertant (le mouvement de trajectoire associé au mot « end ») et le contexte. L'isotopie dans cette création est constituée d'abord par le thème général, « the last performance » ; le mot « last » a plusieurs sens en anglais. En tant qu'adjectif, il signifie « dernier », et en tant que verbe, « durer, perdurer ». L'une des significations du mot « last » (durer, perdurer), le motif du dôme ainsi que la métamorphose incessante des formes décrites par tous les autres mots, suggèrent une connotation religieuse du mot « end » qui pourrait rendre le mouvement de la trajectoire en partie cohérente : cette connotation se formulerait comme « point de départ de métamorphoses incessantes ». En effet, *The Dance* est un poème animé en métamorphose incessante, où chaque cycle de transformations reboucle indéfiniment sur un nouveau cycle. Le prochain mot mis en exergue par des trajectoires est d'ailleurs « begins ».
Comme dans toute métaphore réussie, le sens de ce couplage non conventionnel ne peut pourtant être complètement épuisé par ce processus de superposition de traits. Reste un écart, ou plutôt un « jeu » – mot qui, comme le rappelle Marie-Laure Ryan, désigne aussi l'espace entre deux pièces qui se forme lorsque l'une des deux n'est pas proprement cousue (189). Ce « jeu » peut être interprété soit comme une métaphore de l'instabilité du signe linguistique à laquelle l'iconicité du mouvement ne peut qu'improprement remédier ; soit comme une métaphore de l'espace de liberté où, entre vide et plein de sens, se joue le potentiel littéraire de l'animation métaphorique, et où l'imagination du lecteur est pleinement convoquée.

## 7 En guise de conclusion

En guise de conclusion de ce bref parcours à travers le large champ d'investigation concernant les « figures d'animation média » dans le discours numérique, deux remarques :
La première concerne le regard synchronique adopté dans cet article, qui ne prend pas en compte l'instabilité du dispositif numérique. En réalité, cette instabilité a une influence capitale sur l'actualisation des animations numériques, et donc sur les traits signifiants mobilisés. Dans les « petites formes » de la publicité numérique, la prise en compte de cette instabilité a certes moins d'importance, car la « durée de vie » des bannières dépasse rarement quelques semaines. Dans la littérature numérique, cette instabilité peut en revanche mettre en question le projet esthétique de l'auteur en transformant la vitesse d'exécution de l'animation : un mouvement entrant initialement dans la catégorie de l'unité sémiotique « par vagues », peut par exemple devenir un « obsessionnel » en étant actualisé sur un ordinateur plus performant[18].
La deuxième remarque concerne l'identification des « traits signifiants possibles ». En les déduisant d'une analyse des objets du corpus, nous postulons que les pratiques créatives du discours numérique n'anticipent pas seulement sur les attentes des lecteurs, mais les créent. Comme l'affirme pourtant Jean-Claude Passeron, « aucun texte, ou icône ne sont jamais si contraignants ou si parlants qu'ils

18 Pour une discussion des conséquences de l'instabilité du dispositif pour l'œuvre de poésie numérique, voir par exemple Alexandra Saemmer, Philippe Bootz (références dans la bibliographie).

puissent suffire à imposer en tout contexte un pacte de réception assurant la rencontre des attentes du récepteur inscrites dans le texte ou l'icône » (425). Même si les objets forgent les usages par un certain nombre de structures récurrentes, il sera important de mettre en place des dispositifs d'observation afin de mettre la description sémiotique des signes et figures du discours numérique à l'épreuve des usages réels. Le projet Egide-Utique franco-tunisien USET[19] poursuivra cet objectif en combinant une analyse sémiotique de sites web d'entreprises à des observations d'usages avec les publics cibles.

## 8 Bibliographie


1. T. Baccino et T. Colombi T. « L'analyse des mouvements des yeux sur le Web », *Interaction Homme-systèmes : perspectives et recherches psychoergonomiques*, éd. Alain vom Hofe, Paris/Londres, Lavoisier/Hermès Science Publishing, 2001, www.eyegaze.com/doc/PDFs/Baccino.pdf
2. J.M.C Bastien et D.Scapin, *Ergonomic Criteria for the Evaluation of Human-Computer interfaces. Institut National de recherche en informatique et en automatique*, France, 1993.
3. P. Bootz, « Une poétique fondée sur l'échec », *Poésie: numérique*, éd. A. Gherban et L.-M. De Vaulchier, revue *Passages d'encres* 33, 2008, p.119-122.
4. S. Bouchardon, « Hypertexte et art de l'ellipse », http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/03/58/sic_00000358_02/sic_00000358.html
5. J. Clément, « Du texte à l'hypertexte : vers une épistémologie de la discursivité hypertextuelle », *Hypertextes et hypermédias : Réalisations, Outils, Méthodes*, éd. J.-P. Balpe, A. Lelu et I. Saleh, Paris/Londres, Lavoisier/Hermès Science, http://hypermedia.univ-paris8.fr/jean/articles/discursivite.htm
6. U. Eco, « Vegetal and mineral memory : The future of books », 2005, http://weekly.ahram.org.eg/2003/665/bo3.htm
7. Groupe µ, *Rhétorique générale*, Paris, Larousse, 1970.
8. J.-M. Klinkenberg, *Précis de sémiotique générale*, Louvain-la-Neuve, De Boeck, 1996 ; Paris : Seuil, 2000.
9. G. P. Landow, *Hypertext 2.0*, *The Convergence of Contemporary Critical Theory and Technology*, Baltimore and London, The Johns Hopkins University Press, 1992 et 1997.
10. E. Landowski, La société réfléchie. Essais de socio-sémiotique, Paris, Seuil, 1989.
11. J. Le Marec, « L'analyse des usages en construction : quelques points de Méthode », *Comprendre les usages de l'Internet*, éditions ENS, 2001.


19 Projet financé dans le cadre de l'appel d'offres pour les Projets Hubert Curien – Utique. Titre du projet : « Construction de la signification par l'usager des sites d'entreprises dans un contexte économique franco-tunisien ». Responsables du projet : Raja Fenniche, Brigitte Simonnot, Alexandra Saemmer. Début du projet en février 2010.

12. J. Nielsen J. et H. Loranger, *Site web : priorité à la simplicité*, Pearson Campuspress, 2008.

13. J.-C. Passeron, *Le raisonnement sociologique : un espace non poppérien de l'argumentation*, Paris, Albin Michel, 2006.

14. N. Pignier, « Analyse sémiotique de la webpublicité », Semiotica 156 – 1/4 (2005), p. 521-538.

15. F. Rastier, *Arts et sciences du texte*, Paris, puf, 2001.

16. M.-L. Ryan, *Narrative as virtual reality*, Baltomore : Johns Hopkins University Press, 2001.

17. A. Saemmer, « Aesthetics of surface, ephemeral and re-enchantment in digital literature How authors and readers deal with the lability of the electronic device », *Cyberliteratures of the world*, Neohelicon 36, 2009, p. 477-488.

18. A. Saemmer, *Matières textuelles sur support numérique*, Publications de l'Université de Saint-Étienne, 2007.

19. A. Vandendorpe, *Du papyrus à l'hypertexte*, Paris, Éditions de la Découverte, 1999.

# Pratiques documentaires et construction d'exemplarité : le déni des médiations

**Sarah LABELLE (1), Aude SEURRAT (2)**

sarah.labelle@sic.univ-paris13.fr / aseurrat@hotmail.com
(1)*PARIS 13 – Villetaneuse*
(2) *GRIPIC, CELSA – PARIS SORBONNE*

**Résumé.** Notre article se penche sur certaines pratiques documentaires destinées à faire circuler des « initiatives exemplaires » et analyse comment la réécriture d'actions singulières permet de les constituer en modèles à imiter. Nous nous sommes intéressées à deux projets de société contemporains, le développement de « la société de l'information » et la promotion de « la diversité » pour observer les modes d'affichage de ces « bonnes pratiques » au sein de bases de données. Dans un premier temps, nous proposons de décrire et d'analyser les médiations sous-jacentes à la valorisation de ces pratiques. Puis, nous proposons de mettre au jour les différents modèles de la communication qui sont convoqués implicitement (diffusion, relation, interaction, transparence…). En d'autres termes, comment en voulant construire de l'exemplarité, les pratiques documentaires témoignent de certaines conceptions de la communication ?

**Mots-clés.** Base de données, pratiques documentaires, exemplarité, médiation

## 1 Introduction

« Bonnes pratiques », « pratiques de références », « initiatives exemplaires ».... sont quelques expressions qui acquièrent une certaine fortune pour désigner des exemples d'actions promus. Des agences, des observatoires et plus largement des institutions enquêtent et collectent sur leur secteur d'activités et identifient, pour leurs usagers, publics et partenaires, des pratiques qu'elles qualifient et répertorient sous forme de documents numériques. Comment des actions situées sont-elles érigées en « bonnes pratiques » ? Notre interrogation porte sur les processus de qualification et de transformation de ces actions répertoriées. Nous étudions plus précisément comment une pratique discernée pour son originalité peut, grâce à ces processus, devenir une pratique préconisée. Comment, en d'autres termes, des actions singulières deviennent-elles exemplaires ? Nous nous sommes intéressées à deux projets de société contemporains, le développement de « la société de l'information » et la promotion de « la diversité » pour observer les modes d'affichage de ces « bonnes pratiques » au sein de bases de données. Notre article se penche plus précisément sur certaines pratiques documentaires destinées à faire circuler des « initiatives exemplaires » et à proposer leur « dissémination » dans

d'autres organismes ou institutions. Notre hypothèse est que ces pratiques documentaires jouent un rôle structurant : elles contribuent à la transformation de pratiques situées correspondant à un lieu, à une organisation, à une temporalité, etc. en des modèles d'action à promouvoir en tout lieu, dans toute organisation, à tout moment. Nous pensons que ces processus de communication et cette production de documents numériques déploient certaines conceptions de la communication que nous souhaitons discuter par la même occasion.
Dans un premier temps, nous proposons de décrire et d'analyser les médiations sous-jacentes à l'écriture et à la valorisation de ces pratiques. Nous nous appuyons pour cela sur l'analyse de bases de données en repérant la disposition des informations et leur valorisation. Puis, nous proposons de mettre au jour les différents modèles de la communication qui sont convoqués implicitement (diffusion, relation, interaction, transparence…). En d'autres termes, comment, en voulant construire de l'exemplarité, ces pratiques documentaires témoignent de certaines conceptions de la communication, telles que la dissémination et la transparence ?

## 2 Base de données et idéal de « dissémination » des pratiques

La « gestion de la diversité », « la société de l'information » ou encore le « développement durable » sont des cadres d'action contemporains extrêmement exigeants pour les institutions et les organisations. L'incitation – voire l'injonction – à développer certaines pratiques et à mettre en place certaines actions, se traduit notamment par des dispositifs de médiation, qui reposent sur le traitement documentaire. La circulation de « bonnes pratiques » joue un rôle important dans la promotion de ces modèles de société. Nous nous penchons sur des bases de données qui médiatisent ces « bonnes pratiques » et analysons leur format éditorial et leur organisation documentaire. En effet, face au foisonnement des pratiques et devant l'enjeu de médiatisation, la base de données devient un espace qui atteste de l'action et la promeut. Nous verrons que cet idéal de « diffusion » de « bonnes pratiques » s'articule très bien avec une conception de la base de données comme espace de partage.
Nous avons réalisé une analyse croisée de deux bases de données qui publicisent des pratiques présentées comme exemplaires. Notre première base de donnée est promue par l'association Villes Internet sur son site web (http://www.villes-internet.net), défini comme « centre de ressources » alimenté par et à destination des acteurs locaux. Quant à notre deuxième base, elle a été élaborée par une association interentreprises (dont nous ne pouvons pas mentionner le nom) qui vise à promouvoir la « lutte contre les discriminations » et la « gestion de la diversité » dans les entreprises françaises. Ces deux bases de données produisent des documents numériques qui s'affichent sous forme de fiches et de tableaux ; elles ont pour mission de mutualiser les expériences en matière de « gestion de la diversité » et de politiques numériques locales.
Ces « bonnes pratiques » ont pour but de servir d'exemples à suivre. Il s'agit de « cas » d'entreprises ou de politiques de la ville qui sont présentés comme des modèles qu'il conviendrait d'imiter. Nous chercherons à démontrer comment des pratiques situées sont constituées en savoirs mis en circulation et dans quelle mesure la base de données comme média fait l'objet d'imaginaires de la communication.

Pour Philippe Marion, « appréhender cette singularité différentielle d'un média, c'est tenter d'en saisir la « médiativité »[1]. L'auteur emploie le concept de *médiativité* pour qualifier l'articulation entre « d'une part, des conditions de diffusion et de circulation attachées au média observé ; par exemple, le type de périodicité de la presse écrite ou le type de distribution d'un dépliant publicitaire », et d'autre part, « en fonction des matériaux sémiotiques d'expression mobilisés par le média ». L'intérêt de ce concept est qu'il souligne l'importance de prendre au sérieux toutes les dimensions qui font « qu'un média n'est pas l'autre »[2]. Le choix par ces associations de faire circuler ces « bonnes pratiques » au sein de bases de données n'est pas anodin : ces bases constituent une forme de médiatisation de l'action. L'intérêt d'une base de données est en effet de fournir des configurations formelles qui orientent et balisent l'écriture des informations. Nous retiendrons de l'idée de *médiativité* que la singularité d'un média s'appréhende en articulant ses dimensions techniques, sémiotiques et sociales, auxquelles nous pourrions ajouter documentaires. Or, les dimensions sémiotiques ne sont pas uniquement un « potentiel expressif », par leurs configurations singulières, elles permettent de se pencher sur les attentes et promesses communicationnelles du média observé. Comme le souligne Yves Jeanneret : « l'espace des positionnements sur les logiques de la communication et les rôles attendus dans l'échange ne se limite pas aux déclarations les plus explicites, mais repose sur un ensemble très complexe de traits formels et sémantiques, liés à l'utilisation des formes, aux figures énonciatives, aux modalités de présentation de la réalité, etc. »[3].
Une rencontre d'imaginaires s'effectue entre les conceptions d'usages de la base de données et l'idéal de « dissémination » de pratiques exemplaires. Afin d'analyser en quoi nos bases de données sont présentées comme les « outils » les plus adéquats pour mettre en visibilité et en circulation ces « bonnes pratiques », nous prenons en compte leurs configurations sémiotiques. Nous partons de l'idée que la matérialité des documents numériques influe non seulement sur la nature du contenu donné à lire, mais encore sur les modalités d'usage. Nous nous attarderons donc sur les formes sémiotiques qui président à la médiation documentaire, en interrogeant le modèle de saisie et la mise en forme des informations. Notre objectif est de mettre en évidence la teneur politique de ces objets documentaires anodins.

## 3 Du modèle de la saisie à l'analyse des mises en forme

Pour permettre la médiation de savoirs pratiques présents dans les organisations, collectivités ou entreprises, les associations étudiées proposent nécessairement un canevas. Ce dernier s'appuie sur une forme, constituée par une organisation éditoriale et une sélection d'informations : cette forme n'est pas un réceptacle impuissant, neutre et interchangeable. Nous allons voir comment ces médiations documentaires participent pleinement à la construction de l'exemplarité de certaines pratiques. Nous nous sommes penchées sur les formulaires de saisie qui sont des techniques de production documentaire. Ils supposent un certain mode

[1] P. Marion, « Médiagénies de la polémique. Les images « contre » : de la caricature à la cybercontestation », *Recherches en communication*, n°20, 2003, p.22
[2] P. Marion, « Narratologie médiatique et médiagénie des récits », *Recherches en communication* n°7, 1997, p.37.
[3] Y. Jeanneret, *Penser la trivialité* - Volume 1 : *la vie triviale des êtres culturels*, Hermès, Lavoisier, Collection Communication, médiation et construits sociaux, 2008, p.153

de présentation des pratiques mises en place dans les villes ou les organisations et impliquent en amont leurs modes de valorisation et de généralisation.
Nos exemples permettent de mieux comprendre le rôle de la fiche en tant qu'unité documentaire et l'enjeu de la mise en série de fiches dans des bases de données. Nous avons pour cela observé les pratiques d'écriture et d'enregistrement que les plateformes associatives mettent en œuvre. L'écriture des pratiques situées au sein des bases de données consiste à segmenter et à catégoriser ces pratiques au sein de fiches et la mise en série des fiches par sections. Or, ces opérations permettent de donner de la lisibilité à des pratiques hétérogènes et à les rendre comparables entre-elles. Les catégories thématiques servent notamment à organiser le stock d'informations récoltées par le biais de formulaires d'inscription. Elles sont des médiations documentaires qui permettent à la fois le classement et l'accès.
L'inscription de ces pratiques au sein d'un même support matériel, des fiches, leur attribue un univers de cohérence qui n'est pas déterminé *a priori*. Par exemple, certaines entreprises dont les pratiques font l'objet de fiches en « gestion de la diversité », peuvent recruter dans les quartiers dits « difficiles » ou aménager leurs espaces pour les handicapés sans pour autant identifier leurs actions comme étant en faveur de « la diversité ». De la même manière, l'association Villes Internet en proposant aux collectivités de témoigner sur leurs « initiatives » incite les acteurs à présenter action par action. Cependant, ce format court (sur lequel nous reviendrons en détail plus loin) concoure à diviser ce qui peut pourtant appartenir à des ensembles (Schéma Directeur, politique dite numérique...). Cela transparaît dans la liste chronologique des initiatives : les « initiatives » d'une même collectivité s'enchaînent, s'empilent les unes sur les autres. Le dispositif induit une fragmentation en action précise et identifiable. Seule la consultation de la fiche de la collectivité accorde la possibilité de matérialiser le système complet de fiches correspondantes. Ce morcellement permet de favoriser l'abondance de fiches et la requalification des actions en fonction des entrées prédéfinies par le système documentaire. Ainsi, l'opération documentaire de saisie dans la base de données permet aux expériences situées d'acquérir leur portée générale et leur lisibilité dans un même cadre d'écriture. Ces pratiques, bien qu'hétérogènes, vont être mises en listes et segmentées par le truchement de l'architexte[4] qui induit un certain formalisme de l'écriture.

## 4 Les pratiques exemplaires en promotion de « la diversité » dans les entreprises

Les deux copies d'écran ci-dessous sont issues de la base de données de notre association d'entreprises.

[4] « Nous nommons architextes (de *arkhè*, origine et commandement), les outils qui permettent l'existence de l'écrit à l'écran et qui non contents de représenter la structure du texte, en commandent l'exécution et la réalisation. Autrement dit, le texte naît de l'architexte qui en balise l'écriture » (Sous la direction de E. Souchier (Emmanuël), Y. Jeanneret, J. Le Marec, *Lire, écrire, récrire*, op.cit, Bibliothèque Centre Pompidou, collection Etudes et recherche, 2003, p.23). La récurrence d'un même cadre d'écriture présidant à l'écriture des « bonnes pratiques » se retrouve bien dans la fonction de l'architexte dont « le principe consiste en cette forme particulière de l'écriture permise par l'informatique, qui se place en amont de toute écriture particulière pour en définir le cadre et les conditions » (Sous la direction Y. Jeanneret et C. Tardy, *Métamorphoses médiatiques, pratiques d'écriture et médiation des savoirs*, rapport final de recherche, Université Paris Sorbonne Paris 4- CELSA, février 2005, p.16

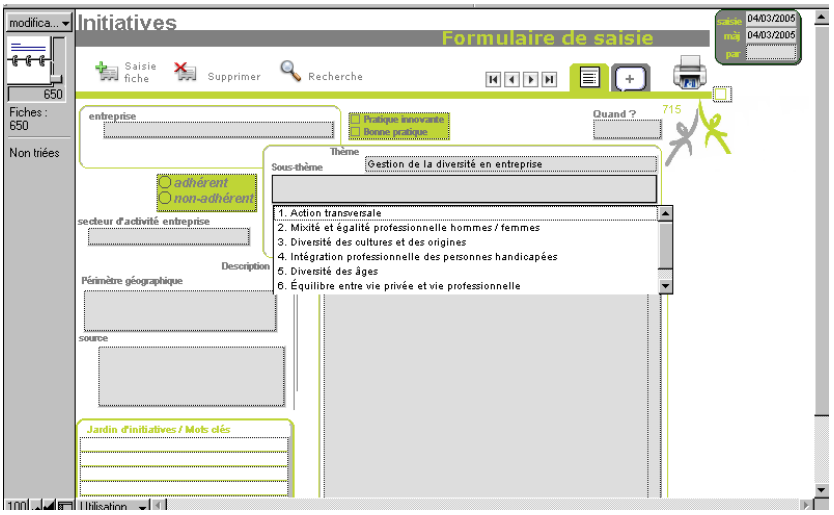


**Figure1**. *Formulaire de saisie de la base de données de l'association interentreprises*

:

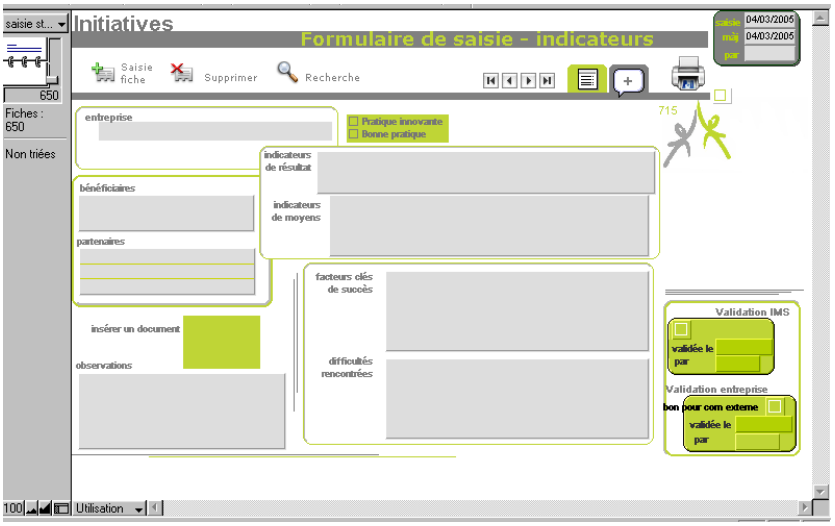


**Figure 2** : 2ème page du formulaire de saisie

Sur la première page, en haut à gauche, l'usager-scripteur est invité à remplir le nom de l'entreprise et en haut à droite la date de mise en œuvre de la pratique. Ces indications permettent d'identifier, rapidement, des éléments de contexte de la pratique et permettront, par la suite, de faire une recherche par entreprise ou par date. Afin de décrire la pratique, l'architexte propose ensuite une série de menus déroulants dans lesquels l'usager-scripteur est invité à sélectionner une information préenregistrée. Avant de passer à l'écriture de la « bonne pratique » ou « initiative » dans la base de données, il doit lire ces menus déroulants pour choisir dans quelle catégorie cette pratique s'inscrit. Le choix des cases « thème » et « sous-thème » inscrit la pratique dans un univers d'autres pratiques prédéfinies. En effet, il faut comprendre que cet archivage formalisé des pratiques doit permettre, à celui qui la

consulte de chercher des pratiques sur la base d'une thématique et de considérer le corpus issu de la recherche comme un ensemble de pratiques « similaires »[5]. Lors de la mise en place du dispositif, les choix formels ont été en grande partie déterminés par la volonté de « faciliter » la consultation future : le fait d'indexer les pratiques autour de thématiques larges et partagées devait permettre une plus grande « efficacité » de la recherche[6]. « La gestion de la diversité » est ainsi découpée en plusieurs champs : « la mixité professionnelle », « diversité des cultures et des origines », « intégration professionnelle des personnes handicapées », « diversité des âges », « équilibre vie privée et vie professionnelle ». Cette segmentation nous donne déjà une idée sur la manière de penser « la diversité » en la segmentant par catégories de personnes : « les handicapés », « les femmes », « les personnes âgées », « les personnes issues de l'immigration ». Formalisme d'écriture, elle est donc aussi un cadre de pensée. Cette organisation des thématiques révèle les enjeux politiques sous-jacents à ce type de documentation et pose, comme pré-requis, la catégorisation des personnes selon une variable identitaire unique. Il s'agit d'une conception éminemment politique du social fondée sur l'idée de « communauté ».

Dans l'endradré central, l'usager est invité à décrire la pratique concernée. Or, il nous semble bien que le choix et l'organisation des rubriques des fiches produisent un certain type de mise en intrigue. Cette mise en intrigue présiderait donc à l'écriture au sein de la base de données. L'entreprise a fixé un objet à sa quête, des « objectifs » ainsi que des destinataires, les « cibles ». Pour y parvenir, elle se dotera de « moyens » et aura aussi des adjuvants, des « partenaires », qui vont l'aider pour mener à bien son projet. Mais, la mise en œuvre de l'action présente des « difficultés », toute action méritante n'étant pas simple à effectuer. Nous trouverons alors des opposants à l'action qui peuvent être des personnes qui posent un « refus »[7], des « résistances internes ». Nous trouverons aussi des difficultés matérielles de « financement » ou d'« infrastructures ». La quête est donc rythmée par des péripéties. Mais, en « bon héros », l'entreprise saura surmonter ces difficultés. Il semble, d'ailleurs, que cette réussite aurait pu être prédite puisque les « facteurs-clés de succès » l'emportent sur les « difficultés ». Des « Bénéfices » pour les destinataires et aussi pour le sujet héros ressortiront alors de cette quête. En effet, ces actions auront permis « d'intégrer de nouvelles collaboratrices compétentes », « d'améliorer l'ambiance », voire même de « réduire des coûts ». L'action aboutit alors à des bénéfices qu'il convient d'évaluer et de mesurer par des « indicateurs ». Schématiquement, nous avons donc un problème de départ, une mise en œuvre avec des péripéties, et enfin, un « happy end ».

---

[5] Pour Delphine Gardey les fiches éditées par les organisations ont : « pour caractéristique d'être le plus souvent des documents préformatés qui conduisent à la restriction et à la standardisation des informations portées à chaque étape d'une procédure ». D. Gardey, *Écrire, calculer, classer. Comment une révolution de papier a transformé les sociétés contemporaines* (1800-1940), Éditions de la découverte, 2008, p. 162.

[6] Une observation participante de 5 mois a été menée au sein de cette association lors de la mise en place de la base de données. Les « bonnes pratiques » étaient auparavant mises en liste au sein d'un « cahier de bonnes pratiques » mis en page sous Word. Ce document présentait un classement par grandes thématiques que l'on retrouve dans la base de données, mais il ne proposait que ce mode de classement. La base de données a ainsi été mise en place dans le but d'éviter les « déséquilibres » entre les fiches – celles-ci étant de tailles très variables dans le cahier, mais devant se plier à un nombre pré-déterminé de caractères dans la base– et « d'optimiser » la recherche d'informations.

[7] Ces citations sont issues d'un corpus de 56 fiches qui a été analysé.

Ce type de lecture nous amène à dégager une sorte de scénario idéal, exemplaire de l'action réussie. Cela nous permet de souligner que l'exemplarité de la « bonne pratique » n'est pas juste liée au statut qui lui est attribué mais qu'elle est construite lors de cette mise en texte.

## 5 Les « initiatives » de *Ville Internet*

La fiche est le résultat d'un travail de saisie effectué par renseignements de champs prédéfinis dans nos deux exemples. Son mode de constitution implique son formatage et la définit comme un objet clos et déterminé. Chaque fiche contient un nombre restreint de renseignements qui permettent de classer et structurer les pratiques sociales ainsi prélevées. Publiée sur un site web, les fiches dépendent alors d'une configuration médiatique particulière : elle se traduit par l'organisation tabulaire et par un appareil paratextuel fourni par les signes-passeurs. Nous allons examiner cela dans le cadre du site de Ville Internet.

Cet exemple est composé de quatre zones textuelles distinctes : la première et la seconde sont composées d'informations de classification (titre et catégories, informations géographiques, administratives, descriptives) ; la troisième d'informations spécialisées qui présentent l'action en trois parties préformatées (Actions, Résultats, Recommandations) ; et en dernier lieu, la fiche contient (de façon facultative) un signe-passeur vers le dispositif en ligne de la collectivité. Cette organisation laisse transparaître la capacité d'un tel document à réunir des informations requalifiées par leur appartenance à un même espace de sens. Une telle fiche « initiative » ambitionne de *vraiment* donner à voir ce qui est réalisé par les acteurs de terrain. Les données de classification permettent de faire système avec les autres fiches grâce aux signes-passeurs qui leur sont associés (villes...) ; ces éléments classificatoires rendent possibles le transfert et la consultation (alphabétique pour les villes, thématiques pour les « initiatives »).

Le triptyque d'informations spécialisées s'appuie quant à lui sur des textes qui *racontent* l'action. Les trois entrées supposent que le « correspondant » puisse évoquer chaque point (ou sinon l'entrée n'apparaît pas dans la fiche). Ces textes sont soumis à une norme narrative qui contraint la répartition des informations. L'entrée « recommandation » présume tout particulièrement de la transférabilité de l'action en fournissant des mises en garde, des alertes connues par l'expérience. Mais les deux autres entrées (Action et Résultats) contiennent aussi des traits narratifs qui suggèrent la capacité de l'action à être reproduite : ces traits relèvent d'une technique intellectuelle (énoncé de l'action) et d'une technique matérielle (fiche tabulaire). Leur répétition dans chaque fiche met en relation et prolonge chaque « initiative » avec la suivante. L'organisation de la fiche constitue dès lors une promesse de reproductibilité de l'action par le partage de l'expérience.

La plateforme de *Villes Internet* n'existe que par l'engagement des acteurs de terrain qui viennent rendre compte et enregistrer leurs démarches. L'association devient alors le réceptacle informationnel et documentaire de l'activité sur les différents territoires. Elle est en mesure de notifier l'évolution de l'activité dans un secteur précis (services publics, démocratie...). Sa plateforme devient le lieu de l'accumulation des actions, qui, sans avoir la prétention de l'exhaustivité, constitue un paysage ordonné et standardisé des pratiques situées sur le terrain. Ainsi, chaque démarche est constituée comme exemplaire, et l'accumulation sur la plateforme donne toute sa teneur au projet de société. C'est en ce sens qu'il y a médiation politique dans ces dispositifs.


**Figure 3** : « Initiative » issue de Villes Internet – 9 juin 2010

Il s'agit bien, dans nos deux exemples, d'appréhender et de « penser la banque de données en amont, comme étant, non pas seulement un réceptacle, mais un formalisme d'écriture, en anticipant les conditions de lecture favorisées par le calcul et la transversalité propres aux dispositifs informatiques »[8]. Comme Dominique Cotte, nous insistons sur « la nécessité de traiter les objets qui servent eux-mêmes à déposer, à sédimenter la connaissance acquise ou en train de se

[8] D. Cotte, « représentation des connaissances et convergence numérique : le défi de la complexité » *Documents numériques* Volume 4, n°1-2, 2000, *L'indexation,* p.180

faire : les documents »[9]. L'analyse de ces dispositifs d'enregistrement met en évidence leur capacité à faire affleurer des scénarios d'usage, sous couvert de ne proposer que des informations. Or, leur mode de constitution et les processus de qualification constituent les objets produits en documents. C'est pourquoi nous voulions mettre en évidence comment se constitue leur matérialité documentaire et comment elles peuvent dès lors acquérir une portée de modèle.

Dans ces plateformes, s'observe à la fois l'absence de détermination (toute démarche peut trouver sa place) et certaines marques qui traduisent des formes d'inspection de l'action. La légitimité de l'association, en lien avec des acteurs ministériels ou des instances européennes, confère au dispositif la capacité de jouer un rôle de garantie sur la nature et la validité des informations collectées et publiées. Le dispositif est agencé de telle manière que l'action peut être abordée en tout point, intégrable, unie aux autres. La série apparaît dans nos exemples sous forme de liste. L'écriture de ces pratiques par le biais de l'architexte opère une opération de décontextualisation, puis de recontextualisation qui les rend comparables. La mise en liste a pour effet de créer une certaine homogénéité et la segmentation une certaine comparabilité. Pour Jack Goody, « la liste implique discontinuités et non continuités. Elle suppose un certain agencement matériel, une certaine disposition spatiale ; elle peut être lue en différents sens, latéralement et verticalement, de haut en bas comme de gauche à droite, ou inversement ; elle a un commencement et une fin bien marqués, une limite, un bord, tout comme une pièce d'étoffe. Elle facilite, c'est le plus important, la mise en ordre des articles par leur numérotation, par leur son initial ou par catégories. Et ces limites, tant externes qu'internes, rendent les catégories plus visibles et en même temps plus abstraites »[10].

Dans cet exemple issu de la plateforme Villes Internet, se dévoile l'enjeu politique de l'intégration progressive de toutes les démarches dans un même ensemble. L'organisation chronologique au fil des enregistrements laisse au système d'information le soin de produire l'objet liste. Cette dernière permet cependant de créer une certaine abstraction qui favorise la capacité à produire un effet de généralisation. Elle correspond à un genre communicationnel ordinaire qui permet l'empilement et la mise en continuité d'objets. Ce système, tout comme le passage de l'oral à l'écrit étudié par Jack Goody, « rend possible d'examiner autrement, de réarranger, de rectifier des phrases ou des mots isolés »[11], ici des actions et démarches isolées. Elle efface les lignes de partage ou les écarts entre les « initiatives », les plaçant dans un ensemble affichant sa cohérence. Le système d'information opère une mise en série des documents et provoquent la « décontextualisation ». Ainsi, ce processus de décontextuation permet à la société d'entretenir un rapport différent avec ce qu'elle sait : cela crée un tout homogène et cela permet une appréhension analogique en effaçant les disparités. Ces pratiques de sélection et de classement inscrivent des pratiques hétérogènes dans une même sphère de discours : « la promotion de la diversité » ou celle des « villes internet ».

---

[9] *Ibidem,* p.168

[10] J. Goody, *La raison graphique, la domestication de la pensée sauvage,* Les éditions de Minuit, Paris, 1979, p.149

[11] J. Goody, *op.cit.,* p.145

▶ LES INITIATIVES [1717 fiches] Trier par Choisissez...

| Titre de l'initiative | Thème | Collectivité |
|---|---|---|
| ▶ Gestionnaire de projets (09/06/2010) | Equipement > Le personnel | HARNES (62) |
| ▶ Dématérialisation des démarches liées à l'état civil (09/06/2010) | Administration > Téléservices | EPINAL (88) |
| ▶ Lancement du nouveau site Internet de Rive de Gier (08/06/2010) | Information > Sur la collectivité | RIVE DE GIER (42) |
| ▶ Forum d'entraide informatique sur le site de l'EPN (08/06/2010) | Dialogue > Instaurer le débat | VEAUNE (37) |
| ▶ Rencontre avec un représentant du Trésor Public à l'Espace EPN de Veigne pour vous aider à déclarer vos impôts en ligne (08/06/2010) | Promotion > Rencontres, Débats | VEIGNE (37) |
| ▶ L'informatique pour les seniors (08/06/2010) | Formation > Aux outils | VANNES (56) |
| ▶ Contact Elus - Messagerie (08/06/2010) | Dialogue > Avec les élus | VANNES (56) |
| ▶ Site : Annuaire des élus et services (08/06/2010) | Administration > Services aux usagers | VANNES (56) |
| ▶ La Brie des Templiers met en ligne une revue de presse (07/06/2010) | Information > Sur la collectivité | COMMUNAUTÉ DE COMMUNES DE LA BRIE DES TEMPLIERS (77) |
| ▶ La Brie des Templiers dématérialise ses marchés publics (07/06/2010) | Information > Sur la collectivité | COMMUNAUTÉ DE COMMUNES DE LA BRIE DES TEMPLIERS (77) |
| ▶ Projet Ecole Numérique Rurale (03/06/2010) | Equipement > La collectivité | SERQUIG (59) |
| ▶ Les Grands Travaux et Projets (03/06/2010) | Information > Sur la collectivité | ANNEMASSE (74) |
| ▶ Information par SMS (03/06/2010) | Information > Sur les services aux citoyens | COMMUNAUTÉ DE COMMUNES DE LA BRIE DES TEMPLIERS (77) |
| ▶ Accès à un nouveau portail (02/06/2010) | Administration > Services aux usagers | LE CHESNAY (78) |
| ▶ Hot-spots WIFI municipaux (02/06/2010) | Equipement > En infrastructure | VANNES (56) |
| ▶ Agenda partagé (01/06/2010) | Information > Sur les acteurs locaux | PLEYBEN (29) |
| ▶ 1ère journée Médi@lture le samedi 12 juin à Plouarzel (01/06/2010) | Mutualisation > Les compétences et savoirs | PLOUARZEL (29) |
| ▶ Défi Internet 12 (01/06/2010) | Formation > Le public scolaire | NOYELLES LES VERMELLES (62) |
| ▶ Assises de l'environnement 2010 (31/05/2010) | Dialogue > Consultation des habitants | VERRIÈRES-LE-BUISSON (91) |
| ▶ Le site internet fait peau neuve (29/05/2010) | Information > Sur la collectivité | SAINT-PIERRE - SAINT-PIERRE ET MIQUELON (97) |

1 2 3 4 5 6 7 8 ... >>

**Figure 4 :** Liste des « initiatives » sur Villes Internet – 9 juin 2010

L'une des caractéristiques qui nous semble les plus notables dans l'étude de ces bases est l'imbrication très étroite des logiques de lecture et d'écriture. Par exemple, les champs non remplis du formulaire d'inscription n'apparaissent pas à la consultation de la base. L'exemple, pour être exemplaire, ne peut souffrir le vide. Ce passage de l'empirique au normatif est lié aux modes de médiatisation des pratiques, ainsi qu'aux formalismes d'écriture dans lesquelles elles sont réécrites. Autrement dit, la médiatisation est favorisée par la capacité des pratiques à « s'adapter médiagéniquement »[12] à la base de données. Et les formats tabulaires facilitent l'effacement ou le crédit d'informations. Ainsi, l'inscription dans une base de données de ces expériences situées permet d'ordonner, de rassembler, de reconstruire ce qui, dans la pratique, était disparate et fragmentaire. Par ce jeu de réécritures, nous notons le passage des pratiques situées à la mise en circulation de savoirs qui leur sont liés. Pour que ces pratiques deviennent *exemplaires*, il convient d'en faire des exemplaires dont la matérialité documentaire ne peut être considérée comme annexe.

Cependant, les acteurs de cette construction et de cette mise en circulation revendiquent leur rôle de simples agents de passage facilitant la « dissémination » des pratiques dans la société. Une posture de déni des médiations qui témoigne de certaines conceptions de la communication.

## 6 Déni des médiations et trivialité des conceptions de la communication

Les bases de données cristallisent des imaginaires sur la « dissémination » des savoirs et leur mise en partage, objectif qui est au cœur de nos deux organisations. Cette rencontre entre un projet et les imaginaires liés aux configurations techno-sémiotiques des bases de données, nous pourrions, en employant la terminologie de Philippe Marion, la qualifier de « médiagénique ». « La médiagénie est [...] l'évaluation d'une « amplitude » : celle de la réaction manifestant la fusion plus ou moins réussie d'une narration avec sa médiatisation, et ce dans le contexte-interagissant lui aussi- des horizons d'attente d'un genre donné. Évaluer la médiagénie d'un récit, c'est donc tenter d'observer et d'appréhender la dynamique d'une interfécondation »[13]. Ce terme de *médiagénie* nous semble intéressant car il permet de souligner que certaines idées semblent particulièrement bien correspondre à certaines formes sémiotiques. Mais, nous ajoutons à cette approche le fait que les formes sont elles aussi sujettes à des imaginaires et ce sont aussi ces imaginaires qui entrent en correspondance lors de la production d'objet médiatique. Ainsi, il ne s'agit pas tant d'une rencontre réussie entre des projets et un média, mais plutôt d'une *interfécondation* entre une conception de la base de donnée comme espace de partage et de connectivité et un idéal de mise en circulation de modèles d'action.

Nos deux bases de données visent à donner à lire les pratiques pour les donner à partager. Elles sont présentées par les acteurs comme le meilleur moyen de « mutualiser » les connaissances. Nos terrains se situent dans une certaine

[12] P. Marion, « Médiagénies de la polémique. Les images "contre" : de la caricature à la cybercontestation ». *Recherches en communication*, n°20, 2003, p.25

[13] P. Marion, « Narratologie médiatique et médiagénie des récits » *Recherches en Communication* n°7,1997, p.83. Nous nous inspirons des notions developpées par Philippe Marion sur la médiagénie (potentiel expressif….), en décalant le point de vue en fonction de la spécificité de nos objets.

conception évidente de la communication qui postule implicitement un partage qui se ferait sans médiation aucune et où l'outil serait efficace en lui-même. Les termes « échange, mutualisation, partager, inscrire » sont récurrents au fil des discours sur les « outils » et sont mis en relation avec la possibilité de consulter et la dimension de « centre de ressources ». Les bases de données participent de cette idéologie des nouvelles technologies qui permettent une meilleure circulation, une plus grande fluidité de l'information. Cependant, cette approche néglige l'épaisseur sémiotique et documentaire de ces dispositifs.

Différents modèles de la communication émergent de cette analyse. Le premier est celui d'une communication qui ne serait que simple logistique. Le support technique semble ici investi d'une certaine utopie car il lui est attribué le rôle de rendre la communication plus directe, plus efficace et plus large. D'autre part, les acteurs qui mettent en forme ces « bonnes pratiques », se présentent comme des intermédiaires neutres qui donnent accès à la connaissance de pratiques exemplaires. Nous sommes en présence d'une certaine conception de la médiation que Jean Davallon qualifie « d'usage ordinaire ». La médiation signifie, dans cette acception, surtout sociologique, le fait de servir d'intermédiaire, « facilitant la communication [elle] est censée favoriser le passage à un état meilleur »[14] . Or, cette conception de la mise en partage facilitée par un tiers participe à gommer le fait que les médiations impliquent « une transformation de la situation ou du dispositif communicationnel, et non une simple interaction entre éléments déjà constitués, et encore moins une circulation d'un élément d'un pôle à l'autre. »[15]

Enfin, le dernier modèle qui se dégage est celui d'une certaine épidémiologie des représentations. Afin de changer les pratiques sociales, il suffirait de diffuser les connaissances au plus grand nombre : un idéal proche de celui des Lumières, mais qui ne permet pas d'appréhender la complexité de la circulation sociale des savoirs. Cette complexité nous semble tenir dans ce qu'Yves Jeanneret nomme la trivialité. La notion de trivialité, du latin trivium, ne désigne pas ce qui serait banal, mais le fait que tout être culturel circule. Elle permet de sortir d'une logique linéaire, d'un amont vers un aval et d'appréhender les pratiques de communication comme créatrices de culture. Cette notion permet, d'autre part, de penser l'altération des êtres culturels en dehors des modèles qui y voient une simple dégradation. Pour l'auteur, analyser la trivialité nécessite de prendre en compte les médiations à l'œuvre dans les pratiques de communication mais aussi de s'interroger sur la manière dont la circulation des êtres culturels fait l'objet d'imaginaires, de normes. « Enfin, si précise qu'elle soit, cette approche par les processus de communication effectifs ne suffira pas à constituer théoriquement l'analyse de la trivialité. En effet, toutes les pratiques de communication qui affectent les êtres culturels se doublent d'un plan imaginaire et normatif, qui est constitué par les représentations de ce qu'est la trivialité et de ce qu'elle devrait être »[16].

## 7 Conclusion

Nos bases données nous semblent être un exemple éclairant de « représentations de ce qu'est la trivialité et de ce qu'elle devrait être ». La circulation des idées est conçue comme une mise en partage effectuée par un dispositif technique envisagé

[14] J. Davallon, « La médiation : la communication en procès ? », *Médiation & information*, n°19. p.40

[15] *Ibidem*, p.43

[16] Y. Jeanneret, *Penser la trivialité* - Volume 1 : *la vie triviale des êtres culturels* Hermès, Lavoisier, Collection Communication, médiation et construits sociaux, 2008, p.7

comme un simple adjuvant transparent. Cet article aura de permis de voir que derrière ces pratiques sociales et techniques normées se cachent certaines représentations de la technique comme outil d'optimisation et de la communication comme dissémination. Questionner les formes concrètes, analyser comment elles sont investies par les acteurs, permet alors de comprendre que derrière des formules circulantes comme celle de « bonnes pratiques », se déploient des imaginaires sur la communication qui peuvent avoir une force instituante dans les phénomènes sociaux. La structure des bases de données n'est pas uniquement une manière de transmettre ces savoirs pratiques, c'est aussi et surtout une manière de les penser et de les transformer. Il n'y a donc pas le social d'un côté et le sémiotique de l'autre, des pratiques sociales et des techniques de communication mais bien des formes, investies dans le social, qui participent à le structurer.

## 8 Bibliographie


1 D. Cotte, « Représentation des connaissances et convergence numérique : le défi de la complexité », *Documents numériques* Volume 4, n°1-2, 2000.

2 J. Davallon, « La médiation : la communication en procès ? », *Médiation & information*, n°19. p.40.

3 M. Despres-Lonnet, « Contribution à la conception d'interfaces de consultation de bases de données iconographiques », thèse à l'université de Lille, 2000

4 *Documents numériques* Volume 4, n°3-4, « L'archivage », 2000

5 M. Foucault, *L'archéologie du savoir,* bibliothèque des sciences humaines, nrf, éditions Gallimard, Paris, 1969.

6 D. Gardey, Écrire*, calculer, classer. Comment une révolution de papier a transformé les sociétés contemporaines* (1800-1940), Éditions de la découverte, 2008.

7 J. Goody, *La raison graphique, la domestication de la pensée sauvage,* Les éditions de Minuit, Paris, 1979

8 Y. Jeanneret, *Penser la trivialité* - Volume 1 : *la vie triviale des êtres culturels* Hermès, Lavoisier, Collection Communication, médiation et construits sociaux, 2008.

9 S. Labelle, La ville inscrite dans « la société de l'information » : formes d'investissement d'un objet symbolique, thèse à l'université de Celsa Sorbonne, 2007

10 P. Marion, « Médiagénies de la polémique. Les images "contre" : de la caricature à la cybercontestation ». *Recherches en communication*, n°20, 2003

11 P. Marion, « Narratologie médiatique et médiagénie des récits » *Recherches en Communication* n°7, 1997.

12 A. Seurrat, La construction de l'exemplarité, mise en forme et en circulation de « bonnes pratiques » en « gestion de la diversité », Master Recherche Celsa Sorbonne, 2005

13 C. Tardy, Y. Jeanneret (dir.), *Métamorphoses médiatiques, pratiques d'écriture et*

*médiation des savoirs*, rapport final de recherche, Université Paris Sorbonne Paris 4- CELSA, février 2005

14 E. Souchier, Y. Jeanneret, J. Le Marec (dir.), *Lire, écrire, récrire,* Bibliothèque Centre Pompidou, collection Etudes et recherche, 2003

# Apports de psychologie du travail pour caractériser l'activité de gestion de l'information

**Orélie Desfriches DORIA, Manuel ZACKLAD**

orelie.desfriches_doria@cnam.fr / orelie.desfriches_doria@cnam.fr

*Laboratoire Dispositifs d'Information Communication à l'Ere du Numérique (Dicen) – CNAM - Paris*

**Résumé.** Dans cet article, les auteurs présentent les apports de la psychologie du travail pour l'élaboration d'une méthodologie d'analyse croisée de l'activité et des documents, associée au développement d'un instrument de documentarisation collective multi-facettes. Ils abordent les notions de surcharge pour initier une réflexion sur les fondements du développement de l'outil. Ils décrivent ensuite le cadre socio-organisationnel dans lequel l'approche de l'outil se situe. Puis l'articulation des facettes individuelles et collectives est abordée à travers le concept de métier. Enfin ils abordent les apports potentiels de l'outil en termes de support à l'innovation.

**Mots-clés.** Systèmes d'Organisation des Connaissances, Méthodologie, Psychologie du Travail, Métier

## 1 Introduction

L'introduction des TIC et la croissance du volume des documents électroniques à gérer par chaque individu à son poste de travail, et en particulier dans le travail tertiaire, conduit à ce que certains auteurs qualifient de « surcharge informationnelle » (Isaac, Campoy, Kalika). En effet, une étude réalisée par Autissier et Lalhou (1999) démontre que dans l'entreprise étudiée « les managers consacrent un tiers de leur temps en moyenne à ces tâches de manutention de l'information ».

La manutention de l'information est définie par Vacher (1998) comme « l'ensemble des tâches de manipulation, manuelles ou mécaniques, des supports d'information : tri de documents, classement dans des dossiers, indexation de bases de données, mise en forme de comptes –rendus, recherche de papiers, disquettes, adresses électroniques, etc. ». La surcharge évoquée plus haut « est directement liée à la notion de maîtrise du temps et au fait que les TIC contribuent à augmenter le temps de traitement de l'information au détriment des activités liées à l'exercice du métier. » (Isaac, Campoy, Kalika)

D'après l'étude menée par Isaac, Campoy, et Kalika entre 2001 et 2005 sur les salariés d'une grosse entreprise, il semble que la perception par les opérateurs eux-mêmes de ces tâches de gestion de l'information soit perçue comme envahissante

au regard de leurs activités métier. En effet, dans leur enquête « la surcharge cognitive, estimée au travers du temps requis pour classer l'information augmente et passe de 43 % en 2001 à 57% en 2005 ». La perception des salariés enquêtés, concernant le temps estimé à effectuer des tâches de gestion de l'information, confirme l'impact prévisible de l'expansion du volume des documents à traiter, en même temps que les systèmes d'information n'apportent pas de solution simple et efficace pour faciliter cet aspect du travail, et que le temps alloué pour ces tâches diminue. D'après ces auteurs, Meier, en 1963 identifie déjà « la surcharge d'information comme source de stress chez les employés, productrice de dysfonctionnements opérationnels et de pertes d'efficacité ».
La diversité des systèmes de gestion de l'information et des Systèmes d'Organisation des Connaissances (SOC) (thésaurus, taxonomies, classifications, folksonomies, index…) intégrés dans les premiers ne facilite pas toujours les tâches de traitement des documents (tri, formatage, nommage, indexation, classement, stockage et recherche) car ils requièrent l'appropriation des outils par les agents et leurs affordances ne sont pas toujours suffisantes, même si leur complémentarité pourrait apporter des éléments de solution aux problèmes évoqués dans cet article.
L'objet de cet article est de présenter un nouveau type de SOC, conçu, selon les principes de l'ergonomie cognitive, pour soutenir et faciliter les activités de traitement des documents d'une organisation, dans le cadre d'un projet soutenu par l'Agence Nationale de la Recherche. Dans un premier temps nous étudieront les approches de la psychologie du travail et de l'ergonomie cognitive, en approfondissant la notion de surcharge cognitive, puis nous aborderons les transformations récentes des métiers à travers une approche socio-organisationnelle. Ensuite, nous présenterons rapidement un instrument de documentarisation collaborative multi-facettes. Enfin, nous aborderons les réflexions méthodologiques qui accompagnent le développement de cet outil en particulier des réflexions sur l'articulation entre les aspects individuels et collectifs de l'exercice d'un métier.

## 2 Approches issues de la psychologie du travail et de l'ergonomie cognitive

### 2.1 Surcharge cognitive

Notre méthode s'inscrit dans le champ des études menées en psychologie du travail et en ergonomie cognitive. Pour bien comprendre les phénomènes de surcharge cognitive, informationnelle, et communicationnelle, il nous faut d'abord définir quelques notions.

#### *Charge mentale de travail*

On distingue la charge de travail, de la charge mentale de travail. P. Falzon et C. Sauvagnac analysent la notion de charge de travail en termes de contrainte et d'astreinte. La contrainte « est définie par la tâche, et est formulée en termes d'objectifs à atteindre, de résultats attendus, de qualité à obtenir, etc… » tandis que l'astreinte « est définie en référence à l'activité. Elle est fonction du degré de mobilisation (physique, cognitive, psychique) de l'opérateur. » L'analyse de la charge consiste donc d'après ces auteurs « à identifier les contraintes de la tâche-objectifs, procédures, cadence, équipements, etc.- et des descripteurs plus ou moins directs de l'astreinte. » Il peut s'agir pour ces derniers de mesures physiques

(quantité d'acide lactique, consommation d'oxygène) mais pour un travail moins physique, comme le travail de bureau, d'autres concepts sont nécessaires.
Le concept de charge mentale a été très discuté, et correspond selon la définition proposée par Damos (1991) cité par Falzon et Sauvagnac à « une construction hypothétique, induite par la réalisation d'une tâche et provoquant la réduction de la capacité mentale à réaliser d'autres tâches. » Cela implique la théorie du canal unique qui suppose l'existence « d'une capacité limitée de traitement de l'information, capacité mobilisée proportionnellement à la difficulté de la tâche à réaliser ». (Falzon & Sauvagnac, 2004). Cette théorie a beaucoup été critiquée mais nous considérerons qu'elle garde une certaine justesse au regard des constatations empiriques des études évoquées dans cet article.

***Surcharge informationnelle***

L'enquête menée par Autissier et Lahlou (1999) révèle que « les acteurs ont l'impression de passer beaucoup de temps à des tâches de manipulation d'information qui ne font pas partie du travail pour lequel ils sont rémunérés », et qu'ils « déplorent le temps passé à la manipulation de l'information pour lequel ils ne voient pas la rentabilité immédiate et les solutions envisageables ».
D'après les mêmes auteurs, « les acteurs déclarent recevoir des volumes d'informations de plus en plus importants […] dont la difficulté d'appréciation crée des 'no man's land informationnels' » qui engendrent ensuite des coûts supplémentaires de recherche. Les auteurs de cette étude évoquent un obstacle face à cette masse croissante d'informations à gérer : les limites des capacités de traitement des acteurs, qui aboutissent à un traitement des informations « par ordre d'urgence », ce qui confine la gestion de l'information à une « gestion purement conjoncturelle ». Tout ce processus illustre le concept de surcharge informationnelle qui peut déclencher le « Cognitive Overflow Syndrome » (COS ou syndrome de débordement- ou de saturation cognitif) qui se caractérise ainsi selon Autissier et Lahlou :

- Une production croissante d'information
- Un stress des individus lié au manque de temps et au retard dans l'exécution du travail de leur cœur de métier dus aux opérations de traitement de l'information
- La difficulté à imputer ce problème à un facteur identifié et à trouver les acteurs à qui le confier
- La perte de sens : les difficultés des acteurs à créer du sens à partir des nombreuses informations à gérer aboutissent à des stratégies de gestion de l'information à court terme « sans souci de cohérence avec le fonctionnement global ou les objectifs globaux de l'organisation », ce qui peut entrainer des pertes d'informations, de connaissances, et compromet toute démarche de knowledge management.

Il semble que la surcharge informationnelle, d'après Barki et Sauders (1990), cités par les mêmes auteurs, puisse avoir des conséquences sur la qualité du travail ou causer des dégradations des conditions de travail, comme des manques de communication, de l'hostilité ou de la jalousie entre les groupes, des frictions

interpersonnelles, l'escalade hiérarchique des conflits, la prolifération de règles et de règlements, un manque d'entrain.
La surcharge informationnelle est souvent le corollaire de la surcharge communicationnelle qui correspond à « un excès de sollicitations non pertinentes pour la tâche principale ». La surcharge cognitive est définie par Rouet et Tricot (1995) comme « une excès de traitements à réaliser ou d'informations à retenir ».
La gestion de la surcharge cognitive et informationnelle peut être traitée soit dans le cadre d'approches « Systèmes de traitement de l"information » qui proposerait des outils de filtrage pour réduire le flux d'information excessif, ou des outils de traitement automatique des informations, soit dans le cadre d'approches plus socio-cognitives qui renvoient à des dimensions managériale, psychique, organisationnelles que nous adoptons.

## 3 Cadrage sociologique et organisationnel

### 3.1 Le concept de métier à l'épreuve des changements organisationnels

Les modèles contemporains d'organisation du travail, caractérisés par le développement de la flexibilité pour s'adapter à la fluctuation de la demande du marché, qui s'accompagne de l'externalisation de certains services de l'entreprise, de l'organisation par projet au sein des organisations, de la décentralisation des pôles de décision, et du développement de l'organisation en réseau aboutissent à des mutations des formes de travail qui nécessitent de la part des acteurs des organisations adhérant à ces phénomènes, des adaptations notoires. Ainsi une plus grande polyvalence est nécessaire, pour le développement de leur employabilité, une autonomie accrue, qui se traduit par le développement de l'auto-contrôle, une disponibilité croissante, due à l'exigence d'être joignable à tout moment et en n'importe quel lieu, donc d'être connecté en permanence (organisation en réseau). Ces évolutions des modèles d'organisation du travail impactent aussi la gestion des ressources humaines, qui s'individualisent et deviennent de nouveaux lieux de négociations (salaires, fiches de postes, définition des compétences mobilisées). (Boltanski & Chiapello, 1999).

#### *L'évolution de l'organisation du travail*

La recherche de la flexibilité qui trouve son origine dans des motifs économiques s'est traduite par « des techniques d'organisation et de gestion efficientes de type zéro stock, juste-à-temps, gestion en flux tendus » (Pesqueux, 2005). Il s'agit de la flexibilité appliquée à la production des biens, principalement. La flexibilité se caractérise aussi par la capacité d'apprentissage de l'organisation nécessitant de la mémoire, si bien qu'elle est liée aux performances des hommes et des systèmes d'information. La flexibilité se matérialise en prenant pour variable d'ajustement les effectifs en recourant à l'externalisation pour répondre aux fluctuations de l'activité, phénomène soutenu par les réseaux de télécommunication. (Pesqueux, 2005).
L'organisation en réseau repose elle aussi sur les infrastructures de communication qui impactent l'économie, la société et l'organisation du travail. L'organisation en réseau se singularise par les idées suivantes selon (Pesqueux, 2005) : le changement permanent, (qui engendre la réactivité face à la concurrence), un espace-temps en constant remodelage, (par l'immédiateté des communications et la connexion permanente de ses acteurs), et l'innovation organisationnelle permanente (par le foisonnement des outils de coordination, de transmission, partage, communication instantanée, et de réseaux sociaux qui évoluent continuellement).

L'organisation par projet « évoque une entreprise dont la structure est faite d'une multitude de projets associant des personnes variées dont certaines participent à plusieurs projets. La nature même de ce type de projets étant d'avoir un début et une fin, les projets se succèdent et se remplacent, recomposant au gré des priorités et des besoins, les groupes ou équipes de travail. » (Boltanski & Chiapello, 1999). Cette organisation par projets modifie la nature des rapports de pouvoir et de hiérarchie à l'intérieur des organisations, l'exercice d'un métier par un acteur ne prenant plus effet seulement dans la sphère d'un seul service d'une organisation et les acteurs n'étant plus subordonnés uniquement à l'autorité d'un responsable de ce service.
Dans le cadre de ces évolutions de l'organisation du travail le concept de métier est-il encore pertinent pour l'analyse de l'activité des personnes dans les organisations ?

### 3.2 La notion de métier réhabilitée

La notion de métier est ancienne et renvoie aux corporations de métiers, au compagnonnage, au syndicalisme. On pourrait penser que l'éclatement des structures du travail évoquées plus haut est propice à la disparition de cette conception de l'activité rémunératrice. Pourtant, on observe, d'après F. Osty, qu'un « phénomène social d'affirmation des métiers pourrait bien s'amplifier, sous le coup des nouveaux enjeux de production, réhabilitant le métier comme une configuration sociale et organisationnelle de la modernité ». Elle fonde cette affirmation qu'elle résume dans l'expression de l'émergence d'un « désir de métier » sur l'observation de trois processus dans des milieux de travail variés (Osty, 2002) :

- l'élaboration d'une compétence spécifique à travers la production, l'actualisation et la transmission de savoirs au sein d'un collectif
- la construction d'une identité collective, le collectif assurant la socialisation et l'identification à une communauté de métier
- l'édification de règles sociales, à travers la régulation, le groupe professionnel, stabilisant les règles, cherche à obtenir une reconnaissance légitime

On constate donc que le métier ne se conçoit que dans une dimension collective, d'une part à l'intérieur du collectif métier mais aussi d'autre part au sein des organisations.

#### *Définition de la notion de métier*

Commençons par souligner qu'il est possible de retrouver des éléments communs aux métiers que leur domaine d'exercice soit l'industrie ou le tertiaire : « la même revendication d'un groupe à faire reconnaitre sa différence et sa compétence sur un périmètre de techniques » et la « quête d'une stabilité mêlée au souci de statuts garantis » (Clot, 2008). Clot souligne aussi que « autrui » joue un rôle essentiel dans le sentiment d'appartenance à un métier par le biais de la reconnaissance par la hiérarchie et par les pairs. Le métier vu comme un ensemble de personnes partageant des pratiques, des savoirs et savoir-faire « devient synonyme de communauté ou de collectif d'appartenance » et par extension implique que « le jugement des pairs octroie l'appartenance au métier ».
De la même manière, d'après les trois processus observés par F. Osty, avoir un métier c'est tout d'abord « s'inscrire dans une dynamique de production d'un savoir opératoire érigeant l'expérience en mode d'apprentissage privilégié ». En effet, on assiste à un glissement de la conception des savoirs correspondant initialement aux connaissances explicites, formelles, acquises et sanctionnées par un diplôme attestant des qualifications, vers un modèle plus axé sur les

compétences, lié à la reconnaissance de l'activité réelle des individus dans le travail, dont les connaissances implicites et les savoirs pratiques sont fondés sur l'expérience. Ces compétences reconnues au sein des organisations ont vocation à être partagées et évaluées par les pairs.
Le deuxième processus, qui permet de définir la notion de métier, le sentiment d'une identité collective, se caractérise par le partage de valeurs, normes, système de représentations, et d'une culture professionnelle (règles techniques, comportements, langage spécifique) qui marque l'appartenance à une communauté métier. La quête de reconnaissance sociale accompagnant cette identification à l'activité de travail se double d'une identification plus ou moins prononcée à l'organisation.
Enfin, à travers l'édification de règles et de compétences pour la régulation, et sa capacité à les transmettre et les diffuser assurant ainsi leur pérennité, une communauté métier joue sa reconnaissance au sein de l'entreprise ainsi que la légitimité de sa professionnalisation.
Ainsi le soutien des collectifs de travail à l'aide d'outils adaptés à leurs configurations organisationnelles, pour la transmission des compétences apparait un enjeu fort de la transformation actuelle des métiers. De ce fait la gestion des documents, utile à la transmission des connaissances et compétences d'un métier se révèle cruciale. Nous proposons donc dans la suite de l'article de présenter un outil de documentarisation collaboratif multi-facettes qui nous apparait adapté à cet objectif en même temps qu'il contribue à réduire la charge cognitive que cette tâche représente.

## 4 Tâche de gestion des documents et système multi-facettes collaboratif

La surcharge cognitive peut survenir à cause de différents facteurs, une quantité de travail mal évaluée, un changement des modalités organisationnelles, un manque d'équipements adaptés, des difficultés des agents à s'adapter à leur poste de travail, une distorsion trop importante entre le travail prescrit et le travail réel… En effet, Leplat (1997) rapporte le constat suivant : « Les analystes ont fait très tôt la constatation que la tâche effectivement réalisée par l'agent (…) ne coïncidait pas toujours avec la tâche prescrite. » Ce dernier facteur nécessite de préciser quelques notions.

### 4.1 Une tâche non reconnue et des outils limités

***La gestion des documents de travail, une tâche non prescrite***
Le travail prescrit correspond à l'ensemble des activités et tâches demandées à un opérateur sous forme de consignes ou d'objectifs à atteindre dans des conditions données en échange de sa rétribution. De nombreuses théories tentent de définir les concepts d'activité et de tâche sans les épuiser. Dans cet article, nous nous intéresseront à la conception de l'activité et de la tâche proposée par Leplat (1997). Selon lui, l'activité ne peut être regardée qu'en interaction avec l'opérateur et avec la tâche prescrite, dans un environnement de travail particulier et dans lequel les autres agents conditionnent l'activité du premier. Il écrit par ailleurs : « l'activité dépend de la tâche et des caractéristiques du sujet, mais elle peut contribuer, en retour, à la définition de la tâche et à la transformation du sujet ».
L'activité se décompose en étapes, des tâches et des sous tâches, ou tâches intermédiaires, selon des degrés de raffinement variables. « La tâche est communément définie comme un but à atteindre dans des conditions

déterminées. » Leplat (1997). La tâche prescrite est une « formulation », le fruit du travail, en amont, des concepteurs de la tâche, elle peut « être conçue comme le modèle de la tâche à réaliser » destiné à un opérateur. Elle comporte donc une part d'implicite et nécessite une interprétation de la part de celui qui va la réaliser, une redéfinition qui va consister « à opérationnaliser celle-ci en fonction des conditions présentes » pour aboutir selon Leplat (1997) à une tâche redéfinie. La tâche redéfinie prend en compte les objectifs de la tâche prescrite mais aussi d'autres buts, notamment des buts personnels (promotion, intégration sociale, santé...). Elle constitue un compromis et va guider l'action qui caractérisera la tâche effective correspondant au travail réel.

Il est également possible, de hiérarchiser les tâches et sous tâches, en fonction du prescrit. En effet, la tâche d'indexation et de gestion de l'information plus généralement, peut dans certaines situations de travail être considérée comme une tâche redéfinie par les opérateurs. En effet, si en matière d'archivage et de gestion des documents à valeur probante, « engageante », (Chabin, 2009) ou patrimoniale, bien souvent les organisations élaborent des prescriptions pour la conservation et l'indexation de ces documents, le champ des documents intermédiaires (mails, rapports en cours de rédaction, mémo…), ou DopA (Documents pour l'Action) (Zacklad, 2006) n'est pas traité par ces prescriptions. Leur gestion n'est pas prescrite, et peu outillée, laissant les utilisateurs livrés à eux-mêmes quant à l'organisation, au classement et à l'indexation de leurs documents de travail.

Cette gestion des documents de travail n'apparait donc pas nécessairement dans les prescriptions tout en restant une tâche incontournable pour la réalisation des missions principales des agents. Si cette tâche de gestion de l'information, comme nous l'avons vu occupe de plus en plus les agents mais n'est pas formulée dans les prescriptions du travail, alors elle n'est pas non plus reconnue dans l'activité qui détermine les rétributions. Elle souffre donc d'un manque de visibilité et le travail réel effectué dans ce sens souffre d'un manque de reconnaissance par les prescripteurs.

Ainsi, la gestion des informations considérée comme une tâche secondaire, ni prescrite ni formulée, par rapport à l'activité principale des agents, contribue à leur surcharge cognitive et engendre une mauvaise considération de cette tâche par ceux-ci. De plus, la non-reconnaissance de l'empiètement de ces activités de gestion de l'information sur les activités principales des agents par les prescripteurs risque bien souvent d'engendrer un équipement mal adapté à cet aspect de leur travail.

Il nous apparait donc qu'il peut être utile de considérer cet aspect du travail réel effectué par les opérateurs et d'outiller les stratégies qu'ils mettent en place pour réduire le coût cognitif de ces tâches effectives de gestion de l'information qui semblent trop les mobiliser avant que ces dysfonctionnements n'aboutissent à une surcharge cognitive.

***Des outils de gestion des documents existants peu satisfaisants***

Dans le cas des tâches de gestion de l'information, des outils d'indexation et de recherche sont mis à disposition des opérateurs mais nous faisons le constat que bien souvent ces outils d'indexation et de classement des documents tels que les langages documentaires (thésaurus, taxonomies, classifications…) se révèlent trop formels et trop difficiles à utiliser par les acteurs, non documentalistes. (Mas, *et al.,* 2008). De plus en plus, la gestion des documents dans les organisations est confiée non plus comme traditionnellement à des documentalistes mais aux acteurs, occupés par d'autres métiers, notamment en matière d'archivage des documents. Plus globalement les acteurs, surtout dans le secteur tertiaire se trouvent

confrontés à la surcharge informationnelle étant donné l'évolution de l'organisation du travail vers une organisation en mode projet, dans un contexte économique où l'externalisation des services et l'autonomie des travailleurs sont privilégiées. Ainsi ils sont très sollicités par le biais des TIC par les intervenants externes et par leurs co-équipiers internes dans le cadre des projets, dans une configuration du travail où l'information est considérée comme source de profit et de productivité (Boltanski & Chiapello, 1999).

Ces tâches de gestion de l'information ne représentent donc pas l'activité principale de leurs métiers mais constituent un moyen de l'exercer. Nous faisons l'hypothèse qu'une bonne gestion des documents accroit l'efficacité des acteurs dans leurs tâches principales. En effet, les acteurs dans l'exercice de leurs activités sont confrontés à des manques de connaissances, qui se traduisent par des besoins informationnels qui eux-mêmes déclenchent des recherches d'information.

Dans le cadre de cet article, nous nous intéresserons plutôt aux processus de gestion des documents (tri, formatage, nommage, indexation, classement, stockage et recherche) et plus précisément aux processus d'indexation. Notre approche met l'accent sur cet aspect de la gestion des documents car nous considérons qu'une indexation rigoureuse en amont de la recherche, quel que soit le système de recherche utilisé (moteur de recherche, système à facettes…), la rend plus efficace, car elle gagne en pertinence, rapidité et exhaustivité.

Quelles sont, dans ce cadre, les principales difficultés rencontrées par les agents à utiliser les langages documentaires ?

La rigidité des SOC due à la conception de ces outils, effectuée par d'autres acteurs que ceux qui les utilisent, généralement descendante et en amont du travail des agents, peut aboutir à créer une dichotomie entre l'outil et les besoins des agents en matière de gestion des documents. En effet, l'idée que se font les concepteurs des SOC des besoins des agents ne correspond pas nécessairement à leur besoins réels, de la même manière que le travail prescrit ne correspond souvent pas complètement au travail réel. De plus ces SOC sont élaborés pour privilégier les intérêts de l'organisation et non celle des agents qui y fournissent un travail. (Mas, *et al.,* 2008).

Les modalités d'évolution de ces SOC sont souvent peu flexibles, et l'ajout de nouvelles valeurs est souvent décidé par les mêmes concepteurs ou responsables des SOC ce qui ne permet pas aux agents de gérer leurs documents selon leur propre représentation de leur activité. A l'opposé, si les modalités d'alimentation de valeurs dans les SOC sont trop souples, cela aboutit à des SOC trop volumineux et à des redondances qui compliquent l'usage de ces outils.

Le niveau de précision des SOC peut s'avérer insuffisant pour la gestion des documents de travail des agents, qui reflètent leur activité à un grain plus fin que celui présents en général dans les systèmes d'indexation.

Enfin, l'investissement en temps pour connaitre les SOC proposés par l'organisation pour la gestion des documents et les utiliser se révèle être un frein à l'usage de ces outils dont il est parfois nécessaire de déplier les branches arborescentes, par exemple pour les thésaurus, ou de passer en revue les différentes catégories présentées pour trouver la plus adaptée, dans les classifications, pour des usagers non professionnels de la gestion de l'information.

Il nous apparait que l'utilisation d'un système de documentarisation collaborative multi-facettes, proposant un compromis entre une approche exclusivement formelle (et contraignante) et un système informel, peut contribuer à la réduction de la charge cognitive due aux tâches d'indexation des documents.

### 4.2 Documentarisation et système à facettes

***Présentation de l'outil***

Dans le cadre d'un projet soutenu par l'Agence Nationale de la Recherche (ANR), associant des chercheurs de l'équipe Tech-CICO de l'Université de Troyes et de l'équipe Dicen du Conservatoire National des Arts et Métiers (CNAM), ainsi que la société Cogniva Europe, un instrument de documentarisation collective multi-facettes intuitif et ludique est en cours de développement.

Le principe de cet instrument est d'intégrer les opérations de nommage/ classement/ indexation/ stockage en une seule et même opération réalisée à l'aide de facettes et de valeurs de facette, représentées pour ces dernières sous forme d'icônes appelées des Sémiotags. Un Sémiotag est donc une valeur de facette, représentée sous forme d'icône associée à un tag, qui la qualifie.

La documentarisation est l'opération qui consiste à pérenniser les documents en les dotant d'attributs permettant leur ré – exploitation. Elle désigne le nom du document, les mots-clés et les autres types de métadonnées. Cette documentarisation est d'autant plus efficace quand elle intègre une réflexion, une anticipation sur les usages futurs ou les potentielles réutilisations des documents, dans un autre contexte d'activité ou par d'autres acteurs.

Le système à facettes est inspiré de la classification à facette élaborée en 1924 par le mathématicien et bibliothécaire Ranganathan qui permettait d'indexer en fonction de catégories de base (les facettes), mais aussi d'exploiter les possibilités combinatoires. Le modèle de facette de Ranganathan comprenait 5 facettes qui pouvaient selon lui constituer « la syntaxe fondamentale (absolute syntax) sous-jacente à tout sujet » (MANIEZ). Un des chercheurs du Classification Research Group, Brian C. Vickery, «récuse l'ambition d'une syntaxe fondamentale des sujets qui serait applicable à la médecine, à la psychanalyse, ou à la linguistique et à la métallurgie. Pour lui, les facettes sont d'autant mieux adaptées à une indexation de qualité que leur domaine d'application est plus étroit, et il recommande la construction de schémas spécialisés ». (MANIEZ)

C'est aussi ce que propose notre approche en adaptant les facettes au contexte de l'organisation qui souhaite l'utiliser. Un jeu de facettes est quand même proposé par défaut avec par exemple les valeurs suivantes (présentées dans l'application sous forme d'icônes paramétrables):

Genre : Note, Compte- rendu, Rapport, Documentation, etc.

Projet : Nom du projet X, Nom du projet Y, Nom du projet Z, etc.

Sujet : Comptabilité, Partenariats, etc.

Statut : Brouillon, En cours, Validé, Archivé, etc.

Le principe de l'outil est donc de proposer à l'utilisateur un jeu de facettes et de valeurs de facettes (les Sémiotags) capables de décrire ses documents de travail, qu'il sélectionnera dans le flux de son activité pour qualifier ses documents, leur donner un titre, leur ajouter des métadonnées et le classer dans le dossier approprié, en intégrant ces différentes opérations dans l'opération de sélection des Sémiotags dans l'interface d'indexation. L'utilisateur dispose potentiellement de la liberté de créer des facettes personnelles avec des valeurs privées et également des facettes et valeurs partagées avec d'autres membres de l'organisation, par exemple d'autres personnes du même métier. L'interface de recherche reprend les facettes proposées dans celle d'indexation et agit comme un filtre au fur et à mesure que l'utilisateur sélectionne des Sémiotags pour effectuer sa recherche de documents. Ainsi l'utilisateur n'est plus confronté aux difficultés liées au choix d'un nom pertinent pour un document, opération qui interrompt le travail en cours au moment de l'enregistrement du document, en nécessitant de se projeter dans

l'utilisation future ou potentielle du document en cours de rédaction, ni aux difficultés liées au choix des mots-clés à lui associer qui nécessite de parcourir par exemple un thésaurus, ni enfin aux difficultés liées au choix du bon dossier pour classer le document, qui nécessite de parcourir une classification à l'emplacement de stockage approprié.

## 5 Collaboration et partage

Comment caractériser les modes de collaboration et de partage dans le cadre de l'utilisation d'un outil de documentarisation multi-facettes pour la gestion de documents et le partage des connaissances et compétences métier ?

### 5.1 Pratiques collaboratives et types de collectifs

Dans le tableau suivant, les pratiques de gestion de l'information sont considérées en fonction des types de collectifs classifiés selon le niveau de structuration de ces collectifs. Leurs modalités organisationnelles (types d'activités : individuelles, métiers et de coordination), en partie déterminées par la taille des types de collectifs proposés (nombre d'acteurs impliqués) contribuent à déterminer si les facettes doivent être collectives ou individuelles. Les pratiques de gestion de l'information sont aussi classifiées dans ce tableau selon le degré de formalisme du SOC (en fonction des types de SOC utilisés : informel, semi-formels, formels) qui participe aussi à la détermination de la nature (collective ou individuelle) des facettes. Ce degré de formalisme est lui- même dépendant des choix en matière de gouvernance des systèmes de gestion de l'information dans ces collectifs et du degré d'évolutivité de ces systèmes, recherché par les instances décisionnaires.

Les facettes collectives peuvent être partagées au niveau de l'entreprise ou d'une autre forme de collectif comme un groupe d'acteurs travaillant sur un même projet ou exerçant le même métier. On constate donc qu'il y a deux types de facettes collectives : les facettes « organisationnelles » plutôt dédiées à des fins d'indexation du contexte organisationnel de création du document, pour renseigner les futurs utilisateurs sur les activités de coordination associées aux documents, et les facettes « métier » plutôt dédiées au partage d'informations entre acteurs, par exemple du même métier, ou appartenant à une communauté de pratiques, dont les acteurs peuvent être disséminés dans plusieurs départements de la même organisation. Les utilisateurs peuvent aussi créer des facettes personnelles qui ne seront pas visibles par les autres utilisateurs pour l'indexation des documents, selon leur propre appréhension de leur activité.

Les degrés de formalisme qui concernent la contrainte que l'on souhaite exercer sur l'activité de gestion de l'information avec l'outil, impactent notamment les procédures d'alimentation des valeurs de facettes. Soit l'organisation souhaite exercer une pression forte et l'alimentation des valeurs de facettes est déterminée selon une approche « descendante », par exemple administrée par un responsable de l'application. Soit elle est vue comme participative, et les acteurs ont la possibilité de proposer des valeurs qu'ils jugent utiles à la communauté (approche semi-formelle, « ascendante ») pour compléter le paramétrage initial de l'application, issu de l'analyse méthodologique en amont. Un troisième scénario est envisagé pour des utilisateurs autonomes, qui créent leurs facettes et valeurs de facettes individuellement.

| | **SOC Informels** (type folksonomies) | **SOC semi-formels** (approches par point de vue avec contraintes simples) | **SOC Formel** (facettes basées sur des contraintes ou ontologies) |
|---|---|---|---|
| **Individuel** : lié à l'activité métier ou la coordination des activités individuelles | *Tagging rapide par rapport à de nouvelles problématiques personnelles* | *Organisation systématique mais évolutive de l'activité individuelle (métier)* | *Organisation systématique selon les prescriptions en vue de l'archivage dans le cadre d'une production individuelle de type recherche, par exemple, au sein d'une organisation* |
| **Groupe** : différentes structurations de l'activité possibles (réseau, communauté, organisation projet…) impliquant des problématiques liées à l'activité métier ou à la coordination | *Exploration collective d'une nouvelle problématique métier ou managériale* | *Organisation systématique mais évolutive d'une activité communautaire (dans le cadre du collectif métier ou projet)* | *Connaissances ou modalités de coordination formalisées et stable* *(Prescriptions au niveau métier ou projet)* |
| **Entreprise :** Vu la diversité des métiers, plutôt des problématiques liées à la coordination | *Emergence d'une nouvelle problématique (surtout relevant de la coordination)* | *Organisation systématique mais évolutive de processus organisationnels sur de nouveaux projets donnant lieu à des expérimentations* | *Modalités d'organisation formalisées et stables au niveau de la coordination entre les grands services de l'entreprise* |

**Figure. 1.** *Pratiques de gestion de l'information en fonction des types de collectifs considérés et des degrés de formalismes des SOC utilisés*

L'outil de documentarisation collaborative multi-facettes se situe plutôt au niveau des SOC proposés dans la deuxième colonne. Il se rapproche plus d'un SOC semi-formel représentant un compromis entre les approches folksonomiques et les approches formelles.

Il constitue un support pour les évolutions des documents, des pratiques d'indexation, des utilisations diverses qui sont faites des documents et des points de vue qui les sous-tendent. Par ailleurs, la caractérisation des utilisations variées d'un document peut nous renseigner sur les pratiques d'un métier, et participer par ce biais à la capture de la connaissance informelle et à la capitalisation des connaissances.

Par exemple, un rapport réalisé par un expert du vieillissement des matériaux sur un élément particulier d'une installation industrielle peut servir pour un autre acteur comme un ingénieur en charge d'élaborer des procédures de maintenance. Ce même document peut aussi servir à un autre chercheur expert qui travaille sur l'évaluation d'un nouveau type de matériau ayant des composants similaires. Ce rapport va d'abord être indexé par le premier expert, auteur du document, dans le cadre de son activité individuelle au sein d'un métier à l'aide d'un SOC semi-formel, puis sa version finale pourra être archivée à l'aide d'un SOC formel, tandis que la documentation telle que la bibliographie en lien avec ce rapport pourra être indexée à l'aide d'un SOC informel sur le poste de l'expert.

L'ingénieur en charge des procédures de maintenance utilisera ce rapport comme une ressource documentaire éventuellement bibliographique dont il ira chercher la version finale dans la base d'archivage à l'aide d'un SOC formel. Il pourra intégrer ce document dans sa documentation personnelle à l'aide d'un SOC informel ou

bien si cela présente un intérêt pour les autres acteurs de son métier, dans un emplacement partagé à l'aide d'un SOC semi-formel.
Enfin, le chercheur qui travaille sur les nouveaux matériaux ira lui aussi chercher ce rapport dans la base d'archivage et pourra l'intégrer dans sa documentation personnelle à l'aide d'un SOC informel ou bien si cela présente un intérêt pour ses collaborateurs, dans un emplacement partagé à l'aide d'un SOC semi-formel.
Chacun de ces acteurs, l'expert, l'ingénieur, ou le chercheur vont pouvoir produire une indexation différente de ce document qui peut apporter des éléments de connaissances sur le travail réel des opérateurs, des éléments de réflexions sur la capitalisation des connaissances et sur de nouvelles modalités de coordination entre acteurs.

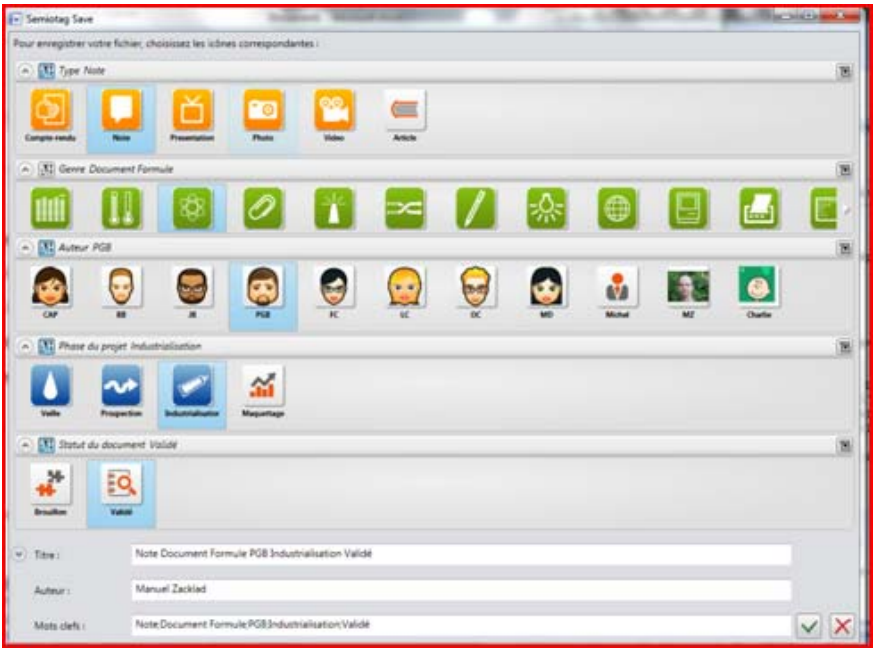


**Figure. 2.** *Interface d'indexation de l'outil de documentarisation collective multi-facettes*

Dans l'articulation entre les facettes collectives et individuelles, nous considérons d'une part les facettes organisationnelles qui permettent de coordonner les acteurs au niveau d'un projet, et de partager les informations nécessaires à la réalisation de ce projet, et d'autre part les facettes métier qui forment un compromis, une dimension de partage intermédiaire entre les aspects organisationnels et individuels et qui permettent de créer un espace d'échange pour les acteurs d'un même métier à l'intérieur d'une organisation. Bien souvent les informations communes à plusieurs acteurs d'un même métier mais éclatés dans différents lieux ou projets d'une entreprise sont amenées à circuler de manière informelle sans possibilité pour l'organisation de profiter de l'émulation et de la richesse des échanges entre pairs du même métier, ni possibilité de capitaliser ces échange, faute d'espace ou d'outil adaptés.

### 5.2 Le passage de l'individuel au collectif à travers le prisme du métier

La dimension partagée, collective, des facettes implique une réflexion sur la nature, la taille et les activités des collectifs partageant un même jeu de facettes, plusieurs jeux de facettes métier pouvant coexister au sein de la même organisation, au

même titre qu'on trouve fréquemment plusieurs métiers dont les acteurs coopèrent pour la production des biens et services proposés par une entreprise.

***Une conception du métier en évolution dynamique***

D'après Y. Clot, lorsqu'un opérateur commence dans son métier, il respecte au maximum la dimension impersonnelle (prescrite) du travail. Petit à petit, il s'approprie cette dimension, notamment à travers les échanges avec ses pairs, puis devient un expert. La dimension personnelle dans le métier atteint alors ses limites, le genre professionnel dans lequel il est exercé est trop restreint.

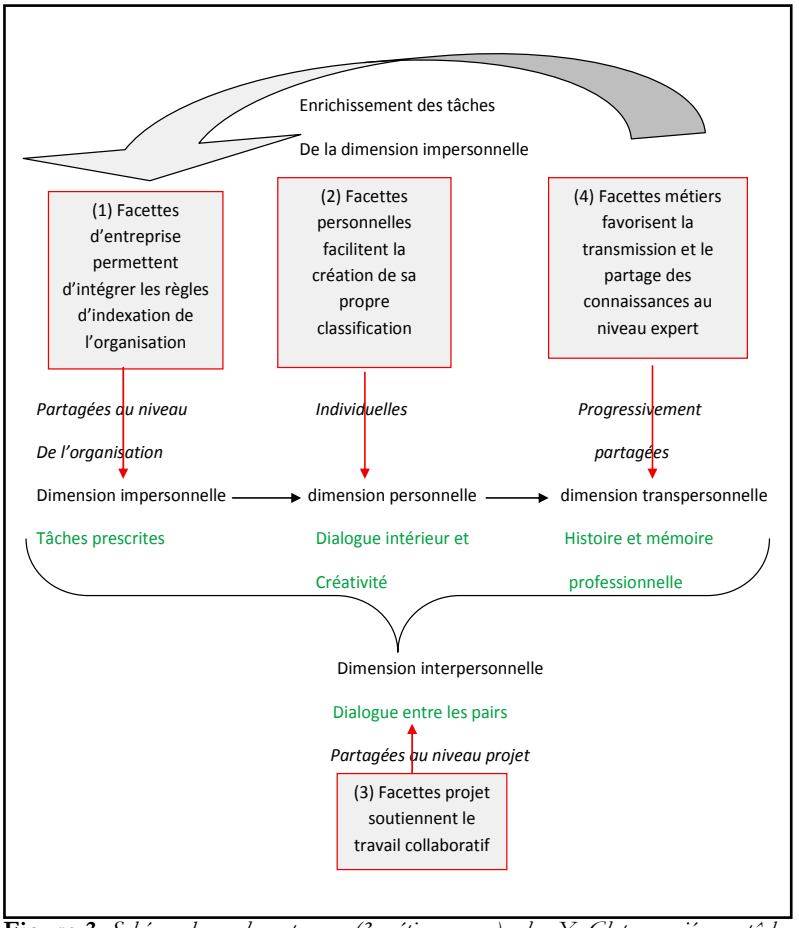


**Figure.3**. *Schéma du cercle vertueux (3 métiers en un) selon Y. Clot, associé aux tâches de gestion de l'information avec l'outil de documentarisation collaborative multi-facettes*

Le sujet va donc devoir inventer, c'est -à- dire faire l'inventaire de ce qui existe déjà, identifier les limites et trouver avec le collectif de nouveaux moyens pour exercer son métier. Cette stylisation du genre professionnel rend l'expert auteur de son métier. La dimension personnelle du métier devient plus présente, et ce processus peut aboutir à un enrichissement des tâches prescrites initiales à partir des tâches réalisées au niveau de maîtrise de l'expert et à une dimension impersonnelle (prescrite) plus forte.

« C'est que le travail collectif soutenu par le collectif de travail a rendu possible ces développements mutuellement indépendants ». (Clot, 2008). C'est aussi par ce biais que la mémoire transpersonnelle du genre professionnel s'affranchit de chacun pour être mieux disponible pour tous.
Ce processus d'évolution de l'exercice d'un métier, qualifié de « Cercle vertueux » des métiers par Y. Clot, est représenté dans la figure suivante, en association avec l'usage des facettes de l'instrument de documentarisation collective multi-facettes présenté plus haut.
Les facettes selon leur aspect individuel ou partagé ont des fonctions différentes, en cohérence avec cette conceptualisation de l'activité d'un métier.
En suivant le schéma proposé ci-dessus, les facettes au niveau entreprise (1) facilitent le respect et l'usage des plans de classement et de nommage qui traduisent l'aspect impersonnel des tâches de gestion de l'information, grâce au paramétrage initial qui respecte ces contraintes.
La dimension individuelle du développement d'un métier (2) est favorisée par le biais de la création de facettes personnelles pour indexer les documents selon ses propres intérêts.
Les facettes projet (3) permettent aux acteurs d'un projet de partager des informations relatives à ce projet.
Les facettes métier (4) sont liées à l'expertise des acteurs qui vont pouvoir contribuer au savoir présent dans l'organisation à travers ses acteurs en partageant des facettes et valeurs de facettes qui pouvaient à l'origine être des facettes et valeurs individuelles.
La souplesse technique d'alimentation des valeurs de facettes permet aussi aux utilisateurs de traiter les documents provenant des activités imprévues et innovantes propres à la réalisation des objectifs. Cela en fait un outil plus proche de la conception de l'activité qui intègre à la fois l'imprévu et les autres objectifs de l'activité à un niveau plus global. Ceci peut favoriser l'innovation au niveau du travail réel des opérateurs.

### *4.2.2 Innovation et capitalisation des connaissances*

Pour Nonaka et Takeuchi (Nonaka & Takeuchi, 1997), l'innovation est conditionnée par l'émergence de nouvelles connaissances organisationnelles. Cette création de nouvelles connaissances se caractérise par la conversion de connaissances. Ils distinguent les connaissances tacites qui sont personnelles, spécifiques au contexte, difficiles à formaliser et à communiquer, des connaissances explicites sont transmissibles dans un langage formel voire systématique. Il existe plusieurs modes de conversion des connaissances (socialisation, extériorisation, combinaison, intériorisation) mais celui le plus efficace est l'extériorisation c'est-à-dire la conversion de connaissance tacites en connaissances explicites, à l'aide de métaphores, analogies, hypothèses, concepts, modèles. Ce dernier processus prend effet dans un dialogue et/ou une réflexion collective.
Pour favoriser la création de nouvelles connaissances organisationnelles, les auteurs décrivent cinq conditions :
- « l'intention » : la stratégie et le positionnement de l'organisation en matière de développement des connaissances, de définition des connaissances à développer et des outils mis à disposition comme supports
- « l'autonomie » : l'organisation en mode projet met en place des structures de travail dont la diversité fonctionnelle des acteurs donne l'opportunité à ces derniers de s'auto-organiser, au sein du collectif projet, et à travers la flexibilité de

cette structure leur permet de faire face en autonomie à des problèmes complexes, en les poussant à faire appel à des solutions nouvelles, en stimulant leur créativité
- « la fluctuation et le chaos » ou « chaos créatif » : les auteurs entendent par là, un « ordre sans récursivité », qui augmente les interactions entre l'organisation et l'environnement extérieur. La fluctuation permet de rompre les routines. Elle conduit à une ré-interrogation du savoir et de nouvelles interactions sociales, dans la recherche d'alternatives. Elle permet de renforcer l'engagement subjectif des individus.
- « la redondance » : ce terme décrit l'existence d'informations allant au-delà des exigences immédiates de l'organisation, qui permettent aux individus d'avoir une visibilité sur le travail produit par les autres, leur apporte une ouverture d'esprit supplémentaire pour de nouvelles perspectives, et facilite la communication à l'intérieur de l'entreprise.
- « la variété requise » : cette expression sert à expliciter le fait que les individus doivent avoir un égal accès à l'information qui concerne l'entreprise pour une meilleure prise de conscience de l'environnement extérieur dans laquelle elle s'intègre, pour une plus grande mobilisation des individus et une créativité stimulée.
Il nous semble que l'utilisation de l'outil de documentarisation collective multi-facettes peut renforcer certaines de ces conditions, notamment l'autonomie et la redondance, dans le cadre de la transformation de connaissances tacites en connaissances explicites.
En effet, les facettes personnelles par la création de valeurs adaptées à sa propre représentation du métier exercé et de ses activités se révèlent un outil de développement de l'autonomie, et donc de la créativité des utilisateurs, tandis que les facettes partagées métier correspondent à une mutualisation des connaissances en vue d'anticiper de nouvelles actions et de nouvelles alternatives de réponses à des problèmes, collectivement. Il semblerait donc que l'outil présenté peut constituer un support pour soutenir et faciliter l'innovation au sein des organisations. Les autres conditions évoquées relèvent plus des choix en matières de stratégies, management et de hiérarchie dans une organisation.

## 6 Conclusion

L'apport de la documentarisation collaborative multi-facettes à l'activité de capitalisation des connaissances consiste à outiller cet aspect de la gestion de l'information dans les organisations. La possibilité pour l'utilisateur de partager des facettes et des valeurs de facettes provenant d'une partie plus globale de l'entité, comme le département, ou d'un autre acteur du même métier favorise le travail collaboratif et le partage des informations, connaissances, compétences.
De la même manière qu'Yves Clot aborde les instruments que constituent par exemple de petits carnets de notes sur l'activité, qu'il qualifie de « prothèses cognitives » ou encore « d'instruments auxiliaires », l'instrument de documentarisation collective multi-facettes permet de mutualiser, dans la base de connaissances, ces informations issues de l'expérience grâce la simplicité d'utilisation de l'outil alors qu'elles étaient jusqu'alors peu exploitables et peu partagées.
Il écrit à propos de ces petits carnets de note : « Ces objets (…) sont précisément les particules anonymes où se trouve consignée la recherche d'efficience des collectifs de travail. Mémoire objectivée du milieu intérieur, c'est en partant d'eux

qu'on peut remonter le cours de l'activité en conservant ses « annales ». (…) Ils ouvrent un accès privilégié aux réserves d'efficience du travail humain ».
Il écrit encore à propos de ces « instruments auxiliaires » que « quand ils servent, ce n'est pas à appliquer les solutions qu'ils renferment, mais à résoudre des problèmes inédits en permettant de réinvestir dans un autre contexte l'expérience acquise. »
L'outil de documentarisation collective multi-facettes s'inscrit dans cet axe de recherche d'économies cognitives. La présence des icones participe aussi de cette démarche, en comptant sur la de la perception facilitée des dimensions d'indexation, par l'aspect graphique des icones.

**Remerciements**
*Cette recherche a bénéficié du soutien de l'ANR MIIPA-Doc n°2008 CORD 014 03.*

## 7 Bibliographie

1 D. Autissier, S. Lahlou, Les limites organisationnelles des TIC : Emergence d'un phénomène de saturation cognitive, *Actes de communication à la 4ème Conférence de l'AIM,* Essec Cergy (95) *1999,* p. 121 – 130

2 L. Boltanski et E. Chiapello, *Le nouvel esprit du capitalisme*, Gallimard, Coll. NRF Essais, Mayenne. 1999

3 M.A. Chabin, Maitriser les documents qui engagent la collectivité, *Les nouveaux territoires de l'information et de la documentation dans les collectivités territoriales*, ENACT/ INTD, Paris, 10 novembre 2009, [en ligne] <http://intd.cnam.fr/servlet/com.univ.utils.LectureFichierJoint?CODE=1258127766226&LANGUE=0> Consulté le 15/09/2010

4 Y. Clot. Le *travail sans l'homme, Pour une psychologie des milieux de travail et de vie*, coll. Sciences humaines et sociales, La Découverte/ Poche, Paris. 2008

5 Y. Clot. *Travail et pouvoir d'agir*, coll. Le travail humain, PUF, Paris. 2008

6 P. Falzon et C. Sauvagnac, Charge de travail et stress, in *Ergonomie*, sous dir. P. Falzon, PUF, Paris. 2004

7 H. Isaac, E. Campoy, M. Kalika, Surcharge informationnelle, urgence et TIC. L'effet temporel des technologies de l'information, *Revue Management & Avenir*, 2007, N° 12, p. 153 – 172

8 J. Leplat. Les automatismes dans l'activité : pour une réhabilitation et un bon usage », in *Repères pour l'analyse de l'activité en ergonomie*, coll. Le travail humain, PUF, Paris, 2008

9 J. Leplat. *Regards sur l'activité en situation de travail, Contribution à la psychologie ergonomique*, coll. Le travail Humain, PUF, Paris. 1997

10 J. Maniez. *Actualité des langages documentaires, Les fondements théoriques de la recherche d'information,* ADBS Editions, Paris. 2002.

11 S. Mas, A. Bénel, J.P. Cahier, M.Zacklad, Classification à facette et modèles de points de vue : Différences et complémentarité, *Actes du 36ème congrès annuel de l'Association canadienne des sciences de l'information (ACSI), University of British Colombia*, Vancouver, 5-7 Juin 2008

12 I. Nonaka et H. Takeuchi, *La connaissance créatrice, La dynamique de l'entreprise apprenante*, De Boeck Université, Coll. Management, Paris Bruxelles. 1997

13 F. Osty. *Le désir de métier, Engagement, identité et reconnaissance au travail*, PUR, Rennes. 2002

14 Y. Pesqueux, *Organisations : modèles et représentations*, Puf, Coll. Gestion, Paris. 2002

15 J.F. Rouet et A.Tricot, Recherche d'informations dans les systèmes hypertextes : des représentations de la tâche à un modèle de l'activité cognitive, *Sciences et Techniques Educatives*, 1995, 2 (3), p. 307 – 33

16 Zacklad, M. 2006. Documentarisation processes. Documents for Action (DofA):the status of annotations and associated cooperation technologies. Computer supported cooperative work, 0925-9724, June 2006, 15(2-3): 205-228.

17 M. Zacklad, H. Zaher, E. Lewkowicz, C. Zhou, G. Bertin, Une Interface Homme Machine combinant approche par facettes et icônes, *Rapport du projet Müppa-Doc*, prochainement en ligne (www.mipadoc.org).

# Structure, historique et évolution des dictionnaires arabes : le cas d'iSPEDAL

**Abd el Salam al Hajjar (1)(2), Mohammad Hajjar (1), Khaldoun Zreik (2)**

abdsalamhajjar@hotmail.com / m_hajjar@ul.edu.lb / zreik@univ-paris8.fr
(1) Institut Universitaire de Technologie, Université Libanaise, Saida, Liban
(2)Laboratoire Paragraphe, Université de Paris 8 - Vincennes - Saint-Denis, France

**Résumé:** Dans cet article, nous présentons un survol des dictionnaires arabes classiques et électroniques. Un dictionnaire qui existe sous forme papiers ou en format fichier électronique plat est considéré comme classique. Ces dictionnaires sont aux alentours des cinquante et ils couvrent plusieurs époques, du VIIIe siècle jusqu'à présent. Ces dictionnaires présentent les mots qui y sont contenus selon plusieurs ordres. Ces ordres peuvent être aléatoire, lexicographique, phonétique, selon la prononciation, selon la racine du mot ou selon les sujets. Les dictionnaires arabes classiques ne sont pas structurés. La consultation de l'information se fait d'une façon linéaire, l'utilisateur peut chercher un mot uniquement à partir de la liste des mots. Ce qui limite l'efficacité de l'utilisation de ce type de dictionnaires. L'avènement de l'outil informatique et l'augmentation des moyens de communication ont changé les modes de stockage et d'accès aux informations. De ce fait, les dictionnaires électroniques à usage humain se trouvent libérés des contraintes de leurs versions papiers et répondent à un certain nombre de besoins pratiques d'actualisation et d'exploitation pour une population hétérogène d'utilisateurs. Plusieurs dictionnaires électroniques sont aujourd'hui disponibles sur le Web ou commercialisés sur des supports électroniques (DVD). Les dictionnaires électroniques sont des systèmes de recherche qui extraient les informations morphologiques utiles à partir des versions électroniques des quelques dictionnaires classiques. Ils peuvent présenter ces informations sous forme exploitable et structurée. Dans cet article, nous avons présentée sept dictionnaires électroniques. Le premier dictionnaire électronique arabe est apparu en 1998. En 2010, nous avons proposé « An Improved Structured and Progressive Electronic Dictionary for the Arabic Language » (iSPEDAL) qui est une version améliorée du « Dictionnaire Electronique Structuré et Evolutif de la Langue Arabe » (DESELA). iSPEDAL est un dictionnaire structuré et évolutif de la langue arabe, qui est présenté sous la forme d'une base de données relationnelle ou d'un document XML, facilement exploitable en utilisant un langage de requête appropriée. Il contient les racines, les préfixes, les suffixes, l'infixe, et les modèles, ainsi que des informations fournies par un dictionnaire standard. En outre, pour un mot donné, il fournit des liens avec ses racines, des affixes associés, et son modèle. Il est enrichi grâce à un système automatique, à partir d'un ou de plusieurs dictionnaires textuels classiques et à partir d'un corpus textuel arabe quelconque. Nous présentons également quelques caractéristiques de la langue arabe ainsi que les éléments essentiels de sa morphologie.

**Mots-clés :** Corpus, Dictionnaire Classique, Dictionnaire Electronique, Langue Arabe, Morphologie.

## 1. Introduction

La langue arabe est une langue orientale sémitique qui s'écrit et qui se lit de droite à gauche. Elle est composée de 28 lettres et de 8 signes diacritiques. L'arabe est la quatrième langue parlée dans le monde [45], [47]. C'est la première langue de plus de 330 millions arabophones qui sont répartis sur 22 pays [16], et par plus de 1.3 billion musulmans à travers le monde [46].

La richesse morphologique de la langue arabe a fait l'objet de beaucoup de travaux réalisés par des linguistes arabes [2], [11], [24], [28], [30],[54]. L'essentiel de ces travaux est la proposition des dictionnaires au départ classique (maintenant disponible aussi sous format des fichiers plats) qui évoluent, de plus en plus, vers les dictionnaires électroniques [39], [49], [29], [22]. Un dictionnaire est un ouvrage de référence contenant l'ensemble des mots d'une langue ou d'un domaine d'activité, généralement présentés sous un certain ordre qui est, en générale, alphabétique. Le contenu des dictionnaires classique et des dictionnaires électroniques est essentiellement le même. La différence principale entre les dictionnaires classiques et électroniques n'est pas seulement le support sur lequel est présentée l'information. Toutefois, cette différence d'ordre technique en entraîne plusieurs autres aux niveaux de l'utilisation, de la présentation, du contenu, des capacités de recherche, des fonctions bureautiques et des aspects techniques [43]. Avec les dictionnaires classiques, la consultation se fait de façon linéaire, tandis qu'avec les dictionnaires électroniques, elle peut se faire de façon fragmentée ou linéaire et l'utilisateur peut effectuer différents types de recherche, notamment recherche alphabétique, recherche dans le texte entier, recherche de formes fléchies et recherche à partir du lemme. La recherche avec les dictionnaires électroniques offre ainsi plus de flexibilité que celle avec les dictionnaires classique [42], [53], [58].

En générale, un dictionnaire indique la racine, la définition, l'orthographe, les sens et les modes d'utilisation d'un mot donné [6], [8], [18], [42]. Par contre, ils ne sont pas directement exploitables puisqu'ils sont aux formats textuels dans le cas des dictionnaires classiques et non structurés dans le cas des dictionnaires électroniques (fichiers plats). Ce contexte nous a amené à construire un dictionnaire électronique structuré, évolutif et informatiquement exploitable langue arabe : iSPEDAL (An Improved Structured and Progressive Electronic Dictionary for the Arabic Language) [38] qui est une version améliorée du DESELA (Dictionnaire Electronique Structuré et Evolutif de la Langue Arabe) [5]. iSPEDAL 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. il 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.

Dans ce papier, nous présentons une cinquantaine de dictionnaires classiques et sept dictionnaires électroniques avec une brève historique des ces dictionnaires, en plus de leurs évolutions vers les dictionnaires électroniques [48]. Nous présentons aussi le cas du dictionnaire iSPEDAL [38].

Le reste de ce papier est organisé de la façon suivante. Dans la première section, nous présentons les éléments essentiels de la morphologie de la langue arabe. La section suivante présente un survol sur les dictionnaires classiques arabes et leurs caractéristiques. Dans la section trois, nous présentons les dictionnaires

électroniques arabes avec leurs évolutions. La section quatre présente le cas du dictionnaire électronique arabe iSPEDAL que nous avons développé. Dans la dernière section, nous terminons avec la conclusion générale les perspectives de ce travail.

## 2. Les éléments essentiels de la morphologie arabe

Le vocabulaire de la langue arabe est essentiellement construit à partir des racines. Un mot arabe est construit à partir de sa racine en deux étapes [14], [33], [34], [35], [44]. La première consiste à appliquer un modèle précis sur la racine, ce qui produit le Stem. La deuxième étape consiste à ajouter des affixes (préfixe ou suffixe) sur le Stem ainsi obtenu [3], [7], [9], [11], [13], [25], [28], [50].
Les éléments essentiels de la morphologie de la langue arabe sont [4], [17] , [19], [25], [26], [31], [37], [40], [46], [55], [41] :
**Les signes diacritiques**: Ils sont les caractères ajoutés au dessus ou en dessous des lettres arabes afin de spécifier la prononciation du mot. Ce rôle phonologique influe aussi sur le sens de ce mot. En effet, deux mots peuvent être écrits de la même manière mais différencier par l'ajout des diacritiques différents. Par exemple, prenons le mot «شعر», il est prononcé (Choor) en l'écrivant sous la forme «شعر» ( au dessus de «ش») et il signifie une fissure, il est prononcé (Chaar) en l'écrivant sous la forme «شعر» ( au dessus de «ش») et il signifie des cheveux, ou il sera prononcé (Chiir) en l'écrivant sous la forme «شعر» ( au dessus de «ش») et il signifie un poème.

| Signe | Prononciation arabe | Transcription | phonologie | Position |
|---|---|---|---|---|
| ◌ْ | سكون | Sokon | -- | Au dessus |
| ◌َ | فتحة | Fatha | a | Au dessus |
| ◌ً | تنوين الفتح | Tanwin Al Fath | an | Au dessus |
| ◌ِ | كسرة | Kasra | i | En dessous |
| ◌ٍ | تنوين الكسر | Tanwin Al Kasr | in | En dessous |
| ◌ُ | ضمة | Dama | o | Au dessus |
| ◌ٌ | تنوين الضم | Tanwin Al Dam | on | Au dessus |
| ◌ّ | الشدة | Chadda | Doubler les lettres | Au dessus |

Table 1 : Les diacritiques arabes, leurs transcriptions, leurs phonologies et leurs positions.

**Les modèles** : Ils sont des déclinaisons du mot «فعل» (Faal) qui sont obtenus en y ajoutant des affixes. Ils servent à produire des Stems à partir d'une racine. Ils sont aux alentours de 900 modèles. Par exemple, on y trouve : «فاعل» (Faael), «مستفعل» (Mostafael).
**Les affixes** : Ils sont des lettres qui s'ajoutent au début (les préfixes) ou à la fin des mots arabes (les suffixes). Ils sont aux alentour de 150. Par exemple, on y trouve: («ال التعريف» (Lam al taerif), «واو العطف» (waw al aatef), «واو الجماعة» (Waw al jammaa), «نون النسوة» (Noon al niswa),…).
**Les racines** : Elles sont à l'origine de la pluparts des mots arabes. Elles sont formées de trois à cinq lettres. Ils sont aux alentour de 7000 racines, la grande

majorité des racines arabes (85%) sont trilatérales. Par exemple on y trouve : «كتب» (Kataba).
**Les Stems**: C'est la dérivation à partir d'une racine donnée selon un modèle. Par exemple on y trouve : «مكاتب» (Makateb), il est obtenu à partir de la racine «كتب» (Kataba) selon le modèle «مفاعل» (Mafael).
**Les mots dérivés**: ils sont construits à partir d'un Stem en y ajoutant des affixes. Par exemple on y trouve: le mot «المكاتب» (Al Makateb) construits à partir du Stem «مكاتب» (Makateb), en y ajoutant le préfixe «ال» (Al).
**Les mots isolés** : Ces sont les mots qui ne possèdent pas une racine. Par exemple on y trouve : «بئس» (Boeose), il signifie minable, «إنسان» (Insan), il signifie homme.

## 3. Les dictionnaires classiques arabes

La structure générale d'un dictionnaire peut être écrite sous la forme : {Clé = Description}, où les Clés sont généralement des mots de la langue et les Description est un ensemble de mots de cette langue qui les définitions, les explications ou les correspondances (synonyme, antonyme, cooccurrence, traduction, étymologie). La langue arabe est une langue très riche en dictionnaires. Dans cet article, nous en avons répertorié une quarantaine (Table 2). Cette richesse est due surtout à des raisons religieuses mais aussi pour assurer la transmission de la langue arabe aux peuples arabisés [6], [8], [18], [42], [52], [53], [57], [58], [59].

Le Table 2 présente une liste des dictionnaires arabe. Dans cette table et pour un dictionnaire donné, les deux premières colonnes donnent le nom du dictionnaire et sa transcription. Les deux colonnes suivantes donnent le nom de l'auteur et sa transcription. Les deux colonnes suivantes présentent les dates Hégire et Grégorienne. En effet, les dates sont présentées de deux façons différentes et ceci selon les dictionnaires. La première présente une valeur qui est la date exacte de publication du dictionnaire, quand celle-ci est connue. La deuxième présente un intervalle qui est, en générale, la période de vie ou l'âge de l'auteur. Les deux dernières colonnes de la table 2 présentent le type du classement utilisé par le dictionnaire et sa traduction.

| الإسم | Transcription du nom | إسم المؤلف | Transcription du nom de l'auteur | Date | | الترتيب | Ordre |
|---|---|---|---|---|---|---|---|
| | | | | Hégire | Grégorienne | | |
| العين | Al ayain | خليل ابن أحمد الفراهيدي | Khalil Ibin Ahmad Al Farahidi | 100-173 | 719-790 | الصوتي | Ordre Phonétique |
| الغريب المصنف | Al Gareb Al Mousanafe | أبو عبيد القاسم بن سلام | Abo Oubayda Al Kasime Ibin Salam | 150-244 | 767-858 | الموضوعي | Par le sujet |
| معجم الجيم | Moajam Al Jime | أبو عمرو الشيباني | Abou Amro Al Chibani | 206 | 821 | العشوائي | Aléatoire |
| الجمهرة | Al Jamhara | أبو بكر بن دريد | Abo Baker Ibin Dourayde | 223-321 | 838-933 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| البارع | Al Bariaa | أبو علي القالي | Abo Ali Al Kali | 280-356 | 893-967 | الصوتي | Phonétique |
| تهذيب اللغة | Tahzibe Al Loga | أبو منصور الأزهري | Abo Mansour Al Azharye | 282-370 | 895-981 | الصوتي | Phonétique |
| المنجد في اللغة | Al Monjide Fi Al Loga | علي بن الحسن الهنائي | Ali Ibin Al Hasan Al Hanaeyi | 310 | 922 | الهجائي النطقي | Alphabétique des prononciations |
| الألفاظ الكتابية | Al Alfaze Al kitabiya | عبد الرحمن الهمذاني | Abd Al Rahman Al Hamazani | 320 | 932 | الدلالي | Sémantique |
| المحيط باللغة | Al Mouhite | الصاحب بن عباد | AL Sahib Ibin Abbade | 324-385 | 935-995 | الصوتي | Phonétique |
| الصحاح في اللغة تاج اللغة وصحاح العربية | Taje Al Loga Wa Sihahe Al Arabia | أبو نصر الجوهري | Abou Naser Al Jawhari | 332-400 | 944-1010 | الألفبائي للأواخر(القوافي) | Alphabétique selon les dernières lettres |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| مختصر العين | Mokhtasar Al Ain | أبو بكر الزبيدي | Abo Baker Al Zoubaydi | 379 | 989 | الصوتي | Phonétique |
| مجمل اللغة | Moujmal Al Loga | أحمد بن فارس | Ahmad Ibin Fares | 395 | 1005 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| معجم مقاييس اللغة | Moajam Makayse Al Loga | أحمد بن فارس | Ahmad Ibin Al Fares | 395 | 1005 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| متخير الألفاظ | Moutakhayare Al Alfaze | احمد بن فارس | Ahmad Ibin Fares | 395 | 1005 | الموضوعي | Par le sujet |
| المحكم والمحيط الأعظم | Al Mouhkam | ابن سيدة | Ibin Sayida | 398-458 | 1008-1066 | الصوتي | Phonétique |
| المخصص في اللغة | AL Moukhsase Fi Al Loga | ابن سيدة | Ibin Sayida | 398-458 | 1008-1066 | الموضوعي | Par le sujet |
| فقه اللغة وسر العربية | Fikeh Al Loga Wa Sire Al Arabia | أبو منصور الثعالبي | Abou Al Mansour Al Zaalabi | 429 | 1038 | الدلالي | Sémantique |
| الفائق في غريب الحديث والأثر | Al Faeik Fi Gareb Al Hadise wa Al Asar | محمود بن عمر الزمخشري | Mahmoud Ibin Omar AL Zamachekhari | 467-538 | 1074- 1143 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| النهاية في غريب الأثر | Al Nihaya Fi Gareb Al Asar | محمود بن عمر الزمخشري | Mahmoud Ibin Omar AL Zamachekhari | 467-538 | 1074- 1143 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| أساس البلاغة | Asase Al Balaga | محمود بن عمر الزمخشري | Mahmoud Ibin Omar AL Zamachekhari | 467-538 | 1074- 1143 | الهجائي الجذري | Alphabétique selon les racines |
| كفاية المتحفظ ونهاية المتلفظ | Kifayate Al Mouhtafaze Wa Nihate Al Moultafaze | ابن الأجدابي | Ibin Al Ajdabi | avant 600 | avant 1204 | الموضوعي | Par le sujet |
| لسان العرب | Lesan Al Arabe | جمال الدين محمد إبن مكرم إبن المنظور | Jamal Al Dine Mohamad Ibin Makram Ibin Al Manzour | 630-711 | 1232-1311 | الألفبائي للأواخر(القوافي) | Alphabétique selon les dernières lettres |
| العباب الزاخر | Al ibabe Al Zakher | حسن بن محمد الصغاني | Hasan ibin Mohammad Al Sagani | 650 | 1252 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| مختار الصحاح | Moukhtar Al Sihahe | محمد بن أبي بكر الرازي | Mouhamad Ibin Abo Baker Al Razi | 666 | 1267 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| القاموس المحيط | Al kamouse Al Mouhite | الفيروزآبادي | Al Fayrouze Aabadi | 729-817 | 1328-1414 | الألفبائي للأواخر(القوافي) | Alphabétique selon les dernières lettres |
| المصباح المنير | Al Mousbah Al Mounir | أحمد بن محمد المقري الفيومي | Ahmad Ibin Mouhamad Al Moufri Al Fayoum | 770 | 1369 | الهجائي الجذري | Alphabétique selon les racines |
| تاج العروس | Taje Al Aarose | مرتضى الزبيدي | Mortada Al Zoubaydi | 1145- 1205 | 1732-1790 | الألفبائي للأواخر(القوافي) | Alphabétique selon les dernières lettres |
| نجعة الرائد وشرعة الوارد في المترادف والمتوارد | Najaate Al Raed Wa Choraate Al Wared Fi Al Motaradife Wa Al Motwared | إبراهيم بن ناصف | Ibrahim Ibin Nassef | 1263-1323 | 1847 - 1906 | الدلالي | Sémantique |
| نجعة الرائد | Nageate Al Raed | براهيم بن ناصف | Ibrahim Ibin Nassef | 1263-1323 | 1847 - 1906 | الدلالي | Sémantique |
| محيط المحيط | Mouhite Al Mouhite | بطرس البستاني | Boutrose Al Boustani | 1286 | 1869 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| أقرب الموارد في فصح العربية والشوارد | Akrabe Al Mawrade Fi Foseh Al Arabia Wa Al Chawared | سعيد الخوري الشرتوني | Saiid Al Khoury Al Chartouni | 1307 | 1890 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| المنجد في اللغة والأدب والعلوم | Al mounjed | الأب لويس المعلوف | Pere: Louis AL Maalouf | 1326 | 1908 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| الإفصاح في فقه اللغة | Al Ifssahe Fi Al Loga | عبد الفتاح الصعيدي وحسين موسى | Abd Al Fatah Al Saadi et Hussein Mousa | 1347 | 1929 | الموضوعي | Par le sujet |
| المعجم اللغوي التاريخي | Al Moajam Al Tarikhi | المستشرق الألماني فيشر | Orientaliste Allemand: Ficher | 1355-1368 | 1936-1949 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| الغني | Al Ganiye | عبد الغني أبو | Abd al Gani Abo Al Aazem | 1360-1431 | 1941-2010 | الألفبائي للأوائل | Alphabétique selon les premières |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | العزم | | | | | lettres |
| متن اللغة | Maten AL Loga | أحمد رضا العاملي | Ahmad Rida Al Aamili | 1377 | 1958 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| الوسيط | Al Wasite : Académie de la langue arabe au Caire: مجمع اللغة العربية بالقاهرة | براهيم مصطفى، وأحمد حسن الزيات، وحامد عبد القادر، ومحمد علي النجار | Ibrahim Moustafa, Ahmad Al Zayate, Hamed Abed Al Kader, et Mohamad Ali Al Najar | 1380 | 1960 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| | | إبراهيم أنيس، وعبد الحليم منتصر، وعطية الصوالحي، ومحمد خلف الله أحمد | Ibrahim Anis, Abd Al Halim Mountaser, Atiya Al Sawalhi, Mouhamad Khalaf Alah Ahmad | 1392 | 1972 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| المعجم الكبير | Al Moajam Al Kabir | مجمع اللغة العربية بالقاهرة | Académie de la langue arabe au Caire | 1390-1421 | 1970- 2000 | الألفبائي للأوائل | Alphabétique selon les premières lettres |
| لاروس: المعجم العربي الحديث | Larose: Al Moajam Al Arabi Al Hadise | خليل الجر | Khalil Jar | 1393 | 1973 | الهجائي النطقي | Alphabétique des prononciations |
| المعجم العربي الأساسي | Al Moajam Al Asasi | لمنظمة العربية للتربية والعلوم والثقافة | ALECSO - Arab League Education, Culture and Science Organization | 1410 | 1989 | الهجائي الجذري | Alphabétique selon les racines |
| الرائد | Al Raed | جبران مسعود | Joubran Masoude | 1424 | 2003 | الهجائي النطقي | Alphabétique des prononciations |
| المعجم العربي الألباني | Al Moajam Al Arabi Al Alban | سليمان تومتشيني | Soulayman Toumitchini | 1431 | 2010 | العشوائي | Aléatoire |

Table 2 : Les dictionnaires arabes classiques.

La table 2 montre que la lexicographie arabe est une discipline très ancienne, les plus fameux lexicographes arabes classiques furent al-Khalil (VIIIe siècle), al-Moubarrad (Xe siècle), al-Zamakhshari (XIIe siècle), Ibn Manzour (XIVe siècle) et al-Firouzabadi (XVe siècle) [18], [52], [8].

En générale, Les dictionnaires arabes classiques présentent les mots selon l'ordre alphabétique des racines. Ce classement par racines a été repris par les auteurs de dictionnaires arabes-français comme « De Biberstein Kazimirski » [27] et « Daniel Reig » [12]. Toutefois d'autres auteurs, comme « Jabbour Abdel-Nour » [20] ont suivi un classement alphabétique de mots. Récemment, « Jean-Pierre Milelli » [23] a inclut aussi les mots pluriels dont les formes sont très variées et parfois imprédictibles. La lexicographie arabe a connu tout le long de son histoire une diversité d'écoles ayant chacune ses spécificités [42], [58], [56]. Ces ordres sont :

- **Ordre alphabétique selon les premières lettres:** Le lexicographe organise ses entrées suivant la première lettre en réunissant l'ensemble des entrées sous la première lettre, puis classés selon la seconde lettre puis la troisième et ainsi de suite.
- **Ordre alphabétique selon les dernières lettres:** Les entrées du dictionnaire sont ordonnées suivant le dernier caractère, dans chaque section les données du dictionnaire sont ordonnées suivant la première lettre puis suivant la deuxième et ainsi de suite.
- **Ordre phonétique:** Il est arrivé après l'arrangement alphabétique, il est fondé sur la base de la voix, en tenant compte de la convergence en termes de voix suivant le mouvement des lèvres. es lettres arabes sont divisées phonétiquement en plusieurs ensembles (Table 3), ils sont :

| 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|---|
| و ا ي | ف ب م | ر ل ن | ظ ذ ث | ط د ت | ص س ز | ج ش ض | ق ك | ع ح ه خ غ |

Table 3: Les sous-ensembles phonétiques des lettres arabes.

- **Ordre alphabétique selon les racines:** Dans cet ordre, le vocabulaire est divisé en ensemble, comprenant chacune d'elles un certain nombre de mots dérivés à partir d'une racine unique. Par Exemple : Sous le mot (Kataba) كتب, on peut trouver les mots : يكتب، مكتبة، كاتب، مكتوب. Les difficultés apparaissent dans l'ordonnancement des mots dérivés sous chaque racine.

- **Ordre alphabétique des prononciations:** Dans cet ordre, le dictionnaire est divisé en sections suivant le nombre et la séquence des lettres de l'alphabet. Puis l'ordre des mots dans chaque section se fait suivant la première lettre, sans tenir compte sa racine, ou bien les affixes ajoutées. Ainsi, le mot apparaît dans le dictionnaire suivant sa prononciation.

- **Ordre sémantique:** Dictionnaire est divisé en sections morales, chaque porte présente la famille sémantique d'un mot, où il est le tire de cette section est le mot de chaque section du titre mot famille sémantique de la présente section. Par exemple, dans la section الخطأ(erreur) : زلة، وهفوة، وعثرة، وسقطة، وفلتة، ونبوة.

- **Ordre aléatoire:** Dans ce cas aucun ordre n'est spécifié.

Donc, un dictionnaire arabe est organisé en tenant compte des points suivants [15]:

- **Contenu**: La description du contenu des entrées varie d'un dictionnaire à un autre. Certains se contentent des exemples pour définir le sens d'un mot, d'autres donnent la définition accompagnée d'exemples. Les informations d'ordre morphologique et syntaxique diffèrent d'un dictionnaire à un autre.

- **Macrostructure**: L'organisation des entrées lexicales diffère d'un dictionnaire à un autre (selon l'école lexicographique). Par exemple, *MaKTaBaTun* « مكتبة» (librairie) représente une entrée lexicale à part dans le dictionnaire *El-Ghani* « الغني» alors qu'il figure comme une sous-entrée de l'entrée lexicale *KaTaBa* « كتب» (écrire) dans le dictionnaire *Al-Wassit* « الوسيط» puisque c'est un mot dérivé de *KaTaBa* [52].

- **Microstructure**: L'organisation des informations linguistiques, au niveau des entrées lexicales, varie d'un dictionnaire à un autre et même au sein d'un même dictionnaire. Par exemple, le pluriel et/ou le féminin d'un Stem de type nom "اسم" ou adjectif "صفة" se trouvent en tête de l'article dans le dictionnaire *El-Ghani* « الغني» et à la fin de l'article, si elles existent, dans *Al-Wassit* « الوسيط».

- **Consultation**: Il n'y a pas une manière unique de consulter ces dictionnaires. Dans la plupart des cas, l'utilisateur est contraint à connaître la structure interne d'un dictionnaire pour pouvoir trouver l'information qu'il cherche.

## 4. Les dictionnaires électroniques arabes

L'avènement de l'outil informatique et l'augmentation des moyens de communication ont changé nos modes de stockage et d'accès aux informations. De ce fait, les dictionnaires électroniques à usage humain se trouvent libérés des contraintes de leurs versions papiers et répondent à un certain nombre de besoins pratiques d'actualisation et d'exploitation pour une population hétérogène d'utilisateurs [21]. Plusieurs dictionnaires électroniques sont aujourd'hui disponibles sur le Web ou commercialisés sur des supports électroniques (DVD). Dans cet article nous nous limitons aux dictionnaires électroniques disponibles sur le Web [10], [51], [56], [57], [61].

| Nom | Transcription | Date de publication | Source |
|---|---|---|---|
| الباحث العربي | Al Bahes Al Arabi | 2007 | [56]مركز الباحث العربي(Centre de recherche arabe) |
| عجيب-صخر | Al Ajeeb -Sakher | 1998-2010 | [51]شركة صخر(Sakher Company) |
| المكتبة اللغوية الإلكترونية | AL Maktaba | 2007 | [10]موقع روح الإسلام (Islam Spirit) |
| البراق | Al Burak | 2008 | http://www.alburaq.net/mukhtar/root.cfm[57] |
| كلمات | Kalimet | 2007 | http://www.kl28.com/lesanalarab.php[60] |
| Lexilogos | lexilogos | 2002-2010 | http://www.lexilogos.com/[32] |
| المعجم العربي التفاعلي | Almaajm Alarbyi Altfaalyi | 2009 | http://www.almuajam.com/ [62] |
| iSpedal | iSpedal | 2010 | [38][5] |

Table 4 : Les dictionnaires électroniques arabes.

La différence principale entre les dictionnaires classiques et électroniques n'est pas seulement le support sur lequel est présentée l'information. Toutefois, cette différence d'ordre technique en entraîne plusieurs autres aux niveaux de l'utilisation, de la présentation, du contenu, des capacités de recherche, des fonctions bureautiques et des aspects techniques [43]. Avec les dictionnaires classiques, la consultation se fait de façon linéaire, tandis qu'avec les dictionnaires électroniques, elle peut se faire de façon fragmentée ou linéaire. En effet, dans le premier cas, l'utilisateur peut chercher un mot uniquement à partir de la nomenclature, alors que, dans le deuxième cas, il peut chercher un mot à partir de la nomenclature ou du texte entier. En effet, les capacités de recherche de ces dictionnaires permettent d'extraire des unités lexicales à partir de parties de mots ou à partir d'un ou de plusieurs mots se trouvant soit dans la nomenclature, soit dans le corps de l'article. De plus la recherche à partir de la nomenclature diffère quelque peu avec les dictionnaires électroniques [43].

La structure d'un dictionnaire classique est différente de la structure nécessaire lorsque la clé est la racine et les mots dérivés sont joints à cette clé. Pourtant, la clé dans un dictionnaire classique est la racine, mais est joint à cette racine, sa synonyme, les mots dérivés et de leurs synonymes. Dans ce cas, ces synonymes ajoutés ne sont pas attachés à cette clé, mais à une autre, et par cela, ils induisent un ensemble de données bruitées. Par ailleurs, les outils de consultation accompagnant ces dictionnaires offrent un seul type de recherche qualifié de primitif (recherche de mots vedettes dans un ou plusieurs dictionnaires). La Table 4 montre certains dictionnaires électroniques arabes. Al Bahes Al Arabi «الباحث العربي», qui a été lancé en juin 2007, contient plus de 4.000.000 mots [56]. Il offre le service de recherche

dans les grands dictionnaires classiques arabes (Lesan Al Arabe «لسان العرب», Makayse Al Loga «مقاييس اللغة», Sihahe Al Loga «الصحاح في اللغة», Al Kamous Al Mouhite «القاموس المحيط», Al ibabe Al Zakher «العباب الزاخر»). Il est encore en développement. Al Ajeeb Sakher «صخر عجيب» offre à l'utilisateur de nombreux services tels que : la traduction arabe-anglais et la recherche linguistique [51]. Il offre aussi le service de recherche dans les grands dictionnaires classiques arabes (AL Mouhite «المحيط», Mouhite AL Mouhite «محيط المحيط», Al Wasite «الوسيط», Al Kamous Al Mouhite «القاموس المحيط», Lesan Al Arabe «لسان العرب», Najaate Al Raed «نجعة الرائد», Al Ganiye «الغني», Taje Al Aarose «تاج العروس»). AL Maktaba «المكتبة اللغوية الإلكترونية» est une bibliothèque électronique des livres spécialisés dans les dictionnaires de la langue arabe [10]. Ces livres sont : Sihahe Al Loga « الصحاح في اللغة», Al Faeik Fi Gareb Al Hadise wa Al Asar «الفائق في غريب الحديث و الأثر», Al Kamous Al Mouhite «القاموس المحيط», Al Nihaya Fi Gareb Al Asar « النهاية في غريب الأثر», Taje Al Aarose « تاج العروس», Lesan Al Arabe «لسان العرب», Moukhtar Al Sihahe « مختار الصحاح». Al Burak «البراق» offre le service de recherche dans certains dictionnaires classiques arabes [57]. Kalimet « كلمات» est un site qui offre plusieurs services, parmi eux les caractéristiques morphologiques et les synonymes d'un mot. Ce dictionnaire se limite au dictionnaire classique Lesan Al Arabe «لسان العرب» [60]. Lexilogos est un site qui offre outil facile et pratique qui permet de consulter les informations morphologiques des mots arabes à partir des dictionnaires classiques arabes (Lesan Al Arabe «لسان العرب», Al Kamous Al Mouhite «القاموس المحيط» [32]). Al moujam al arabi al tafaoli est un dictionnaire arabe interactif encore en developpement. L'interface de ce dictionnaire est un site qui offre un service de recherche des caracteristiques morphologiques et des synonymes d'un mot arabe. Les données de ce dictionnaire sont preparees manuellement a partir des dictionnaires classiques tel que Al Ayain «العين» [Almuajam, 2009].

## 5. Le dictionnaire électronique arabe iSPEDAL

L'élaboration des dictionnaires électroniques arabes est dans une phase qu'on peut qualifier de primaire. Il est évident que la quasi-totalité des dictionnaires électroniques arabes disponibles sur le Web, ne sont rien d'autre qu'une numérisation de fonds dictionnairiques (sous format : ".doc", ".pdf" ou ".html"), non structurées, peu bénéfiques, et ils ne sont pas directement exploitables [1], [18], [36]. La plupart de ces dictionnaires collecte ces données à partir des plusieurs dictionnaires classiques et offre un service de navigation et de recherche dans ces dictionnaires. Ces limites d'interrogation sont dues, principalement, à une faiblesse de structuration des entrées dictionnairiques utilisées.

Pour ces raisons, nous avons déjà proposé DESELA (dictionnaire électronique structuré et évolutif de la langue arabe) [5] qui est ensuite amélioré pour devenir iSPEDAL (An Improved Structured and Progressive Electronic Dictionary for the Arabic Language) [38]. Ce dictionnaire évolutif de la langue arabe est présenté sous la forme d'une base de données relationnelle ou d'un document XML [15] qui est facilement exploitable en utilisant un langage de requête appropriée. Il contient les racines, les préfixes, les suffixes, l'infixe, et les modèles, ainsi que des informations fournies par un dictionnaire standard. En outre, pour un mot donné, il fournit des liens avec ses racines, des affixes associés, et son modèle. De plus, ce dictionnaire est évolutif grâce à un système automatique qui l'alimente à partir d'un ou de

plusieurs dictionnaires textuels classiques ou à partir d'un corpus textuel arabe quelconque.

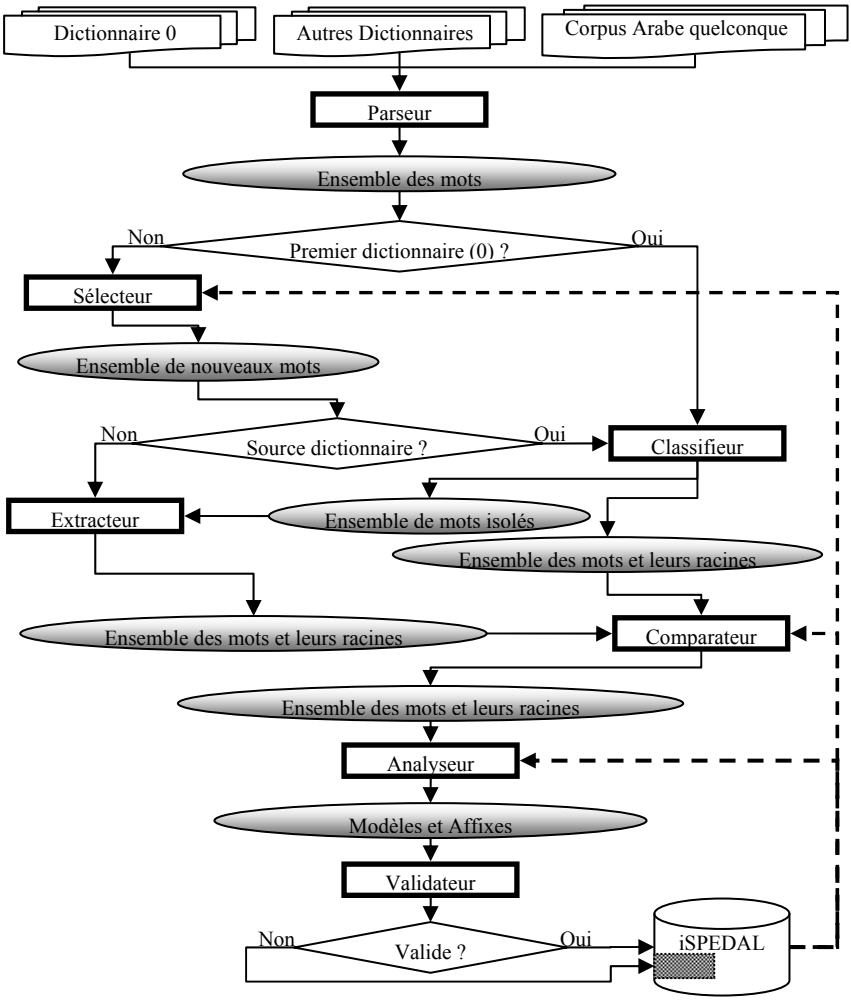


Figure 1 : Schéma général d'iSPEDAL.

La figure 1 présente le Schéma général selon lequel iSPEDAL fonctionne. Elle montre l'architecture générale du système d'alimentation et d'enrichissement automatique à partir d'un ou de plusieurs dictionnaires classiques ainsi qu'à partir des corpus textuels arabes quelconques. iSPEDAL est constitué de plusieurs composantes qui sont : le Parseur, le Sélecteur, le Classifieur, l'Extracteur, le Comparateur, l'Analyseur et le Validateur. L'entrée de ce système peut être un dictionnaire arabe classique sous format plat (fichier texte, PDF,..) ou un n'importe quel autre corpus arabe textuel (page web, fichier texte, …). La première composante de ce système est le parseur qui permet de transformer le document en un ensemble des mots, selon les séparateurs qui sont généralement des espaces. Si le document est le premier dictionnaire (0), l'ensemble des mots est passé au classifieur. Ce premier dictionnaire est utilisé pour initialiser iSPEDAL. Dans les

autres cas, c'est le sélecteur qui reçoit l'ensemble des mots. Le rôle du sélecteur est d'éviter les doublons en s'assurant que les mots à ajouter à iSPEDAL n'y sont pas déjà. La sortie de cette composante est un ensemble des nouveaux mots qui est soumit au classifieur, si cet ensemble est en provenance d'un dictionnaire, ou à l'extracteur dans le cas contraire. Le classifieur est la composante qui permet de scinder l'ensemble des mots reçus en entrée en deux sous ensemble : d'un coté les racines et leurs mots dérivés qui sont envoyées vers le comparateur, d'un autre les mots isolés qui sont envoyés vers l'extracteur. Cette séparation est basée sur le format du dictionnaire d'entrée, où les racines sont encadrées par des séparateurs spéciaux et les mots, qui sont situés après cette racine et avant la racine suivante, dérivent de la première. L'extracteur utilise la méthode d'extraction détaillée dans [25] pour trouver la racine d'un mot arabe. Les ensembles des mots associés à leurs racines, en provenance du classifier et de l'extracteur, sont soumis au comparateur qui permet d'éviter les doublons, à tous les niveaux, dans l'iSPEDAL. L'ensemble des nouveaux mots et des racines associées sont utilisés par l'analyseur pour produire les affixes et les modèles. La sortie de cette composante est un ensemble des mots, des racines, des modèles et des affixes sont soumis au validateur pour approuver ces résultats ainsi que les liens entre eux. Le validateur utilise les éléments essentiels de la morphologie de la langue arabe pour l'approbation de ces éléments. Seuls les éléments valides sont ajoutés à iSPEDAL, le reste est mis dans une zone tampon en attente d'une validation ultérieure.

## 6. Conclusion et perspective

Dans cet article, nous avons présenté un état de l'art des tous les dictionnaires arabes classiques qui sont aux alentours des 50 dictionnaires. Les plus fameux sont très anciens, ils sont en VIIIe siècle, et la production des dictionnaires arabes est en cours jusqu'à maintenant. Les entrées de ces dictionnaires sont ordonnées selon des écoles ayant chacune ses spécificités. En général, Ces ordres peuvent être aléatoires, alphabétiques, selon la prononciation, phonétiques ou encore selon les racines.

Les dictionnaires arabes classiques, qui sont sous forme de papier ou bien en format électronique (fichiers plats), ne sont pas exploitables. La recherche d'un mot dans ce type des dictionnaires se fait d'une façon linéaire, ce qui limite l'importance de ce type de dictionnaires dans les systèmes d'extraction d'information (IR) et des moteurs de recherche, qui ont besoin des sources de donnée exploitables et structurées. De ce point de vue, il y a plusieurs dictionnaires électroniques qui ont utilisé un ou plusieurs dictionnaires classiques comme source. Dans cet article, nous avons présentée sept dictionnaires électroniques qui se sont développés à partir de l'année 1998 [56], [61], [51], [10], [57]. Notre contribution dans ce domaine apparait dans iSPEDAL [38] qui est une version améliorée de DESELA [5]. iSPEDAL est un dictionnaire structuré et évolutif de la langue arabe qui est présenté sous la forme d'une base de données relationnelle ou d'un document XML [15]. Ce dictionnaire est facilement exploitable en utilisant un langage de requête appropriée. Il contient les racines, les préfixes, les suffixes, l'infixe, et les modèles, ainsi que des informations fournies par un dictionnaire standard [15]. En outre, pour un mot donné, il fournit des liens avec ses racines, des affixes associés, et son modèle. Il est enrichi grâce à un système automatique, à partir d'un ou de plusieurs dictionnaires textuels classiques et à partir d'un corpus textuel arabe quelconque.

La perspective de ce travail est de classifier les dictionnaires arabes classiques et électroniques existent. Cette classification sert à construire repérer les données des ces dictionnaires et à les regrouper dans un dictionnaire unique tel que iSPEDAL [5]. Aussi, travailler au niveau des types de données (mots, racine, mot isolé,…) et de créer des relations morphologique et sémantique entre ces éléments.

## 7. Remerciements

Ce travail est effectué dans le cadre du projet «Arabic speech synthesis from text, with natural prosody, using linguistic and semantic analysis» financé par le « Lebanese- Syrian Scientific Research Cooperation Program ».

## 8. Bibliographie

[1]. Abou Al Azem, *Al Gani* , http://lexicons.sakhr.com/intro/intro.aspx ?fileurl=introduction.asp

[2]. Al Ameed, S. Al Ketbi, A. Al Kaabi, K. Al Shebli, N. Al Shamsi, N. Al Nuaimi, et S. Al Muhairi, *Arabic Light Stemmer: A new Enhanced Approach* , The Second International Conference on Innovations in Information Technology (IIT'05), 2005.

[3]. Al Hajjar, M. Hajjar, et K. Zreik, *A new system for evaluation of Arabic root extraction methods*, The Fifth International Conference on Internet and Web Applications and Services. ICIW, Barcelona, Spain, 2010.

[4]. Al Hajjar, M. Hajjar, et K. Zreik, *Classification of Arabic Information Extraction methods*, 2nd International Conference on Arabic Language Resources and Tools, Le Caire, Egypte, 21-23 Avril 2009.

[5]. Al Hajjar, M. Hajjar, K. Zreik, *Un nouveau dictionnaire électronique structuré et évolutif pour la langue arabe*, In Patrimoine 3.0 : Actes du douzième colloque international sur le document électronique (CIDE.12). Ed. Europia, Paris, 2009.

[6]. Al Hamid, *المعاجم العربية,* http://www.voiceofarabic.com . 2010.

[7]. Al Kharashi, *A Web Search Engine for Indexing, Searching and Publishing Arabic Bibliographic Databases*, 1999.

[8]. *Al-Musba7 Al-Muneer Arabic Dictionary* قاموس المصباح المنير http://www.islamweb.net/ver2/library/ BooksCategory.php?idfrom =1&idto=4974&bk_no=45&ID=1&lang=A.

[9]. Alsifayani, *المعالجة الآلية للغة العربية (التحليل الصرفي)*, King Fahd University of Petroleum & Minerals INFORMATION AND COMPUTER SCIENCE.

[10]. C. Al Mougrabi, *المكتبة اللغوية الإلكترونية,* http://www.merbad.net/vb/showthread.php?t=6603.2010.

[11]. Chen, A. Chen, et F. Gey, *Building an Arabic stemmer for information retrieval*, TREC-11 conference, 2002.

[12]. D. Reig, *Dictionnaire arabe français*, Paris, 1983.

[13]. F .Douzidia, et 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.

[14]. F. Ahmed and A. Nürnberger, *N-grams Conflation Approach for Arabic*, ACM SIGIR Conference, Amsterdam, 2007.

[15]. F. Baccar, A. Khemakhem, et B. Gargouri, *Modélisation normalisée LMF des dictionnaires électroniques éditoriaux de l'arabe*, TALN 2008, Avignon, 9-13 juin 2008.

[16]. H. Al Ameed, S. Al Ketbi, A. Al Kaabi, K. Al Shebli, N. Al Shamsi, N. Al Nuaimi , et S. Al Muhairi, *خطوات عملية باتجاه تطبيق مفهوم البحث بالمفاهيم لمحركات البحث والاسترجاع العربي*, UNITED ARAB EMIRATES UNIVERSITY – Faculty of Information Technology, Arabic Search Engines Improvement: A New Approach using Search Key Expansion Derived from Arabic Synonyms Structure. 2008.

[17]. H. Suleiman Mustafa, *Character contiguity in N-gram based word matching: the case for Arabic text searching*. Information Processing and Management.41 (4), 2004, pp. 819-827.

[18]. Ibn Manzour, *Lisan Al-Arab*. www.muhaddith.org, 2009.

[19]. Isbihani, S. Khadivi, O. Bender, et H. Ney, *Morpho-syntactic Arabic preprocessing for Arabic-to-English statistical machine translation,* Proceedings of the Workshop on Statistical Machine Translation, New York City, 2006 .

[20]. J. Abdel-Nour, *Dictionnaire Abdel-Nour Al-Moufassal*, Beyrouth, 1983.

[21]. J. Dendien et J-M. Pierrel, *Le trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence*, TAL, Volume 44 - n°2/2003, 28 p. Le Dictionnaire de l'Académie française : histoire et nuances de la langue française (1694- 1935).2000. éditions Redon, 2003.

[22]. J. Micher, et C.Voss*, Buckwalter-based Lookup Tool as Language Resource for Arabic Language Learners Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, USA, 2008, pp. 66–67.

[23]. J.-P. Milelli, *20 001 Mots*, Paris, 2006

[24]. K. Darwish, *Building a Shallow Arabic Morphological Analyzer in One Day*. The ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, USA, 2002.

[25]. K. Taghva, R. Elkoury, et J. Coombs, *Arabic Stemming without a root dictionary*, International Conference on Information Technology: Coding and Computing (ITCC'05) – Vol. I, 2005, pp. 152-157.

[26]. Kanaan, R. Al-Shalabi, J. Jaarn, M. Al-Kabi, et A. Hasnah, *A New Stemming Algorithm to Extract Quadri-Literal Arabic Roots*, 2004.

[27]. Kazimirski , *Dictionnaire arabe-français, contenant toutes les racines de la langue arabe, leurs dérivés, tant dans l'idiome vulgaire que dans l'idiome littéral, ainsi que les dialectes d'Alger et de Maroc*, Maisonneuve et Cie, 2 volumes, 1860, 1392 et 2369 pp. (réédition 1944, Beyrouth, éditions du Liban, et 2005, édition Albouraq).

[28]. Khemakhem, B. Gargouri, A. Abdelwahed, et G. Francopoulo, *Modélisation des paradigmes de flexion des verbes arabes selon la norme LMF - ISO 24613, Traitement Automatique des Langues Naturelles*, Toulouse, France, 2007.

[29]. 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, 2006, pp. 78-82. 2006.

[30]. L. Larkey, L. Ballesteros, et 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.

[31]. L. Larkey, L. Ballesteros, et 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, 2002, pp. 275 – 282.

[32]. *Lexilogos* : http://www.lexilogos.com/arabe_langue_dictionnaires.htm. 2002-2010

[33]. M. Attia, M. Rashwan, A. Ragheb, M. Al-Badrashiny, et H. Al-Basoumy: *A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields*. LREC. 2008.

[34]. M. Attia, M. Rashwan, A. Ragheb, M. Al-Badrashiny, H. Al-Basoumy, et S. Abdou, *A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields,* GoTAL. 2008.

[35]. M. Attia, M. Rashwan, M. Al-Badrashiny, *A Semi-Automatic Visual Interactive Tool for Morphological, PoS-Tags, Phonetic, and Semantic Annotation of Arabic Text Corpora,* IEEE Transactions on Audio, Speech & Language Processing 17(5): 916-925. 2009.

[36]. M. Ben Abderrahmen, B. Gargouri, et 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.

[37]. M. El-Halees, *Arabic Text Classification Using Maximum Entropy*, The Islamic University Journal (Series of Natural Studies and Engineering) Vol. 15, No.1, 2007, pp. 157-167.

[38]. M. Hajjar, A. Al Hajjar, K. Zreik et P. Gallinari, *An Improved Structured and Progressive Electronic Dictionary for the Arabic Language: iSPEDAL*. The Fifth International Conference on Internet and Web Applications and Services, ICIW 2010, May 9 - 15, 2010 - Barcelona, Spain.2010.

[39]. M. Mustafa, H. AbdAlla, et H. Suleman, *Current Approaches in Arabic IR: A Survey*. In Proceedings The Annual International Conference on Asia-Pacific Digital Libraries (ICADL), Bali, Indonesia. 2008.

[40]. M. Sanan, *Etude Des Methodes De La Recherche D'information Et De L'indexation Sur Les Documents Electroniques : Cas De La Langue Arabe*, Doctorat Informatique, UNIVERSITE PARIS VIII – SAINT DENIS, ECOLE DOCTORALE Cognition, Langage et Interaction (CLI). 2008.

[41]. M. Sanan, M. Rammal, et K. Zreik, *Arabic documents classification using N-gram*, Conférence ICHSL6, Toulouse, 2008.

[42]. M. Youssef, *المعاجم اللغوية والمعاجم المتخصصة*, http://www.saaid.net/Doat/hasn/75-4.htm.

[43]. N. Forget, *Les dictionnaires électroniques dans l'optique de la traduction*, Thèse de maîtrise, Ottawa (Ontario), 1999.

[44]. N. Habash, O. Rambow, et R. Roth, *MADA+TOKAN: A System for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging.Stemming*

*and Lemmatization*, Center for Computational Learning Systems, Columbia University, August 6, 2008.

[45]. Nwesri,Seyed, M. Tahaghoghi, et F. Scholer, *Stemming Arabic Conjunctions and. Prepositions*, School of Computer Science and Information Technology, RMIT University, GPO Box 2476V, Melbourne 3001, AUSTRALIE, String processing and information retrieval : ( 12th international conference, SPIRE 2005 ) (Buenos Aires, Argentina, November 2-4, 2005 ) .

[46]. O. Al Dakkak, N. Ghneim, A. Alshalaby, R. Sonbol, et M. Saïd Desouki, *Arabic Language Resources in HIAST*, Medar 2nd International Conference on Arabic Language Resources and Tools 22-23 April 2009.

[47]. S. Abdelhadi, V. Bosch, A. Günter, *Arabic Computational Morphology*, *Knowledge-based and Empirical Methods*, Series: Text, Speech and Language Technology, Vol. 38 2007, VIII, 308 p., Hardcover, ISBN: 978-1-4020-6045-8. 2007.

[48]. S. Ait Taleb, *Dictionnaires électroniques arabes : le modèle des dictionnaires de Sakhr*, revue de l'Association Marocaine des Etudes Lexicographiques, Numéro 3-4, 15-31. 2005.

[49]. S. Mesfar , *Analyse Morpho-Syntaxique Automatique Et Reconnaissance Des Entités Nommées En Arabe Standard*. Thèse en vue de l'obtention du titre de docteur en informatique, Université de France -Comite, Ecole Doctorale « Langage, Espace, Temps, Sociétés ». le 24 novembre 2008.

[50]. S. Mesfar, *Towards a Cascade of Morpho-syntactic Tools for Arabic Natural Language Processing*. CICLing 2010: 150-162. 2010.

[51]. Sakher Company, *Al Ajeeb*, http://lexicons.ajeeb.com 1998-2010

[52]. 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.

[53]. Tahech, *AOUNIات*, http://www.aouniat.com/2009/04/23/arabic-dictionary.html, 2009.

[54]. W. A. George, et 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, 1974, pp. 253-260.

[55]. Y. Kadri, et J. Nie, *Effective Stemming for Arabic Information Retrieval*, proceedings of the Challenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni, 2006.

[56]. *الباحث العربي*: www.BAHETH.INFO.2007-2010.

[57]. *البراق*, www.Al-Buraq.net. 2008.

[58]. الكشف في المعاجم: www.yabeyrouth.com/pages/index2851.htm.

[59]. *إسلام أون لاين نت*, http://www.islamonline.net/Arabic/index.shtml. 1999 - 2010.

[60]. كلمات: http://www.kl28.com/service. 2010.

[61]. المكتبة *للكتب الإسلامية والعربية*, http://www.almaktba.com. 2004.

[62]. المعجم العربي التفاعلي(Almuajam), website: http://www.almuajam.com/ ArabicDictionary /about.jsp. 2009.

# De l'utilisation de WordNet pour l'indexation conceptuelle des documents

**Fatiha Boubekeur (1), Mohand Boughanem (2), Lynda Tamine (2), Mariam Daoud (2)**

amirouchefatiha@mail.ummto.dz / boughane@irit.fr / Lechani@irit.fr / daoud@irit.fr

(1) Université Mouloud Mammeri, 15000 Tizi Ouzou, Algérie

(2) IRIT-SIG/RFI, Université Paul Sabatier, 31062 Toulouse, France

**Résumé .** Ce papier décrit une approche d'indexation sémantique des documents. Nous proposons d'utiliser WordNet comme ressource linguistique afin de retrouver les concepts représentatifs du contenu d'un document. Notre contribution porte sur un double aspect: d'une part, nous proposons une approche d'identification des concepts en utilisant la base lexicographique WordNet, d'autre part, nous proposons une approche de pondération de ces concepts basée sur une nouvelle notion d'importance.

**Mots-clés** : Recherche d'information, indexation sémantique, indexation conceptuelle, WordNet.

## 1 Introduction

Un processus de recherche d'information (RI) a pour but de sélectionner l'information pertinente pour un besoin en information exprimé par l'utilisateur sous forme de requête. Ce processus intègre deux principales étapes, l'indexation et l'appariement. L'indexation consiste à représenter requêtes et documents par un ensemble, l'index, de termes (généralement des mots simples) pondérés, sensés au mieux leurs contenus sémantiques. Les termes sont automatiquement extraits ou manuellement assignés aux documents et aux requêtes, puis pondérés par des valeurs numériques qui traduisent leur importance dans le document. L'appariement consiste à « *matcher* » les représentations des requête et documents pour sélectionner les documents qui correspondent au mieux à la requête. Une caractéristique clé des systèmes de recherche d'information (SRI) est que l'appariement est impacté par la qualité de la description du besoin en information et par la qualité de l'indexation.

Une problématique fondamentale en RI est l'imprécision du besoin utilisateur (une requête est habituellement une description vague et incomplète du besoin en information de l'utilisateur) et l'ambiguïté de l'indexation. A l'origine de cette problématique est la disparité et l'ambiguïté de la langue naturelle.

- La disparité de la langue naturelle traduit la propriété qu'ont certains termes à être représentés par différentes chaînes de caractères, et associés aux mêmes

sens ou à des sens liés. C'est ainsi par exemple qu'un document sur *Linux* pourtant pertinent pour une requête sur les *systèmes d'exploitation*, ne sera pas retrouvé si les mots *système* et *exploitation* sont absents de ce document. En RI, la disparité des termes implique un silence documentaire.

- L'ambiguïté est divisée en homonymie et polysémie [13]. L'homonymie traduit la propriété qu'ont certains termes à être représentés par une même chaîne de caractères, et associés à différents sens. *Souris* (le mammifère) *vs souris* (du verbe *sourire*) est un exemple d'homonymie. La polysémie est liée à la propriété qu'ont certains termes à exprimer différents sens. *Prendre le large vs prendre un thé* est un exemple de polysémie. Dans les systèmes de recherche d'information (SRI) classiques, l'ambiguïté implique que des documents non pertinents sont retrouvés. Ainsi, un document qui traite de la politique en *France* sera retrouvé comme pertinent pour une requête portant sur *Anatole France* si le mot *France* figure dans le document et dans la requête. L'ambiguïté des termes implique un bruit documentaire.

Les SRI classiques présentent ainsi des insuffisances du fait de leur incapacité à traiter avec l'ambiguïté de la langue et l'imprécision sémantique des mots simples. Pour lever ces problèmes d'ambiguïté et de disparité des mots, de nombreux travaux de recherche en RI se sont orientés vers la prise en compte des sens des mots dans le processus d'indexation. L'indexation sémantique, se base sur les sens des mots (entités sémantiques) plutôt que sur les mots simples (entités lexicales) pour indexer les documents. Pour retrouver les sens des mots dans un contenu donné, l'indexation sémantique se base sur des techniques de désambiguïsation contextuelle des mots dans les documents et requêtes. La désambiguïsation a pour objet de retrouver le sens d'un mot dans un contenu donné. Pour ce faire, la désambiguïsation s'appuie :

(1) sur des corpus d'apprentissage [7], [14], [18] : Une manière d'indexer serait par exemple, d'associer aux mots extraits, des mots du contexte qui aident à déterminer leur sens [31]. Une autre manière serait d'apprendre le sens d'un mot à partir de ses usages possibles du mot à désambiguïser [23], ou à partir de règles d'agencement ou règles de fonctionnement des mots à désambiguïser [28].

(2) sur des ressources linguistiques externes telles que les thésaurus [30], dictionnaires automatisés [10], [15], [26], [29], ontologies [20], [24], et autres Wikipédia [17], qui constituent des sources d'évidence pour les définitions et sens du mot cible. On parle alors d'indexation conceptuelle.

Nous proposons, dans ce papier, une approche d'indexation conceptuelle de documents. Le principe de l'approche consiste à extraire les mots du document, puis à leur associer les sens adéquats correspondants. Nous proposons d'utiliser WordNet [19] comme source d'évidence pour l'identification des sens des mots et pour leur pondération. Les sens des mots correspondent alors à des concepts (ou synsets) de WordNet. L'identification des sens des mots se base sur le calcul d'un score de désambiguïsation. La pondération d'un concept s'appuie sur ses relations sémantiques avec les autres concepts du document en tenant compte de leurs importances.

Le papier est structuré comme suit : en section 1, nous posons la problématique de l'indexation conceptuelle, puis présentons une synthèse des travaux dans le domaine, puis nous situons notre contribution. En section 2 nous présentons notre approche d'indexation conceptuelle. La section 3 présente quelques résultats expérimentaux. La section 4 conclut le papier.

## 2 Contexte des travaux et problématique

### 2.1 Problématique

*La plupart des approches d'indexation conceptuelle s'appuient en général sur des ontologies pour déterminer les différents sens du mot mais aussi pour désambiguïser les sens des mots.* L'indexation conceptuelle se base sur des concepts extraits d'ontologies, de taxonomies, et autres ressources lexicales pour indexer les documents contrairement aux listes de mots simples. Pour ce faire, l'indexation conceptuelle soulève deux principaux problèmes: l'identification des concepts et leur pondération.

*(1)* L'identification des concepts a pour objectif d'extraire l'ensemble des mots ou collocations de mots du document à indexer, et à leur associer leurs sens correspondant dans le document. L'extraction des mots simples est un problème d'indexation classique. Les approches utilisées se basent le plus souvent sur des techniques linguistiques (tokénisation, lemmatisation, élimination de mots vides) et statistiques pour identifier les mots clés du document. *Etant donnés ces mots clés, le problème crucial de l'indexation sémantique est d'abord d'identifier, pour chacun de ces mots clés, les entrées correspondantes dans l'ontologie, puis de sélectionner parmi ces entrées le sens adéquat du mot clé considéré dans le document : c'est la désambiguïsation des sens des mots (WSD*[1]*).*

(2) La pondération des termes d'indexation a pour objet d'associer à chaque terme d'index son poids d'importance dans le document. La pondération est un problème crucial en RI. La qualité de la recherche dépend de la qualité de la pondération adoptée. Dans les SRI classiques basés sur une indexation par mots clés, la pondération dite *tf*idf* et ses variantes sont largement adoptées. En indexation conceptuelle, le problème est alors de définir une pondération adéquate pour les entités sémantiques que sont les concepts.

*Une fois les termes d'i*ndexation désambiguïsés et pondérés, la représentation des textes indexés se fait soit à partir des seuls sens (ou concepts) identifiés lors de l'étape de désambiguïsation, soit à partir d'une combinaison des mots-clés et sens corrects associés. *Les approches d'indexa*tion *[1], [2], [12], [18], [25], [27] sont basées sur ce principe.* Nos travaux se situent dans ce même contexte, et consistent en l'utilisation de WordNet tant pour l'identification des concepts que pour leur pondération. Dans ce qui suit, nous présentons la base lexicographique WordNet, puis quelques travaux d'indexation conceptuelle principalement basés sur cette ressource. Nous situons enfin notre contribution par rapport à ces derniers.

### 2.2 Préliminaires: WordNet

WordNet est un réseau lexical électronique qui couvre la majorité des noms, verbes, adjectifs et adverbes de la langue Anglaise, qu'elle structure en un réseau de noeuds et de liens.

- Les noeuds sont constitués par des ensembles de termes synonymes appelés *synsets*.
  - Un synset représente un concept.
  - Un concept est une entité sémantique, lexicalement représentée par un terme.

[1] *Word Sense Disambiguation*

- Un terme peut être un mot simple ou une collocation (mot composé).

- Les liens représentent des relations sémantiques entre concepts, dont par exemple les relations d'hyponymie-hyperonymie suivantes:
  - la relation de subsumption entre noms, (relation *is-a*) qui permet d'associer un concept classe (l'hypernyme) à un concept sous-classe (l'hyponyme). Par exemple, le nom *tower#1* a pour hyponymes *silo*, *minaret, pylon*…Cette relation permet d'organiser les concepts de WordNet en une hiérarchie.
  - la relation d'instanciation (*instance*) qui permet d'associer un concept et son instance. Par exemple, le nom *tower#1* a pour instance hyponyme *tour Eiffel*.

Un exemple de hiérarchie de synsets correspondant au mot « dog » est donné dans la table 1.

| **Noun** |
|---|
| UUUUS : (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"* |
| S : (n) frump, dog (a dull unattractive unpleasant girl or woman) *"she got a reputation as a frump"; "she's a real dog"* |
| S : (n) dog (informal term for a man) *"you lucky dog"* |
| S : (n) cad, bounder, blackguard, dog, hound, heel (someone who is morally reprehensible) *"you dirty dog"* |
| S : (n) frank, frankfurter, hotdog, hot dog, dog, wiener, wienerwurst, weenie (a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll) |
| S : (n) pawl, detent, click, dog (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward) |
| S : (n) andiron, firedog, dog, dog-iron (metal supports for logs in a fireplace) *"the andirons were too hot to touch"* |
| Verb |
| S : (v) chase, chase after, trail, tail, tag, give chase, dog, go after, track (go after with the intent to catch) *"The policeman chased the mugger down the alley"; "the dog chased the rabbit"* |

Table1 : Les concepts de WordNet correspondants au concept dog

### 2.3 Synthèse des travaux sur l'indexation conceptuelle

L'indexation conceptuelle représente les documents par des concepts. Ces concepts sont extraits d'ontologies et autres ressources linguistiques. Pour ce faire, le processus d'indexation s'appuie en générale sur deux étapes : (1) l'identification des concepts et (2) leur pondération.

(1) L'identification des concepts : Les termes d'indexation (généralement des mots clés) sont d'abord extraits du document par une approche classique d'indexation (tokenisation, élimination des mots vides, puis lemmatisation) [1], [3], [4], [12], [25], [27]. Ces termes (non vides) sont ensuite projetés sur l'ontologie afin d'identifier les concepts (ou sens) correspondants dans l'ontologie. Un terme ambigu correspond à plusieurs entrées (sens) dans l'ontologie. Il faut le désambiguïser. Pour désambiguïser un mot ambigu, Voorhees [27] classe chaque synset de ce mot en se basant sur le nombre de

mots co-occurrents entre un voisinage (Voorhees l'a appelé *hood*) de ce synset et le contexte local (la phrase où l'occurrence du mot apparaît) du mot ambigu correspondant. Le synset le mieux classé est alors considéré comme sens adéquat de l'occurrence analysée du mot ambigu. Dans une approche différente, Katz et al [25] proposent aussi une approche basée sur le contexte local. Le contexte local d'un mot est défini comme étant la liste ordonnée des mots démarrant du mot utile le plus proche du voisinage gauche ou droit jusqu'au mot cible. L'hypothèse de Katz et al., est que des mots utilisés dans le même contexte local (appelés *sélecteurs*), ont souvent des sens proches. Les sélecteurs des mots d'entrée sont extraits des contextes locaux gauche et droit, puis l'ensemble *S* de tous les sélecteurs obtenus est comparé avec les synsets de WordNet. Le synset qui a le plus de mots en commun avec *S* est sélectionné comme sens adéquat du mot cible. Dans l'approche d'indexation de Khan [12], pour désambiguïser un mot à partir des concepts correspondants (dans une ontologie de sport), on détermine le degré de corrélation des concepts sélectionnés, sur la base de leur proximité sémantique. La proximité sémantique de deux concepts est calculée par un score basé sur leur distance minimale mutuelle dans l'ontologie. Les concepts qui ont les plus hauts scores sont alors retenus. Dans une approche similaire, Baziz et al. [1] se basent sur le principe que, parmi les différents sens possibles (dits concepts candidats) d'un terme donné, le plus adéquat est celui qui a le plus de liens sémantiques [15], [16], [21] avec les autres concepts du même document. L'approche consiste à affecter un score à chaque concept candidat d'un terme d'indexation donné. Le score d'un concept candidat est obtenu en sommant les valeurs de similarité qu'il a avec les autres concepts candidats correspondant aux différents sens des autres termes du document. Le concept candidat ayant le plus haut score est alors retenu comme sens adéquat du terme d'indexation associé. La désambiguïsation est ici globale contrairement aux approches précédentes. Dans notre approche de désambiguïsation proposée dans [3], [4], ce score est basé sur la somme des valeurs de similarité qu'il a avec les concepts candidats les plus fréquents dans le document.

(2) La pondération des concepts : La pondération des concepts se décline en deux principales tendances : (1) la pondération des concepts en tant qu'entités lexicales et (2) la pondération des concepts en tant qu'entités sémantiques. Dans l'approche de pondération des concepts en tant qu'entités lexicales, les concepts sont considérés à travers les termes qui les représentent. La pondération des concepts consiste alors en la pondération des termes correspondants. Les approches de pondération de Voorhees [27] et de Baziz et al. [1] sont basées sur ce principe. En se basant sur le modèle vectoriel étendu introduit dans [9], dans lequel chaque vecteur est composé d'un ensemble de sous-vecteurs de différents types de concepts (appelés *ctypes*), Voorhees [27] propose de pondérer les concepts en utilisant un schéma de pondération classique *tf*idf* normalisé. L'approche proposée par Baziz et al. [1], étend la pondération *tf*idf* pour tenir compte des termes composés. L'approche proposée dite approche *Cf*idf*, permet de pondérer un terme *t* composé de *n* mots, par la fréquence d'occurrences du terme lui-même, et par celles des sous-termes qui le composent. Dans l'approche de pondération des concepts en tant qu'entités sémantiques, il s'agit d'évaluer l'importance des sens (concepts) dans le contenu du document indexé. L'importance d'un concept dans un document

est évaluée en tenant compte du nombre de ses relations sémantiques avec les autres concepts du document [5], [8], [11]. Ces relations sémantiques sont en outre pondérées dans [11]. Dans l'approche proposée par Boughanem et al. [5], le nombre de relations d'un concept avec les autres concepts définit une mesure dite de centralité du concept. Les auteurs combinent centralité et spécificité pour évaluer l'importance des concepts d'un document. La spécificité du concept définit son degré de « spécialité » (par opposition à généricité). .En combinant pondération sémantique et pondération lexicale des concepts, notre approche définie dans [3], [4], propose de pondérer les termes composés sur la base d'une mesure probabiliste tenant compte des sens possibles du terme par rapport aux sens de ses sous-termes, et de ses sur-termes en tenant compte de leurs fréquences d'occurrences respectives.

### 2.4 Positionnement de nos travaux

Notre approche proposée dans ce papier est une version revue de notre approche d'indexation conceptuelle dans [3], [4]. L'objectif est de représenter le document par un noyau sémantique composé de concepts pondérés. Les concepts sont extraits de WordNet à l'issue d'une identification-désambiguïsation. Puis les concepts sont pondérés. Dans notre approche d'indexation conceptuelle proposée dans ([3], [4]), les termes d'indexation sont d'abord extraits en se basant sur des étapes d'indexation classiques. Puis chaque mot non vide identifié est projeté sur WordNet. L'objectif est de recenser toutes les entrées de WordNet contenant ce mot. Ces entrées sont utilisées pour définir le contexte local du mot dans le document. Ce contexte permet d'identifier dans le document la collocation (suite de mots) la plus longue correspondant à un synset de WordNet. Lorsqu'un terme est ambigu, la désambiguïsation est appliquée. L'approche se base sur le calcul d'un score, basé sur la somme des valeurs de similarité qu'il a avec les concepts candidats les plus fréquents dans le document. Une approche de pondération des concepts est proposée. La pondération d'un terme *t* est basée sur une mesure probabiliste des sens possibles de *t* (notés *Sens(t)*) par rapport aux sens de ses sous-termes (*Sub(t)*) et de ses sur-termes (*Sur(t)*), en tenant compte de leurs fréquences d'occurrences respectives (*tf*). La probabilité qu'un terme *t* soit un sens possible d'un terme *t'*est mesurée comme le rapport entre le nombre de sens possibles du terme *t* incluant le terme *t'*, sur le nombre de sens possibles du terme *t*. Formellement :

$$P(t \in Sens(t')) = \frac{|\{C \in Sens(t') / t \in C\}|}{|Sens(t')|}$$

Le poids d'un terme *t* est alors défini par la formule suivante, où *N* représente le nombre total de documents dans le corpus et *df(t)* la fréquence documentaire inverse :

$$W_{t,d} = \left( tf(t) + \sum_i tf(Sur_i(t)) + \sum_j \left[ P(t \in S(Sub_j(t))) * tf(Sub_j(t)) \right] \right) * \ln\left( \frac{N}{df(t)} \right)$$

Dans ce papier, nous redéfinissons l'approche d'identification des entrées de WordNet correspondant à un mot donné ainsi que l'approche de pondération des concepts.

- L'approche d'identification des concepts est basée sur le degré de recouvrement des entrées de WordNet et du contexte local (la phrase) dans

lequel le mot apparaît dans le document. Contrairement à l'approche proposée dans [3], cette approche présente l'avantage de permettre la détection de collocation de mots indépendamment de leur ordre d'apparition dans le contexte.

- L'approche de pondération des concepts est basée sur une nouvelle mesure de l'importance d'un concept dans un document. Cette mesure tient compte d'une part des proximités sémantiques entre le concept à pondérer et les autres concepts du document, et d'autre part des fréquences d'occurrences de ces concepts. Dans cette approche, l'apport des sous-termes n'est pas considéré. Notre précédente approche dans [3] peut être combinée à la présente approche pour en plus tenir compte de cet apport.

Dans ce qui suit, nous décrivons les différentes étapes de notre approche d'indexation conceptuelle.

## 3 Indexation conceptuelle des documents

L'indexation conceptuelle vise à représenter un document par un noyau sémantique composé de concepts pondérés qui décrivent au mieux son contenu.
Le processus d'indexation du document s'effectue en trois étapes: (1) l'identification des termes d'index, (2) la désambiguïsation des termes d'index et (3) la pondération des concepts.

### 3.1 Identification des termes d'index

Le but de cette étape est d'identifier l'ensemble $T(d)= \{t_1, t_2,\ldots, t_n\}$ des termes $t_i$ du document $d$ qui correspondent à des entrées dans WordNet. L'identification des termes se base sur le degré de recouvrement du contexte local du mot analysé avec chaque entrée correspondante dans WordNet. L'entrée qui a le plus haut degré de recouvrement est retenue comme sens possible du mot analysé. Le principe de l'identification des termes est décrit à travers l'algorithme de la Table 2.

**Algorithme de détection de concepts**
**Entrée** : document *d*.
**Sortie** : index *T(d)*
**Procédure** : Soit $mot_i$, le prochain mot, non vide, à analyser dans *d*. On appellera contexte $\zeta_i$ du mot $mot_i$ dans le document *d* la phrase courante de *d* qui contient le mot $mot_i$.

1. Calculer $S= \{C^i_1, C^i_2,\ldots C^i_n\}$ l'ensemble des synsets contenant le mot $mot_i$. *S* est composé de mono et de multi-mots.
2. Ordonner S comme suit : $S= \{C^i_{(1)}, C^i_{(2)},\ldots, C^i_{(n)}\}$ où $(j)_{1..n}$ est une permutation d'indices telle que $|C^i_{(1)}| \geq |C^i_{(2)}| \geq \ldots \geq |C^i_{(2)}|$, où $|C^i_{(j)}|$ est la longueur exprimée en nombre de mots, de la chaîne de caractères représentant le concept $C^i_{(j)}$.
3. Pour chaque $C^i_{(j)}$ dans *S*, faire :
4. Calculer l'intersection des deux chaînes de caractères $\zeta_i$ et $C^i_{(j)}$.
5. Si $\left|\zeta_i \cap C^i_{(j)}\right| < \left|C^i_{(j)}\right|$ (le concept $C^i_{(j)}$ n'apparaît pas dans le contexte $\zeta_i$), le concept suivant, $C^i_{(j+1)} \in S$ est analysé,
Si $\left|\zeta_i \cap C^i_{(j)}\right| = \left|C^i_{(j)}\right|$ le concept $C^i_{(j)}$ est identifié comme concept associé au mot $mot_i$, et le terme représentatif correspondant $t_i$ est ajouté à l'index du document *d* ;

Table 2: Algorithme de détection des termes d'index

### 3.2 Désambiguïsation des termes

Les termes d'index sont associés à des sens (synsets) correspondants dans l'ontologie. Chaque terme extrait pouvant avoir plusieurs sens possibles, le but de cette étape est de sélectionner le meilleur sens du terme dans le document. L'approche de désambiguïsation utilisée est celle proposée dans [3]. Pour désambiguïser un terme $t_i$ donné, on associe à chacun de ses sens possibles $C_j^i$ un score basé sur:

- les distances sémantiques $Dist(C_j^i, C_k^l)$ entre ce concept et les autres sens possibles associés aux autres termes dans le document,
- les fréquences d'occurrences des termes associés.

Formellement :

$$Score(C_j^i) = \sum_{\substack{l \in [1,..,m] \\ l \neq i}} \sum_{1 \leq k \leq n_l} tf(C_j^i) * tf(C_k^l) * Dist(C_j^i, C_k^l) tf(C^i) \quad (1)$$

Où $Dist(C_j^i, C_k^l)$ est la distance sémantique entre les concepts $C_j^i$ et $C_k^l$.

Le concept $C_j^i$ ayant le plus grand score est alors retenu comme sens adéquat du terme $t_i$ dans $d$. L'ensemble des concepts retenus constituera le noyau sémantique $N(d)$ du document $d$.

### 3.3 Pondération des concepts

Partant de l'idée qu'un concept est d'autant plus représentatif du contenu du document qu'il est fortement corrélé avec les concepts des termes les plus importants (au sens fréquents) du document compte tenu de sa propre importance dans le document, nous proposons de pondérer un concept avec un poids basé sur :

- sur les distances sémantiques entre ce concept et les autres concepts dans le document,
- et sur les fréquences d'occurrences des concepts associés.

Formellement, le poids $W(C^i)$ d'un concept $C^i$ est défini par :

$$W(C^i) = \sum_{i \neq l} tf(C^i) * tf(C^l) * Dist(C^i, C^l) \quad (2)$$

Le noyau sémantique de $d$ est alors construit en gardant seulement les concepts dont les poids sont plus grands qu'un seuil fixé. Nous proposons, dans un premier temps, de garder tous les concepts dont le poids est différent de zéro.

Evaluation expérimentale

L'objectif de ces expérimentations est de mesurer l'efficacité de notre approche de RI sémantique. On présente dans ce qui suit la collection de test et l'approche d'évaluation utilisées.

(1) Collection de test : La collection de test utilisée est la collection Muchmore[2]. Le corpus MuchMore est un corpus parallèle de résumés médicaux scientifiques

[2] http://muchmore.dfki.de/

anglais-allemands obtenus à partir du site web de Springer. Seule la collection des textes anglais non annotée a été utilisée. Cette dernière collection est composée de 7823 documents et de 25 requêtes. Les documents et les requêtes sont composés de textes simples.

(2) Approche d'évaluation : L'approche est évaluée en utilisant le système Mercure [6]. L'évaluation est effectuée selon le protocole TREC. Chaque requête est soumise au système de RI avec les paramètres fixés. Le système renvoie les 1000 premiers documents pour chaque requête. Les valeurs de précision P5, P10, P20 et MAP (précision moyenne) sont calculées. La précision au point *x* (x=5, 10, 20,), *Px*, est le ratio des documents pertinents parmi les *x* premiers documents restitués. R-Prec et MAP sont les précisions exacte et moyenne respectivement. Nous comparons ensuite les résultats obtenus à partir de notre approche à un système de référence (ou baseline).

### 3.4 Evaluation de l'approche d'indexation par les concepts

Les premières expérimentations menées concernent l'approche d'indexation par les concepts. Il s'agit alors d'évaluer l'impact de la qualité de l'index sémantique du point de vue de l'efficacité de la recherche. Pour atteindre cet objectif, nous comparons plus précisément deux index :

- Le premier constitué par les concepts détectés par notre approche (décrite en section 2.2) et désambiguïsés (approche décrite en section 2.3). C'est l'approche notée *Concepts-TF* sur la Figure 1.

- Le second constitué par les concepts détectés par notre approche et désambiguïsés, combinés aux mots clés. Les mots clés font référence aux mots du document qui n'ont pas d'entrée correspondante dans WordNet. C'est l'approche notée *Concepts-Fusion* sur la Figure 1.

Les résultats de ces deux index sont d'abord comparés par rapport à ceux de deux baselines :

- La première est une baseline classique fondée sur une indexation basée mots clés pondérés par *tf*idf*. Cette approche est désignée par *Classic-TFIDF* sur la Figure 1,

- la seconde est une baseline classique fondée sur une indexation basée mots clés pondérés par la BM25 [22]. Cette approche est désignée par *Classic-OKAPI* sur la Figure 1.

Les résultats de l'évaluation de ces approches sont donnés en Figure 1. De ces résultats, il ressort que :

- l'approche d'indexation *Concepts-TF* par les seuls concepts désambiguïsés est meilleure que la baseline *Classic-TFIDF* avec des taux d'accroissement respectivement de 61% pour P5, de 51% pour P10, de 54% pour P20 et de 51% pour la MAP

- l'approche d'indexation *Concepts-Fusion* est nettement meilleure que l'approche *Concepts-TF* avec des taux d'accroissement respectivement de 20% pour P5, 19% pour P10, 15% pour P20 et de 23% pour la MAP. Ces résultats nous confortent dans l'idée qu'une indexation combinée concepts+mots-clés est plus performante qu'une indexation par les concepts seuls.

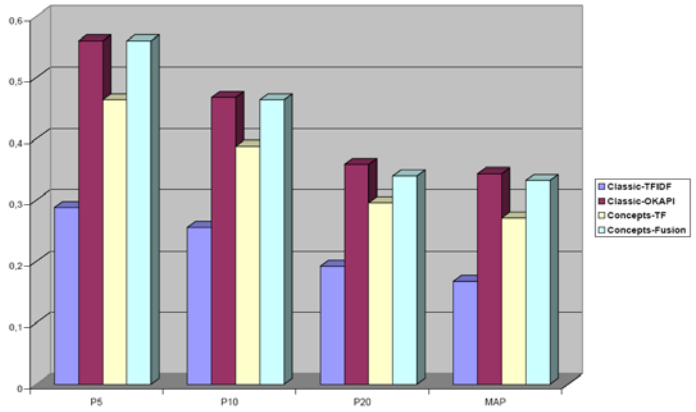


Figure 1 - Résultats d'évaluation de la méthode de détection de concepts par rapport aux baselines

Par ailleurs, notre approche combinée *Concepts-Fusion* présente des résultats nettement meilleurs qu'une baseline *Classic-TF*, avec des taux d'accroissement de 94% pour la P5, de 45% pour la P10, de 77% pour la P20 et de 77% pour la MAP. Néanmoins, comme le montre la Figure 1, l'approche *Concepts-Fusion* présente des résultats moins bons que ceux de la baseline *Classic-OKAPI* avec des taux de décroissement de 0% pour P5, -1% pour P10, -5% pour P20 et de -3% pour la MAP. La cause la plus probable à l'origine de ce problème pourrait être l'imprécision de la désambiguïsation. En effet, dans un contexte de désambiguïsation précise, on s'attend à ce que l'indexation par les concepts apporte au moins autant qu'une indexation classique.

### 3.5 Evaluation de l'approche de pondération de concepts

La deuxième série d'expérimentations menées concerne l'évaluation de notre approche de pondération des concepts introduite en section 2.3. Concrètement, il s'agit alors d'évaluer l'impact de la qualité de la pondération proposée (en section 2.3) du point de vue de l'efficacité de la recherche. Pour atteindre cet objectif, nous comparons plus précisément deux index :

- le premier est l'index composé des concepts détectés par notre approche proposée en section 2.2, pondérés par la fréquence. Cette approche est notée *Concepts-TF* sur la Figure 2.

- Le second est l'index composé des concepts détectés par notre approche proposée en section 2.2, pondérés par le poids proposé en section 2.3. Cette approche est notée *Concepts-Score* sur la Figure 2.

Les résultats de ces deux index sont comparés mutuellement. L'objectif est de mesurer l'apport de la pondération proposée par rapport à une pondération classique des concepts.

La Figure 2 présente les résultats obtenus. Il apparait que les résultats de la pondération par le poids proposé sont moins bons que ceux basés sur la fréquence des concepts, avec des taux de décroissement de -5% pour la P5, -6% pour P10, -12% pour P20 et -6% pour la MAP. Les résultats obtenus sont bien en deça de ce

qui était attendu. Le problème à l'origine de cette insuffisance vient probablement du score de *ranking*, utilisé par Mercure pour évaluer la correspondance d'un document pour une requête. Ce score est basé sur *tf*idf* (ou une de ses variantes). Ce qui explique que l'approche *Concept-TF* réponde favorablement à cette évaluation. A contrario, dans l'évaluation de l'approche *Concept-Score*, le poids du concept remplace *tf* dans le score de *ranking* et est combine donc à une mesure non corrélée, *idf*, provocant ainsi une baisse de la précision.

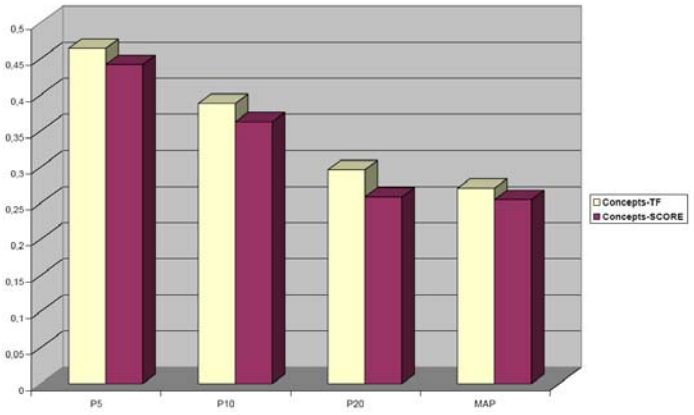


Figure 2 – Résultats d'évaluation de la méthode de pondération des concepts

## 4 Conclusion

Nous avons présenté dans ce papier, une approche d'indexation conceptuelle basée sur l'utilisation de WordNet. Le formalisme basé concepts est susceptible de résoudre les problèmes de disparité et d'ambiguïté des termes en RI.

En vue de palier à ces problèmes, nous avons présenté une approche en vue d'une indexation conceptuelle des documents. Notre contribution porte sur deux aspects principaux. Le premier consiste en l'indexation conceptuelle basée sur l'ontologie WordNet. L'approche n'est certes pas nouvelle mais nous avons proposé de nouvelles techniques pour identifier les concepts et pour les pondérer. Des résultats préliminaires ont montré que l'approche d'identification des concepts est plus performante qu'une baseline *Classic-TFIDF*, et apporte des taux d'accroissement appréciables par rapport à cette dernière, que les concepts soient utilisés seuls ou combinés aux mots clés. Cependant, l'approche d'indexation par les concepts n'a pas apporté les résultats escomptés par comparaison à une baseline *Classic-OKAPI*, probablement du fait de l'imprécision de la désambiguïsation. Par ailleurs, l'approche de pondération a produit des résultats mitigés. La cause probable de ces insuffisances inattendues serait l'inadéquation du score de *ranking* utilisé par rapport à l'index sémantique. Pour lever ces insuffisances, nous nous proposons, dans un premier temps, de parfaire le score de désambiguïsation, et dans un second temps de réfléchir un schéma de *ranking* pour des index sémantiques qui tienne compte des poids sémantiques des concepts et qui s'affranchit de la mesure classique *idf*. Des réflexions sont en cours dans ce sens.

## 5 Remerciements

Le présent travail a pu avoir lieu grace au soutien de l'A.U.F (Agence Universitaire de la Francophonie) et de l'U.M.M.T.O. (Université Mouloud Mammeri de Tizi-Ouzou) avec l'aimable collaboration de l'IRIT (Institut de Recherche en Informatique de Toulouse).

## 6 Bibliographie

[1] M. Baziz, M. Boughanem, N. Aussenac-Gilles. A Conceptual Indexing Approach based on Document Content Representation. Dans : CoLIS5 : Fifth International Conference on Conceptions of Libraries and Information Science, Glasgow, UK, 4 juin 8 juin 2005. F. Crestani, I. Ruthven (Eds.), Lecture Notes in Computer Science LNCS Volume 3507/2005, Springer-Verlag, Berlin Heidelberg, p. 171-186.

[2] M. Baziz, M.Boughanem, N.Aussenac-Gilles. The Use of Ontology for Semantic Representation of Documents. Dans: The 2nd Semantic Web and Information Retrieval Workshop(SWIR), SIGIR 2004, Sheffield UK, 29 juillet 2004. Ying Ding, Keith van Rijsbergen, Iad Ounis, Joemon Jose (Eds.), pp. 38-45.

[3] F. Boubekeur, M. Boughanem, L. Tamine. Exploiting association rules and ontology for semantic document indexing. Dans: 12th International conference IPMU08, Information Processing and Management of Uncertainty in knowledge-Based Systems, Malaga, 22- 27, june 08, Spain.

[4] F. Boubekeur, M. Boughanem, L.Tamine. Semantic Information Retrieval Based on CP-Nets. Dans : IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2007), London, 23/07/07- 26/07/07, IEEE, (support électronique), juillet 2007.

[5] M. Boughanem, I. Mallak, H. Prade. A new factor for computing the relevance of a document to a query (regular paper). Dans : IEEE World Congress on Computational Intelligence (WCCI 2010), Barcelone, 18/07/2010-23/07/2010, 2010 (à paraître).

[6] M. Boughanem, C. Soulé-Dupuy: A Connexionist Model for Information Retrieval. DEXA 1992: 260-265.

[7] M. Cuadros, JM., Atserias, J., M. Castillo, M., & G. Rigau, G. (2004). Automatic acquisition of sense examples using exretriever. In *IBERAMIA Workshop on Lexical Resources and The Web for Word Sense Disambiguation*. Puebla, Mexico.

[8] D. Dinh, L.Tamine. Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients (short paper). Dans : Conférence francophone en Recherche d'Information et Applications (CORIA 2010), Sousse, Tunisie, 18/03/2010-21/03/2010, Hermès, Mars 2010.

[9] E.A. Fox. Extending the boolean and vector space models of information retrieval with p-norm queries and multiple concept types. PhD thesis, Ithaca, NY, USA, 1983.

[10] J.A Guthrie, L. Guthrie, Y. Wilks, H. Aidinejad (1991). Subject-

dependant cooccurrence and word sense disambiguation. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkley, CA. 146-152.

[11] B.Y. Kang and S.J. Lee. Document indexing: a concept-based approach to term weight estimation. In Journal of Information Processing & Management. Volume 41, Issue 5, September 2005, Pages 1065-1080

[12] L.R. Khan, D. Mc Leod, E.Hovy. Retrieval effectiveness of an ontology-based model for information selection. The VLDB Journal (2004)13:71–85.

[13] R. Krovetz. Homonymy and polysemy in information retrieval. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (A CL-97}, pages 72-79.

[14] C. Leacock, G.A. Miller, and M. Chodorow. Using corpus statistics and WordNet relations for sense identification. Comput. Linguist. 24, 1 (Mar. 1998), 147-165.

[15] M.E. Lesk, Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a nice cream cone. In Proceedings of the SIGDOC Conference. Toronto, 1986.

[16] D. Lin. (1998) An information-theoretic definition of similarity. In Proceedings of 15th International Conference On Machine Learning, 1998.

[17] O. Medelyan ; D. Milne ; C. Legg ; I.H. Witten. Mining meaning from Wikipedia. In International Journal of Human-Computer Studies archive, Volume 67 , Issue 9 (September 2009). Pages: 716-754. Year of Publication: 2009. ISSN: 1071-5819

[18] R. Mihalcea and D. Moldovan. Semantic indexing using WordNet senses. In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October 2000

[19] G. Miller (1995) WordNet: A Lexical database for English. Actes de ACM 38, pp. 39-41.

[20] P. Resnik. Disambiguating noun groupings with repect to WordNet senses. 3thWorkshop on Very Large Corpora, 54–68. (1995).

[21] P. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, Journal of Artificial Intelligence Research (JAIR), 11, 1999, (p. 95-130).

[22] S.E. Robertson, The probability ranking principle in IR. Journal of Documentation 33, 294-304 (1977). Reprinted in: K. Sparck Jones and P. Willett (eds), *Readings in Information Retrieval*. Morgan Kaufmann, 1997. (pp 281-286).

[23] H. Schütze and J. Pedersen. Information retrieval based on word senses. In Proceedings of the 4th Annual Symposium on Document Analysis and

Information Retrieval, pages 161-175.

[24] M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. 2nd International Conference on Information and Knowledge Management (CIKM-1993), 67–74.

[25] O. Uzuner, B. Katz, D. Yuret: Word Sense Disambiguation for Information Retrieval. AAAI/IAAI 1999 : 985

[26] J. Véronis and N. Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. 13th International Conference on Computational Linguistics (COLING-1990), 2, 389–394. 1990.

[27] E. M. Voorhees. Using WordNet to disambiguate word senses for text retrieval. Association for Computing Machinery Special Interest Group on Information Retrieval. (ACM-SIGIR-1993) : 16thAnnual International Conference on Research and Development in Information Retrieval, 171–180. (1993).

[28] S.F. Weiss. Learning to disambiguate. Information Storage and Retrieval, 9, 33_41. (1973).

[29] Y. Wilks & M. Stevenson. Combining independent knowledge source for word sense disambiguation. Conference « Recent Advances in Natural Language Processing », 1–7.

[30] D. Yarowsky. "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora" Proceedings of the 14th International Conference on Computational Linguistics (COLING-92). Nantes, France, August, 454 – 460.

[31] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods, In 33rd Annual Meeting, Association for Computational Linguistics, Cambridge, Massachusetts, USA , 1995, (p189-196).

# Indexer des parcours thématiques pour valoriser les collections de presse numérisée

**Viviane Clavier**
Viviane.Clavier@u-grenoble3.fr

Université Stendhal, Laboratoire Gresec, Grenoble3

**Résumé** : Notre étude se situe dans le cadre d'une réflexion sur la valorisation des collections de presse numérisée du XIXème siècle et sur les modes d'accès à ce patrimoine, témoignage inestimable de notre passé. Nous nous intéressons aux parcours thématiques, dispositifs qui peuvent permettre au grand public de découvrir les collections. L'élaboration de parcours nécessite une indexation thématique du texte intégral. Notre propos est de définir ce qu'est un thème dans la presse, notion qui fait intervenir le typage des unités rédactionnelles ainsi que des marqueurs linguistiques qui rendent compte du positionnement éditorial du journal sur les événements.

**Mots-clés :** collections de presse numérisée, valorisation du patrimoine, indexation thématique

## 1 Introduction

Fin 2006, le journal hebdomadaire lyonnais le *Progrès Illustré (1890-1905)* est numérisé et mis en ligne par la bibliothèque municipale de Lyon sur un site expérimental, considéré comme pilote en France[1]. Une partie de la collection, notamment les *Causeries*, a fait l'objet d'une reconnaissance optique de caractères (ou océrisation). Début 2010, la collection numérisée a migré sur le site officiel de la bibliothèque municipale de Lyon : le *Progrès Illustré* figure à présent parmi de nombreux titres de presse régionale lyonnaise du XIXème siècle[2]. Représentatif d'une période qualifiée de « l'âge d'or de la presse » (Bellanger, 1972 : 22), le *Progrès Illustré* était le supplément littéraire du *Progrès* et paraissait le dimanche. La singularité et la richesse de cette collection régionale a été décrite par Jean-Pierre Bacot (2005). Vaste fourre-tout qui s'inscrit dans la lignée des *magazines*, la presse illustrée mêle littérature populaire, gravures, faits-divers, chroniques de jardinage, actualités régionales.
Actuellement, la plupart des institutions en charge de mettre en ligne leurs collections patrimoniales s'orientent vers une numérisation de masse qui conduit à « digitaliser » des millions de pages. A l'inverse de ces palmarès volumétriques, la bibliothèque municipale de Lyon reste plus modeste dans ses objectifs et souhaite mettre l'accent sur la constitution et la valorisation d'un patrimoine numérique de presse à destination du grand public. Isabelle Westeel rappelle à ce titre que le *« patrimoine » au sens d'héritage commun et de propriété collective est le bien de tous les publics,*

[1] (Landron, 2010)
[2] http://collections.bm-lyon.fr/presseXIX/showObject?id=PER003&date=00000522

*tous légitimes pour se l'approprier, il faut donc que les bibliothécaires pensent « usages et publics » dès la conception des projets de mise en ligne* (Westeel, 2004). C'est à l'occasion de la mise en ligne du *Progrès Illustré* que des contours plus précis ont été donnés à la notion de valorisation dans le cadre du programme de recherche CaNu XIX[3]. La valorisation est à entendre à plusieurs niveaux : a) permettre au lecteur de construire ses propres documents à partir des sources ; b) permettre aux professionnels des bibliothèques de construire des parcours thématiques ; c) offrir au lecteur une reconstruction du contexte spatial et temporel dans lesquels ces textes et gravures ont été produits. C'est le deuxième objectif qui nous intéresse, *i.e.* la construction de parcours thématiques. Nous conviendrons d'appeler *parcours thématiques* un dispositif de mise en exposition d'une collection numérique suivant un ensemble de *sujets* prédéfinis. La construction de ces parcours repose sur le principe de l'indexation thématique. Les parcours sont susceptibles d'être utilisés par le grand public pour découvrir les collections, ils suivent le même objectif que les dossiers thématiques qui sont des documents de synthèse rédigés par les professionnels. Dossiers et parcours sont conçus pour éveiller la curiosité, donner accès aux contenus et favoriser le passage d'un mode découverte à un mode d'interrogation de la base. Les parcours ne sont donc pas une fin en soi, mais un moyen de s'approprier les connaissances nécessaires pour devenir autonome dans la consultation.

La conception de dossiers thématiques est une pratique largement répandue sur les sites web des bibliothèques numériques (Gallica) ou des agrégateurs de contenus numériques (Europeana) : ils permettent de faire vivre le site, de l'animer et également de mettre en valeur un patrimoine numérique suivant un choix de thèmes attractifs. Cependant, les dossiers supposent un travail humain important puisqu'il s'agit de dépouiller la base, d'indexer le contenu au fil de la page, de collecter, trier, organiser l'information et, *in fine,* de rédiger une synthèse. Les parcours sont conçus comme une alternative aux dossiers thématiques, ce qui permettrait de diversifier les modes d'accès aux collections et d'épargner les professionnels des étapes de mise en forme et de rédaction abouties des documents de synthèse. Il faut comprendre cette étude comme une réflexion plus générale sur l'indexation thématique des journaux du XIXème siècle qui s'inscrit dans une optique de mise en valeur de ce type de fonds.

Si la question de l'indexation thématique n'est pas nouvelle en soi, ce qui est plus original en revanche, c'est de l'appliquer à des collections de journaux. Dans la presse, la définition de la thématique ne peut, selon nous, faire l'impasse d'une analyse qui s'attache à décrire les discours « au prisme de la ligne éditoriale » pour reprendre les propositions de Roselyne Ringoot, qui stipule que « l'analyse éditoriale est en quelque sorte un concept textuel [et que] c'est l'analyse du journal qui permet de la dégager » (2004 : 88). Cet objectif conduit à redéfinir les contours de notions généralement convoquées dans le domaine documentaire tels que les *sujets* et les *thèmes* que nous avons croisés avec une notion mobilisée dans les médias, les *événements*. Le principal objet de cet article réside dans la définition de ces notions qui déterminent trois niveaux d'indexation. Au-delà de ces aspects, l'étude pose également la question des ressources à mobiliser aux différentes étapes du processus d'indexation : quelles sont les contributions respectives des corpus, des collections, des connaissances des spécialistes, des documents

[3] Projet de recherche sur les *corpus numériques,* CaNu XIX (Canards Numériques du 19ème siècle, resp. Geneviève Lallich-Boidin), financé par la région Rhône-Alpes http://cluster13.ens-lsh.fr/spip.php?article117

historiques et des attentes des usagers ? Que peuvent apporter les outils et méthodes d'indexation automatique, sachant que les textes océrisés de collections numériques présentent entre 15 et 20% d'erreurs ? Enfin, quelle est la place des langages contrôlés dans ce genre d'application ? Les langages contemporains sont-ils adaptés à la description de documents anciens, peuvent-ils répondre à des états de langue différents, celui des collections du XIXème siècle, celui des usagers du XXIème siècle ? Sont-ils tout simplement adaptés au grand public, plus familier de requêtes sur les moteurs de recherche que sur les bases de données ?
Après une présentation des objectifs liés à la valorisation du patrimoine numérique et une analyse du rôle que peuvent jouer les parcours thématiques, nous dresserons un bref état de l'art sur les thèmes pour en venir à notre propre définition de l'indexation. Une étude des sciences et des techniques dans les *Causeries* illustrera la proposition d'indexation à trois niveaux.

## 2 Le parcours thématique, un enjeu de valorisation du patrimoine numérique ?

### 2.1 Missions pour les bibliothèques numériques : conservation, diffusion et mise en visibilité des collections

Parmi les innombrables programmes de numérisation du patrimoine écrit, la presse ancienne connaît depuis quelques années un grand succès aussi bien en France qu'à l'étranger. Les programmes de numérisation sont pour l'essentiel dévolus aux bibliothèques, plus rarement aux organismes de presse. Pour ces derniers, la numérisation est liée à des objectifs de réédition éditoriale qui permettent de « donner une seconde vie » aux collections (par exemple, *Le Progrès, 150 ans d'actualités à la Une*, novembre 2009), alors que pour les bibliothèques, le processus de numérisation est lié à des objectifs de conservation et de diffusion (Mezzasalma, 2009).
Dans son rapport sur la « Numérisation du patrimoine écrit », Marc Teissier préconise toutefois une stratégie plus offensive destinée à rendre les bibliothèques numériques visibles sur le web. Il évoque trois actions susceptibles de favoriser l'accès à une bibliothèque numérique (ici Gallica) et d'en accroître la visibilité : a) la multiplication des accès, depuis la base, à des contenus variés (stratégie dite de « liens fins ») ; b) l'amélioration du signalement et du référencement ; c) et un meilleur accès pris en compte par les moteurs de recherche des métadonnées et de l'indexation de l'ensemble des contenus (indexation « plein texte ») (Teissier, 2010 : 28)
Ces incitations doivent conduire à indexer massivement le contenu des bibliothèques numériques, dans le but d'améliorer le référencement auprès des moteurs de recherche et de favoriser un accès en mode texte. Certains professionnels de l'information affichent cependant une certaine prudence à l'égard du mode de recherche plein texte et évoquent d'autres formes d'accès au contenu. Ainsi, Isabelle Westeel (2009) affirme que ce mode de recherche, tout en étant indispensable, ne peut se passer d'une découverte préalable des collections :

> *On touche là à la nécessaire réflexion sur l'accès aux documents, qui ne saurait calquer les catalogues des bibliothèques : il faut donner à voir une collection numérique avant d'offrir une recherche précise, les promenades libres ou guidées ont une signification.* (Westeel, 2009 : 30)

Cette réserve bien légitime concernant la mise en ligne d'un patrimoine numérique culturel sans réflexion préalable sur les modalités d'accès a été maintes fois soulignées. Récemment, Marie Després-Lonnet indique au sujet d'Europeana qu'il y a une confusion entre la capacité technique de « rendre accessible » des

données, par leur « mise en ligne » et l'accès aux savoirs, c'est-à-dire, la possibilité effective donnée à un individu de s'approprier de nouvelles connaissances (Miège, 1997, cité par Després-Lonnet, 2009 : 19). L'auteur précise en effet que la numérisation du patrimoine relève davantage d'une *textualisation* que d'une *digitalisation,* et que l'accès au patrimoine numérique suppose de franchir la barrière de la langue, « barrière autrement plus redoutable que celle qui consiste à pousser les portes d'un musée » (Després-Lonnet : 21).

Finalement, la valorisation du patrimoine numérique révèle dans sa mise en œuvre et ses missions des objectifs difficilement conciliables. Pour les bibliothèques, la conservation et la diffusion imposent une politique de numérisation systématique et massive afin de préserver les collections. Cependant, compte tenu des coûts liés à la numérisation[4], la reconnaissance optique de caractères ne pourra s'appliquer aux collections dans les mêmes proportions. Or, l'océrisation est l'étape indispensable pour procéder à l'indexation du texte intégral, traitement qui pourrait assurer la visibilité des collections auprès des moteurs de recherche et l'accès au grand public. La réflexion sur les parcours thématiques se situe donc dans un contexte technique, institutionnel et politique relativement complexe. Il semble qu'elle trouve davantage d'écho auprès d'institutions bénéficiant d'une marge de manœuvre plus importante au niveau de leur politique documentaire. Les bibliothèques municipales, en raison de leur relative autonomie vis-à-vis des grands programmes nationaux et européens, de leur proximité avec leurs publics et de leur attachement aux collections ont sans doute plus de latitude pour valoriser leur patrimoine numérique.

### 2.2 Attentes des usagers : l'accès au document en mode texte, l'importance relative des thèmes

Les études consacrées aux usages des bibliothèques numériques sont encore relativement rares (Gallica[5], Europeana[6]). L'enquête réalisée sur Gallica 2 montre que les usagers plébiscitent les bibliothèques numériques dès lors que l'on peut accéder en mode texte aux collections, le feuilletage d'un document en mode image étant perçu comme fastidieux. Certains usagers reconnaissent toutefois que l'accès en mode texte présente des difficultés.

> *« Le perfectionnement sur Gallica 2, c'est la mise en mode texte. On sait ce qu'il y a dans l'ouvrage en quelques minutes. Dans Gallica 1, ça prend du temps de feuilleter, de trouver la table des matières. Mais il faut savoir rechercher. Il y a de l'imprécision dans le texte qu'on recherche.* » (Matharan et al., 2008 : 7)

Par ailleurs, cette enquête, qui s'attache aux pratiques de recherche d'information sur l'internet, indique que l'accès est facilité lorsque l'organisation de l'information est présentée de manière thématique. Si l'on peut supposer que cette assertion s'applique également à des collections fermées, aucune étude, à notre connaissance, ne peut encore le confirmer.

> *Une question ouverte permettait aux répondants d'indiquer quels sites ils fréquentaient (sur Internet) et auxquels ils participaient le plus. 240 réponses ont en tout été récoltées. Une logique thématique nette en émerge : on voit que les répondants mentionnent avant tout des sites traitant d'un thème précis, spécialistes d'un sujet (ibid : 7)*

---

[4] Une page numérisée coûterait environ 1€ auquel il faudrait rajouter 1€ supplémentaire pour la conservation. (intervention orale de Pascal Sanz, Directeur de Département Droit, économie et politique de la BnF, aux journées d'études « Regards croisés sur la mise en ligne et la valorisation de la presse XIX-XXI. » Lyon, les 6 et 7 mai 2010).

[5] (Matharan et al., 2008)

[6] (Lesquins, 2007) ; (Bouvier-Ajam, 2007)

Plus récemment, une enquête sur les usages du *Progrès Illustré* conduite par les participantes au Projet CaNu XIX (Paganelli et Mounier, 2010) montre que les lecteurs se répartissent globalement en deux catégories : les spécialistes consultent le support papier, le grand public, le support numérique. Le profil type du lecteur de la version en ligne du *Progrès Illustré* est lettré, souvent à la retraite, peu préoccupé par la pertinence des résultats fournis, parce qu'il ne cherche rien de précis. Ce qui le gêne en revanche, ce sont les problèmes d'affichage, les caractères étant souvent trop petits. Il s'intéresse essentiellement aux dates, aux lieux, aux événements, aux personnes, en dernier lieu aux sujets. Cette enquête ne révèle pas d'engouement particulier pour les parcours ou les dossiers thématiques.
Si l'on résume ces différents points de vue, il semble y avoir un consensus fort autour de l'accès en mode texte. En revanche, si la dimension thématique semble appréciée des internautes, elle ne saurait suffire pour le *grand public lettré*[7], lequel attend des collections de presse, des informations en lien avec l'actualité régionale de l'époque, des lieux et des dates. Concernant les parcours thématiques, on peut pour l'instant supposer qu'ils s'adresseront à un public de non-spécialistes, i.e qui ne sont ni amateurs éclairés ni professionnels, ce que l'on convient d'appeler par méconnaissance, « le grand public ».

### 2.3 Modes d'accès aux collections de presse : l'accès thématique encore marginal

Récemment, Agnieszka Smolczewska-Tona et Geneviève Lallich-Boidin (2008) ont présenté un état de l'art des différents modes de recherche dans les collections numériques de presse : des dispositifs les plus élémentaires, (accès par titre et par date de publication), aux plus élaborées (recherche par mots-clés), certaines interfaces offrent parfois la possibilité d'affiner les critères de recherche. Les auteurs mentionnent que les accès thématiques (*Colorado Historic Newspaper)* ou par sujets (*Brooklyn Daily Eagle Online)* commencent à se développer.
Ce mode d'accès, qui permet de naviguer dans une collection suivant une logique thématique ou par sujet, est considéré comme beaucoup plus performant que le mode de recherche par mot-clé (Abdullah et Gibb, 2009 cité par Da Sylva, 2009). Pierre Zweigenbaum et Benoit Habert (2004) indiquent que le « foisonnement de données textuelles et d'outils » conduit la plupart du temps à une désorientation des usagers, et qu'il est nécessaire de fournir des « boussoles sémantiques » pour naviguer dans les documents. Plusieurs travaux montrent que la navigation dans une structure pré-établie serait une aide considérable pour les usagers. Il existe plusieurs dénominations pour qualifier ces outils : « outil de butinage » pour Da Sylva (2009), il permettrait de guider l'usager et favoriserait « une appropriation graduelle d'un contenu, même si l'utilisateur n'a aucune connaissance préalable de celui-ci. » ; il serait également une « aide à la lecture » qui permettrait l'évaluation de la pertinence de documents. « Système de visualisation de l'information » pour Davis (2006)[8], il permettrait de regrouper les documents semblables, de cerner la pertinence des documents retrouvés et de combiner la recherche par mots-clés.
Si l'on observe le site de la bibliothèque municipale de Lyon, on peut observer deux procédés pour faire découvrir les collections. La rubrique intitulée « Notre sélection d'articles et d'illustrations » renvoie au premier procédé. Il permet à l'usager de choisir des documents dans une liste : les critères de la sélection ne sont cependant pas explicites. Les documents sont issus de différents titres de presse offerts dans la base, il n'y a pas de critères thématiques qui président à la

[7] (Paganelli et Mounier, 2010)
[8] cité par Da Sylva (2009 : 264)

constitution de ces listes. Ce mode de découverte permet de présenter la collection la plus célèbre du fonds (Le *Journal de Guignol*), des anecdotes (*Gazette de la mode*), des événements historiques (*Le canal de Panama*). Le second procédé réside dans la constitution de « dossiers thématiques ». En page d'accueil du site, sept dossiers sont présentés dans un espace dédié, distinct de l'espace kiosque et de l'interface de recherche, comparable à un espace d'exposition, au sens muséal du terme. Les dossiers reposent sur le principe de la synthèse et donnent lieu à un nouveau document placé sous la responsabilité d'un auteur. Les titres des dossiers mentionnent différents sujets : la bicyclette, l'anarchisme, les grandes affaires criminelles, la mode etc. En revanche, la facture des documents varie d'un dossier à l'autre. Par exemple, le dossier *Notre fin de siècle appartient à la bicyclette* s'appuie exclusivement sur des sources extraites du *Progrès Illustré* et des données historiques (naissance du code de la route, nom de l'inventeur du vélocipède). Il s'agit de présenter les points de vue d'éditorialistes du *Progrès Illustré* sur la bicyclette, de rapporter des citations d'hommes célèbres, de retracer des événements sportifs relatés par le journal et d'illustrer le dossier par des gravures sur le sujet. Le dossier *Le Progrès Illustré, témoin de son époque* suit également une logique chronologique et retrace les événements sociaux (grèves), politiques (exécution de l'anarchiste Ravachol), les affaires de corruption qui ont traversé les années 1891-1895. Encore différents se révèlent les dossiers *Surveillance et répression de la presse anarchiste* et *Elégante, suggestive, excentrique, la mode dans tous ses états*. En effet, les sources citées ne se limitent pas au *Progrès Illustré*, mais à d'autres titres du fonds de presse numérisée de la Bibliothèque (*L'émeute, L'étendard révolutionnaire, La mode illustrée*), voire à des sources extérieures comme le roman de Zola *Au bonheur des dames*, une base de données image sur les textiles, des ouvrages sur l'histoire de la mode. S'agissant du dossier sur la répression de la presse, le document s'apparente davantage à un travail de recherche croisant des sources extérieures, que sur une compilation d'extraits tirés de journaux.

En conclusion, bien qu'attractifs, les dossiers sont peu nombreux et couvrent peu de thèmes. Leur réalisation est par ailleurs largement dépendante des connaissances des indexeurs et se révèle très lourde en termes de traitement. Par ailleurs, les dossiers posent d'une part, la question de la nature des connaissances, des types de sources qui président à leur construction et d'autre part, la question de ce qu'est un thème dans la presse.

## 3 Définir, extraire et représenter des thèmes dans la presse

### 3.1 Le thème : une notion polysémique

La notion de thème est abordée par les disciplines littéraires et linguistiques et dans le cadre des pratiques documentaires.

Du côté des sciences du texte, il existe une littérature volumineuse consacrée au *thème* et à la *thématique*. Nombreux sont les travaux qui soulignent l'emploi difficile de cette notion au point même que certains auteurs s'interrogent sur la pertinence du concept.[9] L'une des raisons qui fait obstacle à une définition synthétique du *thème* réside dans l'extrême hétérogénéité des perspectives d'analyse. Le *thème* a été défini dans le cadre de la phrase, du texte et du discours. Les unités thématiques peuvent être diversement approchées. Elles peuvent être identifiées syntaxiquement grâce à des marqueurs de thématisation (Porhiel, 2005). Elles peuvent être assimilées à des unités lexicales structurées en champs

[9] voir Porhiel (2005) pour un état de l'art sur le thème.

sémasiologiques ou onomasiologiques (Trudel, 2009). Elles peuvent consister en un « agrégat des thèmes des phrases qui composent un paragraphe ou un texte » (Goustos, 1997)[10] ou encore ne correspondre à aucun constituant dans un énoncé, mais à un « topic », c'est-à-dire une relation « d'à-propos »[11]. Enfin, François Rastier (1999) indique qu'un thème peut renvoyer « à une structure stable de traits sémantiques, récurrente dans un corpus, et susceptible de lexicalisations diverses. » Il précise en outre que « selon les discours et les genres, les normes de lexicalisation des thèmes varient ». Pour Evelyne Martin le thème connaît plusieurs dénominations : il peut être l'équivalent de *motif* (au sens littéraire du terme) de *leitmotiv* (dans une acception plus musicale) et manifeste le plus souvent un principe de récurrence. (Martin, 1995 : 15-16).

Dans le domaine documentaire, la notion de *thème* est mobilisée pour décrire des outils d'accès au contenu, les index thématiques, ou pour concevoir des documents de synthèses, les dossiers thématiques. La terminologie est flottante entre *sujet* et *thème*, les index thématiques permettant d'accéder aux documents *qui parlent de la même chose*, i.e. qui traitent du même *sujet*. Dans le contexte numérique, Muriel Amar (2004) souligne que de nouveaux enjeux se profilent pour les index thématiques. L'auteur indique que les index doivent à présent être intégrés aux documents primaires afin de servir d'outils de recherche et de lecture « *sous réserve qu'ils relèvent du statut linguistique adéquat (unités nominales référentielles) »*. Par ailleurs, les index doivent permettre *de manipuler non plus l'intégralité d'un document mais aussi des segments pouvant, le cas échéant, être combinés pour produire de nouveaux documents, sous réserve que soient introduites des connaissances contextuelles, externes au document.* (Amar, 2004 : 62-63). La satisfaction de ces objectifs suppose une refonte profonde des objectifs de l'indexation documentaire. L'auteur constate en effet que la construction d'index thématiques rencontre des obstacles que l'indexation contrôlée ne peut résoudre dans le contexte numérique. D'une part, la construction d'un thème est une opération fondamentalement discursive qui nécessite des connaissances extérieures au document, auxquels «l'utilisateur » n'a pas accès. Il en résulte une interprétation difficilement « reconstituable ». D'autre part, les formulations utilisées dans les langages contrôlés sont de nature lexicale et référentielle (seuls les groupes nominaux sont des descripteurs), alors que les thèmes ne sont pas des unités référentielles mais discursives. Ce constat *désespérant* conduit l'auteur à poser la question de savoir comment concilier la thématisation et la référenciation dans les objectifs de l'indexation professionnelle. Elle prône une indexation qui permettrait un accès direct au texte intégral, qui ne s'attacherait pas à une indexation lexicale mais discursive. Elle nomme « indexation discursive » le processus qui consiste à donner à l'utilisateur non pas des mots pour dire les thèmes, mais des documents, regroupés thématiquement, qui sont les contextes « qui rendent intelligibles et interprétables les thèmes des documents » (ibid. 65). Pour l'auteur, indexer consiste alors à permettre la construction des unités d'interprétation que propose le texte, et non les nommer.

Nous adhérons à ce dernier point de vue qui réaffirme l'inadéquation d'une approche lexicologique pour atteindre les thèmes et qui pointe sur l'impossibilité de dénommer des thèmes au sein de catégories référentielles. Bref, ces arguments

[10] Cité par (Porhiel, 2005)

[11] Marandin J.-M., « Thème, topic de discours » *Dictionnaire de sémantique* http://www.semantiquegdr.net/dico/index.php/Th%C3%A8me_%28topic%29_de_discours

montrent que les approches documentaires fondées sur le recours à des langages contrôlés pour décrire les thèmes sont inappropriées. En revanche, mettre à jour des contextes intra-discursifs ou des documents extérieurs pour favoriser l'interprétation des thèmes nous semble une notion prometteuse. Nous allons à présent nous intéresser à la notion de thème dans la presse.

### 3.2 Les thèmes dans la presse

Deux grandes familles de travaux abordent les thèmes dans la presse : la sociologie des médias et la linguistique textuelle. Dans le champ de la sociologie des médias, ce sont les travaux d'inspiration linguistique ou sémiotique sur les documents textuels ou audiovisuels qui mobilisent la notion. La linguistique textuelle, quant à elle, s'attache à théoriser le texte et à hiérarchiser ses composants dans le cadre de grammaires. Le thème participe de l'organisation textuelle, la progression thématique étant notamment liée à la cohésion du discours (isotopie) et, aux marqueurs de connexité (anaphores). Bien que poursuivant des objectifs radicalement différents, ces deux familles de travaux se « re-connaissent ».

Ainsi, Jean-Michel Adam et Gilles Lugrin cherchant à typer les unités rédactionnelles et catégorielles de la presse contemporaine évoquent les *thèmes,* pour désigner *« des objets de discours inséparables des familles d'événements»* (Adam et Lugrin, 2000 : 13). Ils s'appuient sur les travaux de Maurice Mouillaud et Jean-François Têtu (1989) pour qualifier les « familles événementielles » de catégories référentielles apparaissant au sein des rubriques. Les *nouvelles politiques, les catastrophes, les conflits sociaux* sont, pour ces spécialistes des médias, des familles d'événements, notion qui à son tour, fait l'objet d'une littérature titanesque. Ce qu'il faut retenir des événements, c'est que la qualification d'événement dans les médias n'est pas du ressort de la linguistique mais procède d'une reconfiguration de la réalité « *déformée* » par *« l'industrialisation des métiers de la presse, le développement des technologies modernes de communication et/ou les intérêts économiques et financiers des groupes qui les fabriquent* » (Arquembourg, 2006 : 14). Il existe des typologies d'événements dressées dans le cadre de la norme de métadonnées IPTC[12]. Michael Palmer présente et commente des exemples de ces catégories référentielles qui permettent de « ventiler l'actualité » (Palmer, 2006 : 53) Par exemple, la violence présente 21 catégories (*guerres et conflits, actes de terrorisme, rébellions*, etc.). D'un point de vue documentaire, ce genre de typologie peut présenter un intérêt pour enrichir les métadonnées dans les corpus de presse contemporaine. Pour la presse du XIXème en revanche, la difficulté essentielle consiste à typer des notions qui ont pu apparaître comme des événements en leur temps, mais qui, aux yeux de l'histoire n'en étaient pas… et inversement.

L'analyse d'un thème dans la presse nécessite la prise en compte de plusieurs niveaux qui résultent de « l'éditorialisation » des discours. Roselyne Ringoot rappelle en effet que « l'analyse d'un thème informatif nécessite un travail de diagnostic éditorial qui contextualise le traitement d'une information en fonction de la politique éditoriale d'un journal. » (Ringoot, 2004 : 88). Parmi les différentes dimensions qui interviennent pour le typage d'un thème, il y a notamment les éléments qui participent de la morphologie du journal, les rubriques qui permettent d'établir l'identité énonciative des journalistes, le péritexte (titres et

[12] L'IPTC (International Press and Telecommunications Council) est une organisation internationale créée en 1965 pour développer et promouvoir des standards d'échange de données à destination de la presse.
http://www.iptc.org/cms/site/index.html?channel=CH0086

intertitres), les genres, et, pour reprendre la proposition de Maurice Mouillaud et Jean-François Têtu, les « familles événementielles ». Afin de donner une consistance langagière à la notion d'événements, nous pouvons dans un premier temps ramener un événement à un énoncé décomposable en un ensemble de catégories sémantiques et aspectuelles[13], telles que des dates, des lieux, des personnes et des verbes. Le typage des événements se ramènerait à un exercice de catégorisation et de hiérarchisation de classes d'arguments et de prédicats. Pour Maurice Mouillaud et Jean-François Têtu, un thème dans la presse du XIXème serait par exemple, *le patriotisme, le courage, la barbarie, l'anarchisme, la colonisation* –[14] qui révèlent des récurrences sémantiques au sein du discours. On voit donc bien qu'un même événement peut se prêter à une multitude de discours, de cadrages et de thématisations.

Revenons à la presse illustrée. Cette dernière fait feu de tout bois comme l'indique le rédacteur en chef du *Journal Illustré « Le Journal illustré est le journal de tous, comme il est le journal de partout. […] Notre mosaïque illustrée n'a pas d'école. Nos dessins s'inspirent de toutes choses interprétées par tous les crayons. Notre texte est rédigé par des plumes de différentes couleurs. Qui nous en voudrait ? » (Bacot, 2005 : 117).* Les éléments qui participent de la morphologie du *Progrès illustré* sont différents de ceux de la presse actuelle. Les rubriques ne sont pas encore celles que l'on connaît. Ainsi le *Progrès Illustré,* qui se donne pour mission de mettre l'art et la littérature à la portée de tous, présente des titres de rubriques qui ressemblent plutôt à des intitulés de genres (*Feuilleton, Poésie, Roman d'aventure),* à des activités ludiques (*Récréation*), à des moments de divertissement (*Jeux d'esprit, Mots pour rire*), à des rendez-vous familiers (*Causerie*). Les unités rédactionnelles ne sont d'ailleurs pas systématiquement « rubriquées », certains écrits littéraires comportant uniquement un titre *(Le Cuirassier Blanc, Le papillon).* La *Causerie* ressemble pourtant bien à une rubrique au sens contemporain du terme[15, elle] donne au journal son identité populaire et badine. En revanche, le genre de la *Causerie* se rapproche de plusieurs genres actuels, parfois comparable à la tribune, au portrait, à la chronique ou au fait-divers. Les causeries sont rarement sous-titrées, parfois datées (75 causeries sur 389). Elles sont systématiquement signées : les auteurs des causeries sont au nombre de 10. Certains n'ont écrit qu'une seule causerie (Caribet, Arsène Alexandre). Jacques Mauprat est l'auteur de la plupart des causeries (353), Paul Clairfont vient en second. Certains auteurs ont pu être identifiés : ce sont des hommes de lettres, critiques d'art, biographes, romanciers, historiens, chroniqueurs. L'origine littéraire des auteurs donne à voir dans les causeries un curieux mélange d'une presse populaire, qui cultive « l'art de dire » comme dans le journalisme issu de l'Ancien Régime, c'est-à-dire une presse de lettrés qui maintient la primauté des « littérateurs » sur les « informateurs »[16]

En résumé, l'identification des thèmes doit être réalisée dans un cadre contraint qui prend en compte plusieurs niveaux de description qui contribuent à cerner la ligne éditoriale : a) le journal, identifiable par son titre, son numéro, sa date ; b) la rubrique associée au titre et sous-titre éventuels, la date, l'auteur ; c) le genre à défaut de rubrique et d) les familles d'événements, identifiées par des dates, des lieux, des personnes et des verbes. Ces niveaux de structuration peuvent être

[13] L'événement est en linguistique une catégorie aspectuelle des verbes.

[14] Exemples tirés de Têtu (1997)

[15] « La rubrique a plusieurs fonctions, parmi lesquelles celles de classification et de hiérarchisation des informations ; mais elle permet également de donner au journal une identité qui lui est propre. » (Herman et Lugrin, 1999 : 72)

[16] (Ferenczi, 1996 : 21)

décrits par des métadonnées XML, encodées avec un schéma de DTD en TEI[17], recommandations largement utilisées pour décrire les données textuelles en sciences humaines et sociales. Il reste à présent à s'interroger sur les méthodes à utiliser pour extraire les thèmes.

**2.3 Les méthodes d'extraction de thèmes**

Dans le contexte documentaire, les thèmes sont généralement issus d'un seul document, l'indexation thématique consistant à épuiser tout le contenu. Les thèmes peuvent être extraits manuellement ou automatiquement. Les méthodes automatiques représentent actuellement la seule solution pour traiter de grandes quantités de données.

Parmi les méthodes automatiques, Frédérik Bilhaut rappelle que deux familles de méthodes coexistent (2006 : 28). Les méthodes numériques se fondent sur la notion de cohésion lexicale, c'est-à-dire la répétition de mots comme indicateur d'homogénéité thématique. Dénommée *text-tiling* par Hearst, son inventeur, la méthode conduit à une segmentation linéaire du texte en unités continues qui ne se superposent pas, chaque segment étant décrit par un vocabulaire spécifique. Trouver des thèmes consiste alors à identifier les frontières qui révèlent des changements lexicaux dans les segments. La seconde famille de méthodes est linguistique et consiste à exploiter différents marqueurs linguistiques et positionnels porteurs d'indications de la structure thématique ; d'autres approches consistent encore à découper le texte en unité thématiques et rhématiques.

Le problème majeur qui se pose dans la méthode du *text-tiling*, réside dans le procédé de segmentation d'unités non superposables. En effet, dans le cas d'articles de presse, les thèmes peuvent se répéter dans les numéros, un thème pouvant faire l'objet de l'actualité pendant plusieurs semaines, voire plusieurs mois ; ils peuvent être abordés dans plusieurs titres de journaux, et, à l'intérieur d'un même numéro, dans plusieurs rubriques. En ce qui concerne les méthodes linguistiques en partie « automatisables », Frédérik Bilhaut fait remarquer que les marqueurs de thématisation semblent difficiles à caractériser a priori, et que la démarche relève davantage d'une démarche d'observation de corpus que de repérage automatique de thèmes (Bilhaut, 2006 : 100). Quant à la méthode de typage des unités thématiques et rhématiques, l'auteur reconnaît l'intérêt de la démarche, met cependant en doute le degré de généralité des ressources et évoque également la charge de travail nécessaire à leur constitution (*ibid.* : 101).

Récemment, trois auteurs du laboratoire LIRIS ont présenté une méthode de catégorisation de l'information d'actualité dans des dépêches de presse collectées par flux RSS (Laitang et al, 2009). La catégorisation s'appuie sur la sélection et la pondération de « termes » représentatifs d'une thématique (par ex. *le sport, la politique*) et d'un sujet, notion qui tient compte de l'apparition et de la disparition dans le temps de nouvelles (par ex. *élection américaine, tremblement de terre*). Ce qui est très intéressant dans cette démarche, c'est la double prise en compte des dimensions temporelle et thématique pour regrouper les informations d'actualité. La spécificité événementielle des corpus de presse est également mobilisée pour indexer les contenus, puisque les auteurs ont fait le choix de retenir les catégorisations de dépêches de presse (IPTC) évoquées *supra* comme base de référence thématique statique. Ce qui en revanche, semble difficilement transposable à notre travail c'est tout d'abord, le recours à des ressources sémantiques contemporaines pour enrichir les termes (ontologies, thesaurus, bases de données lexicales). En outre, les calculs de proximité entre les termes s'appuient

[17] http://www.tei-c.org/index.xml

sur les corpus lemmatisés, ce qui semble, là encore difficilement applicable aux articles de presse océrisés.
Cette dernière approche présente très clairement l'ensemble des étapes et méthodologies pour extraire, indexer et catégoriser des thèmes récurrents afin de déterminer les schémas de propagation des actualités sur internet. Notre objectif est différent, puisque nous voulons à partir d'un angle d'analyse précis, indexer des thèmes et les structurer en parcours. Dans ce qui suit, nous présentons notre définition du thème et nos choix d'indexation.

## 4 Les sciences et techniques dans le *Progrès Illustré* : étude de cas

Nous proposons de définir trois niveaux d'indexation qui interviennent pour l'identification d'une thématique. Le première niveau consiste à construire un index des *sujets* ; le deuxième, un index des *événements*, et le troisième, un contexte de localisation des *thème*s. Le niveau 2 est un sous-ensemble du niveau 1 : il révèle les sujets qui font l'objet d'un événement ; et le niveau 3 permet de contextualiser les niveaux 1 et 2 : il révèle le positionnement éditorial du journal dans le traitement médiatique d'un sujet, et favorise une interprétation de la thématique. Nous allons à présent détailler chacun de ces niveaux en évoquant le statut de l'unité textuelle, la nature des unités d'indexation et, le cas échéant, le langage d'indexation retenu.

### 4.1 Indexation par sujets

Le choix d'un *sujet* ne repose pas sur une analyse des fréquences « des mots » des textes. Les critères de choix sont extérieurs au corpus : ils peuvent résulter de la connaissance qu'ont les indexeurs des usagers du fonds, de l'analyse des traces de requêtes, de spécificités jugées amusantes ou curieuses de la collection (les anecdotes, les caricatures), de préconisations de spécialistes sur l'apport de ces documents pour l'histoire, la littérature, ou la presse, de choix muséographiques (faire une exposition), etc. La liste des *sujets* n'est donc pas définie *a priori*, elle se construit au fil du temps, en fonction de choix de valorisation. Cette démarche est également celle qui préside à la construction des dossiers thématiques.
Les dénominations des *sujets* peuvent être choisies dans un langage contrôlé, si possible interopérable, ce qui permettrait un moissonnage de métadonnées entre les bibliothèques numériques. Dans l'idéal, le langage de classification devrait être contemporain des collections afin d'être en meilleure adéquation avec les vocabulaires. Les premières classifications Dewey pourraient convenir, mais elles sont en anglais. La question de l'interopérabilité se situe donc à deux niveaux : celui de la langue, et celui des états de langue, puisqu'il faudrait envisager une correspondance entre une classification du XIXème siècle et sa version contemporaine. Il convient d'associer à chaque classe, les mots-clés issus du texte, *i.e* les vocabulaires qui décrivent les *sujets*. Localisés à différents niveaux de la structure rédactionnelle, les mots-clés constitueront l'index des *sujets*. Ce sont des groupes nominaux (entités nommées, noms communs et leurs diverses expansions) qui apparaissent dans des contextes relevant d'un même sujet. Cette perspective onomasiologique ne fournit aucun indice linguistique et quantitatif sur le repérage de ces entités : c'est pourquoi l'indexation est généralement réalisée à la main au fil de la page. Ce qui nous paraît intéressant, c'est de se servir du vocabulaire pour favoriser un accès aux collections. Fournir des « mots » aux usagers leur permettrait d'interroger les collections en mode texte. Par conséquent, un dictionnaire figurant les entrées lemmatisées, auxquelles seraient associées les

familles de mots issues des collections, les parasynonymes, les diverses expansions du lemme nous paraît satisfaire cet objectif. Suivre un parcours de découverte consisterait donc à choisir un sujet parmi une liste proposée, afficher le dictionnaire correspondant à un sujet, cliquer sur une entrée de dictionnaire et lire les textes qui comportent ces entrées. Dans l'idéal, il faudrait afficher des rubriques et des illustrations légendées. L'affichage peut suivre une logique chronologique, celle du numéro du journal.

Pour valider cette approche, nous avons conduit une étude sur les sciences et les techniques dans le *Progrès Illustré* en raison de l'importance historique que représente ce sujet « pour comprendre l'origine d'une mutation globale de la société » (Gille, 1978 : 773). L'extraction du vocabulaire qui dénomme des objets scientifiques, des découvertes, et plus largement, des acteurs ou des institutions en lien avec les sciences et techniques a été réalisée avec Nooj, un outil de description linguistique de corpus[18]. Cette étude a permis de dégager des ensembles lexicaux par année et de les annoter.

Il est bien connu que cette démarche comporte des difficultés importantes dues notamment aux phénomènes de polysémie et d'homographie (Ehrlich, 1995). Toute entrée lexicale doit par conséquent être vérifiée en contexte à l'aide d'un concordancier. Un autre problème de taille liée à la perspective onomasiologique que nous avons retenue est d'apprécier la pertinence des classes lexicales retenues pour décrire les sciences et techniques. En effet, le vocabulaire ne présente pas les propriétés généralement dévolues au lexique scientifique et technique : ce n'est pas un sous-langage. Dans les causeries, il n'y a pas de termes de spécialité puisque cette presse n'est pas de la vulgarisation scientifique. Par conséquent, les critères d'appréciation ne peuvent pas s'appuyer sur des critères morpho-syntaxiques. En outre, on souhaite identifier des unités lexicales qui ont *un lien* avec les sciences et techniques, et qui ne sont pas des termes au sens strict. Par exemple, les inventeurs d'une technique, les chirurgiens célèbres, les instituts de science, etc. Pour valider nos classes, nous avons utilisé des connaissances extérieures (essentiellement des chronologies et des ouvrages sur l'histoire des sciences et des techniques) ainsi qu'une analyse en contexte, puis nous avons utilisé les grammaires d'états finis pour annoter le corpus. C'est le contexte qui nous indique que *ficelle* est la dénomination lyonnaise de *funiculaire*, v*apeur* celle de *tramway à vapeur* et une chronologie qui nous indique à quelles dates ces moyens de transport ont été inventés. Ce travail s'est révélé relativement fastidieux, même s'il aboutit *in fine* à une couverture exhaustive des 389 causeries océrisées : l'annotation, en revanche est très rapide grâce aux grammaires de Nooj.

### 4.2 Indexation par événements

La caractérisation des *événements* pose plusieurs problèmes qui ne pourront être résolus dans le cadre de ce travail. Nous devons considérer qu'un *événement* comporte plusieurs dimensions : linguistique, médiatique et historique. Les dimensions médiatique et historique ne sont pas repérables en langue : ce sont des sources extérieures qui permettent de leur attribuer un statut événementiel. Bien que l'indexation des événements soulève de nombreuses questions, la notion nous semble fondamentale pour caractériser la thématique, puisque c'est à travers les commentaires des événements que transparaît le positionnement éditorial du journal.

Le lien entre les *sujets* et les *événements* peut se faire dans le cadre d'une approche actantielle. Tel *sujet* peut figurer en position d'*actant* dans le cadre d'un procès qui

[18] http://www.nooj4nlp.net

décrit un *événement*. On se retrouve donc au niveau de la phrase, à l'intérieur des rubriques. Par conséquent repérer un *candidat événement* revient à croiser des fonctions dans une phrase et des catégories sémantiques telles que des toponymes, des entités nommées, des dates et des prédicats. Ces *candidats événements* peuvent être indexés à l'aide de deux sources extérieures : d'une part, en associant (ou pas) un *candidat événement* à une « famille événementielle », on valide ainsi la dimension médiatique de l'événement. On peut s'inspirer du standard IPTC pour définir les familles événementielles à l'aide de catégories plus adaptées à la presse du XIXème siècle. D'autre part, il s'agit de valider le *candidat événement* sur un plan historique. Tel assassinat peut, par exemple, relever d'une suite d'événements analysés comme relevant d'une seule et même affaire. Seul le regard critique de l'historien est en mesure de relier des événements épars et de lui donner sens. Cette approche consiste à documenter le corpus par des sources externes, ce qui relève d'une sémantique du lien.

La validation de la proposition a donné des résultats intéressants, en tout cas, pour la connaissance du corpus. Nous avons choisi parmi les *Causeries* qui traitent des sciences et techniques, celles qui évoquent un événement. En nous appuyant sur les repères linguistiques mentionnés, on recueille toutes sortes de récits anecdotiques, qui n'ont sans doute pas la prétention médiatique d'un événement. Jugeons-en : a) *un cas de transfusion sanguine … au sang de chèvre* (1891/02/22, n°10) ; b) *la découverte scientifique d'un physiologiste lyonnais … sur des microbes lumineux* (1891/07/12, n°30), c) *un astronome téméraire … qui s'écrase en aérostat sur une cheminée* (1893/11/19, n°153) d) *un procédé de vieillissement du vin par l'électricité … raconté comme une recette de cuisine (1891/12/13 n° 52)* On observe que dans ces événements, les repères temporels sont souvent très flous : *il y a quelques jours, la semaine dernière, je vous disais l'autre fois…* La structure de l'événement se rapproche plutôt du récit, pour ne pas dire parfois, de la fable.

### 4.3 Indexation par thèmes

Viennent enfin les thèmes que nous proposons d'appréhender comme les traces d'un positionnement éditorial sur l'actualité. Ces traces sont une des manifestations de l'*angle* journalistique, au sens où l'angle peut être appréhendé en tant que « formant » (Ringoot, 2004 : 108). Dans une étude récente, nous avons travaillé la notion de positionnement dans le cadre de la lecture et de l'annotation de documents scientifiques (Clavier et Paganelli, 2009). Nous avons observé que lorsque des lecteurs consultent une thèse, ils sont beaucoup plus attentifs au positionnement scientifique de l'auteur de la thèse qu'à la terminologie. Le positionnement scientifique – qui n'est pas une notion linguistique – a été défini par un ensemble de traces linguistiques qui révèlent une posture surplombante de l'énonciateur et qui se situent dans le métadiscours. Indexer ces marqueurs, révéler le contexte métadiscursif pour favoriser la compréhension et l'appropriation des connaissances, est la conclusion à laquelle nous sommes parvenues.

Nous formulons l'hypothèse qu'il existe également des marqueurs de positionnement dans la presse, et que la mise en évidence de ces traces permettrait une lecture-découverte située de l'actualité. Si l'on considère avec d'autres auteurs[19] que la ligne éditoriale peut être analysée sous l'angle de l'énonciation, cette approche permet d'exploiter des « traces repérables » qui sont à l'articulation de « l'énonciation textuelle » et de « l'énonciation éditoriale » (Ringoot, 2004 : 92). Si le positionnement permet de situer le point de vue éditorial, comment organiser les marqueurs en thématiques ? Dans les *Causeries*, les commentaires des

[19] Voir (Ringoot, 2004 : 92) qui cite divers auteurs dont Annelise Touboul.

chroniqueurs sur les événements sont des lieux privilégiés de l'expression individuelle et éditoriale. Les marqueurs de positionnement peuvent apparaître à plusieurs niveaux : lexical, phraséologique, dans les proverbes et les dictons. Par exemple, le caractère régional de la presse peut transparaître à travers une énonciation fortement ancrée dans des repères territoriaux : personnalités lyonnaises, lieux et institutions de Lyon et de ses environs, événements qui ponctuent la vie locale. Ici, c'est le lexique qui est mobilisé, ce qui transparaît dans les listes de fréquence. La thématique n'en est pas pour autant « régionaliste », sinon, on confond les *sujets* avec les *thèmes*. Ce sont les réseaux de cooccurrences (ou d'association) qui vont permettre de faire émerger une interprétation. Par exemple, Lyon rivalise toujours avec Paris, la thématique pourra être le chauvinisme. Les dénominations des thématiques nous importent moins que la mise en évidence du contexte d'interprétation.
Les marqueurs sont par ailleurs révélateurs de postures énonciatives. Par exemple, l'une des façons de révéler l'approche populaire du journal consiste à montrer que la voix du chroniqueur, dans l'analyse qu'il fait de l'actualité, se fond systématiquement dans celle du plus grand nombre, la *vox populi*. Elle se manifeste le plus souvent par l'ironie : les scientifiques sont de grands hommes, mais leurs découvertes sont bien piètres en regard des grands fléaux de l'humanité. Ainsi la thématique de la dérision peut-elle s'appliquer aux découvertes scientifiques et techniques, et par extension à leurs auteurs. Contrairement au positionnement scientifique, qui s'inscrit dans une dynamique de sur-énonciation, le positionnement éditorial de cette presse illustrée relèverait de la sous-énonciation.
Si la thématique ne peut être installée dans le code, il faut pourvoir circonscrire le contexte, i.e. délimiter les commentaires Dans les *Causeries*, les commentaires sont imbriqués dans le récit des événements, à la manière d'une conversation, mais sont repérables par le jeu des repères énonciatifs. Formellement, l'on passe du *il* (dans l'événement) au *on* doxique, au *je* (dans les commentaires), de l'assertion déclarative aux interrogatives et aux exclamatives, du mode indicatif au mode impératif. L'introduction de guillemets ne renvoie pas à des citations comme dans la presse contemporaine, mais à des commentaires qui sont des aphorismes, des proverbes, des dictons. La présence de traces repérables permet l'annotation des commentaires.

## 5 Conclusion

La création de parcours thématiques constitue l'un des moyens pour diversifier les modes d'accès aux collections, tout en répondant à des objectifs de valorisation et de mise en exposition du patrimoine de presse numérisée. De ce point de vue, les parcours et les dossiers thématiques partagent les mêmes objectifs. A la manière d'un musée qui expose ses collections, les parcours et les dossiers présentent une sélection d'*objets* textuels et iconiques, agencés suivant une certaine logique qui leur donne sens. Cette entrée dans les collections par la voie muséographique est censée favoriser l'accès du grand public à ces ressources, plus que ne le ferait un mode d'accès par affichage des numéros de journaux sous forme de calendrier par exemple. La mise en œuvre des dossiers et des parcours est cependant différente. Les dossiers sont conçus manuellement par des professionnels, ce qui offre une qualité incontestable mais se révèle long et fortement dépendant des connaissances des indexeurs. Inversement, l'idée qui préside à la construction de parcours thématiques est de proposer une méthodologie reproductible et systématique, destinée à sélectionner des *objets*, à les délimiter et à les représenter. Ces différentes

étapes sont envisagées dans le cadre d'une indexation thématique des collections de journaux.

L'indexation thématique a été abordée dans une perspective textuelle et cherche à résoudre la question de l'impossible adéquation entre une chaîne de caractères et un thème, situation qui condamne toute tentative de recours à des langages documentaires contrôlés pour décrire la thématique. Nous avons travaillé la notion de thème dans la presse en mobilisant trois notions : les *sujets*, les *événements* et les *thèmes*. Les *sujets* sont choisis dans le cadre de classifications qui dénomment des classes de connaissances. Les vocabulaires sont issus des collections et alimentent les classes ; leur localisation tient compte de la morphologie du journal. Les vocabulaires sont lemmatisés et organisés à la manière d'un dictionnaire regroupant les parasynonymes et les diverses expansions. Les *événements* sont des entités complexes modelées par un double processus synchronique (la médiatisation d'un événement) et diachronique (sa trace dans l'histoire). Les événements peuvent être regroupés en familles (les catastrophes, les assassinats, etc.). Nous faisons l'hypothèse, qui reste à valider, qu'un événement connaît un ancrage langagier et qu'on peut le ramener à certaines catégories sémantiques figurant dans un cadre actantiel. Les *sujets* instancient des places actantielles. Les *thèmes* sont construits à partir des traces linguistiques qui révèlent le positionnement éditorial sur un événement. Les traces de ce positionnement sont repérables dans les commentaires que font les chroniqueurs de l'actualité. Les commentaires sont donc vus comme le lieu d'expression de la subjectivité du chroniqueur et le lieu de manifestation de l'*angle* du journal. Ces deux énonciations, individuelle et éditoriale donnent à voir la thématique. Indexer signifie alors associer des types de commentaires à des familles d'événements.

La mise en œuvre technique des parcours a été abordée à la marge. Certes, nous avons utilisé un outil linguistique qui offre diverses fonctionnalités utiles à l'indexation : étiquetage, annotation, analyse en contexte. Mais nous n'avons pas exploité le dictionnaire de mots inconnus… bien fournis en noms propres, et regorgeant de mots erronés qui entraveraient toute analyse syntaxique. Du côté des méthodes de classification supervisée qui pourraient permettre de classer les *Causeries* par sujets, il nous semble que le principal problème réside dans la prise en compte des fréquences d'occurrences. En effet, les vocabulaires qui alimentent les classes, se situent dans les basses fréquences et ne peuvent donc constituer un critère de classement. Enfin, reste la question de l'identification des commentaires, qui, bien que présentant formellement des marqueurs, nécessite également une réflexion sur le typage de ces unités textuelles.

Si l'indexation de parcours thématiques dans les collections de presse soulève des problèmes de faisabilité et d'automatisation, elle replace néanmoins la question de l'accès au patrimoine numérique dans une perspective d'interprétation et de contextualisation de l'information.

## 6 Bibliographie

7 Adam J.-M. et Lugrin G., « L'hyperstructure : un mode priviliégié de présentation des événements scientifiques » in Cusin-Berche F. *Rencontres discursives entre science et politique. Spécificités linguistiques et constructions sémiotiques*, *Carnets du CEDISCOR, 6*, Presse de la Sorbonne Nouvelle, 2000, p. 133-149.

8 Adam J.-M., « Unités rédactionnelles et genres discursifs : cadre général pour une approche de la presse écrite », *Pratiques*, *94*, 1997, p. 3-18.

9 Amar M., « L'indexation aujourd'hui », *Les dossiers de l'ingénierie éducative*, *vol. 49*, 2004, p. 61-65.

10 Arquembourg J., « De l'événement international à l'événement global : émergence et manifestations d'une sensibilité mondiale », *Evénements mondiaux regards nationaux, Hermès 46*, CNRS Editions, 2006, p 13-21.

11 Bacot J.-P., La presse illustrée au XIXe siècle : une histoire oubliée, Médiatextes, Limoges : Pulim, 2005. 235 p.

12 Bellanger C., *Histoire générale de la presse française*. T.3. De 1871 à 1940. Paris, Presses universitaires de France, 1972, 687 p.

13 Bilhaut F.**,** *Analyse automatique de structures thématiques discursives - Application à la recherche d'information*, mémoire de thèse de doctorat en informatique, Université de Caen, 2006.

14 Bouvier-Ajam L.*, Europeana. Etude sur les usages et les attentes relatifs à l'interface de consultation de la future Bibliothèque numérique Européenne*, Rapport final, 2007, 53 p.

15 http://www.bnf.fr/documents/ourouk.pdf

16 Clavier V. Paganelli C., « Marqueurs de positionnement et parcours de lecture : un enjeu pour la consultation des thèses en ligne ? » Actes du Colloque *Changements technologiques, mutations organisationnelles et information professionnelle : pratiques, acteurs et documents,* organisé par le GRESEC, les 10-11 décembre 2009 à Echirolles.

17 Da Sylva, L., « Outil de butinage du contenu des documents de collections numériques », *Patrimoine 3.0 : actes du douzième Colloque international sur le document électronique*, 21-23 octobre 2009, Université de Montréal, Canada, sous la direction de Khaldoun Zreik, Paris : Europia productions, p. 263-273.

18 Desprès-Lonnet M., « L'écriture numérique du patrimoine, de l'inventaire à l'exposition : les parcours de la base Joconde », *Culture & musées, 14,* 2010, p. 19-38.

19 Ehrlich D., « Une méthode d'analyse thématique. L'exemple de l'ennui et de l'ambition », *in* Rastier F.(sld), 1995, p. 85-103.

20 Ferenczi T., *L'invention du journalisme en France : naissance de la presse moderne à la fin du XIX*e *siècle*. Paris, Payot, 1996. 275 p.

21 Gille B., « Les techniques de l'époque moderne », in *Histoire des techniques : technique et civilisations, technique et sciences,* Paris : Gallimard, 1978, pp 773-858. (Encyclopédie de la Pléiade, 41).

22 Herman T. et Lugrin G., « La hiérarchie des rubriques : un outil de description de la presse », *Communication et Langages, 122,* 1999, p. 72-85.

23 Laitang C., Egyed-Zsigmond E., Calabretto S., « Diversité de l'Information dans les Sites de Presse », *Patrimoine 3.0 : actes du douzième Colloque international sur le document électronique*, 21-23 octobre 2009, Université de Montréal, Canada, sous la direction de Khaldoun Zreik, Paris : Europia productions, 2009, p. 111-128.

24 Landron, P.Y., « Valoriser la presse illustrée du XIXème : l'exemple de la BM Lyon», Communication aux Journées d'études *Regards croisés sur la mise en ligne et la valorisation de la presse XIX-XXI*, co-organisées par les laboratoires ELICO (Lyon), GRESEC (Grenoble) et la Bibliothèque Municipale de Lyon dans le cadre du Cluster 13 « Culture, patrimoine et création » les 6 & 7 mai 2010 à la Bibliothèque Municipale de Lyon.

25 Lesquins N., *Europeana : rapport de bilan sur les usages et les attentes des utilisateurs*, Bibliothèque Nationale de France, 2007, 60 p.

26 http://www.bnf.fr/documents/europeana_2007.pdf

27 Martin E., « Thème d'étude, étude de thème » in RASTIER F. (sld.), 1995, p. 13-24.

28 http://www.revue-texto.net/Parutions/Analyse-thematique/Martin.pdf

29 Matharan J., Chaguiboff J., Alliot F., *Rapport d'étude sur les usages communautaires et collaboratifs, sur place et à distance, des ressources numérisées de la BnF,* Bibliothèque Nationale de France, 2008.

30 http://www.bnf.fr/documents/rapport_web_communaute.pdf

31 Mezzasalma P., « Conserver la presse », Dossier *La conservation et la numérisation de la presse*, in *Chroniques de la BNF*, n° 47, 2009, p. 5-7.

32 Paganelli C. et Mounier E., « La presse ancienne numérisée : modes d'accès et pratiques de recherche », Communication aux Journées d'Etudes *Regards croisés sur la mise en ligne et la valorisation de la presse XIX-XXI,* co-organisées par les laboratoires ELICO (Lyon), GRESEC (Grenoble) et la Bibliothèque Municipale de Lyon dans le cadre du Cluster 13 « Culture, patrimoine et création » les 6 & 7 mai 2010 à la Bibliothèque Municipale de Lyon.

33 Palmer M., « Nommer les nouvelles du monde », *Evénements mondiaux regards nationaux, Hermès 46,* 2006, p. 47-56.

34 Porhiel S., « Les marqueurs de thématisation : des thèmes phrastiques et textuels », *Travaux de linguistique, 2/51*, 2005, p. 59-88.

35 Rastier F. «La sémantique des thèmes - ou le voyage sentimental», *Texto !* 1999.

36 http://www.revue-texto.net/index.php?id=570.

37 Rastier F., « La sémantique des textes : concepts et applications », *Hermès*, *16,* 1996, p. 15-37.

38 http://www.revue-texto.net/Inedits/Rastier/Rastier_Concepts.html#A.

39 Rastier F. (sld.), *L'analyse thématique des données textuelles, l'exemple des sentiments*, Paris : Didier Erudition, 1995.

40 Ringoot R. et Robert-Demontrond P., *L'analyse de discours*, Editions Apogée, 2004, 222 p.

41 Smolczewska-Tona, A. et Lallich-Boidin, G., « De l'édition traditionnelle à l'édition numérique : le cas de la presse du XIXe siècle. » In Broudoux E., Chartron G. (dir.). *Traitements et pratiques documentaires : vers un changement de*

*paradigme ?* Actes de la deuxième conférence Document numérique et société, Paris : ADBS éditions, 2008, p. 299-316.

42 Teissier M. *La numérisation du patrimoine écrit*, Janvier 2010, *La documentation française*, http://www.ladocumentationfrancaise.fr/rapports-publics/104000016/index.shtml

43 Têtu J.-F., *Le journalisme mis en scène. In*: La presse selon le XIXe siècle. Université Paris III, Paris, France, 1997, pp. 137-154.

44 http://halshs.archives-ouvertes.fr/docs/00/39/73/90/HTML/

*45* Trudel E., «Champ sémantique, champ sémantique lexical ou classe sémantique ?», *Texto!* 2009

46 http://www.revue-texto.net/index.php?id=2277.

47 Westeel I., « Le patrimoine passe au numérique », *Bulletin des Bibliothèques de France*, *1,* 2009, p. 28-35.

*48* Westeel Isabelle, « Patrimoine et numérisation : la mise en contexte du document » [en ligne], in *Colloque EBSI/ENSSIB. Montréal. 13-15 octobre 2004.*

49 http://www.ebsi.umontreal.ca/rech/ebsi-enssib/pdf/westeel.pdf

50 Zweigenbaum P. et Habert B. « Accès mesurés aux sens », *Mots. Les langages du politique*, *74,* 2004*,* p. 93-106.

51 http://mots.revues.org/index4673.html

# Accès multilingue en ligne aux manuscrits arabes numérisés

**Mohammed Ourabah SOUALAH (1), Mohamed HASSOUN (2)**

med_soualah@yahoo.fr / mohamed.hassoun@ennsib.fr

(1) Université Lumière – Lyon 2, ENSSIB – ELICO

(2) ENSSIB – ELICO

**Résumé.** Les manuscrits arabes constituent un trésor universel pour l'humanité. Ils se retrouvent constamment menacés par un effritement à cause de la précarité et du manque de moyens dans les lieux de conservation. La manipulation de ces œuvres constitue un danger supplémentaire de détérioration. Par conséquent, la numérisation constitue une solution idéale qui permettra à la fois, de préserver les manuscrits dans de meilleurs états et de donner la possibilité d'accès distant aux chercheurs. Cette solution est réalisable grâce à la mise en ligne des manuscrits numérisés. Par conséquent, il est impératif de mettre en place un système d'accès adéquat à ces ressources. Ainsi, le catalogage se propose comme une solution salutaire. En effet, constitué d'un ensemble de notices descriptives de manuscrits, le catalogue permet d'une part, d'accéder librement au manuscrit et d'autre part, d'y accéder par vedette matière. A ce niveau se pose le problème d'inexistence de norme de catalogage des manuscrits arabes anciens. Il s'agit de rechercher un ensemble de métadonnées pouvant les décrire d'une manière efficiente. Notre travail, se veut d'apporter une solution pragmatique aux divers problèmes posés par l'accès aux manuscrits arabes numérisés. A travers ce dernier, nous proposons l'adaptation de la TEI P5 Manuscript Description pour le catalogage des manuscrits arabes. Ainsi, des métadonnées spécifiques aux manuscrits arabes ont été ajoutées à ce standard. Par ailleurs, notre solution propose un accès multilingue au catalogue en utilisant l'arabe, le français et l'anglais. Cet aspect peut être élargi à d'autres langues. De ce fait, la requête peut être émise dans n'importe quelle langue et le résultat affiché, sous forme de notice, dans une langue cible. L'accès aux images des manuscrits numérisés se fait grâce à des hyperliens appropriés.

**Mots-clés.** numérisation, manuscrits arabes, accès multilingue en ligne, métadonnées, catalogage.

## 1 Introduction

La numérisation des manuscrits arabes ne constitue pas une fin en soi. En effet, tout projet de numérisation de manuscrits repose sur deux objectifs principaux :

- La préservation des manuscrits, en leur changeant de support de conservation, ce qui limite l'accès aux manuscrits originaux.
- La mise en ligne des manuscrits afin de les rendre accessibles à la communauté scientifique, ce qui valorisera davantage les manuscrits.

Le premier objectif repose sur une procédure classique de numérisation des manuscrits, qui se résume à quatre étapes à savoir, la sélection des manuscrits à numériser, la dématérialisation des manuscrits, la vérification des documents numérisés et enfin, la sauvegarde des manuscrits numérisés.
Le second objectif repose sur le principe de mise en place d'un outil de gestion de la base de données des images des manuscrits numérisés.
Par conséquent, nous proposons, dans cet article, une solution pragmatique basée sur le multilinguisme dont la finalité est la mise en ligne des manuscrits arabes numérisés.
L'accès aux manuscrits numérisés nécessite la mise en place d'un instrument de recherche adéquat. Ce dernier permet l'accès aux manuscrits selon divers critères de recherche (par mot clé et par vedette matière : auteur, titre, sujet,…etc.). Ainsi, apparaît la nécessité d'une indexation efficace et représentative, qui devra décrire d'une manière univoque un manuscrit. Si la recherche par le contenu dans les images des manuscrits numérisés constitue un challenge, la mise en place d'un catalogue est une solution palliative très intéressante.
Le catalogue est constitué d'un ensemble de notices bibliographiques, lesquelles sont formées à leur tour, d'éléments descriptifs des manuscrits. Une notice permet d'accéder à un et un seul manuscrit.
Par conséquent, le catalogue permet d'une part, de décrire le manuscrit et d'autre part, de le référencer. Mais, encore une fois, un problème de taille se dresse devant cette solution : Il s'agit de l'inexistence d'une norme de catalogage pour ce type de fonds documentaire.
Les spécialistes du domaine proposent pour le catalogage d'un manuscrit sa description selon trois aspects distincts :

1. Description codicologique : Description matérielle (codex, reliure, …etc.).

2. Description paléographique : Analyse du contenu (Style d'écriture, Auteur, sujet, …etc.).

3. Histoire du manuscrit.

Dans la littérature peu de travaux ont porté sur la définition d'une norme de catalogage des manuscrits arabes. Nous avons relevé quelques tentatives ayant porté sur l'établissement d'un protocole de description des manuscrits arabes, mais sans aboutir à un consensus de normalisation. Nous citons à titre d'exemple, le protocole mis en œuvre par l'IRHT (Institut de Recherche en Histoire et Texte) et celui mis en œuvre par la fondation du prince "El Saoud", sise au Maroc, sans pour autant qu'il y ait un quelconque rapprochement entre les deux institutions. De ce fait, la détermination des métadonnées descriptives des manuscrits arabes, définissant le catalogue, se pose comme un véritable problème qui nécessite un travail d'investigation conséquent. Il s'agit donc, d'une part, d'apporter une contribution dans le domaine de catalogage des manuscrits arabes et d'autre part, de trouver un moyen efficace d'accès à ces ressources numérisées. Par ailleurs, le besoin de rendre l'accessibilité multilingue ajoute d'un cran la difficulté de mise en œuvre de la solution.

Notre travail s'inscrit dans une perspective de pérennité et de portabilité. Ainsi, le format XML se propose comme outil incontournable pour l'encodage du catalogue des manuscrits arabes anciens. Il faudrait donc, exploiter ce format pour proposer des moyens d'accès fiables et efficaces aux images des manuscrits.
A travers notre article nous allons apporter une réponse à la problématique de catalogage et d'accès aux images des manuscrits arabes numérisés. Dans une première partie nous donnerons un aperçu sur certains travaux déjà mis en œuvre, la seconde partie sera réservée à la description de la problématique du catalogage et de la solution apportée, puis dans la troisième partie nous exposerons les différents outils de mise en œuvre du catalogue multilingue et enfin, dans la quatrième partie nous décrirons la solution globale puis enfin, la cinquième partie sera réservée à la description des différents modes d'accès multilingues aux images des manuscrits arabes anciens numérisés.

## 2 Quelques expériences de mise en ligne des manuscrits arabes numérisés

Les quelques différents projets de mise en ligne de manuscrits arabes numérisés disponibles sur la toile utilisent des outils d'encodage différents les uns des autres.
En effet, la BNF (Bibliothèque Nationale de France), pionnière dans la numérisation et la mise en ligne des manuscrits arabes (disponible sur http://archivesetmanuscrits.bnf.fr), utilise l'EAD (Encoded Archives Description) [11] comme outil d'encodage. Le système propose l'accès aux manuscrits par auteur, par sujet et par titre. Par contre, aucune recherche libre n'est disponible, si bien que l'EAD demeure inadaptée dans sa forme actuelle pour la description des manuscrits arabes. Un des problèmes rencontrés par les encodeurs est l'inexistence d'éléments spécifiques dans l'EAD pour certains aspects descriptifs des manuscrits arabes, exemple l'Incipit [8], ainsi, une mise à jour de l'EAD s'avère nécessaire.
D'autres expériences méritent d'être citées, à l'instar de la Bibliotheca Alexandrina, qui a numérisé et a mis en ligne six catalogues de manuscrits, soient 4071 manuscrits (disponibles sur http://ziedan.com ou sur www.manuscriptcenter.org). Notons par ailleurs que les notices sont décrites d'une manière très sommaire. Elles renferment quatre parties : l'auteur, l'implicit (Awâluhu), l'explecit (Âkhiru-hu) et l'état général du manuscrit (Âl-nuskha); ce qui témoigne de l'insuffisance des métadonnées descriptives utilisées dans le projet.
D'autres projets de catalogage des manuscrits arabes se sont basés sur le format UNIMARC à l'instar de la Bibliothèque Nationale de Tunis (BNT) [16]. Le modèle s'est limité à deux niveaux dans la description des manuscrits, ce qui limite considérablement l'aisance du catalogueur. De plus, la fixation de la taille des différents champs rend le format MARC inadapté pour le catalogage de ce type de document. Ainsi, certaines institutions ont vite compris les limites du format MARC telle que la "Wellcome Library" de Londres, qui après l'utilisation de MARC21 a décidé d'opter pour les métadonnées du format TEI-MASTER pour la mise en ligne de collection "Haddād Manuscripts" (disponible en ligne à l'adresse URI : http://library.wellcome.ac.uk/node273.html).

## 3 Catalogage des manuscrits arabes

Un document non catalogué est un document inaccessible. Il peut être considéré comme un document mort. Le catalogage permet de repérer la disponibilité d'un document pour un lecteur [10].

Deux aspects de catalogage de manuscrits se confrontent. Le premier considère le manuscrit comme une pièce unique et isolée [2], alors que le second considère que le manuscrit est indissociable du fond d'archive [7]. Dans ce cas, il est plus important de rendre compte de la structure de l'ensemble plutôt que de décrire précisément chaque pièce.
Un fonds de manuscrit pose un réel problème de classification à cause de son hétérogénéité. En effet, le fonds est généralement, formé par une grande diversité de nature des documents, ce qui rend complexe la rédaction d'un catalogue. Ainsi, la description d'un manuscrit comme pièce isolée semble plus approprié, afin de pallier au problème de classification du fonds. Par conséquent un manuscrit sera décrit comme une pièce isolée, dissociée du fonds.

### 3.1 Les métadonnées descriptives des manuscrits arabes

L'accès à une ressource numérisée se fait selon un critère d'accès spécifique. En effet, un lecteur matérialise sa recherche selon un besoin particulier. Il peut s'agir d'un titre d'une œuvre, de son auteur ou de toute autre caractéristique relative à l'œuvre : on parle de métadonnées. Il s'agit au fait, d'un ensemble structuré de données décrivant une ressource quelconque [13]. Les métadonnées permettent la gestion et l'accessibilité à la ressource décrite.
Ainsi, une interrogation importante mérite d'être soulevée : Quelles métadonnées faudrait-il utiliser pour la description des manuscrits arabes ? Plusieurs auteurs ont tenté d'apporter une réponse à cette question, mais jusqu'à présent, aucun consensus n'a été établi. De ce fait, il n'existe aucune norme de description des manuscrits arabes. Par conséquent le domaine de catalogage des manuscrits arabes reste ouvert.

### 3.2 Méthode de catalogage des manuscrits arabes numérisés

Notons qu'un manuscrit peut être constitué par un ou plusieurs volumes, traitant le même sujet ou des sujets différents. Ce qui met en évidence l'hétérogénéité d'un manuscrit et la difficulté de mise en place d'une norme de catalogage spécifique.
Par ailleurs, deux méthodes de catalogages restent possibles : Il s'agit du catalogage par volume et du catalogage à l'exemplaire.

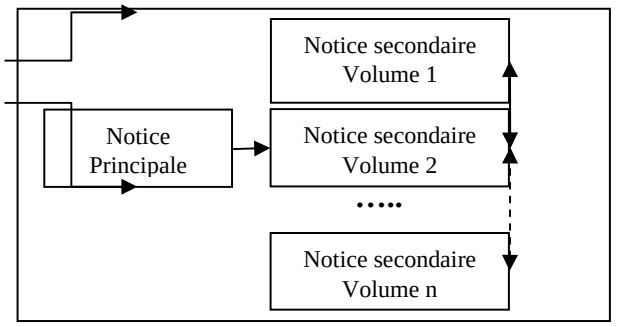


**Figure 1.** Structure des notices : Catalogage par volume

***Catalogage par volume***

Cette méthode consiste à associer une notice bibliographique à chaque volume d'un manuscrit. Un manuscrit formé de plusieurs volumes se verra tributaire de plusieurs notices, qu'il faudrait relier entre elles. De ce fait, la mise en place d'une notice principale est plus qu'indispensable. La notice principale contiendra les informations communes aux différents volumes telles que, les informations de signalement du manuscrit. Cette méthode reste complexe à mettre en œuvre, au risque de

répétitivité d'informations. La figure "Figure 1" illustre le principe de catalogage par volume

***Catalogage à l'exemplaire***

Cette seconde méthode respecte la structure physique du manuscrit. Elle considère le manuscrit dans son intégralité quelque soit le nombre de volumes. Elle est plus souple et plus adaptée au catalogueur, qui considère le manuscrit dans son intégralité quelque soit le nombre de volumes. Cela permet d'associer à un manuscrit une seule notice descriptive. Ainsi, la description globale du manuscrit et celle du premier volume sont faites en tête de notice, alors que celles des autres volumes seront faites dans la partie secondaire de la notice.

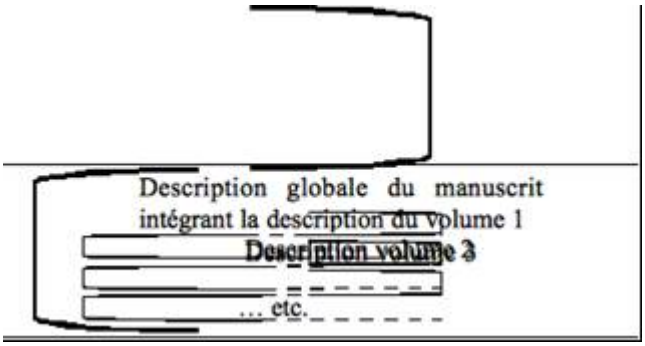


**Figure 2.** *Structure de la notice : Catalogage à l'exemplaire.*

L'unicité du signalement du manuscrit permet d'accéder directement aux informations relatives aux différents volumes en considérant les zones secondaires de la notice. La "Figure 2" illustre le principe de catalogage à l'exemplaire.

## 4 Encodage de la description des manuscrits arabes numérisés

La solution de numérisation des manuscrits arabes est guidée par l'objectif de pérennité, d'interchangeabilité et de portabilité des données. Le format XML se propose en format incontestable. Par ailleurs, le choix de balises et de structure du fichier doit être minutieusement étudié.

Plusieurs formats de description de manuscrits se portent candidats. Nous citons entre autre l'EAD (Encoding Archival Documents), l'EAMMS (Electronic Access to Medieval Manuscript) et le modèle MASTER (Manuscript Access through Standards for Electronic Records).

Il n'est pas dans notre objectif de mener une étude comparative entre ces différents formats, mais affirmons tout simplement qu'après une étude approfondie, notre choix s'est implicitement porté sur la TEI Manuscript Description dans sa version P5 encore notée TEI-ms.

### 4.1 Pourquoi la TEI-ms comme format d'encodage ?

La TEI (Text Encoding Initiative) se propose comme un ensemble de recommandations. Elle se présente sous une forme souple et adaptable. En effet, la structure modulaire de la TEI permet à l'usager de choisir les outils qui lui conviennent.

De plus, la TEI-ms est issue de l'application de la TEI à la description des manuscrits médiévaux. Elle renferme des éléments descriptifs de manuscrits issus

directement de la technique de catalogage de manuscrits à l'exemplaire. Elle présente une structure hiérarchique (arborescente) ouverte pouvant s'adapter à toute évolution des méthodes de description de manuscrits. Par conséquent, la TEI-ms se présente comme un standard idéal qui décrit d'une manière élaborée les manuscrits.

### 4.2 Structure de la TEI P5 Manuscript Description (TEI-ms)

La TEI Manuscript Description dans sa version P5 est née après avoir porté les corrections aux erreurs parues dans la version P4 [3] et a à la fois intégré en son sein les métadonnées MASTER et EAMMS.

La description du manuscrit se fait grâce à l'élément <msDesc> sous élément de <sourceDesc>, élément de la TEI. <msDesc> décrit un seul manuscrit à la fois. Il est formé de sept éléments dont la description sommaire est donnée comme suit :

[1]**<msIdentifier>** : Il définit les informations nécessaires afin d'identifier un manuscrit.

[2]**<head>** : Il s'agit de l'en-tête. Il contient un type d'information le définissant, tel que le titre de la section, le glossaire,…etc.

[3]**<msContents>** : Décrit le contenu intellectuel du manuscrit ou d'une partie du manuscrit.

[4]**<physDesc>** : Il permet la description physique du manuscrit ou d'une partie du manuscrit subdivisé en éléments bien structurés.

[5]**<history>** : Il regroupe les éléments décrivant l'historique du manuscrit ou d'une partie du manuscrit.

[6]**<additional>** : Il regroupe les informations additionnelles telles que la bibliographie du manuscrit, les informations administratives, la disponibilité du manuscrit sur microfilm,…etc.

[7]**<msPart>** : Elément décrivant les manuscrits d'origine rassemblés en un seul manuscrit.

Chaque élément définit ci-dessus est formé d'un ensemble de sous-éléments, qui à leur tour peuvent être constitués d'autres éléments. L'ensemble donnant à la TEI-ms une structure hiérarchique en forme arborescente.

### 4.3 Adaptation de la TEI-ms pour la description des manuscrits arabes

L'objectif assigné à la TEI-ms est la description des manuscrits médiévaux, elle est donc valide pour la description de ces derniers. Notre démarche consiste à apporter un complément informationnel et structurel afin de rendre la TEI-ms valide pour la description des manuscrits arabes anciens.

En effet, dans sa version actuelle, la TEI-ms ne prend pas en compte un certain nombre d'éléments spécifiques à la description des manuscrits arabes. N'étant pas l'objectif de cet article, nous n'allons pas détailler l'enrichissement apporté à la TEI-ms afin de l'adapter à la description des manuscrits arabes anciens. Toutefois, nous citons, dans ce qui suit, les apports les plus significatifs [9].

#### *Nécessité de translitération*

La translittération consiste à transcrire une langue dans un alphabet étranger à la langue. Ainsi, afin de rendre le manuscrit arabe accessible à des usagers qui ne connaissent pas la calligraphie arabe, nous proposons d'introduire cet aspect dans certains éléments de la TEI-ms, tels que les noms de personnes et de lieux.

#### *Prise en compte des deux modes de catalogage*

La TEI-ms procède au catalogage des manuscrits selon le mode de catalogage à l'exemplaire. Par ailleurs, la majorité des fonds de manuscrits arabes étant catalogué

selon le mode de catalogage par volume. L'informatisation de ce fonds, en utilisant la TEI-ms, impose l'adaptation de cette dernière au mode de catalogage par volume.

***Le nom de l'auteur ou du copiste***

Les noms arabes présentent une structure particulière non prise en compte par la TEI-ms. En effet, l'élément <name> présentent en général deux sous-éléments <surname> et <forename>, qui peuvent amplement décrire les noms des auteurs occidentaux, mais restent incomplets pour décrire les noms des auteurs arabes. Ces derniers étant fragmentés en quatre parties, le "*ism*[1]", le "*laqab*[2]", la "*kunya*[3]" et la "*nisba*[4]". D'où la nécessité d'intégration dans la TEI-ms des caractéristiques des noms arabes.

## 5 Description de la solution de mise en ligne des manuscrits arabes numérisés

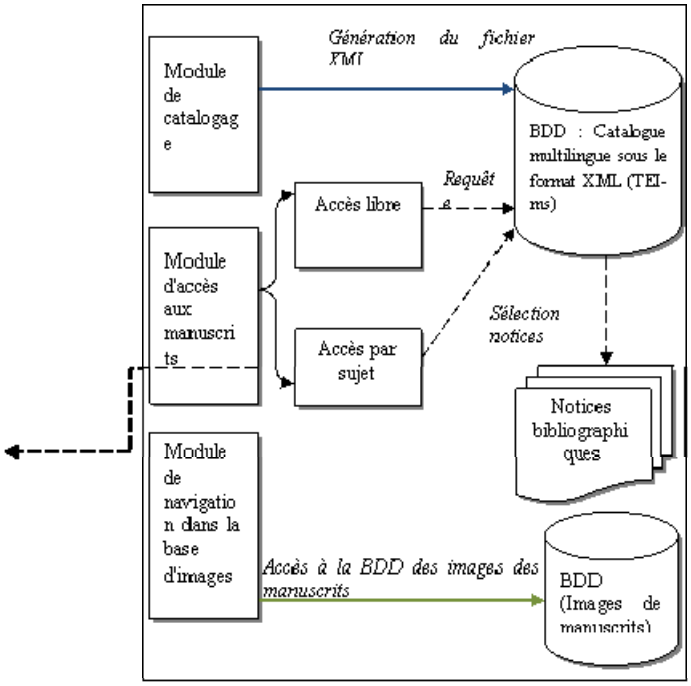


Figure 3 : Structure générale de la solution d'accès en ligne aux manuscrits arabes numérisés.

La solution mise en œuvre est représentée dans la figure "Figure 3". Elle est composée de trois modules :

1 **Module de catalogage** : Il s'agit d'une interface interactive, constituée de l'ensemble des champs décrivant le manuscrit, mise à la disposition du

[1] Nom donné dans l'ordre direct.

[2] Mot ou une expression appliquée à un personnage éminent pour évoquer une qualité réelle ou attribuée.

[3] Marque de distinction appliquée à des personnages de premier plan pour les honorer.

[4] Adjectif formé à l'aide du suffixe *ī* afin d'indiquer l'origine, le lieu de naissance ou de résidence d'une personne.

catalogueur. L'importance de cette interface réside dans l'aide qu'elle apporte au catalogueur, en lui évitant l'apprentissage du formalisme TEI P5 Manuscript Description. En sortie, ce module génère un fichier XML respectant la DTD TEI-ms. L'ensemble des fichiers XML générés constitue la base de données du catalogue.

2 **Module d'accès aux manuscrits** : Ce module met à la disposition de l'utilisateur deux modes d'interrogation du catalogue. Un système de recherche libre (mots clés) et un système de recherche par vedette matière (sujet). En sortie, ce module affiche l'ensemble des notices qui satisfont la requête de l'utilisateur. Les notices sélectionnées permettront d'accéder aux images des manuscrits arabes numérisés.

3 **Module de navigation dans les images** : Ce module permet de décrire la base des images des manuscrits numérisés. La navigation peut se faire par un simple défilement des images ou l'affichage de la liste des images sous forme de miniatures, ce qui permet d'accéder directement à une page donnée du manuscrit.

### 5.1 Génération du catalogue multilingue des manuscrits arabes

L'accès aux images des manuscrits numérisés repose sur le catalogue. Ce dernier, étant sous le format XML conforme à la DTD TEI-ms, est constitué d'un ensemble de notices bibliographiques.

Chaque notice bibliographique fait référence à une ou plusieurs images du manuscrit arabe, préalablement numérisé. A présent, nous allons décrire la procédure de génération du catalogue :

1. Le catalogueur n'est pas tenu à connaître la TEI-ms, mais devrait avoir des connaissances sur le principe de catalogage des manuscrits. Pour ce faire, nous mettons à sa disposition une interface multilingue. Le choix de la langue de travail est à la fois intégré dans le profil utilisateur et peut être choisie à son bon gré.
2. Après chaque saisie, une notice bibliographique, sous le format XML conforme à la TEI-ms, sera ajoutée à la base de données textuelle dans la langue correspondante à la langue de catalogage.

Un premier aspect de la solution est la confiance accordée à la compétence humaine. En effet, nous aurions pu utiliser un traducteur qui permettrait de transcrire le contenu de chaque balise de la langue source vers la langue cible. Mais, au vu de l'imperfection des systèmes de traduction, le contrôle humain s'avère plus que nécessaire, afin d'obtenir un catalogue dont la correspondance est parfaite.

Un second aspect important de l'interface homme/machine est de veiller à l'élémentarité des données susceptibles de faire office de vedette matière. Nous illustrons cette particularité par l'exemple suivant [16]:

La TEI-ms propose de regrouper certaines informations composites et distinctes dans un seul élément à l'instar de <textLang> qui regroupe la langue et l'écriture.

```
<textLang> Old Church Slavonic, written in Cyrillic script.
</textLang>
```

Afin de favoriser une recherche d'information efficace, par vedette matière, il est impératif d'introduire des concepts de la communauté des bases de données tel que l'élémentarité des données (1NF : première forme normale). Ce qui fera correspondre à chaque balise une information dotée d'une équivalence sémantique [9].

```
<textLang>
```

```
        <Language>old Church Slovanic</Language>
        <writingStyle>Cyrillic Script</writingStyle>
</textLang>
```

### 5.2 Accès à la base des images des manuscrits numérisés

La TEI met à la disposition des encodeurs les outils nécessaires pour lier la base de texte à la base d'images. La liaison est effectuée grâce à l'élément <graphic> ayant pour attribut "url" (Uniform Resource Locator). L'attribut indique le chemin d'accès à la ressource numérisée. Ce qui permet d'accéder directement à l'image indiquée du manuscrit numérisé.

La correspondance entre l'image et la notice est mise en œuvre grâce à l'interface homme/machine. Ainsi, deux types de liens peuvent être créés :

- Liens créé par l'index et la table de matières : Ce type de lien permet un pointage par mots clés et par phrase entière vers une image particulière du facsimile.
- Lien vers la première page : Ce type de lien permet d'accéder à la première page du manuscrit. Le parcours de l'ensemble des images pourra se faire par un système de feuilletage (suivant/précédent). Par ailleurs, ce mode de liaison entre la base XML et la base d'images, offre la possibilité d'une visualisation du manuscrit sous forme de miniatures. Ce qui permet un accès direct vers l'image désirée, par un simple clique sur la miniature ou la saisie du numéro de l'image à consulter.

### 5.3 Système de catalogage collaboratif

Le catalogue de base est réalisé en langue arabe. Les autres catalogueurs utilisant les autres langues de catalogage, le français et l'anglais dans notre cas, auront un accès automatique à chaque nouvelle notice. Cet aspect nécessite la mise en place d'un système de suivi de traces des modifications opérées dans les divers catalogues. Par ailleurs, à chaque catalogueur correspond une interface dans sa propre langue de catalogage.

L'opération de transcription du catalogue dans une nouvelle langue se fait par l'affichage dans une info-bulles du contenu du catalogue source, dés l'activation d'un champ particulier. En cas de problèmes de traduction, le catalogueur pourrait faire appel à un outil de traduction (dictionnaire bilingue). L'activité de catalogage se solde par la génération d'un fichier XML respectant la DTD TEI-ms. Nous obtenons ainsi, des catalogues parallèles dont la correspondance est exactement parfaite entre un catalogue dans une langue avec un autre dans une autre langue et ce, quelque soit le couple de langues choisies.

## 6 Accès au catalogue multilingue des manuscrits arabes

Rappelons que l'accès aux images des manuscrits numérisés se fait par l'intermédiaire du catalogue. La mise en place d'un système d'accès fiable, simple et intuitif au catalogue est de ce fait, d'une importance capitale.

L'accès au catalogue des manuscrits arabes numérisés se fait selon deux modes :

- **Accès par sujet ou par vedette matière** : Ce mode permet à l'usager d'accéder au catalogue par un ou plusieurs critères bien définis tels que, l'auteur, le copiste, le titre du manuscrit, …etc.
- **Accès libre** : Dans ce mode d'accès, l'usager ne connaît pas à priori le contenu du catalogue. De ce fait, il est autorisé à émettre sa requête sans aucune restriction. Par conséquent, un outil de recherche d'informations dans le catalogue en format XML s'avère nécessaire.

Le caractère multilingue du catalogue, suppose l'accès à l'information dans une langue à partir d'une requête émise dans n'importe quelle autre langue du catalogue. Le système repose sur le principe de recherche d'information par croisement de langues [6]. De plus, le changement de sens d'écriture dans le catalogue, avec l'utilisation simultanée du caractère arabe et du caractère latin, ne pose pas de problèmes à cause de l'aspect textuel (XML) du catalogue [4].

### 6.1 Interrogation du catalogue par liste d'autorité (Vedette matière)

La puissance du format XML réside dans sa capacité à s'appliquer à tout genre d'application à des fins de représentation et d'organisation de tout type de données ou de documents [10]. La liste d'autorité usuellement utilisée est l'accès par sujet, par titre ou par auteur. Pour ce faire, le processus d'accès par vedette matière est constitué de deux modules (cf. Figure 4):

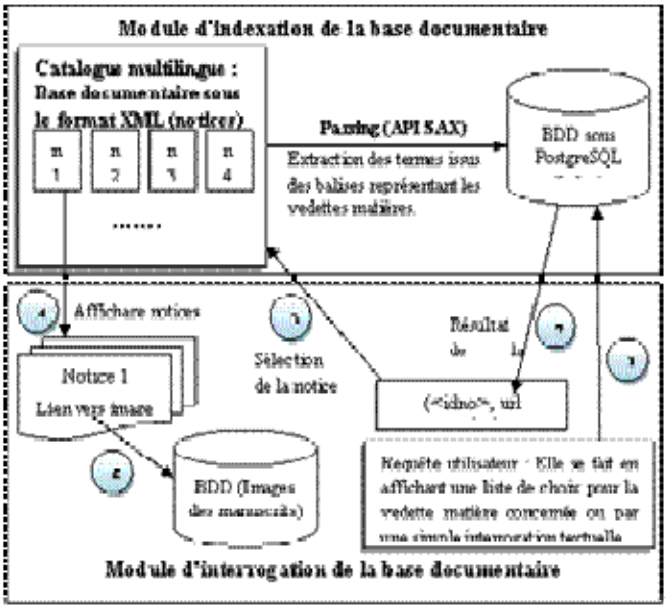


Figure 4 : Processus d'implémentation de la solution d'accès par vedette matière

- **Module d'indexation** : Le rôle de ce module se résume à l'identification des différents termes contenus dans les éléments constituant les vedettes matière. Par conséquent, la base documentaire formée par les notices bibliographiques sous le format XML est parcourue, puis parsée et les balises représentatives des vedettes matières repérées. Le contenu des balises est extrait et rangé dans une base de données sous PostgreSQL.
- **Module d'interrogation** : Ce module est chargé de répondre aux requêtes de l'utilisateur. En effet, le fichier index associe à toute vedette matière une notice bibliographique et son équivalent dans les autres langues de la base documentaire. L'utilisateur interroge la base documentaire via une liste de choix ou une requête textuelle. Après la sélection de la notice, qui est affichée sous une forme standard, l'accès au fichier image du manuscrit se fait par un simple clique sur le lien reliant la notice à l'image correspondante.

L'accès multilingue est assuré par un accès direct à la notice équivalente dans les différents catalogues des différentes langues.

### 6.2 Système de recherche libre dans le catalogue

La représentation textuelle du catalogue peut vite devenir très verbeuse, notamment en introduisant la notion de résumé (<overview>) et l'élément <p> dans le catalogue, qui encourage la formulation de longs textes.
Une requête peut être formulée sous forme de mots clés, de combinaison de mots clés, de phrase, …etc. La solution de recherche d'informations, en texte intégral, repose sur le principe du fichier inverse. Ainsi, elle nécessite une méthode d'indexation particulière.
En effet, le système d'indexation devra tenir compte de la position exacte de chaque terme de la requête (ou de toute la requête) dans le catalogue. Pour ce faire, nous proposons d'associer à chaque terme, la notice dans laquelle il apparaît et l'élément qui le renferme, d'où le triplet : *(**terme, numéro_notice, etiquette_élément**)*.

***terme*** : Désigne un mot clé selon lequel se fait la recherche.
***numéro_notice*** : Contient le numéro de la notice permettant d'accéder à l'image du manuscrit contenant "*terme*".
***etiquette_élément*** : Définit une étiquette par laquelle chaque élément du catalogue est identifié.

De plus, chaque élément doit clairement être identifié dans le catalogue. Cet aspect est mis en œuvre grâce au triplet : *(**numéro_notice, étiquette_élément, etiquette_parent**)*.

***etiquette_parent*** : définit le nœud parent de chaque élément définit par une balise particulière.

L'aspect multilingue du catalogage induit que le modèle d'indexation devra pouvoir non seulement identifier la notice contenant le ou les termes de la requête, mais il devra situer exactement la position du terme dans la notice. La correspondance entre deux termes, se fait par la localisation de la position du nœud où apparaît le terme dans les différents catalogues.
La détermination du nœud permet d'atteindre le terme recherché sans se soucier de sa position dans le texte. Le problème revient alors, à localiser un nœud particulier dans l'arbre de recouvrement correspondant au catalogue en format XML. Une fois fait, l'accès à son équivalent dans le catalogue d'une autre langue est implicite.
Pour ce faire, nous avons mis en œuvre un algorithme de parcours du catalogue en format XML et qui permet de localiser chaque élément par rapport à la structure générale du catalogue.

***Algorithme de parcours du catalogue***

**Entrée** : Catalogue des manuscrits en format XML
**Corps** : Tantque (non fin du catalogue)

{

1 Lire élément ;

2 Attribuer une étiquette à l'élément;

3 Déterminer la position de l'élément relativement à la structure générale du catalogue. C'est-à-dire, définir le triplet (*numéro_notice, etiquette_élément, etiquette_parent*);

4 Relier chaque terme à l'élément où il apparaît. C'est-à-dire, définir le triplet (*terme, numéro_notice, etiquette_élément*);

}

**Sortie** : Fichier définissant la position de chaque terme dans le catalogue.
L'algorithme aura pour résultat la construction de l'arbre de recouvrement associé au catalogue en format XML. L'arbre ainsi associé au catalogue, définit d'une manière précise la position de chaque élément.

***Mise en œuvre du système d'étiquetage***

La phase d'étiquetage est très importante. Elle joue un rôle capital dans l'efficacité du système de recherche d'informations.
L'étiquetage est généralement effectué avec un système de numérotation. Plusieurs méthodes sont décrites dans la littérature parmi lesquelles nous nous limitons à citer les deux méthodes suivantes :

1. La méthode de Dietz [12] qui effectue la numérotation des nœuds en pré-ordre et en post-ordre d'une manière séquentielle. Cette méthode présente l'inconvénient d'une renumérotation totale de l'arbre en cas d'une mise à jour.
2. La méthode XISS [14] qui vient palier au problème de renumérotation instantanée des nœuds en cas de mise à jour, par l'insertion de trous entre les numéros des nœuds d'une même branche. Ainsi, la renumérotation des nœuds se trouve reportée vers un nombre de mises à jour égal au pas de l'intervalle entre deux nœuds consécutifs.

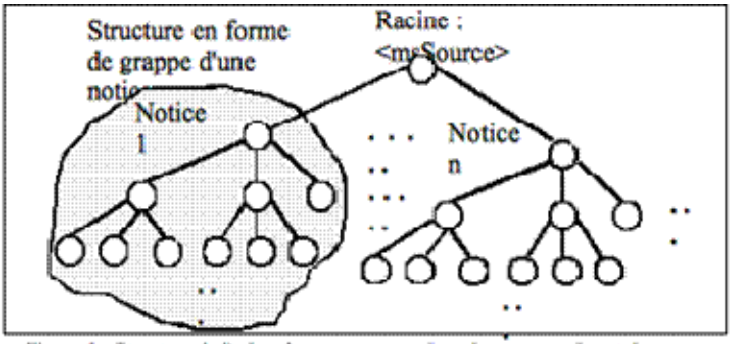


Figure 5 : Structure de l'arbre de recouvrement issu du parcours du catalogue, en format XML.

Le choix d'une méthode d'étiquetage est guidé par la rapidité du système de recherche d'informations, d'une part et la structure du catalogue, d'autre part. Ce dernier étant constitué d'un ensemble de notices. Par conséquent, la structure de l'arbre engendré par le catalogue en format XML, est formé par un ensemble de sous arbres en forme de grappes. Chaque grappe représente une notice descriptive de manuscrit.
La structure de l'arbre de recouvrement est relativement stable. En effet, il arrive rarement d'insérer un élément dans une même grappe ou d'en supprimer un autre. Cette stabilité s'explique par l'activité du catalogage des manuscrits arabes anciens, qui est très minutieuse où les erreurs se font rares. De plus, toute insertion d'une nouvelle notice induit la création d'une nouvelle grappe indépendante des autres. Par conséquent, une numérotation des nœuds selon le modèle de la méthode XISS est très adaptée à ce type de structure. La figure "Figue 5" montre la structure de l'arbre de recouvrement issu du parcours du catalogue en format XML.
La détermination d'un parent d'un nœud quelconque se fait par un simple tri par ordre croissant des valeurs du pré-ordre des nœuds de l'arbre de recouvrement. Un nœud $n_i$ est parent d'un nœud $n_j$ si $n_i$ est le nœud directement supérieur à $n_j$.

***Structure des fichiers index***

Notre approche consiste en la combinaison de deux méthodes d'indexation structurelle à savoir : l'indexation basée sur les champs [1] et l'indexation basée sur les arbres [5].

L'indexation basée sur les champs associe à chaque terme la balise dans laquelle, il apparaît. Par contre, l'indexation basée sur les arbres permet de localiser d'une manière précise la position des nœuds et la relation hiérarchique existant entre eux.

L'accès au catalogue par une recherche libre est basé sur le principe de fichiers inverses. Classiquement, les fichiers inverses se composent de deux fichiers principaux : Le dictionnaire (Lexique des termes) et le fichier "posting" (référence des documents qui, dans notre cas, sont représentés par les notices descriptives des manuscrits). L'introduction du multilinguisme dans notre solution induit la nécessité de correspondance entre les termes d'un catalogue dans une langue et ceux d'un autre catalogue dans une autre langue.

Par conséquent, nous avons introduit un troisième fichier de correspondance, qui permet de donner la position exacte de chaque terme dans le document (notice). L'accès à partir d'un terme d'un catalogue dans une langue vers son équivalent dans un second catalogue d'une autre langue, se fait par une translation dans celui-ci, en considérant le nœud parallèle.

Pour ce faire, nous utilisons le résultat de l'algorithme "parcours du catalogue", définit ci-dessus, pour construire les fichiers index.

Nous définissons trois fichiers :

- **Lexique des termes (mots)** : Il contient l'ensemble des mots du catalogue dans une langue donnée.
- **Références des documents** : Chaque notice correspond à un document. Ainsi, ce fichier contient les termes et les notices dans lesquelles ils apparaissent, soit le triplet (***terme, numéro_notice, etiquette_élément***).
- **Références de positions** : Il contient la référence de la position de chaque élément dans le catalogue, soit le triplet (***numéro_notice, etiquette_élément, etiquette_parent***).

L'accès à un terme particulier se fait grâce au fichier "Référence des documents", qui renvoie l'ensemble des notices où apparaît le terme.

Le fichier "*Références de positions*" trouve son utilité dans le passage d'un catalogue d'une langue $L_1$ vers un autre catalogue dans une autre langue $L_2$. En effet, l'utilisation de ce fichier permet de déterminer l'élément (nœud) d'une langue associé à son équivalent dans une seconde langue. Ce qui permet de sélectionner les termes de cet élément dans la seconde langue sans se soucier de leurs significations. Cette correspondance est rendue possible grâce à la relation bijective existante entre les nœuds des divers catalogues dans les diverses langues.

De plus, il est important de signaler que le formalisme de recherche dans le catalogue est le modèle booléen, qui se justifie par l'équivalence de l'importance accordée à chaque notice (document).

### 6.3 Accès inter-catalogue

La solution multilingue retenue dérive du modèle de corpus parallèle. Ce qui signifie que la quasi-totalité du vocabulaire d'une langue figure dans la seconde langue. Cet aspect de la solution permet d'accéder à l'élément de l'une des langues directement à partir d'un autre élément de la seconde langue.

L'accès au terme équivalent d'une langue $L_1$ dans une autre langue $L_2$ du catalogue se fait par la simple identification du nœud (balise) *ni* du catalogue de la langue $L_1$. L'accès au fichier "Références de positions" retourne sa position dans le catalogue.

Il suffit alors, de faire l'opération inverse dans le catalogue correspondant dans la langue $L_2$.
Dans les deux cas de recherche d'informations dans le catalogue (accès par sujet et recherche libre), il est toujours possible d'identifier une balise à partir d'une autre et ce, grâce au modèle d'indexation proposé, qui identifie la structure et le contenu du document.
Une bijection est établie entre les différents catalogues : L'accès de l'un à l'autre devient de ce fait, implicite malgré la distinction des langues. Par conséquent, aucune traduction n'est nécessaire lors de la phase de recherche dans les catalogues.
La correspondance entre les différents catalogues est illustrée par la figure suivante :

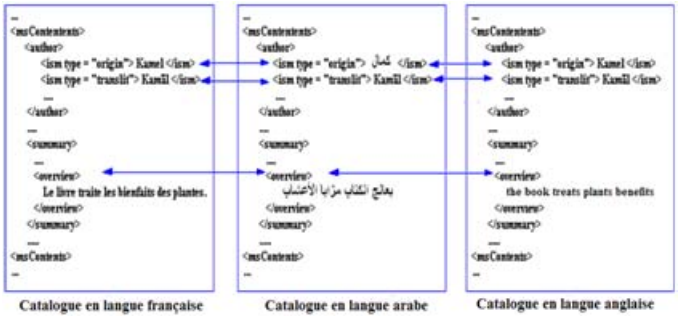


**Figure 6** : Structure Catalogue multilingue des manuscrits arabes numérisés

Les différents catalogues présentent une correspondance parfaite entre eux. Ainsi, l'accès à un élément dans le document d'une langue $L_1$ permet automatiquement de retrouver l'élément équivalent dans le document de la langue $L_2$. Notre solution trouve son originalité dans la possibilité d'intégrer, avec aisance et simplicité, d'autres langues de catalogage et donner ainsi, accès aux manuscrits arabes à une communauté linguistique très variée.

## 7 Conclusion

L'accès en ligne aux manuscrits numérisés constitue un objectif de plusieurs institutions de conservation de ce type de document. A travers notre article, nous avons proposé une solution pragmatique de mise en œuvre d'un tel objectif.
La solution proposée repose sur le principe de portabilité et de pérennité, ce qui justifie le choix de XML comme format d'encodage du catalogue des manuscrits. Ainsi, la TEI-ms a été choisie, enrichie et adaptée au catalogage des manuscrits arabes. Par ailleurs, la mise en place d'une interface de catalogage permet de dispenser le catalogueur de l'apprentissage du standard TEI.
L'implémentation de la solution a mis en évidence l'efficacité de notre solution, qui permet d'une part, un accès par vedette matière : aspect que nous retrouvons dans la majorité des solutions existantes et d'autre part, un accès libre au catalogue : ce type d'accès constitue une nouveauté en terme d'accès aux images des manuscrits numérisés. Toutefois, cela nécessite une description fine et exhaustive du manuscrit.

L'intégration du multilinguisme dans la solution rend les manuscrits arabes accessibles à toute catégorie linguistique, dont la langue est représentée dans le catalogue.
Par ailleurs, la solution a montré ses limites en terme de navigation dans la base des images des manuscrits. En effet, le système de défilement et d'accès par page (miniature) reste insuffisant pour l'utilisateur, qui a besoin de plus d'outils de visualisation. Nous citons entre autre :

- La possibilité d'agrandissement et de réduction de zones particulières de l'image.
- La nécessité de garder la trace des images consultées.
- La mise en place d'un système d'annotation des images consultées.

Ces différents points se présentent comme une intéressante perspective afin d'enrichir la solution. En effet, l'intégration du profil de l'utilisateur et de ses différentes annotations constitue un apport considérable pour le système de recherche et d'accès multilingue aux manuscrits arabes numérisés.

## 8 Bibliographie

[1] A. Gutierrez, R. Motz, and D. *Viera. Building databases with information extracted from web documents*. Chilean Computer Science Society, International Conference of the, page 41, 2000.

[2] Aurélie DELAMARRE – Quel catalogage pour les manuscrits contemporains ? Mémoire d'étude – DCB 2004.

[3] Centre de Ressources Numériques TELMA – *Présentation sur la TEI* – disponible en ligne à l'adresse http://www.cn-telma.fr/documentation/tei. Consulté le 12/09/2010.

[4] Christian Chabillon – *Unicode dans Sudoc* – Agence bibliographique de l'enseignement supérieur – Paris,t 52, n°3 – bbf 2007.

[5] D. Shin, H. Jang, and H. Jin. BUS : an effective indexing and retrieval scheme in structured documents. In The third ACM conference on Digital Libraries, DL'98, pages 235–243, 1998.

[6] Judith KLAVANS, Eduard HOVY et al., *Multilingual (or Cross-lingual). Information Retrieval*, 1999, disponible sur : http://www.cs.cmu.edu/~ref/mlim/chapter2.html, consulté le 12/09/2010.

[7] Marie-Geneviève Guesdon, Nathalie Rodriguez *Les manuscrits arabes, turcs et persans à la bibliothèque interuniversitaire des langues orientales* – MELCOM 27, Alexandrie – Mai 2005.

[8] Marie-Genvière Guesdon – *Arabic Manuscripts* – Communication at TIMA (The Islamic Manuscript Association) conference – available on http://www.islamicmanuscript.org/files/GUESDON_Marie_2008_TIMA.pdf or on http://www.islamicmanuscript.org/conferences/pastconferences/fourthIslamicManuscriptConference.html – Cambridge. Septembre 2008.

[9] Mohammed Ourabah SOUALAH – *Numérisation des manuscrits arabes : Catalogage et accès multilingue* Mémoire de Magistère en Informatique option Système d'Information et Document Numérique, sous la

Direction de M. Mohamed HASSOUN, soutenu Juillet 2008 – Collaboration ENNSIB (Lyon) / ESI (ex INI Alger).

[10] Mokhtar Ben Henda – *Du codage numérique au balisage sémantique des documents électroniques arabes. Approches multilingues et multiculturelles*. ISIC, Université Michel de Montaigne – Bordeaux 3, France – Septembre 2006.

[11] Orélie Bosc – Communication – Journée d'étude – *Manuscrits dans tous leurs états* – BNF – Septembre 2006.

[12] P. Dietz and D. Sleator. *Two algorithms for maintaining order in a list*. In STOC 87 : Proceedings of the nineteenth annual ACM symposium on Theory of computing, pages 365–372, 1987.

[13] Patrick PECCATE - *Les métadonnées* -Table ronde du Campus XML, 28/02/2003 – disponible sur le Web à l'adresse : http://peccatte.karefil.com/Software/MetadataCampusXML.pdf – Consulté le 12/09/2010.

[14] Q. Li and B. Moon. *Indexing and querying xml data for regular path expressions*. In VLDB, pages 361–370, 2001.

[15] Sihem Ghédira Dakhli – *L'usage d'UNIMARC à la Bibliothèque Nationale de Tunisie* – 3ème réunion internationale des utilisateurs d'unimarc – 31 mars 2010.

[16] TEI Consortium – *TEI : Guidelines for Electronic Text Encoding and Interchange* – Disponible sur : http://www.tei-c.org/release/doc/tei-p5-doc/html/MS.html. Consulté le 12/09/2010

# Lecture interactive : accès au contenu d'un document numérique à un niveau d'approfondissement réglable par le lecteur

**Laurence BALICCO, Marc BERTIER**

Laurence.Balicco@iut2.upmf-grenoble.fr / Marc.Bertier@iut2.upmf-grenoble.fr

*(1) GRESEC – Université Stendhal – Grenoble 3*

**Résumé.** La pratique de la lecture linéaire des documents longtemps prépondérante avec les supports traditionnels est concurrencée par d'autres pratiques avec le passage au support numérique. Il reste néanmoins des documents pour lesquels la lecture linéaire demeure l'usage principal. Pour eux aussi, le support numérique permet de proposer des usages nouveaux. Nous présentons ici un dispositif permettant l'affichage des documents sous une forme condensée ou développée à la demande de l'usager. Nous identifions ici quatre modes de condensation du contenu (à l'aide du plan, d'une phrase, d'un extrait ou d'un résumé). Chaque implémentation du dispositif s'appuiera sur l'un de ces modes de condensation en fonction du type de document, du contexte, des attentes du lecteur… Différentes études d'usage de ces formes de condensation sont proposées. Leur objectif principal est la validation du dispositif accompagnée de préconisations.

**Mots-clés.** lecture interactive, document numérique, ergonomie de lecture, lecture numérique, architecture de l'information, usage du document numérique

## 1 Introduction

Indépendamment de la nature traditionnelle ou numérique du support d'inscription, une partie des documents impose au lecteur la consultation du contenu dans un ordre déterminé alors que d'autres documents autorisent de s'affranchir plus ou moins complètement de cette contrainte. Dans de nombreux autres cas, le lecteur peut procéder à une exploitation partielle ou intégrale du contenu informationnel d'un document et dans un ordre plus ou moins libre ou contraint. Bien sûr depuis bien longtemps déjà existent des documents traditionnels conçus pour une consultation non séquentielle (dictionnaire, magazine, etc.), mais il est tout aussi incontestable que l'inscription des documents sur support numérique accroît largement les possibilités de ce mode de consultation.

Cependant la consultation non linéaire impose qu'en amont la conception du document l'ait anticipée et que celle-ci ait conduit à une organisation du contenu adaptée à ce mode de consultation non séquentiel. Autrement dit, dans un cas

extrême, chaque partie constitutive du contenu doit pouvoir être valablement consultée indépendamment du reste. Cette contrainte n'est pas mince pour l'*auteur*[1]. Une publication récente [1] s'intéresse à cette question et décrit diverses méthodologies pour la conception sous forme de multiples pages web de documents numériques présentant pour le lecteur une ergonomie de lecture de meilleure qualité que celle offerte par un document traditionnel. Il apparaît donc que la conception de tels documents occasionne à l'auteur une charge cognitive supplémentaire et spécifique et de nature éditoriale[2]. Nous ne sommes pas en mesure d'évaluer l'ampleur de cette surcharge ni d'affirmer si cette surcharge est liée à des habitudes plus ou moins ancrées dans la culture et/ou la formation des auteurs ou si elle est plus profondément inscrite dans des mécanismes cérébraux. Mais en tout état de cause actuellement le modèle du document séquentiel conserve une prégnance réelle pour certains documents professionnels (textes réglementaires, documents scientifiques, etc.), même si il est concurrencé par le modèle d'hyper document (documentation technique). Par ailleurs la technologie de restitution du contenu étant bien entendu cruciale, nous nous situons ici dans le contexte de la lecture à l'écran de documents destinés à une lecture linéaire, en particulier en contexte professionnel. Cependant nous ne procédons pas a priori à une analyse de besoins correspondant à des usages ou des usagers spécifiques ou prédéfinis. Dans un premier temps, il nous a semblé préférable de laisser l'entière maîtrise du dispositif au lecteur. Il est toutefois envisageable pour des usages définis de particulariser le mécanisme.
A notre sens, comme cela est évoqué ci-dessus, la conception de documents numériques séquentiels reste digne d'étude, de réflexion et de propositions. C'est pourquoi nous nous sommes intéressés ici à un dispositif spécifiquement adapté à ceux-ci. Ce dispositif a donné lieu à la réalisation d'une maquette de faisabilité. Cependant, la présentation technique de la réalisation de la maquette du dispositif n'est pas l'objet du présent texte. Dans la suite, nous commençons par décrire les principes de fonctionnement de notre dispositif, puis nous présenterons les différentes formes de condensation qu'il est possible de mobiliser, avant d'envisager enfin les questions liées aux usages d'un tel dispositif.

## 2 Dispositif

### 2.1 Mécanisme technique de substitution

Partons d'un document[3] produit par un auteur. La version finale de celui-ci est d'une certaine manière intégrale, complète. Laissons de côté comme étant une nouvelle production de contenu la démarche ultérieure dans laquelle l'auteur reviendrait sur un document pour lui ajouter quoi que ce soit. Pour la consultation d'un tel document le dispositif a comme principe fondamental d'offrir au lecteur un mécanisme permettant la simple *substitution* à certains passages, de contenus alternatifs. Le dispositif technique ne présume pas de la nature de la substitution

[1] *Le terme d'auteur est à prendre ici dans un sens extensif de celui qui produit individuellement ou collectivement un contenu.*

[2] *La question peut être rapprochée de celle abordée depuis longtemps en technique journalistique sous le nom d'approche en pyramide inversée, question naturellement soulevée avec une nouvelle acuité avec l'essor du web journalisme [2].*

[3] *La réflexion a concerné au départ les documents majoritairement textuels, mais pourra ultérieurement viser d'autres types de documents.*

mise en œuvre. On pourrait imaginer la traduction du passage dans une autre langue, l'explicitation d'un passage, ou a contrario une formulation plus condensée. C'est ce dernier cas, *condensation* du texte original de l'auteur, qui est envisagé ici. Pour les deux autres cas, la technique de la fenêtre *popup* ou de l'*escamot* [3] est vraisemblablement plus adaptée. Plusieurs modes de condensation de passages sont envisagés plus loin. Dans l'exemple présenté sur la figure 1, le premier paragraphe du présent texte peut être remplacé par un texte plus court, ayant statut de résumé.

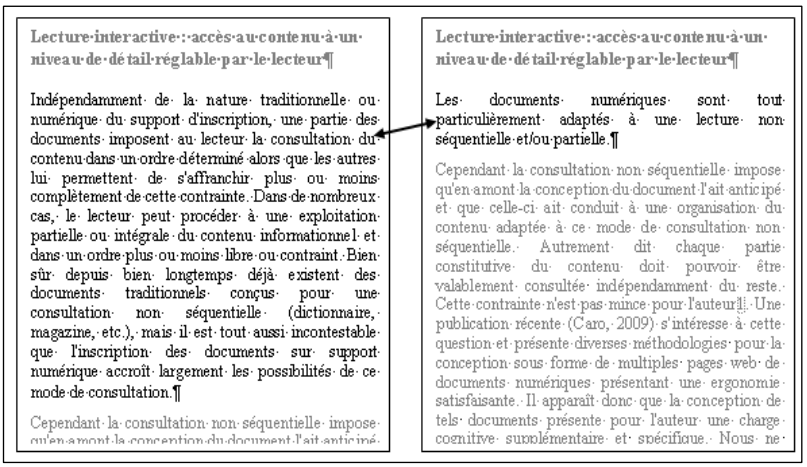


**Figure 1**. Substitution : condensation par résumé.

## 2.2 Interaction

C'est l'usager qui déclenchera ou non la substitution. Il le fera de manière interactive, c'est-à-dire au gré de sa lecture. Ceci a pour conséquence pratique que la mise en forme doit faire apparaître ou *signaler* au lecteur qu'il dispose de cette possibilité de substitution. Dans la figure 1, la couleur de texte (ici le noir) est le seul indice de cette possibilité, ce qui s'avère insuffisant. L'adjonction de divers compléments graphiques plus adéquats est nécessaire. Les interfaces de programmes informatiques offrent de nombreux pistes, comme les boutons + et – des programmes explorateurs. Les exemples dans les figures de cet article sont construits et n'illustrent pas les choix graphiques d'une implantation particulière. De plus, la mise en forme devrait permettre à l'usager de distinguer sans ambiguïté les termes originaux de l'auteur de ceux introduits par un tiers.

Comme déjà rapidement indiqué précédemment, c'est bien l'usager lecteur qui commande interactivement les substitutions. En revanche, le choix des passages substituables et la production des contenus alternatifs pour chacun d'entre eux relèvent quant à eux d'un travail d'adaptation du document pour sa publication en ligne. La question de qui est le mieux placé, de l'auteur ou d'un *éditeur,* pour effectuer ce travail est ouverte. La réponse changera vraisemblablement d'un cadre éditorial à l'autre, comme c'est le cas pour la rédaction des intertitres dans la presse, ou dans la littérature scientifique.

L'opération interactive de substitution -ou de condensation pour nous restreindre au cas envisagé dans la suite- d'un passage à un autre est *réversible*. Ainsi dans l'exemple de la figure 1 le lecteur pourra-t-il revenir à tout moment à la version longue.

### 2.3 Approfondissement "à la carte"

Enfin, dernier aspect fondamental du mécanisme technique, l'opération de substitution / condensation peut être appliquée de manière itérative, c'est-à-dire plus précisément de manière imbriquée. Ceci est facile à imaginer dès lors que l'on envisage le mécanisme de condensation dans le sens inverse, celui de l'*explicitation*. Ainsi un passage court peut-il être dans un premier temps remplacé par un contenu plus développé. Ce contenu alternatif peut inclure (un ou) des passages eux-mêmes susceptibles d'être remplacés à leur tour par des passages plus longs. C'est d'ailleurs l'objectif fonctionnel principalement poursuivi dans la conception de ce mécanisme : permettre à l'usager lecteur un *approfondissement* sélectif du document conformément à son usage courant. Ainsi à chaque instant l'usager aura face à lui une certaine *vue* du document couvrant l'ensemble du contenu original. En effet, même si certains passages sont à cet instant fortement condensés, d'autres plus ou moins explicités et d'autres encore au niveau de détail le plus fin, celui prévu par l'auteur, le document affiché à l'écran représente à sa manière le document intégral. Ceci est illustré sur la figure 2.

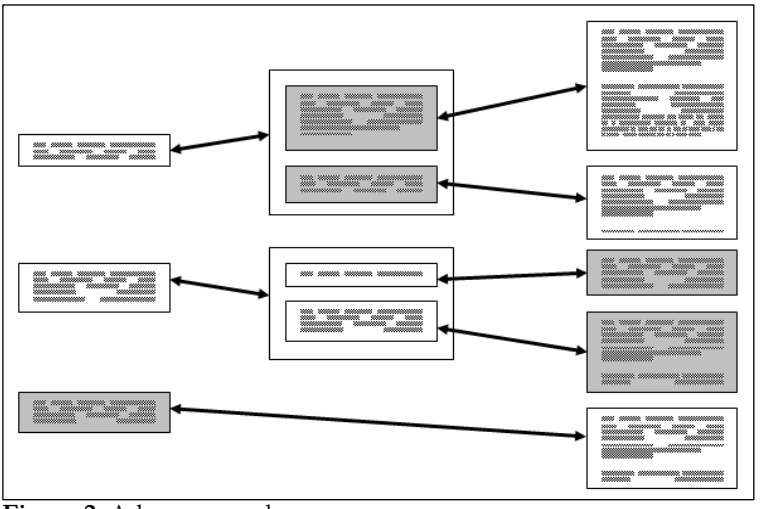

**Figure 2.** Arborescence des passages.

Sur cette figure, le document original fourni par l'auteur apparaît en colonne de droite, formé ici de cinq passages consécutifs. Il constitue la *vue intégrale* du document. La colonne de gauche à l'opposé donne la vue la plus condensée disponible du document mais cette vue couvre toutefois, à sa manière, l'ensemble du contenu. Enfin la séquence des cinq passages en gris constitue une vue du document que le lecteur peut choisir d'afficher. Le niveau d'approfondissement y est inégal d'un bout à l'autre de la vue : intermédiaire au début, détaillé ensuite pour terminer par un niveau condensé à l'extrême. Mais là encore tout le contenu du document original est couvert par cette vue du document.

Il apparaît sur ce schéma que le nombre de vues potentielles différentes peut devenir important lorsque le document initial est long, permettant à chaque lecteur usager de choisir un « profil » de lecture qui convient à son cas, approfondissant de manière sélective les passages présentant une forte pertinence dans le contexte de lecture et ne gardant qu'un rapide *fil conducteur* pour les autres parties (par exemple des généralités bien connues de lui).

Attention, si une *vue* du document prétend toujours afficher un document dérivé qui *couvre* / *représente* la totalité du contenu, cela ne signifie pas pour autant qu'il apparaîtra en totalité dans les limites de l'écran de visualisation, mais qu'il est accessible en totalité par simple *défilement* à l'écran et sans *suivre de lien*, opération cognitivement plus lourde pour le lecteur.
Enfin, si le lecteur souhaitait conserver pour lui-même ou transmettre à un tiers le *profil* d'approfondissement d'une vue particulière, cela pourrait être aisément envisagé, mais n'a pas été implanté jusqu'ici dans la maquette.

## 3 Différentes formes de condensation

Entre un document intégral et une forme de résumé, il y a différents stades de condensation de ce document. Les condensations proposées ici s'appliquent sur des documents majoritairement textuels qui sont les documents les plus fréquents actuellement. Un document majoritairement textuel peut contenir, en plus de texte, d'autres modes, principalement l'image. Nous souhaitons prendre en compte dans le dispositif de condensation ces autres modes porteurs d'information et essentiels lors de la lecture des documents [4].
Nous avons identifié ici quatre possibilités de condensation du document, qui correspondent à des degrés divers de traitement. Il est à noter que les mécanisme de condensations proposés portent sur du texte, mais le résultat de la condensation peut contenir d'autres modes non textuels. En effet, s'il est peu courant d'imaginer une condensation sous forme d'une image ou autre (correspondant à un changement de mode et donc à une forme de traduction d'un mode à l'autre, avec toutes ses difficultés), une image peut largement faire partie de l'élément condensé proposé.

### 3.1 Condensation à l'aide du plan

Dans l'hypothèse d'un document structuré majoritairement textuel, il est intéressant d'utiliser la structure du texte, et ce à moindre coût. Cette forme de condensation consiste en l'utilisation des différents niveaux de plan existants dans le document. Cette proposition présente de nombreux avantages. Tout d'abord, elle ne nécessite que peu de traitements car la structure est normalement présente dans le document, et de plus elle est conçue par l'auteur du texte. Elle acquiert ainsi une légitimité. L'utilisation du plan, avec ces différents niveaux de titres et sous-titres permet facilement de hiérarchiser le document. Il est ainsi possible d'imaginer une première présentation du document avec un premier niveau de titre, puis une possibilité de développement sur un deuxième niveau, puis un troisième et ainsi jusqu'à l'obtention de l'intégralité du document. Cela renvoie à l'usage bien connu d'un sommaire.
Toutefois, cette technique ne peut être utilisée que dans le cas d'un document essentiellement textuel (la notion de plan n'étant pas pertinente ici en l'absence de texte) et surtout structuré. De plus, le document induit son degré de condensation en fonction de son plan et ne laisse pas la possibilité de choisir les différents niveaux mis à disposition du lecteur.
La condensation à l'aide d'un plan est combinable avec toutes les méthodes suivantes. En effet, on peut toujours proposer un plan de plus en plus détaillé avant toute autre forme de condensation (et ce qu'on choisisse une phrase, un extrait ou un résumé). Cela rajoute de fait un degré supplémentaire d'organisation du texte.

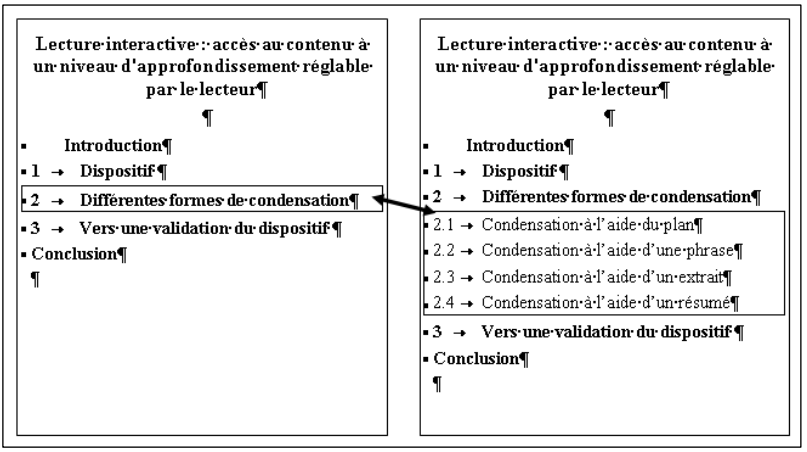


**Figure 3**. Condensation à l'aide du plan.

### 3.2 Condensation à l'aide d'une phrase

Une deuxième forme de condensation consiste à prendre une phrase pour représenter un paragraphe. Généralement, c'est la première phrase qui est utilisée. Dans de nombreux documents, on voit la présentation de la première phrase puis la possibilité de lire la suite en déroulant le paragraphe dans son intégralité.
Cette solution est également plutôt simple à mettre en œuvre. L'unité « phrase » est identifiable, que ce soit par un traitement manuel ou automatique, même si cela nécessite une analyse langagière intermédiaire. De nombreuses formes de présentation, en particulier de résultats de recherches d'information, utilisent cette méthode afin de laisser au lecteur la possibilité d'accéder à l'intégralité de la réponse si celle-ci s'avère l'intéresser.

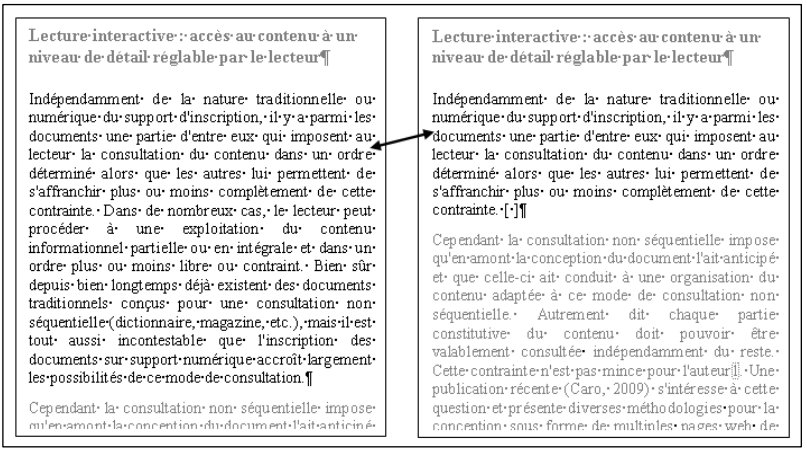


**Figure 4.** Condensation à l'aide de la première phrase.

Cette méthode présente malgré tout des inconvénients. Tout d'abord, elle nécessite, comme la première méthode, un document essentiellement textuel car la présence de phrases à l'intérieur du document ainsi que la structuration de celui-ci en paragraphes sont évidemment indispensables. Par ailleurs, il n'y a aucune

certitude quant à la pertinence[4] de la première phrase. Le rôle de cette première phrase est souvent introductif, ce qui n'est pas forcément adapté à l'usage fait ici. Celle-ci n'est donc peut-être pas suffisamment déterminante ou représentative pour l'usager qui ne déroulera alors pas la suite du paragraphe, même si celui-ci l'intéresserait.
De plus, autre difficulté, pour que cette méthode soit efficace il faut qu'un paragraphe recouvre une unité significative formant un tout. Or, cette règle d'écriture n'est pas toujours respectée dans de nombreux documents dans lesquels nous allons trouver plusieurs idées dans un paragraphe ou à l'inverse une idée s'étalant sur plusieurs paragraphes.

### 3.3 Condensation à l'aide d'un extrait

L'idée de cette condensation est d'utiliser, non la première phrase d'un paragraphe, mais un extrait d'un passage de document. Nous qualifions de *passage* un paragraphe ou une unité plus réduite ou plus importante. Cet extrait peut être constitué d'une ou de plusieurs phrase(s), jugée(s) significative(s) du passage considéré, mais il peut aussi contenir une autre unité du document (une image, un schéma…). Il est à noter que les éléments de l'extrait ne sont pas forcément consécutifs mais peuvent être récupérés dans différents emplacements du passage traité. De la même façon que dans le choix de la première phrase, le reste du passage de document est affiché sur demande de l'usager.

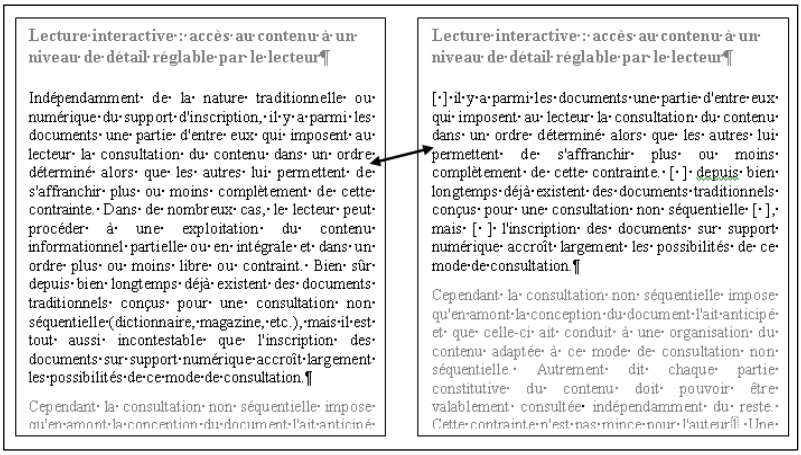


**Figure 5.** Condensation par extrait.

Cette méthode se révèle sans doute beaucoup plus significative que les deux précédentes, car elle utilise un extrait choisi du document (et non imposée par la structure ou l'écriture de l'auteur). Elle ne nécessite pas non plus un document essentiellement textuel.
Par contre, il est extrêmement difficile de sélectionner l'extrait représentatif, et ce que ce soit de manière manuelle ou automatique. En effet, l'extrait pertinent n'est peut-être pas unique. Différents lecteurs peuvent sélectionner des extraits différents (mais tout aussi pertinents pour eux). Manuellement, il est sans doute possible à une personne de déterminer, grâce à l'expertise qu'elle se forge sur la

[4] Nous ne discuterons pas ici de la notion complexe de pertinence, et l'utiliserons comme l'apport au lecteur d'éléments significatifs pour lui.

lecture de documents, un extrait significatif. Mais, d'un point de vue automatique, la sélection d'un extrait passe par une méthodologie de représentation de connaissances, et d'identification de connaissances significatives. Par ailleurs, un passage de document fait parfois référence à des connaissances contenues dans d'autres parties du document, voire externes au document. La détermination d'une unité représentative de la partie considérée peut se révéler particulièrement délicate.

Cette *condensation par l'extrait* permet donc de sélectionner des passages de documents afin d'en proposer l'affichage. Il est également possible d'utiliser une technique similaire afin de proposer, non une sélection, mais une coupure du document. L'idée est alors d'identifier un morceau du document comme étant moins significatif, d'en proposer la coupure, facile à matérialiser […/…], afin que le lecteur puisse réafficher l'intégralité du passage. Cette détermination est d'ailleurs aussi délicate à réaliser que dans le cas de l'extrait. Ces deux techniques d'extrait et de coupure fonctionnent de manière similaire. Nous les distinguons car elles correspondent à des intensions différentes et donnent aux morceaux choisis des rôles opposés.

### 3.4 Condensation à l'aide d'un résumé

La *condensation par le résumé* est la plus évoluée sur un document. En effet, le résumé permet de s'affranchir de la structure, de l'organisation du texte pour construire une représentation du document. Par ailleurs, l'utilisation de résumés permet d'élaborer différents degrés de condensation. Il est en effet possible d'envisager un premier résumé très succinct, développable en différents résumés, de plus en plus importants (de moins en moins résumés en fait) pour atteindre le document intégral. Un exemple de cette condensation est donné dans la figure 1. On peut imaginer un résumé textuel, mais également un résumé constitué exclusivement ou non d'autres modes.

Le résumé peut être élaboré manuellement par l'auteur. Cette élaboration peut être le résultat de contraintes rédactionnelles. Il est en effet tout à fait envisageable de demander à l'auteur du document de proposer un résumé, ou une succession de résumés de plus en plus larges, conjointement à l'élaboration du document. Il est également possible de constituer a posteriori (une fois le document élaboré) différents résumés. Cette construction, si elle est manuelle (avec toutes les difficultés connues du traitement de documents), relève de l'éditeur du document et se base sur la structure et le contenu du document.

Le traitement automatique nécessite, comme dans le cas de l'extrait, la description des connaissances contenues dans le document. Il faut alors décrire le document à l'aide d'un formalisme de représentation de connaissances[5]. Par ailleurs, l'élaboration de résumé automatique met en jeu des méthodes de traitement automatique de la langue, tout d'abord l'analyse de la partie textuelle des documents, puis la génération automatique de résumés textuels, complètement ou partiellement.[6] Cette méthodologie complexe présente de nombreuses difficultés ou limites. D'abord, il est complexe de gérer les connaissances contenues dans les documents, et leur évolution, enrichissement tout au long du document. Ensuite les analyses ou générations de documents donnent des résultats parfois décevants. Enfin, le temps de traitement de documents peut être très long, que ce soit dans la

[5] *Plusieurs formalismes sont utilisables dans ce contexte [5], [6]. Leur description ne fait pas l'objet du présent article.*

[6] *Le résumé automatique fait l'objet de nombreux travaux présentés par exemple dans [7].*

phase d'analyse ou de génération, et il est donc peu envisageable d'imaginer un traitement dynamique. Par ailleurs, les résumés élaborés automatiquement sont souvent préconçus et n'offrent finalement pas une large liberté de contenu [8], [9], [10].

On pourrait imaginer que cette vision du texte sous forme de résumés successifs correspond à une représentation de la genèse du texte, marquant les différents développements suivis par l'auteur, et donc matérialisant son cheminement inverse : une idée développée de plus en plus pour aboutir au document intégral [11]. De notre point de vue il n'en est rien. Les résumés successifs, qui construisent une hiérarchie du contenu du document, correspondent davantage à un modèle pédagogique appliqué a posteriori sur le document.

Cette méthode du résumé est différente des trois autres méthodes proposées sur un point essentiel. Les autres méthodes proposent à chaque fois des éléments de condensation appartenant au document. Or, le résumé va construire un nouvel élément n'appartenant pas au document intégral[7]. Le développement d'un élément condensé dans l'une des trois premières méthodes consiste en une réinsertion, dans la vue du lecteur, du document primaire. Le développement d'un résumé aboutit au remplacement de celui-ci, soit par un résumé plus conséquent, soit par une partie du document intégral. C'est le seul cas, dans notre dispositif, où un contenu nouveau est introduit dans une vue du document.

## 4 Vers une validation du dispositif

Les différents principes de condensation présentés ici ont un impact important. En effet, ils vont permettre la personnalisation des documents par la sélection des parties à détailler, et, dans certains cas, par le niveau de détail adopté. L'utilisation de cette nouvelle forme d'organisation et de présentation du document modifie sa linéarité. Le document majoritairement textuel, surtout s'il est en format papier, est essentiellement basé sur la lecture linéaire. Le passage au document numérique adoptant diverses condensations permet de s'affranchir en partie de cette linéarité ou du moins de la reconstruire en fonction des attentes et besoins des lecteurs usagers du document. Il apparaît donc essentiel de mesurer les diverses réalisations du mécanisme de condensation. Il est à noter ici que les études et observations détaillées sont toutes réalisées dans un contexte professionnel, et ce pour deux raisons. Tout d'abord, ce contexte est celui qui, à nos yeux, est le plus pertinent pour proposer ce mécanisme. Ensuite, l'usage de documents en milieu professionnel s'accompagne de contraintes très spécifiques, dont la nécessité d'accéder de manière rapide et efficace au document.

Nous souhaitons donc valider notre dispositif au travers d'une étude des usages de la lecture de documents numériques. Nous pensons qu'il est important de nous assurer de la solution de condensation la plus appropriée, apportant la meilleure commodité de lecture. Il nous apparaît tout d'abord que la condensation des documents ne doit pas être considérée de manière systématique. En effet, comme nous l'avons déjà mentionné, certains documents n'ont sans doute pas à être condensés, ou partiellement seulement. La décision de condensation et le mécanisme de condensation choisi sont vraisemblablement liés à l'usage fait du document, ainsi qu'aux attentes du lecteur. Il est donc essentiel de déterminer les

[7] *Nous excluons ici la possibilité de créer un résumé en regroupant différents éléments du document, ce qui, dans notre travail, rentre dans le cadre de l'extrait.*

conditions et choix de condensation, en fonction de l'usage. Pour cela, il est nécessaire de procéder à différentes observations.

Tout d'abord, nous pouvons observer un public en situation de lecture à l'écran pour déterminer les zones de lecture sélectionnées et identifier avec ces lecteurs les raisons de leurs choix. Cette première étape doit nous donner des éléments d'expertise des documents, ayant comme résultats la sélection de documents (ou de passages de documents) à condenser, ainsi que le choix d'un mécanisme de condensation [12].

Une fois cette observation menée, il devient nécessaire de proposer des documents condensés afin de valider le mode de condensation le plus approprié en fonction des usagers, de la situation de lecture, du type de texte… Les différents mécanismes de condensation seront alors proposés aux lecteurs. Cette expérience a un double objectif. Nous souhaitons d'abord identifier si certaines formes de condensation se révèlent mieux accueillies que d'autres. Puis, nous voulons tester, au travers des propositions faites à nos lecteurs, si la méthodologie utilisée pour la condensation (choix des parties de texte à condenser, choix du contenu condensé…) s'avère pertinente pour eux.

Les questions du choix du contenu à condenser ainsi que du type et des modalités de condensation peuvent être également abordées en mettant les lecteurs en situation de condenser des documents. Il s'agit alors de leur proposer des documents en version intégrale en leur demandant de décider de la condensation des documents et de la technique utilisée en discutant avec eux de leurs choix. Cette démarche est utile car les lecteurs affichent des contraintes spécifiques. Leurs choix et décisions seront essentiellement guidés par leurs attentes et leurs usages.

Il sera ensuite intéressant de confronter le point de vue et la technique de condensation des lecteurs avec ceux de l'auteur et de l'éditeur du document. L'un comme l'autre seront susceptibles de traiter le document a posteriori. L'auteur a une idée très précise du contenu de son document et a fait des choix éditoriaux en conséquence. Il souhaite parfois en préserver certaines parties jugées essentielles ou imagine un usage ou un public spécifiques. L'auteur peut en particulier s'inquiéter de voir dénaturer certains passages par une formulation simplifiée non pertinente. Mais il reste toujours possible de faire valider par l'auteur le résultat de la condensation. Il peut donc être intéressant de comparer les choix et vœux des lecteurs et de l'auteur.

L'éditeur a également un rôle spécifique. S'emparant du document de l'auteur, il a comme mission sa médiation auprès d'un public de lecteurs. Pour cela, des choix éditoriaux sont faits (mise en avant de certaines parties du document, élaboration de résumés, valorisation d'éléments…). Il est essentiel d'analyser ces choix éditoriaux et de mesurer leur impact sur la lecture du document, sur la façon de l'appréhender, ainsi que sur la représentation que les lecteurs vont se faire de ce document. Une étude de la méthode suivie par un éditeur s'avère donc intéressante, ainsi qu'une validation des résultats de la condensation par les lecteurs usagers.

Toutes ces expérimentations sont en cours. Les résultats doivent nous permettre de valider, mais aussi d'affiner nos propositions en les accompagnant de préconisations de mise en œuvre.

## 5 Conclusion

Le mécanisme présenté ci-dessus vise à rendre plus efficace la lecture à l'écran de documents numériques professionnels longs et complexes. Les documents visés présentent les caractéristiques d'être séquentiels et de contenir du texte, même si d'autres modes sont souvent présents. Pour cela, notre mécanisme s'appuie sur différents modes de reformulation abrégée du document intégral. La contrainte principale que nous nous sommes imposée ici est de toujours fournir au lecteur une version présentant une certaine intégrité du document.
Au-delà, le mécanisme pourrait être utilisé pour la présentation de résultats de recherche d'information en mettant en valeur le détail des parties du texte pertinentes, à l'image des travaux réalisés dans le contexte de l'initiative INEX [13].
Ce mécanisme de lecture interactive à niveau d'approfondissement modulable par le lecteur usager impose un accord de principe entre l'auteur et l'éditeur dont les rôles diffèrent : d'une part l'auteur devra conserver la prérogative de la gestion du contenu et de sa pertinence à son sens, quelle que soit la vue affichable du document (prérogative liée au droit moral) ; d'autre part, l'éditeur doit quant à lui rendre possibles toutes les vues qui pourraient être conformes aux attentes d'un lecteur dans les limites imposées par l'auteur. Enfin, dernier acteur, le lecteur usager dispose d'une réelle capacité de construire interactivement la vue du document complet qui convient à son usage particulier.

## 6 Bibliographie

1. S. Caro Dambreville (sous la dir. de), *Conception, design des documents numériques*, Document numérique, Hermès, Lavoisier, Vol. 12, n°2/2009.fffff

2. C. Scalan, Writing from the Top Down: Pros and Cons of the Inverted Pyramid, http://www.poynter.org/column.asp?id=52&aid=38693, consulté le 13 juin 2010.

3. S. Caro Dambreville, *L'écriture des documents numériques : approche ergonomique*, Lavoisier, Paris, 2007.

4. E Jamet, L'intégration spatiale d'éléments textuels et illustratifs améliore-t-elle la performance ?, In Alain Vom Hofe (sous la dir. de). *Interaction homme-système : perspective et recherches psycho-ergonomiques. Revue d'intelligence Artificiel*, vol. 14, n° 1-2/2000, Hermès, Paris, 2001, pp. 167-187.

5. J Rouault et MC Manes-Gallo, *Le couple sémantique – pragmatique et le calcul des énoncés élémentaires,* Hermes, Paris, 2003.

6. L Balicco, *Génération automatique de documents associant textes et images : Méthodologie de conception de générateurs fondée sur la représentation des connaissances*, Mémoire d'Habilitation à diriger des recherches, Université Stendhal, décembre 2002.

7. JL Minel (sous la dir. de), *Résumé automatique de textes*, Traitement automatique de la langue, Hermes, Paris, Lavoisier, Cachan, Vol. 45, n°1/2004.

8. JP Balpe, Pour une littérature informatique : un manifeste, *Littérature et Informatique, la littérature générée par ordinateur,* Artois Presses Université, 1995.

9. L Danlos et L Roussarie, Génération automatique de textes, In Jean-Marie

Pierrel (sous la dir. De), *Ingénierie des langues,* Hermès Sciences Europe, Paris, 2000, pp. 311-330.

10. C Luc, *Représentation et composition des structures visuelles et rhétoriques du texte. Approche pour la génération de textes formatés,* Thèse, Université Paul Sabatier, Toulouse, 2000.

11. M Fayol, *Des idées au texte : psychologie cognitive de la production verbale, orale et écrite,* Presses Universitaires de France, Paris, 199, 288 p.

12. JP Bronckart, *Activité langagière, textes et discours : pour un interactionisme socio-discursif,* Edition Delachaux et Niestlé, Lausanne-Paris, 1996, 352 p.

13. J Kamps et B Sigurbjörnsson, What do users think of an XML element retrieval system, http://staff.science.uva.nl/~kamps/publications/2006/kamp:what06.pdf, consulté le 5 septembre 2010

# Système d'Information et écritures numériques en entreprise : les mutations du travail informationnel

**Animateur :** Manuel Zacklad

**Participants :** Dominque Cotte (Université de Lille 3), Benoit Habert (EDF R&D), Yves Chevalier ( Université Européenne de Bretagne), Yves Jeanneret (Université Paris4 la Sorbonne)

Le SI semi-structuré des entreprises (privés, publiques, sociales et solidaires) est la partie cachée de l'iceberg numérique : il est devenu dominant et pervasif (messageries, documents bureautiques, smartphone, etc.). Le travail informationnel lié à ce SI prend une part essentielle dans les activités tertiaires en l'absence de tout cadre prescriptif. Il implique une multiplicité d'environnements numériques nouveaux qui sont souvent mal appréhendés par les DSI et les MOA : messagerie, moteurs de recherche, CMS et GED, environnements bureautiques (diapositives numériques, tableurs, traitements de texte). Face à ces évolutions rapides et profondes, les approches théoriques du Système d'Information n'offrent pas de réponses totalement satisfaisantes. Elles peinent, par ailleurs, à rendre compte de l'extrême complexité de l'écosystème informationnel dans lesquels sont immergés les utilisateurs. Plus grave, elles ne soulignent pas assez à quel point la compréhension du travail informationnel et de ses enjeux pourrait être une condition essentielle de la performance des activités au niveau des individus, des groupes et des entreprises. Les travaux en sciences de l'information et de la communication sont susceptibles de renouveler regard porté sur les SI à travers trois types de contributions :

- Capacité à appréhender la signification des pratiques de lecture et d'écriture dans les environnements documentaires numériques (au sens large les bases de données sont aussi des documents numériques) en se dotant d'une épaisseur de vue historique et anthropologique dégagée de la métaphysique cybernétique de l'information qui prévaut dans certaines approches gestionnaire et informatique du SI.
- Capacité à jeter un éclairage nouveau sur les instruments du travail informationnel : écriture et lecture numérique, gestion des documents (métadonnées et plans de classement), recherche d'information, visualisation graphique, etc. en prenant en compte la circulation et les transferts de savoir-faire entre l'ensemble des pratiques lettrées, de l'administration et la gestion des entreprises aux humanités numériques en passant par les différents secteurs de la recherche scientifique et des arts.
- Capacité à situer l'entreprise et le travail informationnel dans le cadre élargi des évolutions de la société dite « de l'information » pour interroger

un certain nombre de postulats : caractère plus ou moins immatériel de ce travail, plus ou moins intellectuel, plus ou moins transparent, recours plus ou moins efficace à la vision « processus d'affaire », etc.

# Médiation numérique et institutions patrimoniales: le site web comme complexe de pratiques

Bernadette Dufrêne

bernadette.dufrene@u-paris10.fr

Laboratoire HAR et Culture et communication. Université de Paris Ouest

**Résumé**. A l'encontre d'une conception de la médiation numérique qui ne serait pas située, l'objet de l'article est de montrer que le site web comme forme éditoriale spécifique est un complexe de pratiques professionnelles et sociales ; tout en soulignant les continuités dans les formes de médiation, il cherche à montrer comment les spécificités de la médiation numérique permettent de formater les rapports des visiteurs à l'institution et à la collection sans pour autant les subsumer. La médiation numérique se superpose à d'autres formes de médiation avec lesquelles se créent des interactions.

**Mots-clefs**. sites web, médiation, médiation numérique, éditorialisation, musée virtuel, institutions patrimoniales

## 1 Introduction

"Collecter, conserver, présenter", la triade qui définit les missions des institutions patrimoniales et par là même leur fonction épistémique a été affectée dans la deuxième moitié du XXème siècle d'une part par l'accélération des transports qui, favorisant la circulation des hommes et des œuvres, l'a considérablement intensifiée, d'autre part, par la révolution numérique qui a conforté cette logique de circulation de l'information : le premier phénomène s'est traduit par la montée de l'exposition temporaire depuis le dernier quart du XXème siècle que permet une rotation accrue des œuvres et des hommes; le second, par la généralisation des sites web que les institutions patrimoniales ont mis en place. Les sites web des institutions patrimoniales ont connu un succès croissant (pour exemple plus de 9 millions de visites pour Ina.fr, 3, 5 pour Centrepompidou.fr en 2007 avec une croissance de 13% par rapport à 2006); le recul est aujourd'hui suffisant pour connaître les usages qui en sont faits : des études en France, aux Etats-Unis et au Canada[1] permettent de dresser des profils et d'appréhender des tendances. Elles

[1] *Voir en particulier les études d'O.Donnat (France), Spadaccini (Canada) et celle, menée aux Etats-Unis par l'Institute of museum and library services, qui montre la forte croissance de la fréquentation à la fois des institutions culturelles et de leurs sites :*

nous amènent à nous interroger sur la fonction de ces sites : constituent-ils un simple prolongement de la médiation des savoirs exercée par les institutions patrimoniales ou bien un nouveau régime de symbolisation?

Dispositif public, le site web des institutions patrimoniales peut être considéré comme une forme de document, c'est-à-dire d'"information communiquée au moyen d'un support"[2] , qui partage avec les autres documents numériques des caractéristiques essentielles, notamment le dynamisme -il n'est pas indifférent que les sites web évoluent, créent de nouveaux usages ou tentent de s'y adapter- et l'hybridation d'éléments empruntés à différents médias; il est enfin un ensemble de pages dont l'organisation globale comme la structure particulière obéissent à des représentations. En tant que document numérique public, le site des institutions patrimoniales est avant tout un complexe de pratiques, professionnelles et sociales. On ne saurait donc l'envisager seulement comme vecteur d'information (celle issue des opérations de patrimonialisation via les opérations de sélection, de description, de documentation) mais il est avant tout une forme éditoriale spécifique. Dans un écrit d'écran interfèrent plusieurs strates : technique, cognitive et sociale (JEANNERET, 2001). L'intérêt d'une approche sous l'angle de la médiation est alors de restituer l'ensemble des médiations de l'édition électronique, de voir comment la technique traduit et/ou reconfigure des formes de médiation existant dans l'espace concret des institutions patrimoniales, en crée de nouvelles, de s'interroger sur la constitution de nouveaux publics : médiation de masse, la médiation numérique renforce-t-elle les hiérarchies culturelles ? L'analyse de l'énonciation éditoriale, des usages potentiels que les sites proposent, permet de prendre en compte les évolutions dans la prise d'information, de voir, comment, notamment avec le passage au web 2.0, les pratiques des professionnels et celles des usagers se recomposent. Interroger la médiation opérée par les institutions culturelles au regard du numérique revient donc à s'intéresser à la fois aux continuités/discontinuités en matière de médiation des savoirs, aux mutations que les caractéristiques du support numérique du point de vue du stockage, de l'hybridation multimédia, de la télécommunication (dans son aspect dynamique de délocalisation/relocalisation)[3] a entraînées mais aussi au document numérique en tant que lieu où se cristallisent des pratiques sociales qui lui confèrent un statut particulier selon les usages qui en sont faits.

## 2 Des infrastructures épistémiques au document numérique, continuités et mutations dans la médiation documentaire

Les institutions patrimoniales, notamment les bibliothèques et les musées, apparaissent comme le lieu par excellence de la médiation des savoirs et non pas

---

*In 2006 remote online access increased adult visits to museums by 75% and to public libraries by 73% (while in-person visits have increased over time).*

| | *Museums* | *Public Libraries* |
|---|---|---|
| *Number of in-person visits* | *701 million* | *762 million* |
| *Number of remote online visits* | *524 million* | *558 million* |
| *Total visits* | *1,225 million* | *1,320 million* |

[2]*Définition donnée par Annette Béguin, Stéphane Chaudiron et Eric Delamotte, "Entre information et communication, les nouveaux espaces du document", Etudes de communication, n°30, 2007*

[3]*Je renvoie à la description du document numérique dans l'introduction du n°30 de la revue Etudes de communication*

seulement comme un lieu de réserves. Par médiation des savoirs, on entendra d'abord les dispositifs mis en place par les institutions pour permettre l'appropriation des œuvres par différents publics : du plan de classement auquel s'apparente le parcours du musée ou de l'exposition au catalogue imprimé via le cartel, la médiation documentaire dans ses aspects concrets, revêt donc différentes formes mais poursuit un seul but : rendre accessibles les œuvres ou les documents sur le plan matériel comme sur le plan intellectuel. Dans cette perspective, la fonction épistémique des institutions se partagerait entre des opérations de patrimonialisation, notamment la sélection, le classement, l'inventaire, la description et des opérations de médiation qui seraient alors du côté de la communication aux publics. Ce serait introduire une ligne de partage entre la constitution de l'information -. l'œuvre, l'objet, le livre prélevés dans un environnement et faisant l'objet d'un traitement à l'issue duquel ils s'intègrent au patrimoine- et sa communication. Or les opérations de sélection, de classement, d'inventaire, loin d'être neutres, sont elles-mêmes médiées par des logiques sociales, les missions que se donnent les institutions, leur projet culturel, social et politique. La médiation des savoirs, ce n'est donc pas seulement un ensemble d'actions "techniques" mais c'est aussi le point de vue selon lequel les objets de connaissance, leur classement et leur présentation ne sont pas indépendants du cadre dans lequel ils s'élaborent; leur sens est relatif à des conditions d'énonciation propres à des types d'institutions et à des situations historiques. De là vient la nécessité de considérer la médiation numérique telle qu'elle s'exerce sur les sites web des institutions patrimoniales à la fois dans la continuité historique et institutionnelle et sous l'angle des mutations qu'introduisent ses spécificités.
Dans son souci d'historiciser les sciences de l'information, J.M. Salaun s'appuie sur un article *Epistemic infrastructure in the rise of the Knowledge Economy*[4] pour montrer que le mouvement de re-documentarisation par le numérique est issu de la médiation des savoirs exercée par les institutions culturelles:
"*Les infrastructures épistémiques se sont développées sur les processus de sélection élaborés par les conservateurs, bibliothécaires, archivistes pour filtrer les connaissances selon les normes et les standards professionnels, les sujets et les domaines, en étant attentif aux besoins des communautés d'usagers. Cette sorte de collecte systématique construit la confiance dans les ressources informationnelles. Une économie du savoir, bâtie sur de l'information numérique dépendra, de la même façon, d'indicateurs clairs de qualité, d'autorité, d'authenticité...*"
Cette référence aux institutions culturelles comme infrastructures épistémiques est d'autant plus pertinente que l'enquête sur les usages menée aux Etats-Unis montre la confiance dont jouissent les sites des institutions culturelles:
"*Among other sources, public trust in cultural institutions is very strong: 86% of public library visitors and 77% of museum visitors rated them equal or higher in trustworthiness than all other sources of information*[5]."

A l'appui de cette idée d'une continuité entre les infrastructures épistémiques que constituent les musées, les bibliothèques ou les archives et l'Internet, on peut naturellement rappeler le rôle joué pour la science documentaire par la création du Mundaneum, l'entreprise de Paul Otlet et Henri La Fontaine, à la fois musée,

[4] *Cité par J.M Salaun, article cité supra, p.21*

[5] *Etude menée par Dr. José-Marie Griffiths, Donald W. King, University of North Carolina at Chapel Hill, The IMLS National Study of the Use of Libraries, museums, and the Internet, February 2008, p.27.*

bibliothèque et centre d'archives; ceux-ci avaient créé en 1895 l'Institut international de bibliographie. Beaucoup ont vu dans le système d'indexation connu sous le sigle CDU mis alors au point une préfiguration des moteurs de recherche.
Si la médiation des savoirs est caractéristique des institutions patrimoniales, dans la mesure où elle accompagne leur production, elle l'est tout autant des sites web. De fait notamment dans la première génération de sites mais aussi dans leur forme actuelle, la mission éducative est très fortement marquée : plusieurs rubriques la portent, la banque de données initialement conçue comme un catalogue en ligne ou encore les dossiers pédagogiques. Tant la demande sociale que les pratiques professionnelles amènent au développement de cette médiation. Les enquêtes menées aux Etats Unis ont fait ressortir deux points intéressants. Le premier concerne la visite en ligne : elle est motivée par des besoins d'information qui relèvent essentiellement de l'éducation informelle et du plaisir dans les musées (motifs de visite en personne: *informal learning and recreation, 94%, formal learning 4% ; motifs de visites en ligne : informal learning and recreation, 83%, formal learning 9%).* D'où la conclusion des chercheurs: "*museums are overwhelmingly used to address informal learning and recreational needs (89% of needs). Remote visits to museums are used more than in-person visits to support or extend formal education (9% versus 4%) and for work-related needs (8% versus 2%*). En revanche, dans les bibliothèques, il s'agit davantage d'une information "utilitaire", ce que les chercheurs décrivent ainsi:
" *Most Important Purpose for Online (Remote and in-library) Visits :*
Personal or family needs : 25%
Recreation or entertainment: 14%
Formal education needs :43%
*Work-related needs : 18%*[6]

Le deuxième point intéressant est que la recherche d'information dans un type d'institution culturelle amène à fréquenter un autre type d'institution :
"*Museums, public libraries and the Internet lead users to different sources of information millions of times. Museum visits led to 65 million library visits and public library visits led to 20 million museum visits in the past 12 months*[7]".
Le fameux rêve du « brassage » des publics fortement souhaité à l'ouverture du Centre Pompidou voit ainsi un début de réalisation ; Mais, même si la fonction épistémique des institutions patrimoniales est bien établie, la médiation numérique qui la prolonge sur les sites web présente deux risques : une amplification de la logique documentaire qui s'apparente à une ivresse documentaire et la fragmentation du savoir. L'amplification de la logique documentaire tient à la nature même du document numérique; l'hypertexte, l'hypermedia ont rendu actualisables dans une apparente immédiateté des savoirs, des images voire des œuvres dont le classement, les modes de présentation ont fait souvent l'objet de débats dans l'espace concret du musée. Un exemple montrera comment, en rendant caducs certains débats, la technique peut gommer des enjeux réels dans le mode de présentation des œuvres. En ce qui concerne le patrimoine artistique, les plates-formes actuelles permettent de combiner une œuvre, sa description, ses commentaires, par l'artiste, par un critique, par un amateur etc. Du fait même de la reproduction de l'œuvre sous forme d'image et de la coexistence au sein d'une même page d'éléments hétérogènes, c'est toute la perception de l'œuvre qui est

[6] *Ibidem* p.10
[7] *Ibidem* p.10

modifiée : non seulement elle n'apparaît plus d'abord et essentiellement comme œuvre avec forme, volume, matière mais elle entre dans"la vie triviale des êtres culturels", destinés à circuler, à être appropriés, et ce, quel que soit l'usage qui en sera fait. Faute d'une culture informationnelle formant l'usager à la critique, les différentes informations se juxtaposent sans hiérarchisation. Au visiteur en ligne de s'orienter dans cette surabondance informationnelle. En revanche, dans l'espace concret du musée, l'œuvre tiendra son sens de la présence d'une médiation écrite et orale -du cartel, du panneau explicatif, de la conférence- ou plus souvent de la confrontation avec les œuvres qui l'environnent ou avec l'espace ou encore le parcours. La présence d'une médiation écrite et/ou orale a longtemps fait débat : de Sandberg, le conservateur du Stedelijk Museum d'Amsterdam qui écrivait " « les catalogues ne servent qu'avant ou après une visite
c'est idiot de consulter le catalogue tout le temps pendant la visite d'une exposition
on ne jouit plus de l'objet que l'on voit
on jouit du fait qu'on sait ce que l'objet représente
heureux sont les myopes : leurs lunettes les
empêchent de regarder alternativement les objets
et les explications du guide "
à Bourdieu qui, au contraire, prônait le développement de la médiation écrite pour démocratiser la culture, le débat a été souvant houleux. Le parti d'une éducation du regard par une acculturation progressive se fonde sur l'idée d'une médiation qui doit s'effacer derrière l'œuvre. Au contraire, les formes de guidage dans la médiation numérique doivent être aussi évidentes que possible pour donner au visiteur en ligne les repères nécessaires pour naviguer; les formes de guidage dites intuitives se rapprochent du parcours muséal en cherchant à créer des surprises mais néanmoins elles prédéterminent les associations que le visiteur peut faire sans qu'il ait à s'interroger. Dans le passage d'une page à l'autre, d'un média à l'autre, ce qui prévaut, c'est, selon la formule de Benjamin, « l'esthétique de la distraction » : l'hyperdocumentation joue contre la documentation.
La fonction épistémique des institutions patrimoniales est par ailleurs mise à mal par une autre tendance suscitée par la technique, la fragmentation; rendre accessible le détail d'une œuvre par la recherche en f*ull text* ou par un zoom est considéré comme un progrès devant en favoriser l'appropriation. On se souvient que dans le débat suscité par la prégnance des moteurs de recherche, R.Chartier déplorait cette possibilité d'extraire un fragment ce qui, à ses yeux, jouait contre la compréhension globale d'une œuvre. De même, dans le domaine des arts plastiques, la possibilité de zoomer sur un détail peut être ambivalente: la surexposition peut aller à l'encontre du sens de l'œuvre, en favoriser une interprétation analytique et faire oublier son caractère de composition. Le goût du pixel va à l'encontre de la signification d'une toile impressionniste.

## 3 Collections et numérique

C'est la collection qu'elle abrite qui donne à chaque institution patrimoniale son identité et définit sa clôture informationnelle. Considérée comme le "trésor", elle est au cœur de l'institution mais le sens qui lui est attribué dépend notamment des conditions de sa présentation. Dans les musées, le sens dépend de son accrochage; à titre d'exemple, l'accrochage de la collection du Centre Pompidou sous le titre Elles a bousculé sinon les catégories artistiques, au moins les représentations du monde de l'art. L'accrochage qui est l'une des formes de médiation repose sur le traitement de l'espace, le parcours, ses séquences et les relations entre les œuvres.

Dans le passage au numérique, l'aspect "trésor" demeure : la collection numérisée est d'abord l'exposition du trésor, la "vitrine" de l'institution. La fonction vitrine", publicitaire de la banque d'images est renforcée du fait que les capacités de stockage sur le support numérique rendent visibles non seulement la collection exposée mais aussi les réserves. Ce que Walter Benjamin écrivait dans la deuxième version de son fameux texte « De l'œuvre d'art à l'ère de sa reproductibilité technique » -« on peut supposer que l'existence même de ces images a plus d'importance que le fait qu'elles sont vues »- peut s'appliquer à la banque d'images qui est la représentation de la collection. En ce sens, la capitalisation des images sur les sites web est partie prenante de l'identité du musée ; la banque de données n'est pas un simple auxiliaire. A côté de la collection réelle, physique, elle est un des éléments de la continuité mémorielle de l'institution, d'autant plus qu'elle intègre le rêve des réserves accessibles.

Mais dans ce complexe de pratiques qu'est un site web, la fonction de vitrine n'est qu'un usage possible ; elle est souvent renforcée par la possibilité technique donnée à l'amateur de réunir un choix de chefs d'œuvres jugés représentatifs de la collection; d'autres usages se superposent, notamment en matière d'inventaire et de documentation. Les systèmes d'indexation les plus courants (par artistes, titres, œuvres) qui prévalent dans les sites du MoMA (Museum of Modern Art de New York) ou du Centre Georges Pompidou à Paris supposent une recherche experte, menée par des spécialistes. Mais le dynamisme même du document numérique permet d'ouvrir ces usages selon " des échelles de conception dépendantes certes d'un moment technique et d'enjeux économiques, mais aussi d'appréciations différentes du rôle que l'indexation peut jouer dans l'appropriation des savoirs" : " D'un côté, elle sera confondue avec le catalogage, de l'autre, on la « ré-écrira » en vue d'ouvrir des accès toujours plus significatifs "(REGIMBEAU, 2007). Il se pose alors la question du partage de l'autorité avec des amateurs avertis (DUFRENE, et IHADJADENE 2010) : outre le problème de la certification de l'information par l'institution - qui explique, comme on l'a vu, la confiance dont elles jouissent- , le problème est celui des degrés de pertinence d'une indexation populaire : "l'évolution des critères de qualification de l'information va de pair avec une marchandisation accrue du lien social (web 2) et d'une monétisation du contexte (géo-référencement)". Pour autant, ces nouveaux modes d'accès peuvent produire des résultats inattendus : la possibilité de recourir à des axes descriptifs hétérogènes, la pluralité auctoriale, la plurifocalité au centre de ce nouveau régime de symbolisation peuvent avoir deux effets : d'une part, sur le plan de la description, créer des classifications croisées, d'autre part, formater une nouvelle relation avec l'internaute : cette possibilité de participation peut être aussi un instrument de fidélisation. Là est la force de la médiation numérique : formater la relation à la collection en favorisant la participation par la personnalisation de l'adresse.

Différentes ressources servent cette participation : d'abord le réinvestissement du concept de musée imaginaire. Le numérique permet de pallier les insuffisances de la collection : ce que le musée réel ne peut rassembler, le musée en ligne pourra le faire. De manière prophétique, Malraux avait annoncé : « *un musée imaginaire s'est ouvert qui va pousser à l'extrême l'incomplète confrontation imposée par les vrais musées* »[8]. La notion même de collection s'en trouve modifiée :elle n'est plus tributaire de l'histoire, de passions individuelles ou collectives, d'un lieu. Quand le Ministère de

8 *cf Malraux, Voix du silence, p 205-6*

la culture en France intitule une des rubriques de son site « musée imaginaire », il utilise le concept de Malraux pour signifier le rassemblement sur le support numérique de reproductions d'œuvres dispersées dans la réalité. Ce même concept est aussi utilisé pour impliquer l'internaute: nombre de sites invitent le visiteur à réaliser leur collection personnelle au sein de la collection.
Délocalisée, la collection se recompose quand elle est organisée par un concept qui en fait un tout intelligible et identifiable et non une simple agrégation d'éléments: la médiation numérique ouvre alors considérablement le champ des possibles d'une collection.
Outre la possibilité du musée imaginaire, la médiation numérique permet de pallier une autre insuffisance du musée physique : la décontextualisation des œuvres. La National Gallery a ainsi mis en place un programme "learn about art" qui s'adresse à tous et offre la possibilité de replacer une œuvre dans son contexte de production; la Tate Gallery offre de son côté le même service, d'autant que les archives sont numérisées, mais accorde en outre une large place au contexte de réception à travers la rubrique "collection blog". La curiosité du visiteur pour une œuvre ou pour la collection est stimulée.
Enfin, ce qui pouvait constituer une spécificité de la médiation dans le site physique, la place tenue par l'oralité et les formes de sociabilité qui y sont liées - d'autant que la plupart des visites ont lieu en groupe- est aujourd'hui largement battu en brèche : les médiations orales tiennent une place non négligeable : cours, conférences, podcasts ... Les sites deviennent des plates-formes d'enseignements et d'échanges, créant de nouvelles formes de liens. Les sites d'institutions qui offrent des liens avec sites de "partage" comme Flickr ou les réseaux sociaux type Facebook ne font que renforcer cette sociabilité. Ces liens formatent aussi une relation qui ne s'arrête pas à la médiation des savoirs : une enquête serait à mener pour appréhender leurs effets sur la politique commerciale de l'institution et sur sa politique des publics. La communication événementielle s'en trouve considérablement fortifiée : via les réseaux sociaux, des rendez-vous se mettent en place dans les espaces des institutions.
La médiation numérique donc, loin de détourner de l'institution, peut apparaître comme un moyen de former la compétence du public, de le familiariser aussi bien avec les œuvres qu'avec l'institution. L'enquête de l'ILMS fournit à ce propos un constat intéressant : les médiations, quelle qu'en soit la nature, se renforcent au lieu de s'exclure :
"*Most museum in-person and remote online visitors browse, view specific exhibits or collections, attend or view a lecture or class, complete a class assignment, among other activities.*"
On rejoint là le constat fait par O.Donnat : les pratiques culturelles se confortent au lieu de s'exclure.

## 4 Conclusion

Au final, considérer la médiation numérique dans les institutions patrimoniales, c'est voir la superposition (HENNION) des médiations. Les sites web de ces institutions constituent un complexe de pratiques qui ont pu se greffer -tout en les développant- sur des médiations existantes, notamment la médiation documentaire. Pour autant, les spécificités de la médiation numérique permettent à l'institution patrimoniale de pallier deux manques des collections : l'incomplétude et la décontextualisation. Loin de se réduire à une numérisation du patrimoine, elle permet de constituer des plate-formes dont les ressources sont suffisamment variées pour faire de cette médiation de masse une adresse néanmoins

personnalisée. Enfin un dernier trait de ce complexe de pratiques est son rapport à l'institution : non seulement le site web ne détourne pas le visiteur de l'institution physique mais il en accroit la fréquentation et les usages. Mieux : il donne à voir ce collectif, ses pratiques, son rapport à l'art ou à la culture. Si l'éditorialisation des collections désacralise l'œuvre d'art, en fait un objet trivial au sens d'Y.Jeanneret, un objet qui circule et se diffuse, en revanche, elle en permet l'appropriation, permettant de dépasser l'enceinte "intimidante" des institutions patrimoniales.

## 5 Bibliographie


1 DALBIN S. et GUYOT B, 2007 , "Documents en action dans une organisation : des négociations à plusieurs niveaux", Entre information et communication, les nouveaux espaces du document , *Etudes de communication* n°30,

2 DAVALLON J.,2007, *Le don du patrimoine*, Paris, L'Harmattan

3 DUFRENE B.,2010, « Le concept de *musée imaginaire* et la question de la reproduction », In Actes du colloque « André Malraux et l'Occident», Akita (Japon)

4 DUFRENE B, IHADJADENE M (2010) « l*e web 2 ou une nouvelle entrée des amateurs dans les institutions culturelles »* Médiations documentaires: entre réalités et imaginaires. Journée scientifique internationale du réseau MUSSI. 15 Mars 2010.Université d'Avignon et des pays du Vaucluse.

5 FALGUIERE P., 1996, "les raisons du catalogue", in *Les Cahiers du MNAM*, p5-19

6 IHADHADENE M & FAVIER L., 2008 « Langages documentaires : vers une crise de l'autorité » *Sciences de la société* n°75, p 11-22

7 JEANNERET Y., 2001, "Les politiques de l'invisible : du mythe de l'intégration à la fabrique de l'évidence", *Document numérique*, Hermès sciences publications, vol.5 n°&-2, 2001

8 JEANNERET Y., 2004, "Forme, pratique et pouvoir.Réflexions sur le cas de l'écriture, Sciences de la société, n°63

9 JEANNERET Y., 2008, *La vie triviale des êtres culturels*, Paris, éditions Hermès-Lavoisier

10 PEROUSE de MONTCLOS, "La description" in Nora P.(dir), 1997,*Science et conscience du patrimoine*, Paris, Fayard, p.193-197

11 PEDAUQUE R, 2007, *La redocumentarisation du monde*, Eds. Cépadues, 212 p.

12 QUERE L., 2000, "Qu'est-ceau juste que l'information?", Communiquer à l'ère des réseaux, *Hermès*, n°100

13 REGIMBEAU G., 2005, "Cas et figures en indexation de l'art contemporain" In Timini(Ismaïl), Kovacs (Susan) (dir.), *Indice, index, indexation,* , Paris, ADBS éditions, p. 95-104.

14 SALAUN J.M.,"La redocumentarisation, un défi pour les sciences de l'information", *Etudes de communication*, n°30, p13-23.

15 SCHAER R., 1994, »Des encyclopédies superposées », *La jeunesse des musées,* Paris, RMN, p.38-51

16 TRANT J, 2009 : "Tagging, Folksonomy and Art Museums: Early Experiments and Ongoing Research". In: *Journal of Digital Information* 10 (2009)

17 VIDAL G., 2008, *Les musées à l'heure du Web 2.0 : nouveaux usages de l'interactivité et évolutions des relations avec les publics* ? http://doc.univ-paris8.fr/je/2008/22mai/

18 WELGER BARBOZA C., 2001, *Le patrimoine à l'ère du document numérique*, Paris, L'Harmattan

# Étude comparative de moteurs de recherche pour le repérage d´images illustrant des objets muséaux

**Elaine Ménard**

elaine.menard@mcgill.ca
*School of Information Studies, McGill University*

**Résumé.** De nombreux facteurs peuvent confondre le chercheur qui tente de repérer des images sur le Web. Le repérage d´images demeure complexe pour la majorité des individus, surtout pour les images associées à un support textuel rédigé dans une langue inconnue des utilisateurs. Cette étude exploratoire compare les comportements de recherche d´individus utilisant différents moteurs de recherche afin de repérer des images illustrant des objets de musée. Elle explore également les fonctionnalités multilingues de recherche mises à la disposition des internautes lors de l´exécution de tâches de repérage d´images. La principale contribution de cette étude pilote est d´améliorer la connaissance et la compréhension du comportement de recherche d´images, de manière à établir une base pour la modélisation d´une nouvelle interface de recherche prenant en compte les besoins et les attentes réels des chercheurs d´images.

**Mots-clés.** Repérage d´images, moteurs de recherche, multilinguisme, comportements de recherche, objets muséaux

## 1 Introduction

La croissance d´Internet a mis en lumière le besoin pressant de développer des outils pour décrire les images afin de faciliter leur repérage puisque celles-ci enrichissent la plupart des ressources disponibles sur le Web : pages personnelles, collections d´objets de musées, bibliothèques numériques, catalogues de produits commerciaux et de services, services d´information gouvernementale, et ainsi de suite. De nombreux facteurs font obstacle à la recherche d´images sur le Web dont la surabondance des images disponibles, de même que leur indexation avec un vocabulaire souvent incompréhensible ou trop spécialisé pour être utile. En outre, le chercheur d´images dispose de moteurs de recherche offrant des fonctionnalités bien conçues, mais qui ne sont pas spécifiquement adaptées à ses comportements de recherche. Trop souvent, les besoins réels des chercheurs d´images sont négligés par les concepteurs de moteurs offrant la recherche d´images.
Par conséquent, afin de minimiser les frustrations et les difficultés susmentionnées, cette recherche propose d´examiner les comportements et habitudes de recherche d´individus devant repérer des images illustrant des objets muséaux. En première partie, le contexte général de la recherche est présenté.

Nous décrivons ensuite la méthodologie utilisée pour cette étude, de même que les principaux résultats de la recherche. En guise de conclusion, nous réfléchissons sur les observations mises en relief par cette étude et élaborons sur les pistes de recherche que nous comptons explorées pour faire suite à ce projet.

## 2 Contexte de recherche

Comme beaucoup l´ont sans doute remarqué, le Web est une source formidable pour le chercheur d´images, que ce soit pour la recherche de matériel d´illustration pour un exposé oral ou simplement pour décorer ou embellir un fond d´écran. Tel que constaté dans nos travaux précédents [1], repérer les images désirées ne s´avère pas toujours un processus facile et sans faille. De nombreuses difficultés ont tendance à compliquer le processus de repérage. Selon Jansen [2], le premier obstacle vient de la « traduction » de la représentation visuelle d´un objet dans une description textuelle. Par ailleurs, compte tenu de la possibilité d´interprétations multiples de la ressource visuelle, Turner [3] souligne qu´il existe un risque sérieux d´ambiguïté et d´erreur. En d´autres termes, les chercheurs d´images ne formulent pas nécessairement leurs requêtes avec les mêmes concepts ou les mêmes mots que ceux choisis au moment de l´indexation manuelle ou automatique de l´image. La deuxième difficulté découle de la grande diversité linguistique que l´on retrouve sur Internet, ce qui signifie que le texte associé à une image est susceptible d´être disponible dans de nombreuses langues différentes. Les chercheurs d´images sont ainsi confrontés à un double défi lorsque vient le temps de repérer des images à l´aide d´une requête textuelle. Premièrement, les termes de la requête doivent correspondre au texte associé à l´image. Deuxièmement, la langue de la requête doit correspondre à la langue du texte associé avec l´image. Par conséquent, le choix de termes d´indexation appropriés pour décrire des images est déterminant pour assurer leur repérage.

Au fil des ans, le processus d´indexation de l´image a fait l´objet de plusieurs études clés. Panofsky [4] a identifié trois niveaux de sens dans les œuvres d´art: pré-iconographique pour la matière primaire ou naturel, iconographique pour les sujets secondaires et iconologique pour le contenu intrinsèque de l´œuvre. Quelques années plus tard, Markey [5] a appliqué ces niveaux à l´identification de thèmes ou de concepts illustrés dans les images. Shatford [6] a pour sa part défini trois groupes d´attributs, « spécifiques de », « générique de » et « à propos de », qui correspondent aux trois niveaux de Panofsky [4]. De la même manière, Krause [7] divise les informations contenues dans une image en « indexation dure » (ce qui peut être observée dans une image) et « indexation douce » (signification subjective et interprétation personnelle de ce qu´elle évoque). Layne [8] suggère que l´image pourrait être « de » et « sur » quelque chose. Alors que « ofness » est rigide et objective, « aboutness » est plus abstraite et subjective. La sélection de points d´accès appropriés, l´indexation et le repérage d´images ont été étudiés sous divers angles (Turner [3], [9]; Armitage et Enser [10]; Jörgensen [11], [12]; Markkula et Sormunen [13]; Goodrum et Spink [14]; Choi et Rasmussen [15], [16]; Enser [17]; Greisdorf et O´Connor [18]; Ménard [19]). Tel que démontré par plusieurs études (Goodrum & Spink [14]; Goodrum, Bejune et Siochi [20]; Jörgensen et Jörgensen [21]; Tjondronegoro et Spink [22]), on remarque une évolution dans la manière dont les requêtes sont formulées lors du processus de repérage d´images. Par exemple, les requêtes composées d´un seul terme sont moins fréquentes qu´auparavant. Cette mutation conduit inévitablement à une réévaluation de la façon dont l´image doit être traitée. En outre, elle laisse supposer que les moteurs

de recherche, traditionnellement employés pour le repérage d´images et pour le furetage, ne sont peut-être pas adaptés à tous les types d´images. L´examen du comportement du chercheur d´images Web paraît donc indispensable pour améliorer les systèmes de repérage. De la même manière, il est crucial de poursuivre l´étude des stratégies de recherche des chercheurs d´images afin de développer de meilleurs moteurs de recherche, c´est-à-dire des outils adaptés réellement aux besoins et comportements actuels des chercheurs d´images.

La revue de littérature menée pour cette recherche révèle deux élément importants. D´une part, peu d´études sur les fonctionnalités de recherche dédiées au repérage d´images dans un contexte de recherche multilingue ont été menées jusqu´ici. D´autre part, le repérage d´image tient une place importante dans les activités des internautes puisque nous trouvons des images dans la majorité des ressources disponibles sur le Web. Parmi les nombreux types d´images disponibles sur Internet, les images illustrant des objets de musée génèrent de nombreuses recherches sur le Web. Par images d´objets de musée, nous entendons des images représentant des objets que l´on trouve dans un musée et, par extension, sur un site Web de musée. L´accès aux sites de musée a fait l´objet de plusieurs études (Müller [23]; Cameron [24]; Knell [25]; Herman, Johnson et Ockuly [26]; Marty [27]). Toutefois, l´examen de la littérature met en relief une lacune fort importante. En effet, peu de choses sont actuellement connues sur les fonctionnalités de recherche utilisées pour le repérage d´images d´objets muséaux. De plus, comme les images offertes par les musées sont souvent indexées en diverses langues, il convient de s´interroger sur la facilité de repérage de ce type d´images dans un contexte de recherche multilingue, c´est-à-dire lorsque la langue de requête est différente de la langue d´indexation. Les organisations de musée avec une présence sur le Web reconnaissent d´ailleurs la nécessité d´offrir un contenu multilingue pour leurs auditoires. Par exemple, MINERVA (2006)[1] a réalisé un projet de recherche ayant pour objectif d´effectuer une vaste enquête destinée à obtenir un aperçu de la situation concernant l´usage des langues sur les sites culturels. Les résultats de ce projet soulignent d´ailleurs la nécessité pour les musées d´être en mesure d´offrir des informations sur les contenus disponibles dans différentes langues afin de rejoindre davantage de visiteurs en provenance de différents pays. De nombreux musées reconnaissent ainsi la nécessité non seulement de préserver les différences locales et nationales, mais aussi de mettre à la disposition des visiteurs virtuels de toutes origines l´information disponible en plusieurs langues.

Le but de ce projet exploratoire est d´étudier les habitudes des chercheurs d´images et examiner comment les individus formulent leurs requêtes pour repérer des images d´objets de musée indexées dans des langues différentes. Cette étude compare également les comportements de recherche d´individus devant utiliser différents moteurs de recherche. Elle vise à mettre en relief les fonctionnalités de recherche multilingue, c´est-à-dire la manière dont celles-ci sont perçues et utilisées lors de l´exécution d´une tâche de recherche d´images dans un contexte de recherche multilingue. L´apport principal de cette étude pilote est d´améliorer la connaissance et la compréhension du comportement de recherche d´images, afin de fournir une base pour la modélisation d´une nouvelle interface de recherche prenant en compte les besoins et attentes des chercheurs d´images et cela, peu importe le contexte linguistique.

[1] Multilingual access to the digital European cultural heritage.
<http://www.minervaeurope.org/structure/workinggroups/inventor/multilingua/documents/Multiligualism_v1_printed.pdf.>

## 3 Méthodologie

### 3.1 Participants

Pour cette recherche, un échantillon non probabiliste de trente participants, où les éléments de la population ont été choisis en raison de la corrélation entre leurs caractéristiques et les objectifs de la recherche, a été utilisé. Avec ce type d´échantillon, il est possible à la fois d´accroître l´utilité de l´information et de limiter le nombre de sujets. En outre, il s´agissait d´un échantillon volontaire puisque que chaque participant devait prendre un rendez-vous pour participer à l´expérience. Les participants ont été recrutés par une publicité affichée sur le site Web de l´Université McGill. Une compensation monétaire de 20,00$ (canadiens) a été attribué à chaque répondant choisi pour l´expérience. La collecte des données a été réalisée dans une période relativement courte, du 30 mai au 12 juin 2009, pour contrer l´effet de contamination des données.

Afin d´explorer les comportements des chercheurs d´images, de même que leurs besoins liés à l´accès multilingue aux images d´objets de musée, l´échantillon de trente participants a été réparti au hasard en trois groupes indépendants correspondant à un des moteurs de recherche utilisés pour cette étude. Pour des considérations éthiques, les participants étaient tous âgés d´au moins 18 ans. En outre, pour assurer l´homogénéité du groupe de participants, et étant donné la nature des tâches à accomplir durant l´expérience, les participants ne devaient avoir aucune expérience professionnelle dans un domaine touchant à l´indexation ou le repérage d´images. Chaque participant devait répondre à ces critères d´inclusion minimale pour être sélectionné. Ces conditions ont permis de contrôler le biais pouvant provenir de participants hétérogènes et les variables étrangères, une variable étrangère étant une variable qui se présente en dehors de la volonté du chercheur et qui a souvent un effet inattendu sur la variable dépendante, risquant ainsi de fausser les résultats de l´étude. Cependant, nous sommes également consciente que la taille de l´échantillon, ainsi qu´une trop grande homogénéité limitent la généralisation statistique des résultats à la seule catégorie de participants retenue pour notre recherche.

### 3.2 Images

Pour cette étude, trois catégories d´objets muséaux (décoratifs, vestimentaires, utilitaires) ont été sélectionnées. À partir de chaque catégorie, une image a ensuite été choisie. Des images illustrant un objet de musée spécifique et offrant une bonne qualité visuelle ont été identifiées en vue de simplifier le processus de repérage d´images par une majorité de personnes. Enfin, il a été vérifié qu´une image de chacune des catégories était disponible dans chaque moteur de recherche utilisé pour l´expérience (Figure 1).

**Figure 1**. *Images à repérer*

### 3.3 Moteurs de recherche

Pour cette étude, deux types d´outils de recherche ont été utilisés. Premièrement, les trente participants ont tous utilisé Google Images[2], un moteur de recherche « multi-usages » afin d´accomplir les trois tâches de repérage prévues. Puis, trois groupes de dix participants ont été aléatoirement définis et affectés à un moteur de recherche spécifiquement dédié à l´accès aux images illustrant des objets de musée : Musée McCord[3], Artefacts Canada[4] et Europeana[5]. Ces outils de recherche offrent la possibilité de rechercher les collections d´objets muséaux, y compris les tableaux, artefacts et autres objets d´art, le tout disponible en format numérique. Les quatre outils (multi-usage et spécialisés) ont également été sélectionnés pour cette recherche parce qu´ils proposent des fonctionnalités de recherche multilingue.

### 3.4 Collecte de données

Initialement, tous les participants ont été invités à formuler la meilleure requête qu´ils pouvaient imaginer pour repérer les trois images montrées consécutivement, en utilisant l´outil de recherche Google Images. Pour chaque image, chaque participant disposait de cinq minutes, puis l´image suivante devant être repérée leur était présentée. Après avoir complété les trois tâches de repérage, les participants ont été invités à remplir un questionnaire sur leur appréciation générale de l´outil de recherche d´images. Ensuite, chaque participant a été assigné aléatoirement à un second moteur de recherche (Musée McCord, Artefacts Canada ou Europeana) et invité à repérer les trois mêmes images, dans le même ordre de présentation. Après avoir utilisé le second moteur de recherche, les participants ont à nouveau été conviés à remplir un questionnaire de manière à recueillir leurs observations générales sur l´outil de recherche utilisé. Finalement, les répondants devaient répondre à un questionnaire général comportant des questions ouvertes et fermées sur les tâches de repérage effectués et leurs comportements de recherche.

## 4 Résultats

### 4.1 Habitudes de recherche

En général, les images sont repérées et utilisées dans des contextes multiples. Pour en apprendre plus sur le rôle des images dans la vie quotidienne des participants, ces derniers ont été invités à énumérer pour quelles tâches le repérage d´images leur est habituellement nécessaire. Tel qu´illustré la figure 2, la majorité des répondants (18) a indiqué considérer les images comme une source d´information générale. Les images sont également utiles pour les arts et l´artisanat (9) et pour le travail (8). En outre, les images sont utilisées pour les loisirs personnels (5) et le réseautage (5). Étrangement, le repérage d´images ne semble pas être un élément déterminant lors des transactions commerciales des répondants (1).

La figure 3 présente une liste des catégories d´images que les internautes recherchent dans différentes tâches quotidiennes qu´ils accomplissent. Les images d´art (7), les images médicales (7), les images en lien avec le cinéma et les acteurs (7) et les images liées aux nouvelles technologies dominent les recherches des internautes, tandis que les images d´objets à acheter (1), de même que les images

[2] Google Images. <http://images.google.ca>
[3] Musée McCord. <http://www.mccord-museum.qc.ca/fr/clefs/collections>
[4] Artefacts Canada. <http://www.pro.rcip-chin.gc.ca/bd-dl/artefacts-fra.jsp>
[5] Europeana. < http://europeana.eu/portal/>

se rapportant au domaine du voyage et du tourisme (1) sont rarement mentionnées par les répondants.

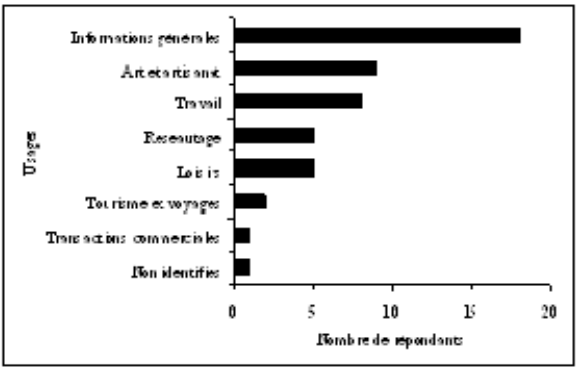


**Figure 2.** Utilisations des images dans la vie quotidienne des participants

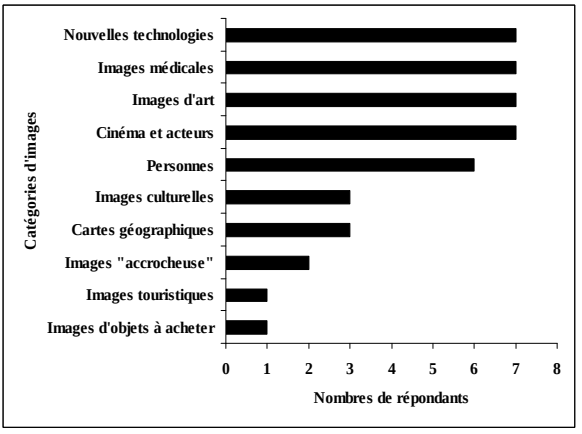


**Figure 3**. Catégories d´images recherchées

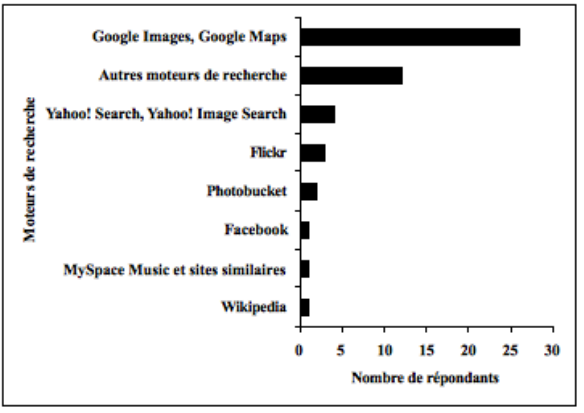


**Figure 4** - Moteurs de recherche habituels

### 4.2 Comportements et méthodes de recherche

Interrogés sur les moteurs de recherche qu´ils utilisent généralement pour le repérage d´images, les outils de recherche de la famille Google (notamment Google Images et Google Maps) ont été mentionnés par la majorité des répondants (26). Yahoo arrive en deuxième place (4), suivi par Flickr (3). Les autres moteurs de recherche sont rarement évoqués par les répondants (Figure 4).

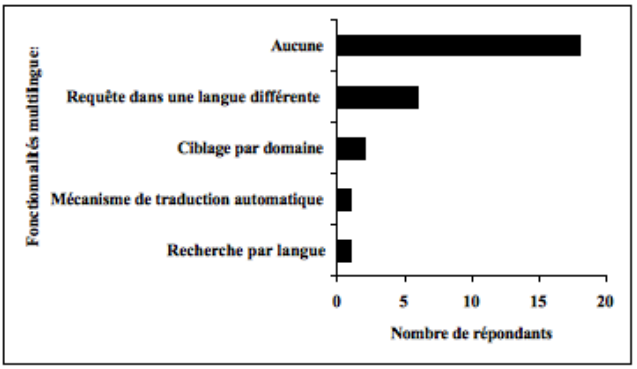


**Figure 5.** *Méthodes de recherche*

Les répondants ont décrit la manière avec laquelle ils recherchent généralement des images (Figure 5). La recherche par mots-clés est la méthode de recherche préférée par la majorité des participants (28). La recherche à l´aide d´un titre ou d´un nom de personne arrive respectivement en deuxième position (7). Étonnamment, seuls quelques répondants confirment l´utilisation des fonctionnalités de recherche avancée lors du repérage d´images (5).

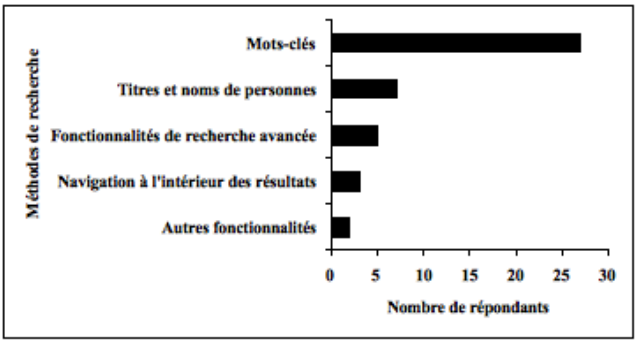


**Figure 6**. Fonctionnalités multilingues

Concernant les fonctionnalités offertes pour faciliter le repérage en contexte multilingue, la majorité des participants (18) affirme ne pas utiliser ces fonctions (Figure 6). Néanmoins, afin de trouver des images indexées avec des langues différentes, certains participants (6) ont indiqué qu'ils formulent régulièrement leurs requêtes dans une langue différente de leur langue maternelle.

Finalement, lorsqu´on leur demande quels critères sont importants lors du choix d´une image, la majorité de participants (16) indique que la reconnaissance visuelle est le critère qu´ils utilisent le plus souvent. D´autres critères divers ont aussi été

mentionnés par les répondants, tel que la qualité (8), la taille et type (6), la similarité (5), la résolution (4) et la couleur (3) (Figure 7).

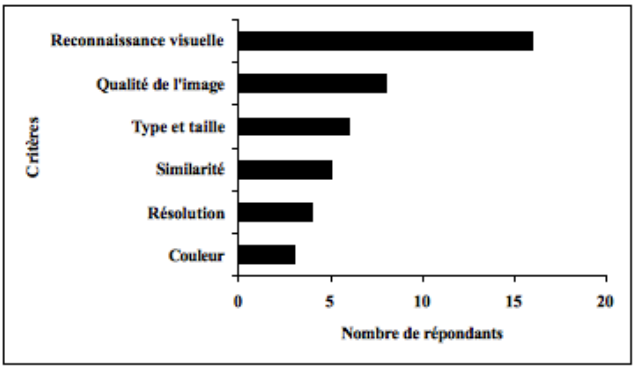


**Figure 7.** Critères de sélection des résultats

### 4.3 Évaluation des moteurs de recherche

***Utilisabilité***

La comparaison entre le nombre moyen d´images repérées par les répondants pour chaque moteur de recherche révèle que le moteur de recherche du Musée McCord a obtenu les meilleurs résultats (une moyenne de 1,9 image par participant), suivi de Google Images (une moyenne de 1,1 image par participant). Les participants associés à Artefacts Canada et Europeana ont respectivement répéré 0,4 image en moyenne.

Une fois les tâches de repérage terminées, les participants ont été invités à évaluer quelques affirmations en ce qui concerne le moteur de recherche utilisé. À l´affirmation « En général, il est facile de récupérer les images exactes », 70 % des participants ayant utilisé Google ou les moteurs de recherche du Musée McCord soulignent être d´accord ou fortement d´accord. Toutefois, les participants ayant utilisé le moteur Artefacts Canada sont majoritairement en désaccord ou fortement en désaccord avec cette affirmation, tandis que 90 % des participants ayant effectué le repérage d´images avec Europeana sont fortement en désaccord.

En termes d´efficacité, les répondants ont été appelés à évaluer dans quelle mesure ils sont capables d´accomplir les tâches de repérage d´images dans un court laps de temps et avec un effort minimal. Selon les données recueillies, 80 % des participants jumelés à Europeana indiquent être en désaccord ou fortement en désaccord avec l´affirmation « En général, il est possible d´accomplir les tâches de repérage d´images rapidement », tandis que les participants ayant effectué les tâches de repérage avec les trois autres moteurs de recherche sont en désaccord ou fortement en désaccord dans une proportion de 60 %.

En ce qui concerne l´effort requis pour accomplir les tâches de repérage, les participants ayant utilisé le moteur de recherche du Musée McCord confirment dans une proportion de 40 % être en accord avec l´affirmation « En général, il est possible d´accomplir les tâches de repérage d´images avec un minimum d´effort ». Toutefois, les participants associés aux autres moteurs de recherche sont majoritairement en désaccord ou fortement en désaccord avec cette même affirmation. Enfin, en termes de satisfaction des participants, 90 % des participants ayant utilisé le moteur de recherche du Musée McCord se disent d´accord ou fortement d´accord avec l´affirmation « En général, je suis satisfaits

des images repérées », alors que 90 % des participants jumelés au moteur Europeana sont en désaccord ou fortement en en désaccord avec cette affirmation.

**Moteurs de recherche**

**Google Images**

Les répondants ont également été interrogés sur les fonctionnalités de recherche offertes par les moteurs de recherche qu´ils ont utilisés. Parmi les options disponibles dans Google Images, la majorité des participants a mentionnée la « Recherche par contenu », tandis que les autres fonctionnalités n´ont pas été réellement utilisées par la plupart des participants. En outre, les fonctionnalités de recherche multilingue proposées par ce moteur de recherche n´ont pas vraiment été utilisées. Par ailleurs, les répondants ont également mentionné quelques options jugées essentielles pour les tâches de repérage d´images, par exemple la possibilité de modifier, ajouter ou soustraire des termes de recherche. En outre, une majorité de participants estime que Google Images est très convivial, facile et simple à utiliser, ainsi que relativement rapide. Un autre avantage mis en relief par plusieurs participants est le grand nombre de résultats de recherche obtenus.
Les répondants ont également décrit les problèmes rencontrés lors de la recherche avec Google Images. Par exemple, ceux-ci ont remarqué que s´ils écrivent seulement quelques mots-clés, le nombre de résultats affichés est très élevé, ce qui signifie que les répondants doivent souvent passer en revue une quantité importante de résultats avant de pouvoir identifier l´image qu´ils doivent repérée. En d´autres termes, les participants affirment être parfois découragés par le manque de précision des images affichées, ce qui signifiait qu´ils avaient besoin de trier de grandes quantités de résultats, y compris des images totalement étrangères à leurs requêtes.
Interrogés sur les limites du moteur de recherche, les participants ont souligné le fait que souvent, ils ne pouvaient correctement décrire avec précision l´image montrées devant être repérée, une conséquence de leur propre manque de connaissances sur les objets muséaux représentés. Autrement dit, le fait que les participants ne connaissaient pas nécessairement les objets illustrés conduit à l´incertitude quant à la façon de formuler les requêtes appropriées ou des requêtes plus spécifiques menant au repérage. Une majorité de participants ne pouvait pas trouver les mots pour nommer ou décrire les images qu´ils cherchaient, afin de différencier ces images à partir des images affichées.
Quelques participants ont également mentionné que la possibilité de rechercher des images ou d´affiner les requêtes de recherche en utilisant la forme et la couleur de base de l´objet représenté est indispensable au succès du repérage.

**Artefacts Canada**

Lors de l´utilisation d´Artefacts Canada, les fonctionnalités les plus utilisées pour effectuer les tâches de repérage ont été les fonctions « Mots-clés », « Nom d´objet » et « Sujet de l´objet ». Les autres fonctions offertes par Artefacts Canada ont été ignorées par la majorité des participants. La plupart des répondants ont également mentionné n´avoir utilisé aucune fonctionnalité multilingue de recherche offerte par ce moteur de recherche. Parmi les options de recherche

préférées, la majorité des participants considère que le « Nom de l´objet » et « Matériel de l´objet » s´avèrent les plus avantageuses.
Lorsque interrogés sur les avantages offerts par Artefacts Canada, la majorité des participants observe que ce moteur de recherche est surtout utile pour les personnes qui connaissent déjà des détails spécifiques au sujet des objets muséaux et constitue un bon outil pour ceux qui connaissent exactement la nature de l´image à repérer. Quelques participants ont également souligné que les fonctions de recherche offertes par Artefacts Canada rendent le repérage relativement facile. Enfin, la majorité des utilisateurs mentionne que les différents menus de repérage offrant diverses catégories prédéterminées les ont obligés à formuler leurs requêtes de manière différente des requêtes qu´ils auraient formulées avec les outils de recherche habituels de type « Google ».
En ce qui concerne les limites, la majorité des participants souligne la difficulté d´identification de certains éléments tels que la période, le créateur ou le lieu d´origine de l´objet muséal illustré par l´image devant être repéré. Ce manque de connaissances sur les objets muséaux limite la capacité d´utilisation de certains filtres de recherche, comme l´origine, la période, la fonction, la technique, la matière, l´artiste et ainsi de suite. En d´autres termes, le manque d´information au sujet de l´objet muséal à repérer a été frustrant pour les utilisateurs qui avaient souvent l´impression de ne pas être en mesure de décrire explicitement et de façon adéquate l´image à repérer, et d´utiliser de manière satisfaisante les filtres de recherche détaillés offerts par Artefacts Canada. En guise de conclusion, quelques participants ont fait observer que la possibilité de chercher avec l´élément « couleur » serait sans doute fort utile, car elle entraînerait moins de conjectures concernant certains critères de recherche tels que « période » ou « créateur » de l´objet muséal.

**Musée McCord**

Parmi les diverses options de recherche offertes par le moteur de recherche du Musée McCord, la majorité des utilisateurs a mentionné avoir utilisé les fonctionnalités « Nom de l´objet », « Titre » et « Lieu ». En outre, les répondants ont majoritairement indiqué ne pas avoir eu recours aux autres fonctionnalités disponibles. À l´instar des autres moteurs, les fonctionnalités de recherche multilingue disponibles n´ont pas été utilisées par la majorité des personnes interrogées.
En ce qui concerne les fonctionnalités de recherche jugées utiles, une majorité des répondants a fait observer que la plupart des options de recherche sont détaillées et pertinentes à la recherche d´images. Les participants ont également mentionné que la recherche avec le « Nom de l´objet » est très utile parce que ce musée ne possède souvent qu´un nombre limité d´images pour chaque type d´objet. Ainsi, quelques participants estiment également que la collection du Musée McCord est si petite qu´il n´est pas si difficile de trouver des objets spécifiques. Enfin, la majorité des participants a indiqué que ce moteur de recherche est facile à utiliser et très utile pour effectuer le repérage d´images. Néanmoins, quelques participants ont estimé qu´il y avait trop de choix de fonctionnalités de recherche, certains d´entre elles étant trop spécifiques pour quelqu´un qui ne sait pas exactement comment décrire les objets recherchés. Ainsi, certains participants ont aussi eu l´impression que les catégories prédéterminées étaient trop spécifiques pour être utiles.

**Europeana**

La majorité des participants ayant effectué les tâches de repérage avec Europeana mentionne qu´elles ont surtout utilisé la fonctionnalité « domaine » et n´ont pas vraiment utilisé les autres fonctionnalités de recherche disponibles. En ce qui concerne les fonctionnalités de recherche multilingue, la possibilité de rechercher par « langue » ou « pays », de même que la possibilité de sélectionner une interface de recherche dans différentes langues ont été rarement mentionné par les répondants. Toutefois, les répondants ayant accompli leurs tâches de repérage et fait usage de la fonctionnalité par « langue » considèrent que celle-ci contribue au succès du repérage. Quelques participants ont fait observer que la possibilité de rechercher des objets muséaux avec l´option « œuvre d´art européen par l´artiste » était un bon avantage offert par le moteur. La possibilité de réduire le nombre de résultats affichés par leurs requêtes avec une période historique est également considérée comme une fonctionnalité avantageuse par plusieurs participants. En ce qui concerne les problèmes éprouvés lors du repérage, les répondants ont identifié quelques défauts du moteur de recherche. Par exemple, la majorité des participants souligne que la recherche avancée ne donne souvent aucun résultat ou encore, affiche des résultats non pertinents. Dans de nombreux cas, les participants ont noté que les images ne correspondaient pas à leurs requêtes, même dans le cas de requêtes simples. En outre, les participants indiquent qu´ils n´ont pas été en mesure d´identifier une solution à l´absence de résultats affichés, les menant à penser que la base de données utilisée par ce moteur de recherche est assez limitée.

## 5 Discussion et conclusion

Les résultats fragmentaires présentés précédemment illustrent parfaitement l´omniprésence de l´image dans la vie quotidienne, de même que le degré de perplexité et d´incertitude qui caractérise toujours l´internaute au moment de repérer des images. Plusieurs facteurs concourent à complexifier davantage le processus de repérage, tels que la spécificité de l´image illustrant des objet muséaux et leur indexation en diverses langues. En outre, les outils de recherche offrant la possibilité d´effectuer le repérage d´images ne sont peut-être pas aussi conviviaux et efficaces qu´on pourrait le souhaiter. Cette étude exploratoire cherche à mieux comprendre les nombreux défis lancés par l´interaction entre les systèmes d´information et les individus et est fondée sur un paradigme qui intègre les aspects technologiques et sociaux.

Les résultats de cette recherche constituent un tremplin pour la conception d´une interface de recherche et de furetage destinée aux grandes collections d´images. Les pratiques actuelles dans la conception de ces interfaces ne sont pas coordonnées, et aucun cadre commun n´existe réellement. Ainsi, l´étude pilote présentée précédemment constitue la première étape d´un programme de recherche menant au développement d´une interface de recherche dédiée au repérage d´images. Cette nouvelle interface de recherche est destinée à être un outil novateur et puissant pouvant être utilisé par des individus cherchant toutes sortes d´images, y compris des images artistiques (pièces de musée, oeuvres célèbres, etc.), des images documentaires liés à un domaine particulier (sports, actualités, imagerie médicale, etc.) et des images ordinaires (images non artistiques occupant une place importante dans les recherches des internautes). Basée sur les résultats préliminaires de cette étude sur les comportements des chercheurs

d´images et les commentaires reçus par ceux-ci, l´interface de recherche proposée devra prendre en compte les principes fondamentaux suivants : cohérence des éléments et de style, processus conviviaux et flexible de furetage et de repérage, groupement logique des fonctionnalités et hiérarchisation des éléments visuels. L´interface de recherche devra également être aussi intuitive que possible afin de faciliter la réalisation des tâches de repérage d´images et de minimiser les facteurs susceptibles de diminuer la performance du repérage. Le nombre de fonctionnalités de recherche disponibles, ainsi que la structure générale de l´interface seront réduits au minimum afin d´éviter la frustration et d´accroître le degré de satisfaction des utilisateurs éventuels. L´interface de recherche sera organisée en une structure cohérente et intuitive pouvant être utilisée à la fois par les chercheurs d´images néophytes ou chevronnés.

L´élaboration de la nouvelle interface de recherche est réalisée en parallèle avec le développement d´une taxonomie bilingue pour l´indexation d´images numériques, un projet financé par une subvention de recherche (2010-2013) du Fonds Québécois de la Recherche sur la Société et la Culture (FQRSC). La taxonomie fournira aux chercheurs d´images des points d´accès flexibles et novateurs pour les images ordinaires. La modélisation de la nouvelle interface est également la prochaine étape logique pour le projet de recherche subventionné par l´Université McGill (2008-2010). Cette recherche, en cours d´exécution, vise à étudier les facteurs qui influençant les comportements de recherche des chercheurs d´images de différents groupes linguistiques (français, anglais, russes et chinois) et à examiner la manière dont ils formulent leurs requêtes au moment de repérer des images.

Finalement, l´interface que nous proposons de développer facilitera le repérage d´images indexées avec des langues différentes. En effet, jusqu´ici le repérage d´images se fait presque exclusivement en contexte monolingue, c´est-à-dire que les images repérées sont indexées dans la langue correspondant à la langue de la requête formulée pour leur repérage. Puisque l´interface est destinée à être un outil bilingue, nous proposons d´inclure un mécanisme de traduction qui permettra aux utilisateurs éventuels de formuler leurs requêtes dans la langue de leur choix et de récupérer des images indexées dans cette langue. Dans un premier temps, l´interface de recherche et les fonctionnalités de recherche sélectionnées, dont le mécanisme de traduction des requêtes, seront modélisées simultanément en français et en anglais. Par la suite, d´autres langues seront graduellement intégrées. En effet, le multilinguisme joue un rôle stratégique dans la qualité et l´efficacité des services proposés sur Internet. Les fonctionnalités essentielles nécessaires menant au succès du repérage d´image auront pour objectif d´accroître l´efficacité et l´efficience du repérage, en contexte de repérage multilingue. Par conséquent, afin de rendre l´information accessible au plus large public possible il devient essentiel de surmonter les barrières linguistiques en proposant des outils adaptés aux besoins réels des chercheurs de l´image. L´interface de recherche multilingue proposée constituera un avantage certain pour les chercheurs d´images qui ne sont pas toujours familiers avec les images indexées en diverses langues, dont l´anglais qui demeure la langue dominante du Web, en leur donnant la possibilité d´accéder facilement à des ressources visuelles variées, peu importe leur provenance.

## 6 Bibliographie


[1] E. Ménard. Images: indexing for accessibility in a multi-lingual environment – challenges and perspectives, *The Indexer*, 27(2), 70-76. 2009

[2] B. Jansen. Searching for digital images on the web, *Journal of Documentation*, 64(1), 81-101. 2008

[3] J. Turner. *Determining the subject content of still and moving image documents for storage and retrieval: an experimental investigation*. Thèse de doctorat. Université de Toronto, Toronto. 1994

[4] E. Panofsky. *Meaning in the visual arts: papers in and on art history*. Doubleday, Garden City, NY. 1955

[5] K. Markey. Access to iconographical research collections, *Library Trends*, 2(fall), 154-174. 1988

[6] S. Shatford. Analyzing the subject of a picture: a theoretical approach, *Cataloging & Classification Quarterly*, 6(3), 39-61. 1986

[7] M. G. Krause. Intellectual problems of indexing picture collections, *Audiovisual Librarian*, 14(4), 73-81. 1988

[8] S. S. Layne. Some issues in the indexing of images, *Journal of the American Society for Information Science*, 45(8), 583-588. 1994

[9] J. Turner. *Images en mouvement : stockage, repérage, indexation*. Presses de l´Université du Québec, Sainte-Foy. 1988

[10] L. H. Armitage et P. G. B. Enser. Analysis of user need in image archives, *Journal of Information Science*, 23(4), 287-299. 1997

[11] C. Jörgensen. Attributes of images in describing tasks, *Information Processing & Management*, 34(2/3), 161-174. 1998

[12] C. Jörgensen. *Image retrieval – theory and research*. Scarecrow Press, Lanham, Md. 2003

[13] M. Markkula et E. Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive, *Information Retrieval*, 1(4), 259-285. 2000

[14] A. A. Goodrum et A. Spink. Image searching on the Excite Web search engine, *Information Processing & Management*, 37(2), 295-311. 2001

[15] Y. Choi et E. M. Rasmussen. Users´ relevance criteria in image retrieval in American history, *Information Processing & Management*, 38(5), 695-726. 2002

[16] Y. Choi et E. M. Rasmussen. Searching for images: the analysis of users´ queries image retrieval in American History, *Journal of the American Society for Information Science and Technology*, 54(6), 498-511. 2003

[17] P. G. B. Enser et al. Facing the reality of semantic image retrieval, *Journal of Documentation*, 63(4), 465-481. 2007

[18] H. F. Greisdorf et B. C. O´Connor. *Structures of images collections: from Chauvet-Pont d´Arc to Flickr*. Unlimited Libraries, Westport, Conn. 2008

[19] E. Ménard. *Étude sur l'influence du vocabulaire utilisé pour l'indexation des images en contexte de repérage multilingue*. Thèse de doctorat. Montréal, Université de Montréal. 2008 https://papyrus.bib.umontreal.ca/jspui/bitstream/1866/2611/1/menard-e-these-indexation-reperage-images.pdf.

[20] A. A. Goodrum, M. M. Bejune et A. C. Siochi. A. C. *A state transition analysis of image search patterns on the web.* 2003 http://www.springerlink.com/media/6p8qd4b65j3vpne80ecn/contributions/0/e/w/7/0ew7r6lelaafcqv8.pdf.

[21] C. Jörgensen et P. Jörgensen. Image querying by image professionals, *Journal of the American Society for Information Science and Technology*, 56(12), 1346-1359. 2005

[22] D. Tjondronegoro et A. Spink. Web search engine multimedia functionality. *Information Processing & Management*, 44(1), 340-357. 2008

[23] K. Müller. Museums and virtuality, *Curator*, 45(1), 21-33. 2002

[24] F. Cameron. Digital Futures I: Museum collections, digital technologies, and the cultural construction of knowledge, *Curator*, 46, 325-340. 2003

[25] S. Knell. The shape of things to come: Museums in the technological landscape, *Museum and Society*, 1(3), 132-146. 2003

[26] D. L. Herman, K. Johnson et Ockuly, J. What clicked? *An interim report on audience research and media resources. In Museums and the Web 2004*, éd. par D. Bearman et J. Trant. Archives and Museum Informatics, Toronto. 2004

http://www.archimuse.com/mw2004/papers/ockuly/ockuly.html.

[27] P. Marty. The changing nature of information work in museums, *Journal of the American Society for Information Science and Technology*, 58(1), 97-107. 2007

# Quand la préservation passe par la classification : le cas des documents sonores et musicaux

**Bouchra Lamrini, Francis Rousseaux, Raffaele Ciavarella, Alain Bonardi, Jérôme Barthelemy**

*prenom.nom@ircam.fr*
Institut de Recherche et Coordination Acoustique/Musique.

**Résumé :** La création artistique contemporaine fait aujourd'hui largement appel aux technologies électroacoustiques et numériques. Dans la création musicale spécialement, les dispositifs et les outils logiciels permettant de manipuler les sons en temps réel sont apparus voici une trentaine d'années, et notamment les « patchs » processus numériques temps réel utilisés lors de performances ou de concerts en live. Soumis aux difficultés de la préservation, ces modules logiciels de traitement sonore sont souvent considérés comme des véritables documents numériques, ils sont à la fois supports de création et supports de constitution de connaissances dans la création artistique contemporaine. Pour soutenir les échanges et la construction d'une interprétation collective autour de ce document, nous proposons dans cet article une approche d'analyse et de classification, par les techniques du data-mining, de ces processus numériques afin de former une ontologie du domaine voire une organologie des traitements musicaux et audio numériques.

**Mots-clés :** Classification, data-mining, document numérique, musique contemporaine, préservation, processus numérique temps réel, Traitement Automatique des Langues.

## 1 Introduction

La musique est traditionnellement considérée comme l'art consistant à arranger et ordonner sons et silences au cours du temps et parfois dans l'espace (le musicien sur scène ou dans le public) selon 4 critères : le rythme (c.-à-d. la durée des sons dans l'espace temporel) qui est le support de cette combinaison dans le temps, la hauteur celle de la combinaison dans les fréquences (son grave ou aigu), le timbre

(la nature du son qui dépend du spectre produit par la source sonore, l'intensité du son (ce que l'on appelle les effets dynamiques en analyse musicale : piano, crescendo...). Le patch s'appuie sur un ensemble de catégories relevant principalement des critères et des formalismes mathématiques issus des sciences de traitement du signal. Ils permettent donc la réalisation d'intentions musicales particulières, hors des schémas traditionnels, et traduisent un ensemble de savoirs et de savoir-faire relatifs à ces intentions, et à la création musicale contemporaine dans son ensemble. Le patch est donc le résultat d'une activité créatrice, et un élément indispensable de la performance et de l'œuvre musicale.

Dans la perspective d'une pérennisation à long terme de la création musicale, le problème se pose donc de savoir préserver les processus temps réel. Pour répondre à cette question, le projet ASTREE[1] (*Analyse/Synthèse de Processus Temps Réel*), dans lequel s'inscrit ce travail, s'appuie sur une stratégie de développement d'outils permettant de transcrire, documenter, expliciter les processus existants afin d'améliorer leur pérennité, et les rendre indépendants de l'environnement technique sous-jacent. Dans ce contexte, notre contribution consiste notamment à développer une méthodologie d'analyse et de classification par des techniques de data-mining, afin de constituer des bases de connaissances ou des ontologies du domaine en jeu. Le passage progressif des documents existants à l'ontologie prônée par cette méthodologie sera réalisé par des traitements qui relèvent successivement de la terminologie, de la modélisation des connaissances et enfin de la représentation des connaissances. Nous étendons ces traitements en prenant en compte la structure des documents et des éléments porteurs d'information. Pour cela, nous nous basons sur des documents où les traitements et objets sont transcrits au moyen du langage fonctionnel FAUST[2] (*Functional Audio STreams*). Le découpage structurel de chaque document correspond à une caractérisation sémantique de son contenu. L'idée est de tirer profit de la structure et de la hiérarchisation du document pour souligner la complémentarité entre identification de concepts (objets recherchés) et extraction de relations. Les différentes réflexions exposées pour l'analyse et la classification de ces documents, font appel à des concepts et des techniques provenant de plusieurs domaines, parmi lesquels figurent la Représentation de Connaissances (*RC*), le Traitement Automatique du Langage Naturel (*TALN*), et l'Intelligence Artificielle (*IA*). Il s'agit d'une recherche de nature pluridisciplinaire dont l'objectif général est de mettre en relation un certain nombre d'hypothèses théoriques, d'explorer plusieurs concepts et d'appliquer des techniques provenant de différentes disciplines afin de proposer une méthodologie pour construire et ensuite maintenir les ontologies obtenues, et en particulier, la découverte de relations entre objets pertinents à l'ontologie. Ce projet de recherche amène une nouvelle réflexion sur les divers paliers à envisager dans une telle démarche de modélisation de connaissances textuelles pour des objectifs de formation d'ontologie du domaine et préservation de l'œuvre interactive et de son exécution. Certes, pour classifier les éléments matériels et logiciels de la musique électronique interactive, une première possibilité est d'envisager par exemple la famille des classes telles que : instruments électriques, synthétiseurs analogiques, synthétiseurs numériques, modules d'effets, synthétiseurs virtuels, et logiciels temps réel. Cette classification statique est facile à

[1] http://www.ircam.fr/sel.html?tx_ircam_pi1[showUid]=46&ext=1
[2] http://faust.grame.fr/

établir. Elle est uniquement fondée sur la nature technique des dispositifs utilisés et ne prend pas en compte leurs fonctions musicales. Le cadre de notre travail répondra donc à cette problématique en proposant d'établir des extractions et des classifications dynamiques à partir de l'extraction de descriptions numériques de ces dispositifs. L'association de ces deux classifications, complétée par le savoir faire de l'expert du domaine, permettra ainsi d'envisager des classifications 'homme-machine' dont le but final est de faire émerger une organologie des traitements sonores temps réel.
L'article est organisé comme suit. Nous faisons tout d'abord une brève introduction sur le patch et les défis relatifs à sa documentation et sa préservation à long terme dans la création artistique contemporaine. Nous exposons ensuite notre démarche envisagée en termes d'analyse et de classification, pour construire l'ontologie du domaine en question. Enfin, nous discutons les travaux en cours et les perspectives de la méthodologie présentée.

## 2 Le patch

### 2.1 Création et représentation de connaissances

L'apparition du patch date de la fin des années 1980, après deux décennies qui avaient misé sur des solutions hardware propriétaires. De nombreux chercheurs se sont investis pour trouver des solutions permettant d'enrichir les musiques mixtes, associant musiciens humains et sons électroniques. Dans ce contexte, le patch marque le début des approches interactives dans lesquelles la machine est en attente d'informations, acquises par des capteurs, venant de l'artiste, musicien ou interprète. L'une des premières pièces du répertoire IRCAM (*Institut de Recherche et Coordination Acoustique/Musique*) à exploiter ces possibilités était *Jupiter* de Philippe Manoury (1987) d'après [1], associant une flûte (dotée de capteurs et d'un microphone pour obtenir un certain nombre de paramètres du jeu du flûtiste) à un dispositif d'informatique musicale temps réel. Miller Puckette a crée ensuite en 1988 un premier système *Patcher,* dont sont dérivés les deux logiciels les plus utilisés aujourd'hui : *Max/MSP/Jitter*[3] et *PureData*[4].
Les patchs reprennent certains paradigmes des dispositifs électroacoustiques utilisés sur scène depuis plus d'une cinquantaine d'années comme les réverbérations, les processeurs d'effets, etc. À titre d'exemple [Fig.1], nous présentons ci-dessous un module créé avec le logiciel Max/MSP, produisant deux sinusoïdes, à 100 et 200 Hz, et les additionnant, l'amplitude du résultat étant atténuée (multipliée par 0.25). L'aspect du patch fait penser à un montage électrique, comme ceux élaborés sur la paillasse d'un laboratoire de physique, avec des appareils qui sont les modules rectangulaires (exemple de l'objet cycle~ qui produit une sinusoïde, à l'instar d'un générateur) et des câbles, qui sont ici les connections entre objets. Il est extrêmement facile de réaliser ces patchs, sans programmation, simplement en incorporant en quelques clics de souris des objets issus de menus, et en les reliant par des connections.
Ce type de dispositif est largement utilisé [1, 2, 3, 4], dans les arts de la performance tels que la musique, la danse, le théâtre, la vidéo, et les œuvres associant plusieurs arts [Tab.1], pour deux raisons : 1) il fonctionne en temps réel

[3] http://www.cycling74.com

[4] http://crca.ucsd.edu/~msp/software.html

en fournissant une réponse quasi immédiate aux entrées proposées (le délai de calcul des sorties étant considéré comme négligeable par rapport à l'ordre de grandeur temporel des entrées); 2) il permet une construction simple des traitements en s'appuyant sur une représentation graphique. Du point de vue de l'écriture, aussi bien informatique que musicale, le patch est un dispositif interactif offrant à l'utilisateur un mode de travail et de production facile à gérer et maîtriser sans faire appel à la programmation usuelle et aux catégories de la musicologie. Il n'y a plus de code à écrire, modifier et compiler, l'utilisateur adapte les objets utilisés, les connexions et les paramètres des patchs jusqu'à ce qu'il obtienne un résultat satisfaisant. Le mode de travail et de production se fait donc par retouche du patch. En revanche, la démarche créative du compositeur devient dépendante du système et de l'environnement technique utilisé.

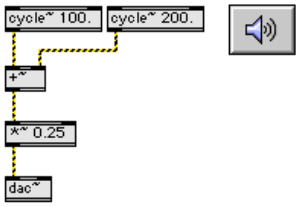


**Figure 1**. *Exemple de patch produisant la somme de deux sinusoïdes.*

| Œuvre | Configuration |
|---|---|
| Anthèmes II de Pierre Boulez, pour violon et électronique temps réel (1997). | Instrument solo et électronique temps réel |
| L'écarlate, performance de danse conçue par Myriam Gourfink, chorégraphe, sur une musique de Kasper Toeplitz (2001) | Performance danse et musique |
| K, musique et texte de Philippe Manoury (2001) | Opéra avec transformations sonores en temps réel |

**Tableau 1**. *Exemples d'utilisation de patchs dans les arts de la performance.*

### 2.2 Problématique et défis de la documentation du patch

Il convient tout d'abord de noter que les actions de préservation concernant les objets numériques mis en œuvre dans une création artistique doivent examiner la nature de l'œuvre originale, et les relations de l'objet numérique avec l'œuvre, avant d'évaluer les actions à effectuer pour le préserver. Il s'agit bien de préserver l'ensemble des connaissances permettant dans le futur d'apprécier le contexte dans lequel l'objet numérique est mis en œuvre : les informations portant sur l'environnement technique numérique, mais aussi l'environnement immédiat (par exemple, les caractéristiques de la salle de concert), les périphériques utilisés (microphone, hauts-parleurs, etc.). On devra aussi tenir compte d'une caractéristique de la démarche artistique, qui est une tendance à repousser les limites d'une technologie donnée, afin de la transcender pour obtenir de nouveaux effets. Il s'agit ici de préserver les intentions qui sous tendent la démarche créatrice. Une approche ou un modèle global et structuré de préservation devra impérativement tenir compte de ces considérations.

Le patch lui-même, document numérique porteur d'informations sur la conception musicale et technique d'une œuvre musicale, est confronté à diverses difficultés (fragilité des supports de stockage, obsolescence des standards techniques, l'instabilité des patchs vis-à-vis de leur complexité, etc.) qui compliquent la tâche de préservation. Le processus de documentation comporte généralement trois activités : la recherche (repérer et identifier les données pertinentes), la préservation (pérenniser les données) et la diffusion (rendre ces données disponibles et les transférer en connaissances utilisables). Une pratique documentaire rigoureuse est d'autant plus essentielle que les sources d'information sur la création contemporaine sont non structurées, et notamment pour les œuvres elles-mêmes. Compte tenu des défis communs aux domaines de la préservation des documents et de l'art numérique, la recherche dans ces documents est bien entendu pertinente. En matière de documents numériques, Margaret Hedstrom [5] fait une recommandation intéressante : « Préserver le contenu, le contexte et la structure tout en préservant la capacité d'afficher, de lier et de manipuler les objets numériques ». La plus grande difficulté réside probablement dans le second aspect, car cela suppose de garder l'accessibilité à une multitude de logiciels et de systèmes d'exploitation. Hedstrom ajoute également : « La préservation de matériel numérique nécessite souvent la transformation complexe et coûteuse d'objets numériques, pour qu'ils demeurent des représentations authentiques de la version originale ainsi que des sources utiles aux fins d'analyse et de recherche ». Le paradoxe de la préservation de documents numériques réside entre les expressions « transformations » et « représentations authentiques de l'original ». Les spécialistes proposent de plus en plus le concept d'émulation [6] et d'enveloppe contextuelle parmi les stratégies d'archivage numérique. Ce concept est une solution potentielle au problème des documents numériques dépendant des logiciels et, ce qui est plus important, du matériel requis pour y avoir accès. Ce concept permet également de conserver l'entière capacité de traitement des données.

Cette approche suppose de placer le document, conservé dans sa forme d'origine, dans une enveloppe virtuelle contenant toutes les instructions nécessaires pour sa récupération, son affichage et son traitement. Les instructions expliquent comment lier le document à un ensemble d'émulateurs qui servent de passerelle entre le document, qui peut demeurer stable, et le contexte technologique en constante évolution. Ainsi, plutôt que de tenter de modifier une multitude de documents, les gestionnaires d'archives ou de collections numériques n'ont qu'à mettre à jour les émulateurs. Une fiche technique renfermant d'importantes spécifications peut être ajoutée au document sous forme de métadonnées. Cela facilite le repérage de documents qui nécessiteront une certaine intervention en vue de leur conservation future. Grâce aux métadonnées, la description du document peut être incluse dans le document même, faisant de ce dernier sa propre fiche de catalogue. Il existe cependant un inconvénient : les métadonnées peuvent être aussi imposantes sinon plus que le document décrit.

Les patchs, comme document numérique de représentation de connaissances, sont source d'intérêt pour la communauté des chercheurs en musicologie et en informatique musicale. Nous citons dans ce cadre l'exemple du projet de collaboration 'Mustica' entre l'IRCAM et l'INA [7] sur la production d'une base de données ouverte aux organismes souhaitant remonter des œuvres contemporaines. En matière d'œuvre, nous citons par ailleurs, quelques travaux [8, 9, 10] (cités dans [1, 3]) menés pour la maintenance des patchs selon quatre actions : préservation,

émulation, migration et virtualisation. L'article de Bullock et Coccioli [11] présente l'exemple de l'œuvre *Madonna of winter and spring* et la tentative d'émulation d'un modèle Yamaha TX816 (pour la synthèse sonore), qui n'existe plus, grâce à des patchs créés sous PureData. A ces exemples s'ajoutent les travaux de [11, 12] dont les auteurs recommandent de sauvegarder les patchs réalisé par MaxMSP au format texte plutôt qu'en codage binaire. Enfin, nous citons le travail d'Andrew Gerzso d'après [2] qui a été réalisé au sien de l'IRCAM sur la recherche des représentations indépendantes d'une implémentation technique, dans le cadre de l'œuvre *Anthèmes II* de Pierre Boulez.

## 3 Approche pour la préservation à long terme des processus numériques temps réel

### 3.1 Cadre du travail

Le travail de recherche que nous présentons s'intègre dans une problématique qui considère le document numérique comme porteur de connaissances et doté d'une intentionnalité qui le construit ou le reconstruit pour une préservation à long terme. Dans cette vision du document comme signe, nous portons plus particulièrement notre intérêt sur l'analyse et la classification d'un document, c'est-à-dire sur les objets qui le constituent. Pour illustrer ce besoin de préservation du document, le projet ASTREE dans lequel s'inscrit notre travail, a permis une réflexion nouvelle sur la documentation et la préservation des processus temps réel vis-à-vis de l'histoire d'une création et l'exécution d'une œuvre de l'art de la performance. Cette réflexion consiste à transcrire les processus existants, conçus dans les environnements tels que Max/MSP ou PureData, dans le langage fonctionnel FAUST [13, 14]. Cette réflexion peut être illustrée par les trois points suivants [Fig.2] :

- Génération automatique de documentation sous une forme indépendante de toute technologie (hardware et software). Cette documentation permettra une réimplémentation manuelle complète du processus originel.
- Application des techniques de data-mining sur les expressions algébriques issues de FAUST et de la documentation afin de constituer des connaissances sur les processus en question.
- Génération d'un code optimisé et indépendant des librairies et des systèmes de traitement de signal.

Au moyen d'un concept familier de bloc-diagramme, FAUST permet de décrire facilement des processeurs de signaux qui traitent et transforment des signaux en entrée pour produire des signaux en sortie. L'utilisateur dispose de modules de traitements élémentaires qu'il combine de différentes manières pour obtenir le traitement souhaité. Le compilateur FAUST utilise ensuite cette description pour écrire automatiquement un programme C++ équivalent. Des techniques d'optimisation particulières permettent d'engendrer un code C++ de qualité dont l'efficacité est généralement comparable à celle d'un programme écrit à la main. La figure 3 présente un exemple de la fonction 'Envelop'. Cette fonction permet de générer un bruit blanc contrôlé par une enveloppe. Plusieurs paramètres sont utilisés pour ajuster la forme de l'enveloppe, la longueur d'attaque, la durée de décadence, le pourcentage de volume soutenu, et la durée de sortie.

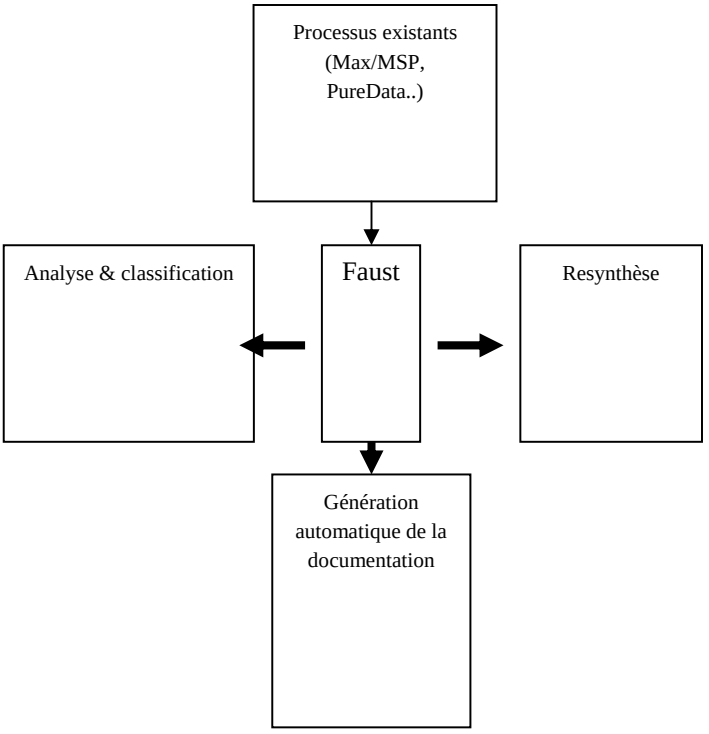


**Figure 2.** *Processus de la transcription des objets existants dans le langage FAUST.*

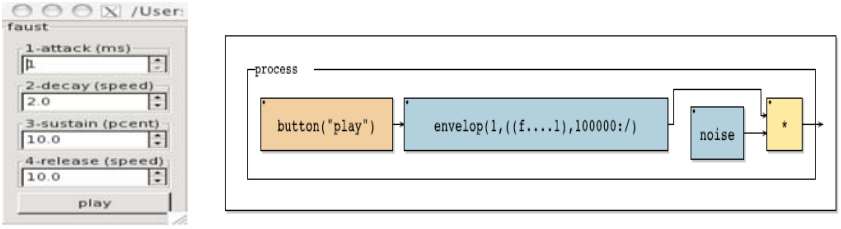


**Figure 3.** *Schéma de représentation du contrôle par une enveloppe d'un bruit blanc au moyen de FAUST.*

### 3.2 Classification des documents numériques et formation d'ontologie

En intelligence artificielle, les ontologies sont apparues comme une réponse aux problématiques de représentation et de manipulation des connaissances au sein des systèmes informatiques. L'importance que revêt aujourd'hui l'usage des ontologies pour le développement de systèmes à base de connaissances n'est plus à démontrer. Les chercheurs du Web ont adopté ce terme ontologie pour référer à un document (ou fichier) définissant d'une façon formelle les relations entre termes [15]. Dans le cadre du Web sémantique, les ontologies sont utilisées comme noyau du système pour accéder à des informations structurées ainsi qu'à

des règles d'inférence supportant le raisonnement automatique. Ces ontologies offrent également la possibilité, pour un programme, de retrouver les différents termes désignant un même concept. Il s'agit pour ce cas spécifique, d'ontologies de type domaine. La génération de l'ontologie et l'extraction des données utilisent plusieurs sources de connaissances telles que les ontologies déjà formées par des concepts arrangés hiérarchiquement avec un haut degré de connectivité. Ces ontologies sont fondées sur une conception de nature générale suivant un héritage des propriétés [16] ; des sources de connaissances lexicales, comme la base de données lexicales Wordnet [17] ; des répertoires des expressions régulières conçues pour convenir des items lexicaux structurés (dates, n° tel, mesures...) et qui peuvent fournir par héritage des informations appropriée aux concepts crées [18] ; et des documents d'apprentissage qui contiennent le contenu textuel du domaine spécifique pour l'intérêt de l'utilisateur (langage spécifique du texte).
Comme nous l'avons mentionné auparavant, cet article propose une nouvelle réflexion de mise au point d'une méthode de classification permettant de former de manière automatique une ontologie conceptuelle du domaine musical en termes de traitements musicaux et audio numériques. De manière générale, la définition d'une méthode de classification conceptuelle repose sur deux éléments : la définition d'une distance permettant de comparer les objets à classer, et d'autre part la définition d'un algorithme de classification qui construit la structure arborescente proprement dite. Dans le domaine de la formation des ontologies conceptuelles, la notion de distance entre mots a été étudiée en Traitement Automatique de la Langue (TAL) pour former les classes sémantiques qui forment les nœuds de ces catégories suivant une approche ascendante ou descendante, tandis que les algorithmes de classification ont été plus largement étudiés en analyse de données et en apprentissage automatique. Dans ce contexte, de nombreux outils destinés à l'acquisition automatique/ou semi-automatique de classes sémantiques visant à regrouper de termes proches, sont élaborés. De point de vue sémantique, cette notion de proximité est généralement fondée sur des mesures de distance entre termes en fonction du degré de ressemblance des contextes dans lesquels ils apparaissent. Les descriptions et les régularités recherchées des contextes des termes dépendent de l'approche menée. Ainsi, les contextes peuvent être purement graphiques, i.e. sous forment de co-occurrences dans une fenêtre de mots [19, 20, 21]. Le choix de la distance appropriée pour un corpus est donc un problème posé et peu étudié [22]. La majorité des travaux ne s'intéressent qu'à la seule évaluation de la tâche finale pour laquelle l'apprentissage est effectué. Les critères de cette évaluation restent quantitatifs et donnent rarement lieu à des études comparatives [23]. En revanche, la caractérisation des résultats des méthodes fournit les outils d'aide au choix de la plus pertinente ou la mise au point de nouvelles méthodologies. Il en est de même pour les algorithmes de classification. Elaborer une méthodologie ou un outil permettant de mettre au point un algorithme de classification conceptuelle pour former des ontologies en fonction de la tâche est une tâche ardue. Par ailleurs, les travaux effectués en classification conceptuelle [24, 25] n'ont pas trouvé de propos en apprentissage à partir de corpus. Certes, il faut souligner que la construction d'une ontologie dans le domaine du TAL exige une phase de choix et d'adaptation des algorithmes existants aux problèmes posés par le corpus autour de la distance, voire développer de nouveaux outils et moyens méthodologiques. Soulignons par ceci et à présent l'importance qu'on doit accorder au choix des outils et des techniques

pour l'extraction et à la sélection des objets (ici les indicateurs des patchs) contenus dans l'ensemble de nos documents transcrits par Faust.
L'ensemble des étapes de notre démarche proposée est élaboré à partir d'une étude bibliographique sur les travaux en TAL (cités auparavant), notamment la formation de classes sémantiques à partir de corpus et à partir de travaux en apprentissage sur la classification conceptuelle et l'analyse des données. Nous avons donc noté que pour effectuer les diverses tâches de classification, recherche et filtrage de documents, il faut d'abord représenter les textes de manière à la fois économique et significative. On sait que le modèle vectoriel est l'approche la plus courante dans lequel le texte est représenté par un vecteur numérique obtenu en comptant les éléments lexicaux les plus pertinents. Ces vecteurs sont fournis par des prétraitements simples. On commence généralement par éliminer les mots grammaticaux (articles, prépositions, etc.) et par réduire les variantes morphologiques à une forme commune (souvent appelée terme). Puis on compte les occurrences des termes les plus importants de manière à représenter chaque document par un vecteur dans l'espace des termes. Un corpus de documents génère donc une matrice *Document-Terme* [Fig.4] qui permet ensuite d'appliquer les opérations vectorielles usuelles avec des résultats sémantiquement pertinents dans l'ensemble.

$$\begin{array}{c} \\ Doc_1 \\ Doc_2 \\ \vdots \\ Doc_m \end{array} \begin{array}{c} \begin{array}{cccc} Terme_1 & Terme_2 & \dots & Terme_n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \end{array}$$

**Figure 4.** *Matrice Document-Terme ($x_{ij}$= fréquence d'apparition du terme i dans le document j).*
Cependant ce type d'approche produit généralement des vecteurs lexicaux de très grande dimension qui sont coûteux à stocker et à traiter. Des vecteurs qui sont partiellement ou entièrement vides (contenant généralement plus de 90% de valeurs nulles), ayant en plus des termes fortement corrélés entre eux. La détection d'une telle synonymie entre termes a des conséquences négatives pour l'indexation et la recherche. Evidement des documents voisins sémantiquement peuvent très bien ne pas contenir les mêmes termes. Détecter les relations entre termes permettra d'améliorer la recherche de documents. Une représentation à partir de ces vecteurs redondants sera donc difficilement lisible par un utilisateur qui voudrait s'en servir pour évaluer rapidement le contenu d'un document et chercher à voir les relations entre divers documents. Devant ce manque, il serait important de trouver la dimension intrinsèque du domaine, c'est-à-dire la dimension minimale permettant de représenter les données sans perte d'information. Pour ce type de traitement, Il y a une gamme de méthodes permettant de calculer un nombre réduit de dimensions pour un ensemble de données, nous citons la méthode Latent Semantic Analysis *(LSA)* qui n'utilise pas la matrice de covariance, mais extrait de nouveaux axes directement de la matrice document-terme [26], l'Analyse Factorielle, dont la méthode la plus connue est l'Analyse en

Composantes Principales (*ACP*), mais elle ne permet pas en pratique de traiter des vecteurs de très grande dimension comme dans le domaine des documents. En revanche, la plus prometteuse est la version neuronale de l'ACP, appelée Algorithme Hebbien Généralisé (*AGH*), permettant de traiter de tels vecteurs avec de bons résultats.
L'objectif de cette étude bibliographique était d'esquisser un panorama des méthodes de data-mining et de text-mining, notamment les méthodes de catégorisation et de classification non supervisée en se plaçant dans un cadre méthodologique pour identifier les indicateurs, liés par exemple à la représentation des documents, à la sélection des termes décrivant le contenu des documents et aux algorithmes de catégorisation et de clustering, à prendre en compte pour employer l'une plutôt que l'autre et savoir comment les évaluer sur notre problème de classification de ce type de documents numériques.

### 3.3 Présentation du corpus

Les objets pris en compte dans nos travaux sont des objets issus des expressions algébriques contenues dans la documentation transcrite par le langage fonctionnel FAUST. Pour illustrer la nature du problème à résoudre, nous avons choisi un exemple de documentation sur un filtre passe band 'BPF' comme un processus temps réel conçu dans Max/MSP et transcrit dans le langage FAUST. La figure 5 présente un zoom sur les types de structures contenues dans le corpus en question. Les concepts que l'on cherche à repérer dans la documentation se trouvent dans les codes Faust, la documentation générée en latex, les trois librairies importées (maxmsp, music, et math) et les commentaires inclus dans les codes et les librairies.

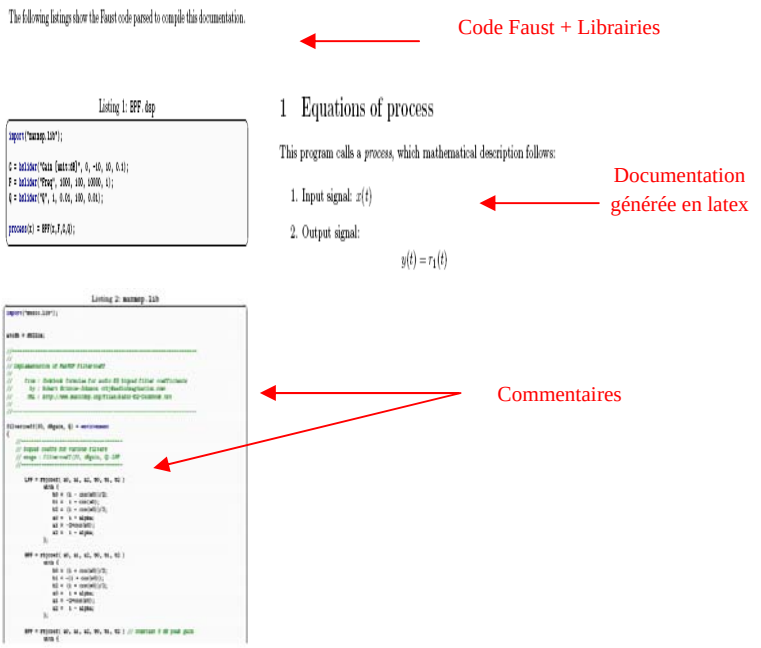


**Figure 5.** *Structures contenues dans le corpus à analyser*

Si nous nous intéressons par exemple seulement aux objets provenant de Max/MSP, les trois fragments du document permettent de fournir une information [Fig. 6] sur les objets clés à savoir :

- e filtre passe-bande 'BPF' présent dans la librairie "maxmsp.lib". Ceci est donné par l'expression décrivant le processus générale "process (x) = BPF(x, F, G, Q)".
- présence d'une entrée x(t), ceci est indiqué dans le texte descriptif et par l'expression algébrique du processus.
- les coefficients du filtre, ici le gain 'G', la fréquence 'F' et le facteur de qualité 'Q'.
- les expressions fonctionnelles qui lient le processus aux coefficients du filtre comme la présence de la fonction de transfert "biquad". La présence de cette fonction entraine également la présence d'autres paramètres de conversion et de fonctions de calcul comme les fonctions trigonométriques.
- et également un domaine de variation de chaque coefficient approprié au filtre en question (Exp : *"Freq", 1000, 100, 10000, 1*).

Les processus dans leur expressions algébriques peuvent être vus comme des objets clés qui génèrent une sorte de connaissances appropriées aux termes recherchés comme 'Filter, BPF, APF, Oscillator, Osc…', et les termes 'Q, G, F,…' comme des attributs de ces termes recherchés. En s'appuyant sur les différents documents transcrits et normalisés à présent, nous avons commencé à constituer une base de connaissances sur les objets clés, les termes associées ainsi que d'autres données permettant de fournir une information sur la proximité, la corrélation des objets, la proximité des classes objets, la distribution des objets autour d'un objet de la même classe (chaque objet a une distance proportionnelle à son degré d'association. En classification conceptuelle ascendante par exemple (car c'est la plus recommandée pour des raisons de complexité algorithmique et 'd'explicabilité' pour la validation par l'utilisateur en cours d'apprentissage) l'algorithme général consiste à réunir successivement par paires, les objets proches, ou les objets et les classes objets proches, afin de former des hiérarchies ou des graphes de classes d'objets.

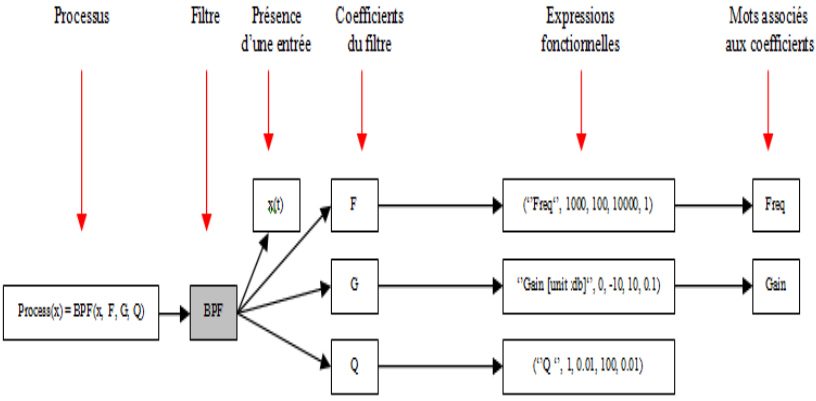


**Figure 6.** *Digramme illustrant les différentes termes clés qui peuvent fournir une connaissance sur le filtre passe-bande 'BPF : Band Pass Filter '*

### 3.4 Méthodologie proposée pour l'analyse et la classification des patchs

Sur notre problème d'analyse et de classification des patchs, un système d'extraction d'information classique seul n'est pas capable d'analyser les différents segments du corpus en question, notamment sa structure qui s'appuient en général sur des expressions algébriques, combinées d'une certaine manière à décrire les entrées et les sorties des processus réalisés d'une part, et les fonctions de transformation effectuée entre ces entrées et ces sorties d'autre part. En effet, pour analyser une phrase, un système d'extraction d'information effectue successivement une analyse lexicale (segmentation de la phrase en chaines de caractères qui représentent des mots), une analyse morpho-syntaxique (étiquetage des mots par leur catégorie syntaxique et association de chaque mot à sa forme canonique..), une analyse syntaxique (analyse de la structure de la phrase), et en fin une analyse sémantique (compréhension du sens des mots et des relations entre les mots). Or, notre document numérique est le résultat d'une transcription à base d'un langage fonctionnel. Sous sa forme observable, ce document peut être considéré comme une suite de mots ordonnés selon des règles qui ne reflètent pas forcément l'ordre dans lequel les mots s'appliquent les un aux autres pour former l'interprétation sémantique fonctionnelle. Plusieurs voies que nous avons citées auparavant donc peuvent être explorées pour résoudre ce type de problème de classification, et par ailleurs celle de la grammaire catégorielle combinatoire applicative [27].

Dans le cadre de ce travail, notre méthodologie envisagée d'analyse et de classification s'appuie sur deux pistes complémentaires, qui s'emboîtent pour donner une classification finale des objets en question, et notamment jugée par l'expert du domaine [Fig.7].

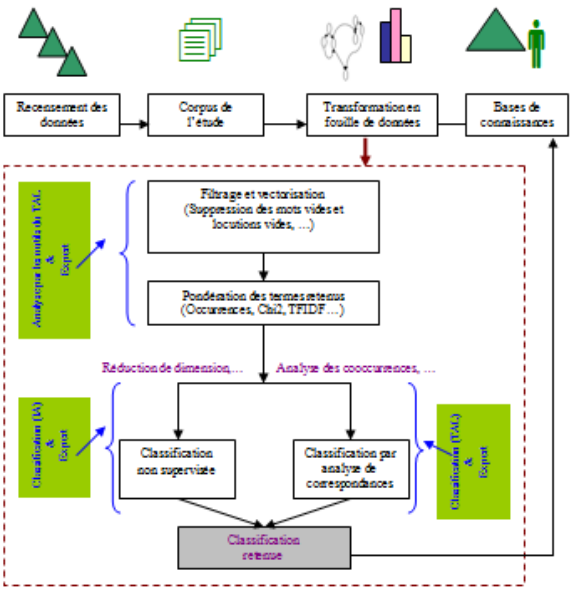


**Figure 7.** *Etapes de la méthodologie proposée pour la classification des patchs.*

Notre idée dans un premier temps est de réaliser une série de prétraitements classiques au moyen des outils du TAL afin d'éliminer les objets non-clés pour notre étude et fixer le type de pondération qu'on doit affecter aux objets retenus pour la classification. Cette phase de prétraitement est guidée notamment par l'expert du domaine. L'intervention de l'expert permettra de valider chaque étape avant la validation finale du la classification attendue par notre étude. Ensuite, on a deux voies à parcourir pour avoir une classification satisfaisante selon nos objectifs fixés au début de l'étude. La première voie est de poursuivre les analyses par les mêmes outils du TAL tels que : les analyses des cooccurrences d'objets-clés, par exemple faire des comparaisons entre les paires d'objets-clés de la famille Filter : 'APF-BPF' et 'APF-HPF', et également l'analyse d'autres objets associés à ces objets-clés ; les analyses thématiques des unités de contexte, autrement dit opérer et modeler les objets qui émergent des différents unités du corpus et qui sont décrits à travers leur vocabulaire caractéristique, c'est-à-dire à travers des ensembles de mots-clés (lemmes ou catégories) co-occurrents. Ces objets émergents peuvent être utilisés pour obtenir de nouvelles variables qui peuvent être utilisées dans des analyses ultérieures. Les analyses comparatives des sous-ensembles du corpus qui donnera lieu à une classification par une analyse des correspondances. Comme toutes les techniques factorielles, les analyses comparatives permettent l'extraction de facteurs qui ont la propriété de récapituler d'une façon organisée l'information significative contenue dans les innombrables cellules des tableaux de données. En outre, cette technique d'analyse permet la représentation graphique, dans un ou plusieurs espaces, des points qui détectent les objets en lignes et colonnes, et qui dans notre cas sont les entités linguistiques (mots, lemmes, segments de texte, textes, etc.). Les résultats d'analyses permettront d'évaluer des rapports de proximité/distance - ou de similitude/différence - entre les objets considérés. La figure 8 illustre un schéma approximatif de ce que nous envisageons de réaliser par la démarche proposée dans ce travail après une pré-étude du corpus transcrits par FAUST ; elle présente un exemple de la famille 'Filter'. La figure 9 présente les distances qui peuvent être analysées entre les classes pour relever et mettre en valeur les différentes corrélations traduisant une certaine connaissance cachée ou mal connue en termes de fonctionnement des patchs comme des processus temps réel.

Pour la deuxième voie, nous envisageons de réduire la dimension de corpus par l'application de l'une des techniques d'analyse factorielle citées auparavant, la plus adaptée pour notre cas est la version neuronale de *l'ACP*, appelée *AGH* (citée dans § 2.2). Cela permettra de projeter les documents dans un espace beaucoup plus compact avec des concepts plus significatifs. En effet ces nouvelles dimensions représentent essentiellement des corrélations entre termes dans un corpus donné, révélant de la sorte les thèmes principaux du corpus presque aussi bien qu'une classification des documents. Ensuite, nous utilisons les vecteurs dans le nouvel espace (donnés par les valeurs de sortie du réseau AHG) pour effectuer une classification non supervisée des documents. Les algorithmes les plus connus sont soit l'algorithme des centres mobiles (k-means) qui emploie une distance euclidienne pour ne pas avoir à normaliser les vecteurs, soit les réseaux de neurones compétitifs par la règle du Kohonen comme les cartes de SOM [28]. Nous citons ici les travaux de Nguyen et Zreik [29, 30, 31] sur le développement du système *Hyperling* pour reconnaitre les langues dominantes dans un site Web

multilingue et dont les deux algorithmes K-means et SOM ont été implémentés et testés avec succès.

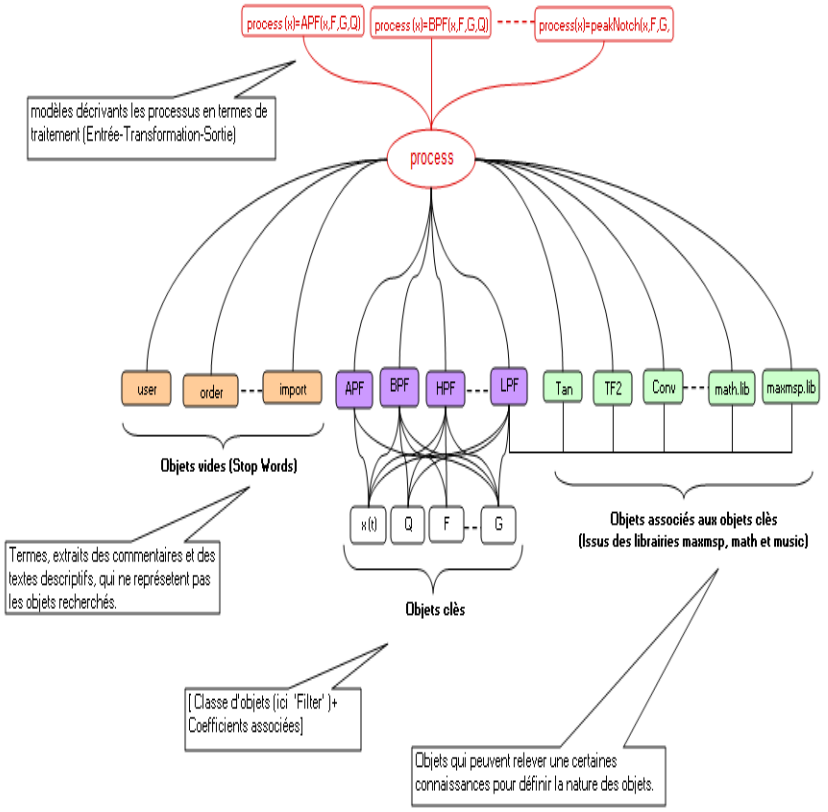


**Figure 8.** *Schéma approximatif illustrant l'un des types d'informations constituées dans un corpus décrivant la famille 'Filter'.*

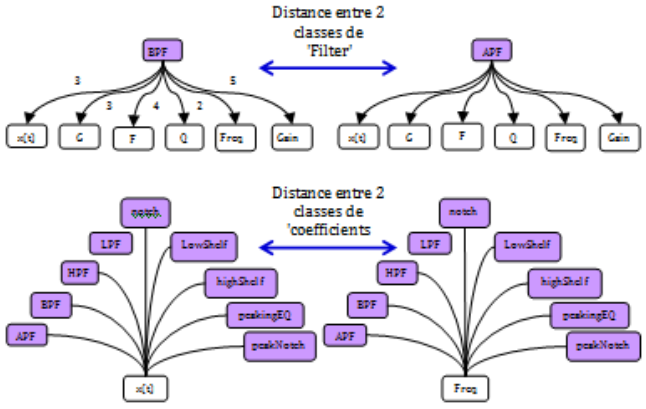


**Figure 9.** *Schéma illustrant les types de distance entre les classes des objets.*

## 4 Conclusion et perspectives

En génie documentaire, les sources documentaires s'accroissent et les applications issues du Traitement Automatique du Langage et deviennent applicables à une large variété de problèmes dans ce contexte. De nombreux travaux ont été proposés pour construire de façon plus ou moins automatique des ontologies de type hiérarchies conceptuelles à partir de corpus analysés. Profitant ainsi des moyens qu'offrent les outils du TAL à savoir : l'extraction d'information, l'indexation de documents, la désambiguïsation syntaxique, etc. A l'opposé, il manque les outils et les méthodologies qui permettent d'évaluer et comparer les points faibles et forts des différentes approches pour un corpus spécifique à un domaine et une tâche donnée. Le travail que nous réalisons dans le cadre du projet ANR ASTREE, permet dans ce sens de proposer une nouvelle réflexion qui permet d'associer les techniques du TAL pour la classification conceptuelle à celles provenant de l'intelligence artificielle pour la catégorisation et la classification non supervisée afin d'analyser la nature de ces processus temps réel en question et d'en relever les propriétés cachées. La spécificité de notre challenge scientifique est donc de constituer, par extraction et classification dynamique, des connaissances permettant de générer et faire émerger une organologie des traitements sonores temps réel qui contribuera ainsi à des pratiques de préservation de l'œuvre de l'art de la performance. Il est à noter que le corpus à analyser dans ce cadre ne demande pas forcément de disposer de connaissances additionnelles (terminologique ou sémantique) pour guider l'apprentissage ou évaluer les résultats. Les travaux en cours et futurs se concentrent sur les moyens qui permettent d'obtenir et d'évaluer les classes attendues par nos objectifs, à savoir les critères de constitution du corpus d'entrée, les distances, et les critères d'évaluation des résultats, nivaux d'abstraction (lié aux objets qui doivent être isolés de la classification, etc.). L'efficacité et la fiabilité de notre approche seront notamment validées par une phase de simulation numérique pour mettre en évidence les limites de l'approche développée. A travers ce travail, nous avons également pu montrer le rôle du patch comme document numérique dans la création artistique contemporaine et les différentes approches mises au point pour leur maintenance, dont la classification est certes l'une des solutions recommandées pour sa préservation à long terme dans la création de l'art contemporain.

**Remerciements**
Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence *2008 CORD 003 01*. Les auteurs tiennent à remercier les partenaires du projet ASTREE, et notamment ceux du Centre National du Création Musicale Grame, à l'origine du langage FAUST, et de l'outil de génération automatique de documentation sur lequel notre travail se base.

## 5 Bibliographie

[1] A. Bonardi et J. Barthélemy, Le patch comme document numérique : support de création et de constitution de connaissances pour les arts de la performance, *10ième Colloque International sur le Document Electronique (CIDE.10)*, 2-4 juillet, Nancy, France 2007.

[2] A. Bonardi, M. H. Serra et M. Fingerhut, Documentation musicale et outils hypermédias*, Deuxième Colloque International sur le Document Electronique (CIDE.2)*, 5-7 juillet, 295-309, Damas, Syrie 1999.

[3] A. Bonardi, J. Barthélemy, G. Boutard et R. Ciavarella, Préservation de processus temps réel. Vers un document numérique. *Document Numérique*, 11 (3-4), 59-80, 2008.

[4] P. Bottoni, S. Faralli, A. Labella et M. Pierro, Mapping with planning agents in the Max/MSP environment: the GO/Max language, *In Proceedings of the 2006 International Conference on New Interfaces for Musical Expression*, Paris, France 2006.

[5] M. Hedstrom, Digital Preservation: A Time Bomb for Digital Libraries, *Computers and the Humanities*, 31, 189–202, 1998.

[6] J. Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation,* Published by ECPA (European Commission for Preservation and Access), Edo H. Dooijes, Amsterdam. 1998. http://www.clir.org/pubs/reports/rothenberg/contents.html

[7] B. Bachimont, J. F. Blanchette, A. Gerszo, A. Swetland, O. Lescurieux, P. Mahoudeaux, N. Donin et J. Teasley, Preserving Interactive Digital Music: A report on the Mustica Research Initiative, *In Proceedings of the Third International Conference on WEB Delivering of Music (WEB'03)*, Leeds, England 2003.

[8] H. Bosma, Documentation and Publication of Electroacoustic Compositions at NEAR, *In Proceedings of the Electroacoustic Music Studies Network International Conference* (EMS 05), Montreal, Canada 2005.

[9] D. Teruggi, Preserving and Diffusing, *Journal of New Music Research*, 30 (4), 403-405, 2001.

[10] V. Tiffon, Les musiques mixtes entre pérennité et obsolescence, *Revue Musurgia,* XII/3, Paris. 2005.

[11] J. Bullock et L. Coccioli, Modernising Live Electronics Technology in the Works of Jonathan Harvey, *In Proceedings of the International Computer Music Conference*, Barcelona, Spain 2005.

[12] M. Puckette, New Public-Domain Realizations of Standard Pieces for Instruments and Live Electronic*s, In Proceedings of the International Computer Music Conference*, Miami 2004.

[13] Y. Orlarey, S. Letz et D. Fober, Multicolore technologies en Jack and Faust, *In Proceeding of the international Computer Music Conference-ICMA,* 2008.

[14] Y. Orlarey, D. Fober et S. Letz, FAUST : an efficient Functional Approach to DSP Programming, *New Computational Paradigms For Computer Music*, Delatour, France 2009.

[15] T. Berners-Lee, J. Hendler et O. Lassil, The Semantic Web, *Scientific American Magazine, 2001.*

[16] K. Mahech, et S. Nirenburg, A situated Ontology for Pratical NLP, *In Proceeding of Workshop on Basic Ontological Issues in Knowledge Sharing* : *International*

*Joint conference on Artificial Intelligence (IJCAI-95)*, August 19-20, Montreal, Canada 1995.

[17] C. Fellbaum, *Wornet : An Electronic Lexical Database*, Combridge, Ma :MIT Press. 1998.

[18] D.W. Embly, D. M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y. K. Ng et R.D. Smith, Conceptual model based data extraction from multiple record Web, *Data Knowledge and Engineering*, 31(3), 227-251, 1999.

[19] K. Sparck Jones et E. B. Barber, What makes an automatic keywords classification effective?, *Journal of the ASIS*, 18, 166-175, 1971.

[20] P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai et R.L. Mercer, Class-based n-gram models of natural language, *Computational Linguistic,* 18(4), 283-298, 1992.

[21] K.W. Church et P. Hanks, Word Association Norms, Mutual Information, and Lexicography, *In proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 76-83, 1989.

[22] B. Habert, A. Nazarenko et A. Salem, *Les linguistiques de corpus*, Ed Armand Collin. 1997.

[23] I. Dagan, L. Lee et F. Pereira, Similarity-Based Methods For Word-Sense Disambiguation, *In proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'96),* 56-63, 1996.

[24] D. H. Fisher, Knowledge Acquisition via Incremental Conceptual Clustering. *In Machine Learning Journal*, 2, 139-172, 1989.

[25] G. Bisson, Conceptual Clustering in a First Order Logic Representation, *In Proceedings of 10th European Conference on Artificial Intelligence (ECAI'92)*, 458-462, Vienna 1992.

[26] S. Deerwester, S. T. Dumais, G. W. Furnas., T. K. Landauer et R. Hashman, Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41(6), 391-407, 1990.

[27] J-P. Desclés et I. Biskri, Logique combinatoire et linguistique: grammaire catégorielle combinatoire applicative, *Mathématiques et sciences humaines*, Tome 132, 39-68, 1995.

[28] T. Kohonen, Self-organization of very large document collections: state of the art, *In Proceeding of ICANN'98*, London 1998.

[29] D. Nguyen, Nouvelle méthode syntagmatique de vectorisation appliquée au Self-organizing map des textes vietnamiens, *JEP-TALN-RECITAL (Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*, 19-22 avril 2004, Fès, Maroc 2004.

[30] D. Nguyen et K. Zreik, HYPERLING : Système de reconnaissance et de classification des hyperdocuments multilingues, *In International Conference in Computer Science « Research, Innovation and Vision of the Future» RIVF05, 21-24 February, University of CANTHO, Vietnam 2005.*

[31] K. Zreik, et D. Nguyen, Catégorisation de documents multilingue : le système Hyperling, *In CIDE.8,* 25-28 Mai*,* Beyrouth, Liban 2005.

# La plateforme d'indexation INVENIO : Une approche MPEG-7 pour la réutilisation des contenus multimédias

**Titus Zaharia, Alain Vaucelle, Thomas Laquet (1)**
titus.zaharia@it-sudparis.eu / alain.vaucelle@it-sudparis.eu / thomas.laquet@it-sudparis.eu
*(1) Institut Télécom ; Télécom SudParis ; Département ARTEMIS*

**Résumé** : Dans cet article, nous proposons une nouvelle plateforme d'indexation de contenus multimédias, dite INVENIO (INdexing Visual ENvironment for multimedia Items and Objects). Fondée entièrement sur la norme ISO/MPEG-7, la plateforme INVENIO offre, dans un système intégré, moteurs d'extraction de métadonnées MPEG-7, outils d'annotation, moteur de requête, outils de gestion de bases de données multimédias, ainsi que des interfaces utilisateurs appropriées et ergonomiques. Pour valider la plateforme INVENIO, nous avons considéré une application industrielle d'indexation et de reconnaissances d'images couleur/texture fixes et animés, liée au projet CapDigital HD3D-IIO. A travers cette plateforme nous répondons à une problématique de réutilisation de contenus numériques au sein d'une chaîne de production audiovisuelle. Les expérimentations réalisées concernent différentes chaînes de production audiovisuelle, incluant des contenus aussi bien naturels que de synthèse (i.e. dessins animés). Les solutions d'indexation et de reconnaissance d'images proposées dans de cet article démontrent notamment que l'exploitation des technologies MPEG-7 à travers la plateforme d'indexation INVENIO permet de réaliser une économie significative de temps de travail, ainsi qu'une réutilisation optimale des contenus numériques en cours de production.

**Mots-clés :** Plateforme d'indexation d'images, norme MPEG-7, indexation par le contenu, moteur de recherche, interface, couleur, texture

## 1 Introduction

La croissance exponentielle des contenus multimédias sur la toile oblige à une remise en question de la pertinence du mode d'indexation des contenus. Nous avons pu voir ces dernières années différentes technologies d'indexation permettant d'indexer un contenu visuel et de le retrouver par similarité (*e.g.* Google). Cependant bien souvent ces technologies restent limitées à un nombre d'images fixes données et fonctionnent dans des environnements fermés. Pour répondre à la diversité (images, vidéos, données graphiques 3D) des productions (contenus autoproduits, personnalisés), et à leur circulation sur l'Internet, il devient nécessaire de disposer de solutions efficaces d'indexation et de recherche, permettant à l'utilisateur d'accéder à ces contenus. Les solutions commerciales existantes s'appuient aujourd'hui exclusivement sur des moteurs de recherche par mots-clés (*e.g.* Google, Yahoo, Exalead…). Dans le cas spécifique de contenus audiovisuels, les limitations d'une

telle approche sont multiples : polysémie et complexité des contenus, barrières linguistiques, subjectivité des annotations, difficulté d'annotation manuelle d'un grand volume de données…

Aujourd'hui les méthodologies d'indexation par le contenu [1] visent à proposer des solutions alternatives pour l'indexation des documents multimédias. Le principe consiste à associer aux contenus des métadonnées non plus textuelles, mais liées intrinsèquement au contenu audiovisuel lui-même. Cela est possible puisque les représentations mathématiques permettent la description de manière automatique (ou semi-automatique) et discriminante du flux audiovisuel. Ce sont des caractéristiques perceptuelles associées classiquement aux attributs visuels comme la forme, la couleur, la texture ou encore le mouvement, qui rendent cette indexation possible. Le principe de la recherche de contenus dans une base de données change alors de façon fondamentale. L'utilisateur présente à l'entrée de son moteur de recherche non plus des mots-clés mais un exemple (*e.g.* une image/vidéo où une partie d'une image/vidéo), ou une ébauche (*e.g.* un dessin réalisé à la main). Les mesures de similarités associées aux descripteurs images permettent alors de réaliser des requêtes automatiques et de retrouver des images similaires.

Le domaine de l'indexation d'images par le contenu a connu une effervescence spectaculaire depuis le milieu des années 1990 et jusqu'à présent, comme en témoignent l'impressionnant volume de méthodes et techniques proposées dans la vaste littérature scientifique [1], [2] consacrée à ce sujet. Dans ce cadre, une étape marquante a été la sortie officielle au début des années 2000 de la norme ISO MPEG-7 [3], [4]. Officiellement appelée « Multimedia Content Description Interface », la norme MPEG-7 propose un large éventail de technologies de description de documents multimédia, intégrant des approches aussi bien textuelles que par le contenu.

En particulier, la spécification ISO MPEG-7 propose aujourd'hui un riche ensemble de descripteurs visuels [5] et de schémas de description, exprimés dans un langage de description de données fondé sur une approche XML Schema (XML Schema). Toutefois, la pertinence de la norme pour des applications industrielles grandeur nature reste à démontrer. Pour cela, il est indispensable d'élaborer et de mettre en œuvre des outils efficaces et ergonomiques permettant de faciliter les processus d'annotation, de consultation, de navigation dans des bases de données et de requête de l'information.

La plateforme d'indexation INVENIO (INdexing Visual ENvironment for multimedia Items and Objects) proposée dans cet article propose une solution unifiée à cette problématique. Fondée entièrement sur la norme MPEG-7, la plateforme INVENIO offre, dans un système intégré, moteurs d'extraction de métadonnées MPEG-7, outils d'annotations, moteur de requête, outils de gestion de bases de données, ainsi que des interfaces utilisateurs appropriées et ergonomiques.

Le domaine d'application ciblé par INVENIO est, de manière générale, celui de l'indexation d'images par le contenu. Toutefois, pour valider la plateforme, nous avons développé une application industrielle d'indexation d'images couleur/texture pour la production de films d'animation. L'objectif est de favoriser la réutilisation de contenus numériques au sein d'une chaîne de production audiovisuelle. Ce développement est lié au projet structurant HD3D-IIO du pôle de compétitivité CapDigital (www.capdigital.com).

La suite de cet article est structurée comme suit. Tout d'abord nous énonçons les objectifs généraux du projet HD3D-IIO en mettant notamment l'accent sur la problématique de réutilisation de contenus au sein d'une chaîne de production

audiovisuelle (Paragraphe 1). L'approche par indexation d'images proposée est ensuite décrite, en précisant notamment les éléments de la norme MPEG-7 retenus pour nos développements, les aspects d'interface utilisateur et d'ergonomie ainsi que les fonctionnalités supportées (Paragraphe 2). Une évaluation expérimentale conduite sur les bases de données grandeur nature constituées au sein du projet HD3D-IIO est présentée dans le Paragraphe 3. Le paragraphe 4 présente les futurs développements envisagés, et enfin le paragrahe 5 conclut l'article.

## 2 Contexte HD3D-IIO : réutilisation de contenus

Le projet HD3D-IIO concerne la création et la fabrication des contenus numériques pour les industries techniques de l'audiovisuel et du cinéma. Son ambition est de doter le secteur de moyens technologiques nouveaux, dans un environnement de travail en mutation permanente et en confrontation avec les exigences de la compétitivité à l'échelle mondiale. Ces moyens doivent être conçus collectivement, dans la perspective d'une industrie ouverte du point de vue des échanges numériques entre diverses entreprises du secteur. Aujourd'hui un producteur de dessins animés peut faire appel à différentes sociétés d'animation pour un même projet, ce qui oblige les créateurs de contenus à travailler dans un environnement homogène (version des logiciels, des plugins, numérotation des séquences, arborescences, vérification de la totalité des données etc…) pour mener à bien le projet.

Le consortium HD3D-IIO regroupe des acteurs professionnels majeurs de la production audiovisuelle (publicité, dessin animés 2D/3D, films de cinéma, effets spéciaux 2D/3D…) en Ile de France, comme les sociétés, Duran Duboi, Eclair, LTC, Mac Guff, Mikros Images, Teamto, ou 2 Minutes.

À long terme, l'ambition de HD3D-IIO est d'augmenter la compétitivité de l'industrie francilienne du film numérique face à des concurrents internationaux. Pour cela, de nombreux verrous technologiques nécessitent d'être levés. Cela concerne la spécification de formats d'échanges de contenus 2D et 3D, l'élaboration de plateformes de production collaboratives avec une maîtrise des développements réalisés sur des sites distants, la spécification de méthodes de réutilisation de contenus ou encore la prise en compte des aspects de protection et de traçabilité des contenus, nécessaires pour leur transmission sécurisée.

Dans ce cadre, un des objectifs majeurs de HD3D-IIO concerne la réutilisation des contenus images en environnement de production. Ce besoin fait écho à la constatation suivante : chaque production (*e.g.* dessin animé, clip vidéo ou film) conduit à des dizaines de milliers d'images. Or, de nombreux éléments (*e.g.* parties de décor d'un film, personnages/accessoires d'un dessin animé 2D ou 3D…), peuvent être ré-exploités lors d'une nouvelle création, à condition toutefois de pouvoir les retrouver facilement dans les collections d'images précédemment produites.

Dans un *workflow* de production, il est très rare d'annoter les images et bien souvent l'indexation manuelle se résume à un nom de fichier qui ne décrit pas le contenu de l'image mais le plus souvent la place qu'elle occupe dans une séquence donnée. Le travail documentaire n'est donc pas effectué le plus souvent par manque de temps et de moyens. Pour retrouver un décor ou une image, on fait donc appel à la mémoire humaine, celle-ci se heurtant au renouvellement des équipes sur chaque nouveau projet.Bien que cette absence d'annotations soit un réel handicap pour la réutilisation des contenus, aucun outil efficace n'est disponible pour accéder à de

telles bases d'images pour faciliter/accélérer les modes de production multimédia en capitalisant sur des créations antérieures. Le seul recours possible est d'exploiter le savoir-faire et l'expérience des professionnels et artistes impliqués dans la chaîne de production. Cela est consommateur en ressources humaines et à un impact sur le temps et donc sur le coût de production.
L'indexation automatique de texture et donc des contenus des bases de données d'images manipulées lors d'une production peut représenter un gain de temps significatif lors de la fabrication d'un dessin animé. En effet, un moteur d'indexation automatique permet la recherche et l'extraction d'images afin d'optimiser la réutilisation de contenus numériques. Aujourd'hui le partage et la réutilisation des contenus dans le cadre d'une production est difficile, les solutions existantes présentent de nombreuses limitations :
– l'indexation manuelle est fastidieuse et peu utilisée à cause des contraintes de temps pour l'indexer,
– la construction et le classement des éléments nécessaires à la production de dessins animés (décors, textures, objets, personnages,…) dans des fichiers à l'intérieur d'une arborescence au sein d'un disque dur partagé sur l'intranet est utile mais peu commode puisque les utilisateurs doivent maitriser la logique de sa construction (classement, arborescence).
– la base de données avec une recherche par mots clés ou par nom ou par extension de fichiers est fastidieuse. Souvent, les mots clés sont absents. De plus, même lorsqu'ils existent, ils sont trop limitatifs. Ainsi, les noms de fichiers suivent la logique définie en début de production et la nomenclature varie fortement d'une production à une autre. Quand à une recherche sur l'extension du nom de fichier sur des images JPEG ou PSD (extension de fichier du logiciel Photoshop) la requête peut conduire à un résultat de plusieurs milliers d'images.
L'indexation automatique par le contenu d'images peut apporter des éléments de réponse à cette problématique. Au minimum trois contraintes doivent être néanmoins respectées pour garantir une solution viable. La première concerne l'efficacité des requêtes (*i.e.* capacité de retrouver l'information pertinente dans des grandes bases d'images). La seconde est liée au temps de calcul associé qui doit être compatible avec des réponses quasi instantanées à des requêtes interactives. Enfin, la troisième concerne l'élaboration de moteurs de recherche adaptés et aisés d'utilisation, supportant un large éventail de formats d'images et intégrables dans les chaînes de production existantes.
La plateforme d'indexation INVENIO proposée dans cet article, décrite dans la section suivante, se propose notamment de répondre à ces différentes contraintes pour assurer une réutilisation optimale des contenus numériques en cours de production.

## 3 La plateforme INVENIO

L'approche d'indexation d'images par le contenu proposée dans la plateforme INVENIO s'appuie entièrement sur les technologies de description de contenus multimédias proposées par la norme ISO MPEG-7.
La section suivante décrit notamment les éléments MPEG-7 retenus pour intégration dans la plateforme en raison de leur pertinence pour les objectifs de réutilisation et d'indexation automatique d'images.

### 3.1 Cadre normatif MPEG-7 : descripteurs retenus

MPEG-7 [3] est une norme de description de documents multimédias. Cette norme spécifie une palette d'outils normalisés pour indexer et décrire syntaxiquement de façon automatique ou semi-automatique tout contenu multimédia. Une même information pourra donc être traitée en fonction des capacités communicationnelles recherchées, allant du spatio-temporel (audio et vidéo traités séparément) à une description sémantique du flux de données. MPEG-7 peut s'associer aux autres descripteurs spécifiant le format, les conditions d'accès, leurs classifications, les liens pertinents en relation avec l'information initiale, le contexte d'enregistrement ou de la diffusion du matériel : c'est la possibilité de naviguer, de chercher, de filtrer et de s'approprier l'information dans un corpus multimédia ouvert.

| **Descripteurs de couleur** | |
|---|---|
| *Nom* | *Principe* |
| Espace de couleur | Spécification de l'espace de couleur de représentation des couleurs (*e.g.* RGB, HSV, Luv, HMMD…) |
| Quantification de couleur | Spécification d'une quantification uniforme de l'espace de couleur (composante par composante) |
| Histogramme de couleur scalable | Histogramme de couleurs indexé et représenté de manière multi-résolution par transformée de Haar |
| Descripteur par couleurs dominantes | Représentation d'une image par un nombre relativement réduit (maximum 8) de couleurs dites dominantes, obtenues par algorithmes de *clustering* |
| Descripteur couleur-structure | Histogramme de couleur enrichi d'information spatiale à l'aide d'un élément structurant |
| Distribution spatiale de couleur | Description de la distribution globale des couleurs dans l'image par transformée en Cosinus discrète |
| **Descripteurs de texture** | |
| *Nom* | *Principe* |
| Histogramme des orientations des contours | Classification grossière des orientations des contours en 5 catégories et construction d'un histogramme par rapport à ces classes |
| Texture homogène | Représentation multi-résolution par transformée de Gabor avec 6 orientations et 5 échelles |

**Tableau 1.** *Descripteurs MPEG-7 retenus pour intégration dans INVENIO.*

MPEG-7 a été développé pour s'harmoniser avec les autres normes utilisées dans les différents domaines d'application préconisés par le W3C. A ce titre, citons : XML, l'IETF (*Internet Engineering Task Force* qui propose les normes concernant Internet), la norme concernant les métadonnées du Dublin Core, celles concernant la terminologie et autres ressources linguistiques de l'ISO TC 37, les métadonnées garantissant les échanges entre les transactions (image, son, données alphanumériques), l'établissement de systèmes ouverts pour des applications de télévision interactive (TV Anytime), la norme ISO/IEC 11179 [6] concernant les registres de métadonnées.

La norme MPEG-7 regroupe un riche ensemble de descripteurs aussi bien par le contenu (audio et visuel) que sémantiques/textuels. Elle propose également une architecture qui combine descripteurs et schémas de description pour décrire des concepts liés à la structure de l'image et de plus haut niveau sémantique. Le langage de description est fondé sur XML Schema.
Parmi les différentes parties de la norme, la partie 3, dite MPEG-7 Visual [3], propose un éventail de descripteurs adaptés et optimisés pour des applications de recherche par le contenu selon différentes caractéristiques visuelles comme la forme, le mouvement, la texture ou la couleur.
Pour l'application HD3D-IIO, seuls les attributs de couleur et de texture ont été jugés pertinents dans la phase actuelle du projet. C'est notamment cette partie qui a été retenue pour intégration dans la plateforme INVENIO.
La norme MPEG-7 propose actuellement quatre descripteurs de couleur et trois descripteurs de texture, présentés succinctement dans le tableau 1 :
L'ensemble des descripteurs retenus sont formalisés à l'aide de représentations aussi bien textuelles (à base de XML Schema) que binaires (pour des raisons de compacité et de transmission). En outre, ils sont munis de mesures de similarité adaptées à base de distance dans les espaces de représentation associées. Cela est indispensable pour pouvoir effectuer des requêtes par similarité.
Pour une description plus approfondie des descripteurs, avec définition mathématique, mesures de similarité associées et propriétés, le lecteur est invité à se rapporter à [7], [8].

### 3.2 Fonctionnalités supportées

Etant essentiellement destiné à des graphistes, l'intégration dans un environnement graphique et ergonomique est déterminante pour son adoption par les utilisateurs. Afin de répondre à ces objectifs, INVENIO propose les fonctionnalités suivantes :
- la requête est faite sous la forme de la recherche par l'exemple. Une image entière ou une région d'une image est tout d'abord spécifiée. La requête en fonction des critères de recherche de couleur/texture spécifiés par l'utilisateur est ensuite effectuée.
- un deuxième mode de requête, cette fois par mots clés sélectionnés à partir d'une liste représentative des textures :
- Matières : métal, bois, pierre, verre, ciel, nuage, liquide, végétale
- Motifs : damier, brique, hexagonal, radial, cubique, cellulaire, graduée, léopard, oignon, matelassé, aplat
- Régularité : ordonnée, désordonnée
- Formes : ronde, ovale, rectangle, carré, autres
- Fréquences : radiale, ondulation, vaguelette
- Couleurs : vert, bleu, jaune, orange, violet
- Types : animale, végétale, humain, objet, paysage

Cette liste permet de constituer une ontologie de type « production », liée à un environnement métier. Les utilisateurs indexent eux-mêmes les images au fur et à mesure de leur recherche. Ainsi cette indexation est réutilisable par l'ensemble de la production et garantie une indexation textuelle à minima pour le futur.
– un système de description personnelle « tags » que l'on peut ajouter permet de représenter l'image de façon personnelle. Lors d'une production, les tags en « langage utilisateur » (*i.e.* annotation en texte libre) permettent d'indexer une image en vue de créer un « dictionnaire terminologique » lié à un projet.

Outre ces éléments de requêtes et d'annotation par descripteurs individuels, le système permet également d'hybrider multiples critères de recherche et de description. La combinaison des descripteurs permet notamment de prendre en compte la polysémie de l'image en croisant un mode recherche lié à la structure de l'image (descripteur de couleurs) et un autre lié à un niveau de représentation visuelle (mots clés et tags). Ainsi ce formalisme descriptif combine différents niveaux de signification de l'image, sa représentation mathématique et sa représentation visuelle. Cette méthode permet donc d'établir des relations entre les différentes composantes d'une description de l'image, cette sémantique donne du sens à partir d'un ensemble de données à priori variées. L'interface déployée offre à l'utilisateur une relation efficace entre une indexation en langage naturel, et un langage de requête structurel afin d'organiser une représentation des modes indexation à partir de critères multiformes.

### 3.3 Interface et ergonomie

La figure 1 illustre la présentation graphique adoptée par INVENIO.
Les différents éléments constitutifs de l'interface, énumérés de 1 à 7 dans la Figure 1 sont explicités ci-dessous :
1 : La représentation graphique de la base de données est symbolisée par une spirale 3D déroulante d'images qui donne à l'utilisateur l'impression de profondeur et qui permet la navigation dans la base. À chaque requête, les images de la base sont triées et présentées par ordre décroissant de similarité avec l'image requête.
2 : Cadre de sélection de l'image sur laquelle on effectue la requête.
3 : Cadre d'affichage/mémorisation de l'image sélectionnée
4 : Barre de navigation dans la spirale des résultats
5 : Liste des critères de recherche structurels : descripteurs de couleurs et de textures (*cf.* 2.1). Le menu utilisateur permet d'indexer l'image à partir de la liste des mots clés : matières, motifs, régularité, formes, fréquences, couleurs, types (*cf.* 2.2), et/ou des tags.

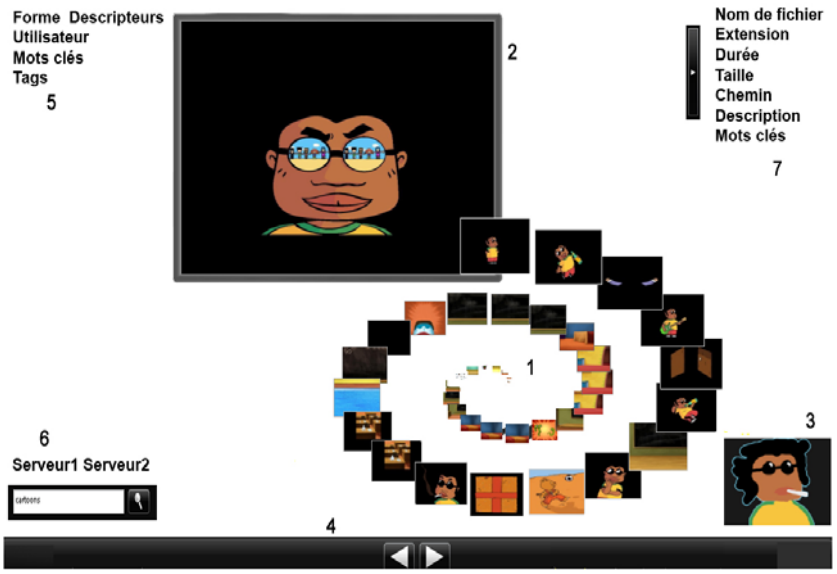


**Figure 1.** L'interface *utilisateur de la plateforme INVENIO*

6 : Menu de sélection des serveurs de calcul : les requêtes peuvent être lancées sur plusieurs serveurs en parallèle afin d'optimiser le temps de réponse aux requêtes. La première requête est prioritaire, la deuxième s'effectue en tache de fond.
7 : Données relatives aux formats du fichier image utilisé et métadonnées textuelles.
Notons enfin que, pour augmenter davantage la sensation d'immersion 3D dans une base d'images, et une interaction plus directe l'utilisation d'un écran tactile est envisagée.

## 4 Evaluation expérimentale objective

Pour évaluer notre approche, nous avons exploité un corpus de deux bases de données. La première contient 2414 images naturelles produites par la société Mikros Image lors du tournage du film Faubourg 36. La taille de ces images est de 3008 x 2000 pixels. Elles représentent dans leur intégralité des scènes urbaines, avec des prises de vue de nuit et de jour, aériennes ou non. La seconde base provient de dessins animés et inclue 4426 images de synthèse (de taille 1024 x 819 pixels pour 90% d'entre elles) produites par la société 2minutes pour différents films d'animation (Yakari, Atomic Betty, Cajou, Step by Step).
Pour évaluer objectivement les descripteurs visuels pour la recherche par le contenu nous avons constitué une vérité-terrain. Pour cela, nous avons construit pour chaque base de données un ensemble de 50 images de test, pour lesquelles nous avons identifié manuellement, avec la participation des experts et artistes, l'ensemble d'images similaires existant dans la base.
Les critères d'évaluation des requêtes par similarités sont les scores FT (First Tier) et BE (Bull Eye ou ST - Second Tier) [9], déjà utilisés pour l'évaluation des descripteurs MPEG-7 lors du développement de la norme. Les scores FT et BE obtenus, ainsi que les temps de calcul associés (calcul des descripteurs et mesure de similarité) sont résumés dans le Tableau 2 :

| **Descripteurs** | **Score FT Mikros/2minutes** | **Score BE Mikros/2minutes** | **Temps d'extraction Mikros/2minutes (s)** | **Temps de calcul de la mesure de similarité (s)** |
|---|---|---|---|---|
| Couleur dominante | 51.53 / **57.21** | 63.86 / **72.58** | 42.12 /10.26 | 0.12 / 0.35 |
| Couleur-structure | **63.21 /** 54.72 | **71.84 /** 67.63 | 1.92 / 0.51 | 0.06 / 0.23 |
| Couleur scalable | 59.24 / 56.30 | 68.69 / 65.11 | 1.87 / 0.42 | 0.07 / 0.25 |
| Distribution spatiale | 52.79 / 40.62 | 64.21 / 53.41 | 0.76 / 0.19 | 0.06 / 0.18 |
| Histogramme d'orientations | 48.58 / 28.93 | 58.43 / 32.09 | 0.32 / 0.07 | 0.09 / 0.25 |
| Texture homogène | 31.49 / 20.52 | 39.87 / 25.29 | 0.2 / 0.18 | 0.2 / 0.48 |

**Tableau 2.** *Performances des requêtes par similarité (bases Mikros et 2minutes).*

Ces résultats montrent que les descripteurs les plus prometteurs en termes de pertinence des requêtes sont ceux par couleur dominante (avec le désavantage d'un temps d'extraction élevé) et celui de couleur-structure, qui offre les meilleurs résultats sur les deux bases. Quant au temps de calcul des mesures de similarité déterminant pour le temps de réponse aux requêtes, ils sont tout à fait acceptables (moins de 0.5 s sur quelques milliers d'images) pour l'ensemble des descripteurs.
Validé par les artistes et les experts graphiques dans le cadre du projet HD3D-IIO, INVENIO offre ainsi un outil performant pour retrouver et réutiliser des images au sein de la chaîne de production.

## 5 Perspectives

La deuxième phase du projet est le projet HD3D² qui s'inscrit dans la continuité de HD3D IIO. HD3D² a pour ambition de mettre en place un outil unique de gestion de la production pour les industries techniques de l'image et du son. Un des volets de ce projet concerne l'indexation de contenus et notamment la fusion des critères de recherche textuels et les critères de structure du contenu dans un environnement métier. Il s'agit donc d'associer les informations liées au système visuel humain et plus précisément la sélection que l'œil opère à partir d'un flux informationnel visuel, et les données qui attirent l'œil. Les cartes de saillance directement liées à la perception visuelle permettent donc de quantifier les différentes zones d'attraction d'une image à partir de critères intrinsèques tels que le contraste, la couleur, le mouvement. Ces critères étant modélisable mathématiquement, le programme de recherche consiste donc à partir du calcul de ces cartes de saillance d'associer automatiquement lors de l'indexation les mots clés préalablement définis dans une ontologie spécifique à un environnement métier.
Cela nécessite dans un premier temps d'identifier les caractéristiques pertinentes à la fois pour l'environnement métier considéré et l'algorithme que nous pouvons associer. Dans un deuxième temps il nous faudra extraire ces données à l'aide de l'algorithme d'extraction et les implémenter dans le moteur de recherche. Ces phases nécessitent l'élaboration et le développement de descripteurs visuels discriminants et dédiés à la fois aux différents types de médias mais aussi aux mots clés associés.
Comment passer du document visuel au sens de sa représentation ? Là se trouve l'enjeu de notre moteur d'indexation et de recherche sémantique. Si pour Marshall McLuhan tout média agit sur le corps social non pas à partir des effets liés à leur contenu mais plutôt des effets culturels que les médias produisent dans le corps social. Notre recherche vise donc à travailler à l'intérieur d'un corps social bien définie (ce que nous nommons l'ontologie métier). Cette approche nous permet d'identifier de façon non ambigüe la relation entre le contenu d'un objet visuel et sa représentation sémantique dans un contexte précis.

## 6 Conclusion

La plateforme INVENIO intègre, dans une interface graphique innovante, des outils d'annotation et de recherche d'images à partir de critères de description d'ordre textuel, sémantique et structurelle.
L'ensemble des outils de description structurelle est fondé sur la norme ISO MPEG-7, dont INVENIO propose une implantation efficace.
Pour valider cette plateforme, nous avons développé une application de réutilisation des contenus au sein d'une chaîne de production audiovisuelle, liée au projet

CapDigital HD3D-IIO. Les expérimentations réalisées concernent différentes étapes de production, incluant des contenus aussi bien naturels que de synthèse (*i.e.* dessins animés). Les solutions proposées dans de cet article démontrent notamment que l'utilisation de la norme MPEG-7 à travers la plateforme d'indexation INVENIO est tout à fait pertinente pour favoriser et accélérer la réutilisation des contenus au sein d'une chaîne de production audiovisuelle.
Dans un deuxième temps, il serait intéressant d'étendre la plateforme, pour le moment dédiée exclusivement à l'indexation d'images fixes, à d'autres types de médias, comme les vidéos et les données graphiques 3D.
A terme, l'objectif de cette plateforme est de démontrer la faisabilité d'une indexation automatique de contenu pouvant aboutir à une génération automatique de mots clés dans des ontologies spécifiques. Cela grâce à des algorithmes d'extraction et d'indexation basés sur le contenu intrinsèque du média considéré.

## 7 Bibliographie


[1] A. Gupta, R. Jain, S. Santini, A. W. M. Smeulders, M. Worring, *Content-Based Image Retrieval at the End of the Early Years*, IEEE Trans. on PAMI, Volume 22, Issue 12, Décembre 2000.

[2] R. Datta, J. Li, J. Wang, J. Z. Wang, Content-based image retrieval: approaches and trends of the new age Proc. 7th ACM SIGMM international workshop on Multimedia information retrieval, p. 253 – 262, 2005.

[3] B. S. Manjunath, P. Salembier, T. Sikora, *Introduction to MPEG-7 : Multimedia Content Description Interface*, John Wiley & Sons, 2002.

[4] ISO/IEC 15938-3 : 2002, MPEG-7-Visual, *Information Technology – Multimedia content description interface – Part 3: Visual*, 2002.

[5] ISO/ IEC 15938-5, Information technology - MultimediaContent Description. Interface - Part 5: Multimedia Description Schemes. 2003

[6] ISO/IEC 11179. (2003). Information technology - Metadata registries. Disponible à: www.iso.org/iso/search.htm?qt=11179&published=on&active_tab=standards

[7] T. Zaharia, F. Prêteux, Descripteurs visuels dans le standard MPEG-7, chapitre dans Gestion des données multimédias (Chapitre 5), *Traité IC2 - Série Informatique et Systèmes d'Information*, Mostefaoui, A., Prêteux, F., Lecuire, V., Moureaux, J.-M. (Eds.), Editions Hermès-Lavoisier, Paris, 2004, p. 85-139.

[8] T. Zaharia, F. Prêteux, Normes de description des contenus multimédias. L'indexation multimédia - description et recherche automatique, *Traité IC2 - Série Traitement du Signal et de l'Image*, P. Gros, Editions Hermès-Lavoisier, Paris, 2007, p. 163-185.

[9] B. Chazelle, D.P. Dobkin, T. Funkhouser, R. Osada, « Matching 3D models with shape distributions », *Proc. International Conference on Shape Modeling and Applications*, Mai 2001.

# Epistémologie du document numérique, pour une approche raisonnée

**Dominique Cotte**

dominique.cotte@univ-lille3.fr
Laboratoire GERIICO, Lilles

**Résumé** : L'étude du document numérique, qui a fait l'objet de travaux importants au début des années 2000 doit se poursuivre en posant un certain nombre de principes, que l'on puisse appliquer aux évolutions constantes de cet objet pour en stabiliser, sinon la nature, du moins le cadre d'analyse. Appliqués de manière systémique, ces cinq critères permettent de rendre compte des tensions qui marquent l'évolution des formes documentaires et leur déploiement dans des univers différents comme la presse en ligne ou les sites de partage vidéos.

**Mots-clés :** Document, document numérique, presse en ligne, épistémologie, sémiotique

## 1 Introduction.

Cette communication vise à s'interroger sur les conditions dans lesquelles encadrer et poursuivre les recherches sur le document numérique, et, au-delà, sur le document tout court.. En effet, l'adjectif « numérique » a été utile un temps pour désigner l'émergence d'un phénomène – et donc d'un objet de recherche – nouveau, mais il nous semble qu'on doive aujourd'hui ré-élargir le point de vue à tout ce qui relève de la production documentaire. En effet, comme nous l'avons souligné ailleurs [6] la caractéristique du document numérique relève moins d'un *statut* que d'un *état*, par définition momentané. C'est toute la sphère du document, au sens large, qui est affectée par l'évolution des technologies numériques. Mais ce retour vers le document ne signifie pas pour autant que l'on puisse repartir de et appliquer tels quels les concepts avec lesquels étaient appréhendés le document traditionnel. Or, sur ce point, malgré les efforts menés par la recherche depuis une dizaine d'années, nous pensons que nous manquons encore à la fois d'un cadre conceptuel général et de concepts particuliers pour caractériser les objets auxquels nous sommes confrontés et les profondes mutations qui les affectent.
Nous aborderons successivement un rapide état de l'art en matière de recherche sur le document, puis une proposition de cadre méthodologique pour l'étude et la définition du document dans ses différents états, et enfin un exemple d'application à un objet documentaire spécifique.

## 2 Un objet dont l'évidence s'évanouit.

« Très peu d'articles scientifiques proposent une définition du document, encore moins la discutent. »
Cette phrase ouvre le premier texte élaboré par le groupe de chercheurs réunis sous le pseudonyme collectif de Roger Pédauque [22]. Ce travail de nature interdisciplinaire mettait le doigt sur un énorme paradoxe qui est le suivant : la recherche sur le document a été stimulée par un mouvement de profonde déstabilisation d'un objet que l'on n'avait pas estimé devoir cerner plus précisément, eu égard à son évidence même. Autrement dit, lorsque le document est un objet de l'évidence, on l'étudie sans le définir, et lorsqu'il cesse d'être évident, on manque de sa définition pour l'étudier dans ses mutations.
Le sociologue Richard Sennett [24] invite à ce qu'il appelle la « micro attention » ; dans un autre registre Georges Perec s'intéressait aux objets de l'infra ordinaire, tellement banalisés qu'on ne les voit plus. Portée sur les éléments décomposables des objets techniques, cette micro attention s'inscrit également dans une logique de l'innovation incrémentale. Il s'agit d'observer comment les objets mutent souvent par le changement ou la métamorphose d'une de leurs parties, qui remet en cause le tout sans forcément bouleverser d'un seul coup la structure ou la morphologie de ces objets. La métaphore biologique, évolutionniste, peut être ici utile, en ce qu'elle décrit une adaptabilité des objets ou structures techniques qui peut procéder par la micro innovation ou la micro adaptation d'une partie au tout. Ce mouvement n'est d'ailleurs pas forcément linéaire ni inscrit dans une logique irréversible. Des formes plus anciennes peuvent réapparaître, métamorphosées. Par exemple l'évolution des appareils téléphoniques a tendu vers une plus grande compaction, en fusionnant dans un seul bloc le combiné, l'écouteur et le socle qui comprenait, entre autres le cadran de numérotation. Un téléphone sans fil ou portable aujourd'hui rassemble ces trois composantes en un seul bloc. Mais on a vu ensuite réapparaître, par dissociation un appareil de locution sous la forme d'un mini-micro détaché de la structure principale. On facilite ainsi un usage rendu parfois difficile par l'évolution précédente, ici la compaction[1].
Dans une approche qui se voudrait Khünienne [17], le numérique interviendrait ici comme l'élément déstabilisant du cycle des révolutions scientifiques. Un objet ou des faits nouveaux soulèvent des questions qui ne rentrent plus dans le paradigme sur lequel vivait la communauté scientifique. Ce paradigme se fissure et l'on rentre dans une période d'instabilité et de controverses à l'issue de laquelle la situation se restabilisera au travers de l'émergence d'un nouveau paradigme. Tel est, grossièrement résumé, le cycle des révolutions scientifiques vu par Kühn. La comparaison avec l'approche khünienne du développement scientifique s'arrête cependant ici, pour deux raisons. La première est que, s'il était peu défini, le concept de document ne constituait pas véritablement un paradigme institué. Par l'antériorité de leurs objets et également par leur origine pragmatique, les sciences de l'information n'ont pas forcément mis à l'agenda une définition stricte de ceux-ci, dans la mesure où ils étaient déjà là. On n'a, a priori, pas besoin de définir ce qui paraît évident. L'autre argument est que, pour être déjà-là, ces objets (le livre, le journal, le document…) étaient également appropriés par la société et objets de pratiques sociales anciennes et culturellement ancrées. Contrairement au paradigme

[1] Il n'est pas rare de voir des personnes utiliser leur téléphone mobile à la manière d'un talkie-walkie, en le portant alternativement à l'oreille et à la bouche, selon la position prise dans l'échange locutoire.

de la révolution scientifique, ce n'est pas la découverte de nouveaux phénomènes naturels, ni la mise en évidence de ces faits en laboratoire (produits ou spontanés) qui viennent mettre en crise le modèle dominant, mais l'apparition de phénomènes qui relèvent à la fois de la technique (évolution des procédés de fabrication des objets documentaires en général) et de l'économie (stratégie des acteurs industriels). Autrement dit, ce n'est pas d'abord dans l'univers d'élaboration des résultats scientifiques que se produit le phénomène déstabilisateur, mais dans la société au sens large, et plus particulièrement dans certaines sphères professionnelles (édition, presse, bibliothèques...) Ce n'est que dans un deuxième temps que les chercheurs se sentiront interpellés par les changements à l'oeuvre en prenant en compte le fait que ces changements ne concernent pas seulement les logiques de traitement ou les modes d'usage des documents, mais la définition même du document comme objet. Ce qui est commun aux deux approches, c'est que nous rencontrons une série de phénomènes déstabilisateurs à l'égard d'un objet et c'est de cette déstabilisation même que surgit une nouvelle façon d'étudier cet objet, autrement dit la nécessité d'un nouveau paradigme. Tout se passe un peu comme dans un mirage où l'image cesse d'être appréhendée au fur et à mesure que l'on s'en approche et que l'on croit pouvoir la saisir; ici ce qui permettrait la prise sur le nouvel objet, à savoir la connaissance de sa forme ancienne, s'avère de peu d'utilité pour appréhender la forme nouvelle, puisqu'il faudrait pouvoir mesurer deux états distants du même objet, alors que celui-ci se trouve remis en cause *à la fois* dans ses anciennes et ses nouvelles formes.

Pédauque poursuit un peu plus loin : « Ce flou fait aujourd'hui problème. En effet, le numérique bouscule profondément la notion de document sans que l'on puisse clairement en mesurer les effets et les conséquences faute d'en avoir au préalable cerné les contours. » (p.29)

Nous avons à faire ici à une véritable mise en abîme et à une injonction paradoxale. Nous nous trouvons en effet devant l'évolution d'un objet de recherche [9] qui suscite un regain d'intérêt et pour lequel il faut mobiliser des concepts inédits, les anciens concepts, jugés eux-mêmes indécis, devenant inopérants en raison même de l'évolution de cet objet. Objet et concepts doivent se (ré)inventer ensemble. Sur le plan de la réalité pratique, il a déjà été relevé que l'univers du document numérique, ou plus largement de l'économie numérique, empruntait au vocabulaire et à l'iconographie de l'ère analogique tous les éléments permettant de décrire ses objets : on parle de *pages* web, d'onglets, de e-book ou *livre* numérique, de *journal* en ligne, de *papier* électronique… Ce recyclage des termes du langage courant, s'il est compréhensible dans une logique d'appropriation des nouveaux objets par le grand public, pose un problème dès lors qu'on se situe dans une logique scientifique.

A ce stade, la communauté scientifique qui étudie le document, a besoin de nouveaux concepts et de nouveaux mots, à la fois pour qualifier pratiquement les objets sur lesquels elle travaille, et pour les penser.

Des jalons pour ce travail ont été posés dans le cadre du collectif Pédauque, avec des concepts comme celui de documentarisation [28]. D'autres collectifs, autour du concept *d'écrits d'écran* [25] et de *médias informatisés* [26] ont proposé de nouveaux concepts comme les *architextes* ou le *textiel*. Des chercheurs travaillent sur les questions de structuration du document [1]. Milad Doueihi [11] évoque une "grande conversion numérique" qui ouvre une ère d'instabilité et de transformation entre les états successifs "du numérique". Des manifestations comme la *semaine du document numérique* [3;4] ou ce colloque international sur le document électronique. Ce travail est à notre sens à poursuivre et à amplifier, par l'élaboration de cadres théoriques, méthodologiques et de concepts adaptés aux nouveaux objets de la recherche.

Jean Davallon [9], à partir d'un constat proche de celui des rédacteurs de Pédauque, aboutit à une proposition qui élargit la problématique à l'ensemble des sciences de l'information et de la communication. En effet, l'une des questions posées est celle-ci :"Comment faire pour que les objets échappent à l'évidence de leur existence (leur aplatissement) comme moyens, leur existence de supports ou de procédures techniques de communication, tout en gardant leur singularité d'objets matériels ou de procédures objectivées ? Comment, dans ces conditions, rendre visible l'invisible de leur organisation en tant qu'objets communicationnels ?"
Ce qui est posé est ici du même ordre que la réflexion qui lance l'entreprise Pédauquienne : comment la recherche doit-elle se situer pour faire exister comme *objets de recherche* ces *objets concrets* qui échappent au regard en fonction de leur évidence même ? Davallon répond en sollicitant une dimension essentielle mais souvent minorée de l'analyse des objets communicationnels : leur caractère inévitablement relié à la technique, considérée dans un sens relativement large et extensif.
"Ainsi, prendre acte de la dimension technique de l'objet, c'est, pour le chercheur en sciences de l'information et de la communication, d'abord et avant tout reconnaître qu'il a affaire à des *complexes* et non à des objets unitaires." [9:34] L'objet est *technique* autant par sa relation à un environnement, sa relation à d'autres objets et en définitive à une filière, pour reprendre un des concepts de Bertrand Gille [14], que par son fonctionnement intrinsèque. L'apport de démarches définitionnelles comme celle de Davallon, est qu'elles permettent de réhabiliter la part de la technique dans la recherche en sciences de l'information et de la communication, sans encourir le reproche de déterminisme qui cache, souvent, de fait, un mépris pour le fait technique trop présent en sciences humaines et sociales. Les objets concrets de la recherche en sciences de l'information et de la communication ont donc une dimension techique importante, qu'il convient de prendre en compte lorsqu'on les construit comme objets de recherche.
De cette première approche, nous pouvons tirer une première conclusion : il n'y a plus de rapport d'évidence à cerner un objet d'étude comme le document. Or, cette évidence distribuait en partie les modalités d'études des objets de communication, dans une approche que nous avons qualifiée de verticaliste, selon une logique de spécialisation par média.
Par exemple le journal comme support est un objet de l'histoire des médias ; comme contenu un objet des études de médiatisation ou des études de discours, comme marchandise un objet de l'économie des médias ou des stratégies médiatiques, comme institution un objet de la sociologie des médias. On a donc à faire à une distribution relativement ordonnée et « logique », qui apparie un objet d'étude et une discipline ou à tout le moins une spécialité. Mais comment doit se redistribuer l'étude du journal en ligne, celle de formes hybrides qui empruntent à la logique médiatique sans pour autant relever de l'univers traditionnel du journalisme [21]?
La déstabilisation de son objet questionne donc la science documentaire, mais au-delà elle constitue un défi pour les « sciences de l'information et de la communication » toutes entières.
Si donc, l'évidence du document comme objet s'est évanouie, un premier travail est de requalifier l'objet de recherche, à partir de l'analyse de l'évolution des objets concrets, pour reprendre le vocabulaire proposé par Jean Davallon. Il reste à définir quels sont les objets scientifiques qui, dans le domaine du document numérique, peuvent servir au renouvellement de la problématique d'étude.

## 3 Proposition de cadrage théorique.

Nous proposons, pour cadrer les études sur l'objet *document*, quelque soit son état, de recourir systématiquement à cinq principes encadrant la recherche. Dans un premier temps nous énoncerons ces principes dans les grandes lignes, et nous les commenterons plus avant par la suite, en les resituant dans une histoire longue du document et de la science documentaire.

Le premier principe consiste à ne jamais dissocier l'étude de la complexion interne du document de son insertion dans un cadre externe, que l'on peut également appeler contexte. Le document ne peut plus être étudié en dehors de son contexte, ce qui revient, à dire, dans le cas du document numérique, de son cadre d'affichage (ou cadre de réalisation au sens où c'est là que le document se "réalise"). Ce cadre est lui-même le plus souvent une métaphorisation des « lieux » d'exercice spécifiques du document (lieux physiques, contextuels, sociaux…). Nous traduisons cette logique de contexte par les concepts d'endogénéité et d'exogénéité du document. Les facteurs endogènes et exogènes se déterminent mutuellement. Le livre appelle l'étagère et l'étagère appelle le livre. Nous faisons ainsi appel à une dialectique qui va plus loin que la simple dialectique de la forme et du contenu. Elle évoque le fait que les objets évoluent dans des environnements et que, tout comme il existe des phénomènes de frottement, il se produit des séries d'ajustements dont l'objectif est de rendre fluide les relations entre facteurs internes et facteurs externes.

Le deuxième principe s'applique à l'analyse de l'objet lui-même. Ce dernier (le document, mais donc aussi son environnement) peut être décomposé en sous-objets, et ceci à plusieurs niveaux de granularité. Par sous-objets nous entendons ici non seulement les composants visibles (par exemple la page dans le livre) mais aussi invisibles, comme le code de mise en page dans le fichier numérique sous-jacent à l'objet [7]. Il convient de rechercher les plus petits éléments identifiables, dans la mesure où ils pourront être à leur tour recomposés et retrouvés dans d'autres contextes (dans une vision en quelque sorte "atomique")

Le troisième principe emprunte à la théorie de la trivialité exposée par Yves Jeanneret [16] : les formes, concepts, idées, circulent en permanence dans la société, s'échangent, se recomposent, se dissocient, migrent... Cela signifie, dans notre cas, mais nous y reviendrons, que nous ne pouvons pas dissocier de manière ferme, ni l'*univers analogique* de *l'univers numérique*, ni même les différents médias entre eux. En dehors même des opnions émises (surtout dans les milieux industriels et professionnels) sur la *convergence*, qui n'est jamais aussi certaine que celà, il existe des passages et emprunts permanents entre les univers communicationnels, documentaires, médiatiques, dès lors qu'ils se fondent sur les mêmes substrats techniques et les mêmes apapreillages logiciels et matériels.

Cette option qui vise à prendre systématiquement en compte le mouvement renvoie aux thèses des grands naturalistes comme Buffon [19] : " tout s'opère, parce qu'à force de temps tout se rencontre, et que dans la libre étendue des espaces et dans la succession continue du mouvement, toute matière est remuée, toute forme donnée, toute figure imprimée ; ainsi tout se rapproche ou s'éloigne, tout s'unit ou se fuit, tout se combine ou s'oppose, tout se produit ou se détruit par des forces relatives ou contraires, qui seules sont constantes, et se balançant sans se nuire, animent l'Univers et en font un théâtre de scènes toujours nouvelles, et d'objets sans cesse renaissans."

Notre quatrième principe implique de ne jamais abstraire l'analyse des formes et contenus documentaires de leurs processus de fabrication, mise en lecture,

diffusion, transport, archivage... autrement dit du substrat technique qui leur permet tout simplement d'exister, et de perdurer. Jean Davallon [9] évoque de son côté, le principe d'une "prise en compte du lestage technosémiotique qui résulte de l'attache de l'objet de recherche aux objets concrets techniques."
Enfin, le cinquième principe fait appel à une théorie de l'innovation qui considère les éléments préalablement dissociés dans leur inter-relation et dans leur mouvement. Les formes et objets nouveaux doivent être regardés comme de possibles reconfigurations d'objets existants, dans une logique de filière, et le caractère innovant ne doit pas forcément être recherché à travers le surgissement ex nihilo de caractères nouveaux, mais dans les formes qui reconfigurent des objets eux-mêmes hérités. Sans que cela en constitue une dimension exclusive, l'innovation comprend souvent une dimension relevant du ré-agencement de formes existantes. "L'invention" de l'imprimerie au milieu du 15° siècle peut être vue comme la combinaison dans une nouvelle configuration d'éléments pré-existants, tels que le caractère mobile, la vis de pressoir ou la pierre à graver. De cette combinaison surgissent des éléments innovants partiels (l'usage de nouveaux alliages pour la fusion des caractères métalliques) ou globaux (la chaîne de fabrication dans l'atelier).
On voit comment ces cinq principes sont associés dans une démarche globalisante et que la méthodologie d'analyse et d'observation qui en découle se déploie à travers des allers-retours dans les différentes dimensions évoquées. N'étant pas abstrait de son environnement (principe d'exogénéité, n°1), le document est étudié dans son tout ou dans ses parties (principe de décomposition, n°2), dans des univers qui ne sont pas figés ni étanches entre eux (principe de circulation, n°3) et en prenant en compte d'une part les infrastructures techniques (dimension technosémiotique, principe N°4) et les phénomènes de recombinaison d'éléments déjà existants (principe d'innovation, n°5).
Pour en revenir à la phrase d'introduction du travail de Pédauque, citée au début de ce chapitre, on comprend alors mieux pourquoi l'objet document est si difficile à cerner en environnement numérique. Analysé par décomposition dans ses éléments premiers, il s'évanouit. Replacé, dans une vision englobante, dans son environnement sémiotique, il perd ses contours et se fond dans un paysage beaucoup plus vaste. La question cruciale reste bien : qu'est-ce qu'on analyse, et à partir de quels critères peut-on délimiter cet objet d'analyse ?
Un retour sur un ouvrage classique, mais méconnu, de la documentation, le "Traité de documentation" de Paul Otlet [20] permettra de voir que ces principes ne sont pas entièrement nouveaux et donc de replacer notre réflexion dans une histoire longue de la documentation. Au moins les deux premiers relèvent explicitement de l'entreprise Otléienne. En cela ils sont structurants parce qu'ils assurent une continuité dans la réflexion au-delà des avatars techniques. Il ne s'agit cependant pas de nier les changements. Otlet avait comme projet de fonder la documentation comme science ; nous ne trancherons pas ici pour dire s'il a réussi, ni même si l'entreprise est légitime. Toujours est-il que le rôle de la science est de fournir des concepts qui sont précisément capable de rendre compte, dans la durée, de la variation des phénomènes. Revenons donc à nos principes, à partir du texte original d'Otlet, pour les reformuler de manière plus approfondie.
Dès le premier chapitre de son texte (0 – Fundamenta), Otlet évoque une idée générale qui constitue, à notre sens, un point fort de son œuvre : la non séparation entre le document lui-même et ses objets/lieux de description et de conservation. Le terme documentation évoque ici plus qu'une somme de documents (la documentation réunie sur un sujet) ou une pratique (la documentation comme activité professionnelle) ; il s'agit d'un ensemble qui relie aussi bien la matière

travaillée (les documents, eux-mêmes considérés comme des produits matériels fruit d'une activité), les outils d'organisation (dossiers, rayonnages), les instruments d'analyse, les lieux (bibliothèques, « offices de documentation »), le tout couronné par une organisation universelle qui organise la documentation en réseau mondial.
Dans cette idée, nous retrouvons nos deux premiers principes, celui de non-séparation du document de l'environnement dans lequel il évolue, et celui de la décomposition du document dans des éléments premiers dont il convient de faire une analytique et qu'il faut nommer, en élaborant une nomenclature spécifique, propre à la science documentaire, afin d'aboutir à un « emploi raisonné des éléments » qui constituent la documentation. L'analyse du document proprement dit, et des éléments qui le composent relève, pour Otlet de la *documentologie*, le terme documentation étant, par conséquent, plus général. Il est à noter que Otlet emploie indifféremment documentologie et bibliologie, le livre étant ici considéré comme désignant de façon générique, universelle, tout type de document. « Livre (Biblion ou Document ou Gramme) est le terme conventionnel employé ici pour exprimer toute espèce de documents. Il comprend non seulement le livre proprement dit, manuscrit ou imprimé, mais les revues, les journaux, les écrits et reproductions graphiques de toute espèce, dessins, gravures, cartes, schémas, diagrammes, photographies, etc. La Documentation au sens large du terme comprend : Livre, éléments servant à indiquer ou reproduire une pensée envisagée sous n'importe quelle forme. »
L'objectif, pour Otlet, est rien moins que de réaliser, pour les objets de la culture et de la vie intellectuelle, la même opération que la biologie a réalisé pour les objets du vivant en unifiant « l'anatomie, la physiologie, la botanique, la géologie » (p.11)
Un autre élément fort de l'entreprise Otléiénne est de ne pas dissocier, sans pour autant les confondre, les éléments matériels et intellectuels, les uns et les autres s'organisant dans un système de références réciproques.
« L'unité physique, matière du document, est marquée soit par la continuité matérielle de sa surface (ex. : la surface d'une lettre, d'un journal), soit par un lien matériel entre plusieurs surfaces (ex. : les feuilles reliées d'un livre) ; soit par un lien immatériel (ex. : les divers tomes d'un même ouvrage) ».
L'une des complexités de l'analyse du document numérique est que son insertion dans un univers marqué par les mêmes ressorts sémiotiques (la page-écran, l'architexte…) rend difficile cette distinction entre éléments matériels et intellectuels puisque les deux sont ramenés sur le même plan à l'écran et que les seconds, pour apparaître, doivent faire l'objet de marques sémiotiques.

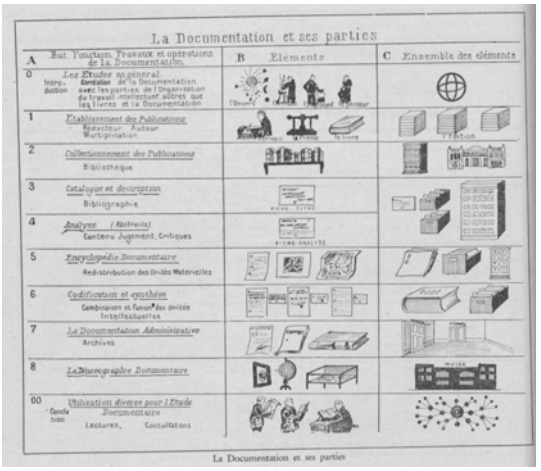


Figure 1. Tableau extrait du "Traité de documentation" (p.42)

On voit, dans le tableau ci-dessus que l'univers documentaire selon Otlet, mêle aussi bien des parties de documents, les outils qui servent à les fabriquer, des lieux pour les stocker, des instruments pour les décrire. Il mèle aussi ce que les théories de l'acteur réseau appellent des "humains et non humains", et, comme Otlet l'écrit plus haut, des éléments matériels et des éléments immatériels (symboliques). Cette question, que nous traduirons plutôt comme étant celle de l'expression symbolique d'une matérialité non tangible, prend une importance particulière avec les techniques contemporaines d'organisation de l'information et de la communication.
Notre deuxième principe est également cohérent avec l'approche d'Otlet. Ce dernier, lorsqu'il cherche à fonder une science du document, ce dernier travaille à une décomposition systématique de tous les éléments qui rentrent dans la fabrication, au sens large, de l'objet.
Ainsi nous trouvons, sous le chapitre 2 ("Le livre et le document"), un sous-chapitre 22 ("Eléments composants du livre et du document"), qui distingue à son tour les éléments suivants :

- 221 "Eléments matériels" (par exemple le papier, la reliure)
- 222 "Eléments graphiques : les signes"
- 223 "Eléments linguistiques : les langues"
- 224 "Eléments intellectuels : les formes d'exposés"
- 225 "Eléments scientifiques ou littéraires du livre : les données de l'exposé"

Puis un sous-chapitre : "23 – Structures et parties du livre", qui détaille à son tour

- 231 Titres et indications externes
- 232 Préface, introduction
- 233 Corps de l'ouvrage
- 234 Tables, index

La même démarche est ensuite répétée pour le journal, les cartes, l'image, le film, bref pour toutes les formes documentaires connues à l'époque.
Une telle démarche systématique et raisonnée n'a pas encore, à notre connaissance, été adaptée au document numérique ou plus exactement au document dans sa configuration numérique, qui peut migrer du numérique à l'analogique, d'un

support informatique à un autre, d'un univers médiatique à un autre. C'est précisément à cause de cette instabilité, considérablement plus forte qu'à l'époque d'Otlet, qu'il conviendrait d'entamer un tel recensement, par décomposition des éléments "atomiques" du document contemporain, préalable à l'analyse de leurs reconfigurations et réagencements successifs.
Les trois autres principes que nous défendons sont moins directement affichés dans la somme Otléienne. Nous ne cherchons pas, en tous les cas, à les relier artificiellement à des éléments qui pourraient figurer dans l'oeuvre d'Otlet, la question ici n'étant pas celle d'une recherche de filiation. Malgré tout, Otlet s'intéresse aux déclinaisons techniques des supports de communication autres que le livre. Achevé en 1934, le Traité de documentation fait état de la photographie, du cinématographe, de la radio (TSF) et des débuts de la télévision.. L'entreprise relève moins, cependant, d'une théorie de l'innovation que d'un inventaire raisonné de ce qui existe et de ce qui pourra se développer eu égard au potentiel technique de l'époque. Quant à la circulation des formes entre divers univers (quatrième principe), c'est une problématique qui est surtout apparue avec l'informatisation. En efet, l'homogénéisation du substrat technique autorise, de manière contradictoire, l'extrême hétérogénéité des formats de présentation ou d'expression des documents.
Nous pensons qu'à partir de ces cinq principes peut être fondée une théorie du document ou du moins un appareillage épistémologique qui permette d'envisager les évolutions contemporaines de cet objet. Cet article ne fera que poser quelques jalons. Il s'agit, dans les pages qui suivent, d'éprouver l'hypothèse conceptuelle évoquée ci-dessus en la confrontant à des éléments déjà étudiés concernant le document numérique.

## 4 Repenser dynamiqument les cadres de déploiement du document

Une vision du document qui se centrerait uniquement sur son organisation et son ingéniérie interne ne permettrait pas de comprendre les mouvements de fonds qui affectent la matière documentaire et qui sont faits de croisements, d'emprunts, de déplacements. Les sciences de l'information et de la communication analysent trop souvent leurs objets de manière cloisonnée ; elles encourent par conséquent le risque de ne repérer suffisamment tôt ni de manière suffisamment aigüe les vrais changement à l'oeuvre. Nous proposons ici de réfléchir, à partir de deux exemples, à la façon dont se *déploient* en se croisant les éléments constitutifs de l'ensemble documentaire formé à la fois par les éléments internes (endogénéité) et les éléments de contexte (exogénéité). La question de la clôture, qui marque un moment important de la réflexion sur le document sera abordée à partir de la presse en ligne, et la question de l'héritage et de l'innovation à partir des sites de partage vidéo sur internet.

### 4.1 La presse en ligne entre dilatation et clôture

Les travaux sur le document numérique – et, partant, sur le document en général – ont pu énumérer, depuis une dizaine d'années, un certain nombre de grandes caractéristiques qui permettent de cerner un peu mieux l'objet. Le relevé de ces caractéristiques s'est naturellement fait en comparant, voire en opposant, les nouvelles caractéristiques du document numérique à celles du document traditionnel. Si cette démarche est méthodologiquement fondée au départ (il faut bien partir d'un déjà-là), elle ne peut rendre compte du profond entrelacement qui caractérise l'évolution du document en général.

Parmi les traits d'opposition entre les deux natures de document on trouve souvent la référence à une logique de fixité, de finitude, de clôture, de globalité qui caractériserait le document traditionnel, opposée à une logique d'ouverture, de mouvance, de transformation, de fragmentation qui caractériserait le document numérique.

L'opposition n'est pas si tranchée puisque, par exemple un quotidien peut connaître plusieurs éditions, avec un contenu qui varie. Autrement dit, dans une unité de temps qui participe de la définition de l'objet (la journée, pour un quotidien), la clôture n'est pas totale ; mais la variation ne peut s'exprimer que par la production physique d'un nouvel objet, partiellement différent du premier (alors que l'édition du lendemain ne conserve aucun des contenus de celle de la veille). Dans le cas du document numérique la variabilité est constamment possible, tant sur le plan spatial que sur le plan temporel. Un site web peut être rafraîchi régulièrement, et ouvert en permanence sur d'autres sites, dans une configuration non stable. Un bon exemple en est Google-Actualités [10] : l'afflux permanent de nouvelles sources d'information reconfigure en permanence l'espace de consultation puisque chaque mention d'une nouvelle source se traduit par un lien renvoyant vers un autre site.

Nous avons à faire ici à une tension entre deux logiques : l'une qui ramène le document vers une unité compacte, organisée, stable, comme condition de sa prise de connaissance, et l'autre qui le tire vers une dispersion, un éclatement en micro-unités qui se retrouvent ensuite recombinées selon des spécificités propres aux différents supports d'accueil.

L'insertion systématique du document dans son contexte (documentaire, spatial, social, institutionnel) permet d'échapper à la dichotomie, puisqu'elle *déplace* la question de l'ouverture ou de la fermeture, de la fixité ou de la variation au-dehors de l'espace du document proprement dit. Ce principe d'exogénéité permet de penser l'insertion du document dans un ensemble plus vaste, ainsi que son déploiement[2].

Sans même reprendre les considérations théoriques empruntées à la sémiologie ( "l'oeuvre ouverte" pour Eco [12]) ou à l'analyse littéraire (l'hypertextualité chez Genette [13]), on ne peut envisager aucun type de document comme parfaitement et définitivement isolé. Considéré dans sa dimension éditoriale, on trouvera, du côté de l'éditeur un projet, l'insertion dans un plan de développement, dans une stratégie éditoriale qui peut s'exprimer, concrètement, par l'appartenance de l'ouvrage à telle ou telle collection [23]. Qu'on le considère dans sa dimension bibliographique, on trouvera la logique de classification qui apparie une unité documentaire à d'autres dans une logique d'approche encyclopédique des contenus. Intellectuellement *et* physiquement, le document, même autonome dans sa logique de finitude, est aussi éclairé par son contexte. La chose devient évidemment encore plus claire pour le document numérique puisque par définition celui-ci ne peut pas être abstrait d'un contexte de lecture, ni d'un appareillage descriptif, auquel il est nativement associé (alors que cette dimension intervient après coup dans le cas du document traditionnel.

Dans la presse ou dans l'édition, l'encart détachable, la fiche à découper, les feuillets à mise à jour, la reliure offerte pour conserver les collections sont autant d'éléments simples qui sont à la fois insérés dans un tout (la livraison périodique) et recomposables dans d'autres univers. Le document traditionnel n'est pas un bloc

[2] Le concept de *déploiement* nous paraît fécond pour l'analyse du document numérique, en ce qu'il contient l'idée d'un mouvement de translation d'une sphère à l'autre, et d'une abolition des frontières trop rigides entre terrains d'étude.

monolithique, qui devrait se laisser saisir d'un seul tenant, sans que l'on puisse repérer, parmi ses composants physiques ou intellectuels, des éléments en migration. La citation analysée par Compagnon pour ne parler que d'elle, est un parfait exemple d'inclusion d'éléments de textes ou de fragments appartenant à d'autres ensembles dont ils ont été préalablement extraits, par découpage.

On a opposé à la clôture du volume papier, dont nous venons de voir qu'elle n'était pas si nette, l'ouverture et la plasticité du document numérique. Mais des contre-tendances peuvent se lire dans les tentatives répétées des éditeurs de presse par exemple, de revenir *enclore* les contenus de presse dans des ensembles qui, tout en restant ouverts (à la mise à jour en ligne notamment), s'insèrent dans un espace circonscrit visant à redonner au lecteur la maîtrise d'une édition globale. C'est l'exemple du *New York Times* qui propose la mise à disposition de ses contenus à travers un lecteur réalisé par la forme Adobe et qui présente ainsi son produit[3] :

« Nous avons construit *Timesreader 2.0* en réponse aux retours que nous avons reçus de vous, notre communauté d'abonnés. Un thème constant dans ces retours voulait que l'expérience de lecture renoue avec les meilleurs aspects de l'imprimé (*the best aspects of print*). Nous vous avons entendu. »

Ce texte renvoie constamment au régime de l'imprimé, en argumentant su plusieurs points :

- la clôture physique (un seul volume livré en lecture, contenant l'information à une place bien déterminée)

« Sur le web, où les lecteurs peuvent ne pas visiter toutes les sections, nous démultiplions les articles à travers les sections. Par exemple une histoire sur une équipe sportive peut apparaître aussi bien dans notre rubrique économie que dans la rubrique sportive. Dans l'imprimé évidemment, on ne le retrouve qu'une fois ».

Le *reader* retrouvant, à nouveau, l'unité de la page comme cadre de lecture et autorisant un feuilletage virtuel à l'écran, permet au lecteur de parcourir séquentiellement l'ensemble de la matière du journal, ce qui permet à l'éditeur de rompre avec l'obligation de redondance qui caractérise le web.

- la clôture temporelle (*un* exemplaire « publié » par jour)

« Sur le web, où les lecteurs peuvent ne pas venir nous visiter tous les jours, nous laissons quelquefois des articles que nous avons publiés la veille ou même l'avant-veille, en section centrale. Dans l'imprimé, évidemment on n'a accès qu'aux nouvelles du jour. Avec Times reader 2.0 vous ne verrez désormais que les articles du jour, et seulement les rubriques qui ont été publiées dans l'imprimé. Cela fait de Times Reader 2.0 un outil plus efficace pour lire les nouvelles du jour. »

- la hiérarchisation de l'information

"Comment les rubriques sont-elles présentées ? Les rubriques d'actualité sont présentées avec la même sélection et la même validation éditoriale que dans l'imprimé. Les suppléments apparaissent le jour de leur publication, avec Science Times les mardis, la gastronomie les mercredis, le magazine le dimanche et ainsi de suite. La maquette cherche à reconduire la même validation éditoriale que celle exprimée dans le journal papier. »

- La lecture séquentielle

« Il existe une fonction très intéressante de balayage (browsing) qui vous permet de survoler les pages du journal. C'est utile pour passer en revue l'ensemble et accroître la sérendipité et l'effet de surprise que l'on trouve souvent dans l'imprimé ».

Ce mouvement de balancier entre dilatation des formes et clôture du document n'est certainement pas achevé. Les éditeurs de presse comptent fortement sur les

[3] http://timesreader.nytimes..com/timesreader/index.html

tablettes électroniques, et notamment l'Ipad d'appel pour recomposer un objet et surtout lui redonner une pertinence économique. Dans le même temps, cependant la matière journalistique se dilate à travers les blogs, les commentaires, l'intervention du lecteur, la multimédiatisation des contenus. Cette tension entre deux mouvement ne peut s'observer, selon nous, qu'en sortant de l'observation du document dans son périmètre étroit d'objet, et en faisant intervenir le contexte dans lequel il est inscrit.

En effet un des points sur lesquels, à notre sens, la recherche sur le document numérique est restée relativement en retrait est l'étude de ce que nous appelons l'exogénéité du document. Il découle en effet de la logique d'appareillage et de la perte d'autonomie du document consultable, que celui-ci n'est activable (au moins dans certains de ses états, l'impression ou la réimpression finale lui restituant son statut d'autonomie) que dans un contexte techno-sémiotique donné. Or, les éléments qui composent ce contexte peuvent eux aussi relever de cette logique de fragment. Les bandeaux, boutons, barres de menus et barres d'outils, systèmes de recherche, et autres éléments qui « encadrent » le document, rentrent dans un rapport d'endo-exogénéité, car ils sont à la fois *en-dehors* et *au-dedans* du document, lorsque celui-ci est dans son état *écranique*. Le texte, dans son sens étymologique de tissu, est bien fait de l'entrelacement de l'ensemble de ces éléments, qui acquièrent alors un sens particulier, et qui changent de sens avec les variations de format.

Dagiral et Parasie [8], à partir de l'analyse de corpus de sites de presse enligne (sites-titres et *pure players*), identifient quatre *régimes* d'insertion des vidéos dans le texte écrit, et montrent que certains d'entre eux se coulent parfaitement dans des moules déjà identifiés dans la presse classique (par exemple au moment de l'introduction de la photographie), tandis que d'autres sont plus directement importés du monde de la télévision.

Deux exemples tirés de sites de presse, montreront ici comment des formes médiatiques ayant pour origine différents supports technologiques s'entremêlent en se citant les unes  les autres.

Dans le premier exemple (figure 2), le journal explique comment, par le biais de Twitter, un député UMP fait « sortir » d'une réunion à huis clos des informations sur le climat de cette réunion. Pour illustrer cette vision moderne de la « fuite » organisée, il copie/colle dans l'article mis en écran sur le site web du journal, des écrans des messages « twittés ».


Fig.ure 2. Extrait de l'article « La colère UMP en direct sur Twitter », Libération du 23 mars 2010

Le deuxième exemple est tiré du site lemonde.fr, qui utilise une technologie héritée du *chat*, pour permettre aux internautes d'interpeller en temps réel le journaliste qui suit, depuis le terrain, un événement. Ici, il s'agit de la manifestation organisée contre la réforme du régime de retraite, le 23 mars 2010 (fig.3). En temps réel, le texte des questions et celui des réponses se succèdent, mais d'autres médias sont sollicités, comme par exemple la vidéo.


Figure 3. Suivi en temps réel de la manifestation, par lemonde.fr, au moyen de l'application CoverIt Live (premier écran)

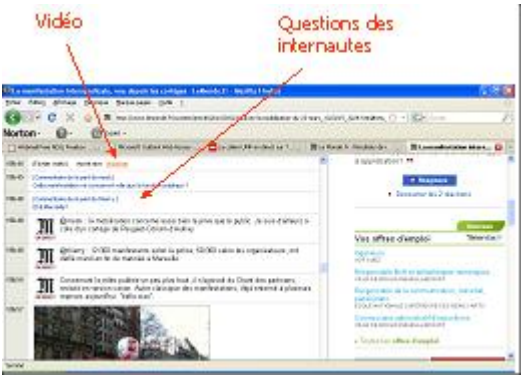


Figure 4. Suivi en temps réel de la manifestation, par lemonde.fr, au moyen de l'application CoverIt Live (deuxième écran)

L'objet, donc, se fait et se défait. Deux questions viennent alors à l'esprit : comment ? et à partir de quoi ? L'objet document numérique est en fait le produit de collages, d'emprunts. Cela implique un modèle particulier d'innovation qui emprunte en les recyclant des bribes à des structures techniques déjà existantes et les agrège en les recomposant. Il est ainsi relativement difficile d'analyser l'innovation véritable. En dehors des grandes ruptures radicales on a plutôt à faire à une logique de patchwork qui entrelace de manière fine le nouveau et l'ancien.
Dans le cadre d'un autre univers, nous abordons cette question dans le paragraphe qui suit, à propos des sites de partage de vidéos sur le web. Ce champ de recherche est, en ce qui nous concerne, en début d'exploration. Nous ne livrons ici que des hypothèses de travail issu d'une première analyse partielle de ces sites.

### 4.2 Les sites de partage vidéos, entre recyclage et innovation.

Nous avons évoqué comme cinquième principe la recherche des filières d'innovation et la façon dont de nouveaux contextes reconfiguraient en les

associant des éléments déjà existants. A titre d'illustration, dans la dernière partie de cette communication, nous analyserons brièvement de nouveaux objets médiatiques, apparus autour de 2005 dans la foulée des sites de réseaux sociaux, et voués au partage de séquences vidéos sur internet. Il s'agit pour partie d'un usage nouveau de techniques existantes et on assiste à la génération de nouveaux modes de consommation de l'image animée, notamment télévisuelle. Nous n'insisterons pas ici ni sur les usages, ni sur les nouvelles discursivités qui s'expriment à travers ces sites. Notre intention d'aborder la question de la (dé)composition de ces objets, en analysant le statut des documents qui y sont présentés et leur cadre de présentation.

En termes d'innovation, ces sites s'inscrivent dans une logique d'assemblage ; ils recyclent des formes existantes, qu'il s'agisse de formes techniques, sémiotiques ou documentaires sans véritablement construire un ensemble nouveau.

Nous cherchons donc à identifier à partir de quels éléments déjà connus sont construits ces sites.

D'un point de vue sémiotique et documentaire, nous retrouvons ici des agencements de contenu qui empruntent pour partie à l'univers de la vidéo, pour partie à celui de la télévision[4] et pour une partie plus large encore à celui de l'internet et notamment des sites de réseaux sociaux (« web 2.0 »). La décomposition de l'objet en éléments simples (en prenant uniquement en compte ici ce qui apparaît à la surface des écrans, c'est-à-dire ce qui relève du *visible*), montre cette conjonction d'univers.

Au plan général nous observons une prédominance de la représentation de l'objet final porté par ces sites, le *gramme* qui est ici une séquence vidéo, de durée variable. En tant que telle, cette unité documentaire première est insécable, mais son indice est donné sous forme de vignette, de dimension calibrée. Contrairement au mur d'images ou à la mosaïque de présentation des chaînes câblées ou satellitaires, il s'agit ici d'images fixes. Leur rassemblement par genre (ici : divertissement, musique, actualités et politique…) renvoie à la logique des annuaires du début du web (et, au-delà aux logiques de classification documentaire).

Chacune de ces unités documentaires est accompagnée d'un court texte structuré qui emprunte à la notice bibliographique quelques uns des éléments de description nécessaires pour caractériser un document (ici son titre, et son auteur ou du moins le « posteur » de la vidéo). L'objectif du partage s'exprime dans l'affichage du nombre de visionnages de la séquence.

L'ensemble du contexte est organisé pour favoriser l'accès à cette première unité de contenu, disponible sous plusieurs formats d'affichage. En ce qui concerne les outils de visualisation et de manipulation de l'image, nous trouvons d'un côté des parties d'écran de taille variable, adaptés aux modes choisis de visualisation (vignette, format *player*, plein écran). Les fenêtres dans l'écran sont équipées des codes de manipulation de la vidéo (avance, recul, pause, play…) qui utilisent une signalétique (flèches, chevrons) elle-même héritée de l'univers analogique.

Un autre élément hétérogène, rapporté ici à la vidéo à partir d'univers différents relève de la logique du *social tagging*, et de l'injonction à participer qui caractérise le « web 2.0 ». Tout en se présentant comme un lieu de recherche et de consommation d'images au premier abord, les sites comme Youtube ou Dailymotion sont avant tout des lieux de chargement de productions vidéos, de la part de particuliers ou

[4] En retour, les chaînes de télévision, en créant leurs sites de vidéo à la demande ou de télévision de rattrapage, ou encore des *web TV*, empruntent à l'organisation des sites de réseaux sociaux.

d'institutions. D'un point de vue sémiotique, ce caractère « participatif » est reconnu par la présence d'icônes de partage qui invitent à qualifier ou évaluer (outils de *rating*, votes…), à classer (« télécharger dans »), à commenter, à partager. De ce fait, le gramme pourra migrer d'une plate-forme à l'autre, ce qui est le cas lorsque des producteurs de vidéo utilisent dailymotion comme plate-forme de stockage pour proposer leur visionnage dans un *player* sur leur propre site.

Cependant, la nature particulière de ces sites est de représenter plus qu'un simple outil de consultation d'images. Leur valeur ajoutée réside dans le *posting*, le partage, la recherche appariée (les captures de pages-écran telles que montrées ci-dessus ne donnent pas le rendu d'une consultation ni surtout le mode de consommation de l'image). Par approximations successives, la proposition incessante de nouvelles images (nouveaux grammes) induit une sorte de frénésie. Leur valeur ajoutée réside dans le regroupement par univers de sens, à travers le descriptif des sujets, mais surtout à travers les « chaînes » réalisées par les internautes inscrits, et les communautés qu'ils constituent. Lange [18] relève que « les participants à youtube utilisent des mécanismes techniques et symboliques pour tenter de délimiter différents réseaux sociaux », et estime après enquête que les inscrits à ces réseaux sociaux jouent sur les collections de vidéo*grammes* pour construire et gérer des systèmes de relations sociales.

Notre thèse est que ces objets sont des assemblages de formes techniques qui sont en même temps des formes éditoriales et des formes documentaires. L'analyse des interfaces de chargement des vidéos, offertes aux internautes inscrits, renvoie à la question des rôles dans la chaîne éditoriale que soulève Valérie Jeanne-Périer à propos des blogs [15].

Les sites de partage vidéo recyclent les formes de trois univers : la vidéo, la télévision, l'internet (web 2.0).

Au premier, ils empruntent les techniques de création et de montage, facilitées, y compris pour les amateurs [5] par la diffusion d'outils grands publics, caméras numériques, appareils photos et logiciels de montage simples. Dans le domaine de l'actualité, la mise en ligne rapide de témoignages amateurs sur les événements donne à la vidéo une ampleur et un relief inédits.

Au deuxième, ils empruntent surtout au plan sémiotique ; l'emboîtage des écrans, les signes utilisés pour manipuler les images (lecture/pause/stop…), la désignation des univers de visionnage (on y parle de « chaînes » créées par les internautes) sont importés du monde de « la » télévision.

Le troisième univers est bien plus qu'un terrain d'emprunt puisqu'il est la condition et le terrain de l'existence même de ces sites ; on y retrouvera donc tout le dispositif techno-sémiotique formé les architextes, les signes passeurs, les logiques de circulation dans les « pages » et à travers les liens dits hypertextes… C'est plus particulièrement un modèle qui s'impose, celui du partage. Ce qui se joue, dans la mise en place de ces nouvelles formes médiatiques, c'est une question de *prégnance*, par rapport à des modèles hérités de médias traditionnels et qui se trouvent remise en cause ou contestés par les nouvelles formes portées par les nouvelles techniques.

Pour faire le lien avec la partie précédente, nous pouvons analyser la façon dont la plupart des titres de presse (qui, comme nous l'avons vu, mettent sur leurs sites des séquences vidéo pour commenter l'actualité), utilisent ces sites de partage de vidéos pour rassembler ces éléments dans ce qu'on appelle, par mimétisme, des « chaînes ». Sur dailymotion, Le monde, le Nouvel observateur, l'Express et bien d'autres constituent ainsi leur « chaîne », qui emprunte à la sémantique de l'audiovisuel, mais ces médias n'ont rien à voir avec l'organisation classique d'une chaîne avec ses grilles et ses programmes. Il s'agit bien plutôt d'une collection d'objets rassemblés

dans un ordre chronologique et qui sont détachés de tout contexte éditorial. Mis à part le logo des entreprises concernées, le graphisme de la mise en page est hérité de dailymotion qui se révèle, à ce titre, le vrai « éditeur » de l'ensemble et qui d'ailleurs propose ses propres liens, et l'accès aux autres « chaînes » selon une classification documentaire qui n'a rien à voir avec ce que pourrait offrir le journal s'il créait lui-même son propre univers télévisé.

## 5 Conclusion

Ce travail s'inscrit dans une recherche d'une quinzaine d'années sur les conditions de production du document numérique ; il consolide une partie des hypothèses émises depuis lors et s'ouvre sur de nouveaux objets, notamment les productions vidéos en ligne dans le domaine de la presse, sur lesquels il reste tout un travail de collecte et d'analyse à mener. A ce titre, il cherche surtout à poser des jalons pour mener à bien des recherches ultérieures sur ce thème.
Étudier aujourd'hui le document, dans tous ses avatars, numériques ou non, revient à étudier un objet en mouvement, instable, en perpétuelle recomposition. Son analyse relève donc plus de la dialectique du vivant que de la confrontation de modèles statiques. Il se compose un paysage totalement bigarré, fait d'emprunts, de recyclages, de transformations d'éléments par ailleurs reconnus dans leurs univers d'origine, mais qui changent en partie de sens et de formes en franchissant la « barrière des médias », comme les virus mutent en franchissant la barrière des espèces. Il nous semble crucial, dans ses conditions, d'établir un cadre épistémologique stable permettant, par le croisement des cinq principes d'analyse que nous avons définir, d'analyser ce cadre mouvant, et qui n'a pas fini de bouger.

## 6 Bibliographie

[1] BACHIMONT B., CROZAT S., Réinterroger les structures documentaires : de la numérisation à l'informatisation, *Revue I3* vol 4, num 1, pp:59-74. 2004

[2] BEGUIN-VERBRUGGE, A. *Images en texte, images du texte, Dispositifs graphiques et communication écrite*, Septentrion, 2006

[3] BROUDOUX E., CHARTRON G., (sous la direction de).Traitements et pratiques documentaires, Vers un changement de paradigme ? Actes Deuxième conférence Document numérique et société, Paris, CNAM, 17-18 novembre 2008, ADBS Editions

[4] CHARTRON G., BROUDOUX E., Document numérique et société - Actes de la conférence organisée dans le cadre de la Semaine du document numérique à Fribourg (Suisse) les 20 et 21 septembre 2006, Adbs Editions

[5] CARNEL J.-S., Des images « plus vraies que vraies » pour les journaux télévisés, la vidéo amateur et son insertion dans le discours institutionnel de la télévision, *Actes, Eutic 2007*, Médias et diffusion de la société de l'information, vers une société ouverte, Athènes, 2007 pp.206-214

[6] COTTE D., Le concept de document numérique, *Communication et langages*, n°140, 2004

[7] COTTE D., Ecrits de réseau, écrits en strates, *Hermès,* n°39, 2004

[8] DAGIRAL E., PARASIE S., Vidéo à la une ! L'innovation dans les formats de la presse en ligne, *Réseaux*, n°160-161, 2010

[9] DAVALLON J., Objet concret, objet scientifique, objet de recherche, *Hermès* n°38,2004 pp.30-38

[10] DESPRES-LONNET M., COTTE D., L'émergence des médias en ligne, *Communication et langages*, n° 154, 2007

[11] DOUEIHI M., *La grande conversion numérique*, Seuil, 2008

[12] ECO U. *L'œuvre ouverte*, Seuil, 1999

[13] GENETTE G., *Palimpsestes*, Seuil, 2000

[14] GILLE B., *Histoire des techniques***,** Gallimard La Pléiade, 1978

[15] JEANNE-PERIER V.*,* L'écrit sous contrainte : les systèmes de management de contenu (CMS), *Communication et langages,* n°146, 2005

[16] JEANNERET Y., *Penser la trivialité, I – la vie triviale des être culturels,* Hermès Lavoisier, 2009

[17] KHÜN T., *La structure des révolutions scientifiques*, Flammarion, 2008

[18] LANGE, P. G., Publicly private and privately public: Social networking on YouTube. *Journal of Computer-Mediated Communication*, 13(1), article 18, 2007

[19] LECLERC, Comte de BUFFON, *Histoire naturelle des minéraux*, disponible sur http://www.buffon.cnrs.fr

[20] OTLET P., *Traité de documentation*, *le livre sur le livre*, Editions Mondaneum, Bruxelles, 1934

[21] MONTETY C., BERTHELOT-GUIET K., PATRIN-LECLERE V., Hybridations des media-marques, *Actes*, Eutic 2008, Dynamiques de développement au Carrefour des mondes, pp.398-408

[22] PEDAUQUE R.T. *Le document à la lumière du numérique*, présenté par J.M. SALAÜN, C&F, 2006

[23] SCHUWER P., *Traité pratique d'édition*, Electre-cercle de la librairie, 2002

[24] SENNETT R., *Ce que sait la main, la culture de l'artisanat*, Albin Michel, 2010

[25] SOUCHIER., et al. *Lire, écrire, récrire. Objets, signes et pratiques des médias informatisés,* BPI, 2003.

[26] TARDY C., JEANNERET Y., *L'écriture des médias informatisés,* Hermès 2007

[27] TOUBOUL A., VERCHER E., Médecine 2.0 : quand communauté, rime avec rentabilité, Eutic 2008, Dynamiques de développement au Carrefour des mondes, pp.275-290

[28] ZACKLAD M., Documentarisation processes in Documents for Action (DofA): the status of annotations and associated cooperation technologies, *Computer Supported Cooperative Work*, Volume 15, Numbers 2-3 / June, 2006, pp. 205-228.

# Webdesign, normalisation & stratégie des firmes

**Hervé Le Crosnier (1) (2) Jean-Marc Lecarpentier (1)**

herve@info.unicaen.fr / jml@info.unicaen.fr
(1) GREYC (Groupe de recherche en Informatique, Image, Automatique et Instrumentation de Caen) – CNRS UMR 6072 – Université de Caen Basse-Normandie
(2) ISCC (Institut des Sciences de la Communication du CNRS, Paris

**Résumé** : Les conséquences pour les webdesigners de la première « guerre des navigateurs » de la fin des années 90 avaient montré l'importance de la normalisation, aboutissant à la création du W3C. L'indépendance des documents vis-à-vis des firmes verticales qui produisent terminaux, outils de composition et plate-formes de diffusion est-elle encore garantie dans la nouvelle période ouverte avec HTML5 et le web mobile ?

**Mots-clés :** documents numériques, diffusion, internet, stratégie, vidéo, web mobile, eBook, HTML5, webdesign

## 1 Introduction

Étant donné la multiplicité des moyens d'accès au web et aux documents, l'acte de création ne suffit plus pour construire un document numérique afin de le présenter au public. Il convient de l'accompagner d'une ingénierie de composition, de lui associer diverses formes de contenu (images, vidéos, animations) et le décliner pour de multiples dispositifs de lecture, depuis les écrans des mobiles jusqu'aux ordinateurs de bureau, et dorénavant les tablettes. Une fois rédigé, avec toutes ces complexités, le document numérique doit encore trouver un chemin parmi les divers modèles de diffusion, entre l'exposition web, le streaming, la circulation de livres numériques ou les Apps pour terminaux mobiles.
Engagé au début des années 90, le travail de normalisation avait pour objectif de réduire la pression sur les créateurs, les compositeurs et les éditeurs de contenus numériques afin de permettre l'accès le plus universel à leurs productions. Cette nécessité relevait de deux objectifs : en finir avec la « guerre des navigateurs » et redonner au document numérique une capacité à construire la mémoire collective, en garantissant l'archivage, et la diffusion la plus étendue. Ces objectifs restent l'apanage de toutes les tentatives de normalisation autour du document numérique, on les retrouve actuellement dans l'adoption du format ePub pour le livre

numérique par le consortium IDPF[1], ou la mise au point de normes de métadonnées pour les bibliothèques numériques telles METS[2].
Cependant, le numérique est aussi un vaste champ de bataille économique pour les firmes technologiques. Dans ce cadre, les documents numériques deviennent des produits d'appel pour toute une série de produits (mobiles, smartphones, tablettes, liseuses, nouvelles consoles de jeux, télévision numérique,...). Proposer une large gamme de contenus devient essentiel pour lancer de nouveaux appareils sur le marché. On pourrait penser que cela favoriserait la normalisation, rendant les documents indépendants des outils de lecture. Mais c'est sans compter avec la tendance de l'économie des réseaux à favoriser les monopoles autour de l'effet de réseau : maîtriser toute la chaîne de production-diffusion-lecture et avoir les moyens d'en évincer les concurrents reste la stratégie principale des firmes du secteur. Elles créent pour cela des systèmes fermés d'achat (les Apps stores) et valorisent des formats propriétaires, notamment en les défendant par des brevets puisque certains pays, en particulier les États-Unis et le Japon, acceptent le dépôt de brevets sur le logiciel, les algorithmes et les formats de données.
L'ouverture des échanges dans un réseau universel et normalisé, l'usage élargi des logiciels libres, la définition de tests reconnus pour juger l'adéquation des outils aux normes (test ACID3[3]), la rédaction de guides de meilleures pratiques (notamment les règles pour l'accessibilité – WCAG[4]-, les règles mobileOK[5] et la logique des cool-URIs[6]) penchent vers la mise au point de solutions collectives, sur lesquelles les webdesigners peuvent construire les documents numériques et les lancer sur le grand terrain de l'économie de l'attention. À cette ouverture et ces solutions s'opposent les stratégies commerciales, les guerres d'influence, la soumission des choix techniques aux rivalités de groupes hyperconcentrés, et la volonté des acteurs industriels de capter à leur profit les développeurs et les designers, en les incitant à se focaliser sur une technologie spécifique pour emballer les contenus dont ils ont la responsabilité. Une stratégie de segmentation du monde documentaire qui fait porter sur les webdesigners les effets de la recomposition permanente des rapports de force industriels, mais aussi les risques concernant la pérennité des documents. Sans parler de l'augmentation du « ticket d'entrée » pour produire des documents qui pourront circuler largement, multiplate-formes et lisibles avec divers dispositifs.

## 2 Vers une nouvelle guerre des navigateurs

« Browser war » est le terme communément utilisé sur le web pour décrire les conflits entre entreprises qui empoisonnaient la vie des webdesigners à la fin des années 90. Netscape possède alors 80% des parts de marché, Internet Explorer 3 environ 10%. Inquiet de sa faiblesse dans le domaine, Microsoft se lance dans le développement de la version 4 de son logiciel. Son lancement en octobre 1997 déclenche la guerre. Fort de son avance, Netscape imagine pouvoir tenir tête au géant. L'installation automatique de IE4 avec Windows98 signe l'arrêt de mort de Nestcape et assure l'écrasante domination de Internet Explorer qui atteignit jusqu'à 90% de parts de marché. Racheté par AOL, le code source du navigateur Netscape est passé en open source via la création de la Fondation Mozilla, créant ainsi

1 http://www.idpf.org/
2 http://www.loc.gov/standards/mets/
3 http://fr.wikipedia.org/wiki/Acid3
4 http://www.w3.org/TR/WCAG20/
5 http://www.w3.org/TR/mobileOK/
6 http://www.w3.org/TR/cooluris/

l'ancêtre du navigateur actuel Firefox. C'est grâce à ses qualités techniques, et notamment à son strict respect des normes du W3C que Firefox doit son retour sur la scène. Les webdesigners ont besoin d'un outil ayant un comportement cohérent, les formateurs et les rédacteurs de tutoriels aussi, qui peuvent alors expliquer des principes et non lister des « trucs et astuces ».
Cependant, l'installation automatique de IE4, puis celle du lecteur multimédia Media Player, avec Windows déclenche une autre bataille, juridique cette fois. Si la procédure menée aux États-Unis par treize États contre Microsoft est abandonnée sur pressions fédérales, la Commission Européenne décide de poursuivre Microsoft pour abus de position dominante. Ce procès, aboutissant à des amendes record qui firent la Une des médias, contraint Microsoft à présenter une version différente de son système d'exploitation en Europe. Pour autant, le navigateur Internet Explorer reste « pré-installé » sur les ordinateurs tournant sous Windows, continuant ainsi à représenter « l'internet » pour les usagers néophytes.
C'est ce qui a conduit l'éditeur norvégien du navigateur Opera à déposer en 2007 une autre plainte [1], moins médiatisée, mais qui va raviver la nouvelle « guerre des navigateurs ». Opera reproche à Microsoft la vente liée du système d'exploitation et du navigateur Internet Explorer. Pour éviter une nouvelle bataille juridique et le risque de nouvelles amendes record, Microsoft négocie une solution alternative. L'accord conclu entre la Commission Européenne et la firme de Redmond prévoit un « écran de choix » (*ballot screen*) du navigateur qui est proposé à l'installation de Windows7 et pour les mises à jour de Windows Vista et XP. Cet écran met plus de 100 millions d'utilisateurs européens devant le choix entre 5 navigateurs : Explorer, Firefox, Opera, Safari et Chrome[7].
Fort de ses récents succès, on pouvait imaginer que la Fondation Mozilla et son produit phare Firefox seraient les grands gagnants de ce nouveau mode d'installation du navigateur. Mais c'est compter sans les nouveaux acteurs, en particulier Chrome, le navigateur de Google. Au départ motivés par ce nouveau concurrent, et fiers de leurs prouesses informatiques, les développeurs de Firefox ont un peu oublié que la bataille ne se joue pas principalement sur le plan technique. Car Google a sorti l'artillerie lourde dans des campagnes de publicité massives et impressionnantes pour son navigateur Chrome dès la fin 2009. Combien d'utilisateurs sélectionneront Chrome au moment du choix du navigateur, simplement parce que Google fait « forcément de bons produits » et que les annonces Chrome, partout dans le métro ou dans la presse, auront inscrit la « marque » dans l'esprit des débutants ? En tout cas les statistiques de navigateurs montrent que les quelques 7% de parts de marché perdues par Internet Explorer depuis janvier 2010 ont bénéficié uniquement à Google Chrome, les autres navigateurs ayant aussi perdu des parts de marché ou au mieux stagné (Opéra).[8]
Le choix du navigateur devient donc un moment stratégique pour les éditeurs de logiciels. Cette nouvelle opportunité pour l'évolution des navigateurs renforce par ailleurs le changement de statut de cet outil au sein du système informatique. Avec le développement d'une approche répartie de l'internet et de l'accès à distance aux logiciels et aux documents (Software as a service, Cloud Computing), le navigateur est en passe de devenir le logiciel central pour l'utilisateur. Certains pensent même qu'il pourrait détrôner les systèmes d'exploitation graphiques. C'est en tout cas l'option choisie dans les développements d'un « Google OS », un système

[7] Ces cinq navigateurs sont proposés sur le « premier écran » du ballot screen... et cinq autres dans un second écran, au grand dam de leurs éditeurs.
[8] Statistiques mensuelles http://www.w3schools.com/browsers/browsers_stats.asp

d'exploitation qui combinerait la machine « locale » et les services répartis sur le web, les données et les programmes étant stockés (et sauvegardés) sur le réseau ou dans des caches temporaires. Une conception en phase avec le développement de HTML5 comme une API sur les pages web et avec la construction de data centers redondants et hyperpuissants [2].
Les produits de Google sont dans ce domaine les plus représentatifs de cette tendance. Microsoft étant difficilement détrônable sur le terrain des systèmes d'exploitation, la société de Mountain View déplace la bataille sur Internet. La suite logicielle Google Docs est pensée comme une alternative simple et gratuite à Microsoft Office. Couplée au navigateur de Google, elle est aussi utilisable hors connexion (grâce à Google Gears). Un type d'applications mixte qui va se généraliser lorsque les navigateurs implémenteront en standard les possibilités de travail hors ligne prévues avec HTML5. Car HTML 5 est aujourd'hui le nœud autour duquel s'articulent tous les affrontements entre les firmes de l'internet, mais aussi tous les espoirs des webdesigners pour enrichir l'expérience utilisateur. HTML5 est accompagné par de nombreuses API qui transforment la page web en une application web qui travaille sur un ou plusieurs documents. C'est un changement radical d'approche [3] qui donne une nouvelle place à la fois aux webdesigners (construction de pages et de services appuyés sur des documents) mais aussi aux programmeurs (les « informations » ont tendance à être envoyées directement au navigateur, par exemple sous forme d'objets JSON, charge à un programme local de les organiser dans des pages présentées au lecteur). Les « documents » ne précèdent ainsi plus la circulation de l'information, mais sont recomposés à l'arrivée à partir de données plus ou moins brutes. La rapidité d'exécution de Javascript d'une part et la mise à disposition d'API sur le contenu des pages HTML5 d'autre part sont des facteurs clés de ce bouleversement. On voit ainsi les test de comparaison des navigateurs inclure de plus en plus un benchmark sur des opérations javascript, autant que la conformité avec les tests ACID3 pour le respect des normes HTML et CSS.

## 3 Navigateurs et Vidéo

La diffusion de la vidéo est le domaine où la dépendance au logiciel utilisé est la plus marquante. Le passage de la diffusion télévisuelle sur internet (convergence) et l'arrivée de la vidéo à la demande s'accompagnent d'une délinéarisation : l'internaute décide du moment, du lieu et de l'outil avec lequel il va accéder au programme. Une pratique renforcée par l'accès au travers des mobiles. La technique d'encodage vidéo et l'usage au travers du navigateur deviennent des questions clé. Jusqu'à présent, les technologies Flash (rachetées par Adobe en 2005) étaient la seule solution pour mettre en ligne des vidéos, via divers « players ». Avec HTML5, le navigateur prend directement en charge la lecture de l'audio et la vidéo. Pour que ces médias puissent être lus sur tous les navigateurs, la norme HTML5 doit définir le codec vidéo qui sera utilisé en standard. La guerre des navigateurs se déplace ici sur ces questions de format et de codec. Un domaine où pullulent de nombreux brevets. Fin 2009, restaient dans la course les formats H.264 sur lequel Apple Microsoft et Adobe possèdent des brevets, Ogg Theora en format libre, et FLV qui reste d'usage principal pour la vidéo en basse définition. Au sein du groupe de travail HTML5, chacun défend donc « son » format. Pas étonnant dans ces conditions que Safari, le navigateur d'Apple, choisisse le codec H.264. Confrontée à une licence annuelle [4] de 5 millions de dollars pour utiliser le format mpeg4 et le codec H.264, il est logique que la Fondation Mozilla, qui refuse ce système de royalties en contradiction

avec les principes du logiciel libre, défende le format Ogg Theora. Jusqu'à présent, Google n'avait pas pris parti et Chrome utilise les deux formats. Mais les relations entre Google et Apple se dégradant depuis le lancement du système d'exploitation Android pour les smartphones, Google acquiert début 2010 la société ON2 Technologies qui développe des codecs de compression vidéo, en particulier le codec VP8. Et le 19 mai 2010, Google, Mozilla et Opera lancent le projet WebM[9] dont l'objectif est de créer un format vidéo optimisé pour le web, ouvert et sans royalties. Celui-ci est en fait le codec VP8, que Google s'engage à rendre public au sein de WebM. En mettant sa force de frappe dans la bataille, notamment en assurant que les vidéos YouTube seront disponibles au format WebM, Google fait d'une pierre deux coups en mettant Apple et Adobe plus ou moins hors jeu dans ce domaine. Reste à voir quel choix fera le W3C pour la norme HTML5. Dans une note faisant le point sur HTML5 et la vidéo [5], Philippe Le Hégaret, *Interaction Domain Leader* au W3C, explique que le problème n°7 (ISSUE-7) du choix du codec associé à l'élément `<video>` de HTML est retiré de la discussion faute de consensus. Dès l'annonce WebM faite, P. Le Hégaret estime que ce nouveau codec est susceptible de faire avancer le problème n°7, tout en restant compatible avec la politique « Royalty-Free » du W3C [6]. En effet il semble essentiel qu'un codec intégré à la norme HTML5 soit ouvert et libre de droit, de façon à assurer un accès pérenne et équitable aux données. Quels webmestres utiliseront les possibilités audio/video de HTML5 sans ces garanties ?

## 4 Web, documents & terminaux mobiles

Apple est devenu l'acteur central dans le domaine du web mobile et des livres électroniques, grâce à ses produits phares iPhone, iTouch et iPad. Et dans ce domaine son avance lui permet de pousser hors de la route un concurrent, et non des moindres, en l'occurrence Adobe, en refusant d'intégrer un lecteur Flash dans le navigateur Safari et sur tous ses terminaux mobiles. Steve Jobs estime que le logiciel d'Adobe est trop gourmand en ressources et finalement aussi dépassé que les disquettes. Pire pour Adobe et les outils de Microsoft .NET ou C#, Apple modifie la licence de développement et stipule que les applications destinées à être distribuées sur iStore ne doivent pas être créées par adaptation d'applications ou en ajoutant une surcouche [7]. Cette règle vise directement la suite logicielle Adobe Flash CS5 qui inclut un outil pour exporter des applications Flash en applications iPhone, tuant définitivement les espoirs de Adobe de s'imposer dans la téléphonie mobile, même si une très récente modification de la licence du iPhone redonne un peu d'ouverture pour les applications externes, en raison de l'accentuation de la concurrence avec Android. [8]

La consultation de sites via des dispositifs mobiles est en pleine explosion et risque de rapidement devenir le premier moyen d'accès au web. Ceci donne aux mobiles un caractère central dans la guerre des navigateurs, que l'on retrouve dans l'affrontement entre Google/Android et Apple/iPhone. Une situation qui contraint Opera et Mozilla à proposer des versions mobiles de leurs navigateurs, profitant du retard de Microsoft dans ce domaine.

Une guerre du mobile que l'on retrouve au centre du modèle de distribution de contenu et d'applications. Le choix d'un contrôle strict par Apple sur tout ce qui peut être installé sur ses dispositifs de lecture, ou celui d'Amazon de lancer un format propriétaire pour sa liseuse Kindle, sont des modèles loin de faire

[9] WebM, an open web media project, http://www.webmproject.org

l'unanimité. Non content d'éliminer Adobe, Apple veut désormais exclure Google du terrain de jeu des produits Apple. Le refus en 2009 de deux applications pour iPhone proposées par Google, en particulier Google Voice, déclenche les hostilités. Le 7 juin 2010, Apple présentait les règles régissant la publicité sur iPhone. Des règles qui excluent Google et AdMob, sa dernière acquisition spécialisée dans la publicité pour mobiles, du marché de la publicité sur les terminaux Apple. Règles pour le moins discutables puisque leur but principal semble en effet d'être l'élimination de la concurrence en stipulant que les données sur le comportement des utilisateurs d'applications iPhone ne peuvent être transmises qu'aux seules agences indépendantes dont l'activité principale est de diffuser des publicités sur mobile, et qui ne soient pas liées à un constructeur mobile ou fournisseur de système d'exploitation mobile. Règles qui excluent de fait Google et Microsoft. L'agence de publicité iAd créée le 1er juillet 2010 devient alors la principale source de publicité sur les produits mobiles Apple, et tente d'amadouer les développeurs en leur promettant 60% des revenus publicitaires produites par leurs applications. Cependant, ces règles risquent de déclencher une enquête de la Federal Trade Commission ou du département de Justice américain pour abus de position dominante.

Le problème du modèle de distribution d'Apple ne s'arrête pas à la publicité. En choisissant le modèle d'agence pour son service de livres électroniques iBook Store (liberté du prix laissé à l'éditeur et partage des revenus en pourcentages), Apple a choisi de mettre les éditeurs de son côté pour contrer Amazon. Ce dernier, leader sur le marché des livres numériques, avait mis les éditeurs en grogne en imposant un « prix unique » de 9,99 dollars pour tout livre numérique. L'arrivée d'Apple et son modèle d'agence donne aux éditeurs une alternative à Amazon et donc des arguments contre le prix unique jugé trop faible. Une reprise temporaire, car dans le déroulé de cet affrontement entre deux modèles de distribution, on a bien vu que le contenu (les livres disponibles) pouvait être mis délibérément en danger par les entreprises de distribution, comme lorsque Amazon a menacé de retirer l'éditeur Macmillan de son catalogue... de livres imprimés comme numériques [9] ; ou quand Apple estime qu'une bande dessinée réalisée à partir d'Ulysse de James Joyce est « pornographique » et ne peut être distribuée sur son réseau [10].

## 5 Le document numérique pris au piège des stratégies commerciales

On voit ainsi que les créateurs de contenus numériques se retrouvent dépendants des stratégies commerciales des grandes firmes, et cela jusque dans la définition même des compositions numériques et des formats de données. Une tendance qui incite ces firmes à proposer des SDK adaptés à leurs visions, sans tenir compte de l'interopérabilité. C'est en ce sens qu'il faut lire l'affrontement entre Apple et Adobe dont nous avons parlé plus haut qui touche au processus de production lui-même. À l'autre bout de la chaîne, le plugin Google Chrome Frame[10] permet d'installer un équivalent de navigateur Chrome à l'intérieur d'Internet Explorer. Une satisfaction évidemment pour les webdesigners qui rejettent profondément les multiples manquements à la normalisation d'IE, mais un marché faustien, car les designers deviennent alors ceux qui pilotent, à partir de leur site, l'installation de Google, et de ses cookies de traçage, chez l'usager.

10 http://www.chromium.org/developers/how-tos/chrome-frame-getting-started

Car au fond, tout ces combats entre firmes n'auraient guère d'importance si des organismes forts pouvaient imposer des normes solides, pérennes, garantissant l'indépendance des producteurs de contenu et des webdesigners. Or c'est au contraire que l'on assiste. Les firmes souhaitent gagner à leur service les producteurs de contenu, avec des arguments tantôt techniques (qualité, rapidité), tantôt commerciaux (partage des revenus, de la vente ou de la publicité), tantôt de facilité d'usage (les nombreuses API proposées aux développeurs de sites qui les lient ensuite à leur fournisseur, notamment dans le domaine de la cartographie en ligne). Le rêve d'une instance se positionnant au delà des guerres commerciales a certainement été celui présidant à la création du W3C, qui associait à côté des entreprises du secteur des universités et des centres de recherches. Devons-nous admettre que ce rêve n'est plus d'actualité, et qu'un faisceau convergent d'intérêts, de complicités et de marchés vont balkaniser la production de contenu elle-même ? Il ne semble plus gère possible aujourd'hui pour les webdesigners et les développeurs d'application de rester en dehors de ce questionnement. Les nouveaux sujets, comme les polices de caractères embarqués (webfont), l'évolution des CSS, l'évolution du format ePub pour les livres numériques multimédias, les métadonnées embarquées (microdata et RDFa),... viennent reposer à chaque fois ces questions. Dans tous ces domaines, on assiste aussi à la volonté des grandes firmes de maîtriser la chaîne de production/diffusion/lecture de bout en bout.
La guerre à laquelle nous assistons aujourd'hui est à l'échelle du volume de données et des moyens de diffusion actuels. L'affaire des négociations secrètes entre Google et Verizon au cours de l'été 2010 [10] montre bien la volonté des grandes firmes de contrôler la chaîne de bout en bout. Les difficultés des instances à réguler ces processus devrait inciter les chercheurs, les web designers et les éditeurs à réagir, à l'image de ce qui a été fait lors de la création du W3C pour en finir avec la nouvelle « guerre des navigateurs ». L'indépendance réelle du contenu vis-à-vis des plate-formes de diffusion, des outils de lecture, et des modèles d'affaire des grands acteurs du secteur devient une extension indispensable de la normalisation et de l'interopérabilité. La recherche en ce domaine ne peut plus se limiter à l'aspect technique, mais inclure les approches issues des sciences économiques, humaines et sociales.

## 6 Bibliographie


7 Opera Press Releases, Opera files antitrust complaint with the EU. , 2007 http://www.opera.com/press/releases/2007/12/13/ Accédé le 16 juin 2010

8 Hervé Le Crosnier, A l'ère de l'informatique en nuages, Le Monde Diplomatique, Août 2008 http://www.monde-diplomatique.fr/2008/08/LE_CROSNIER/16174 Accédé le 19 juin 2010

9 J-M. 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

10 Summary of AVC/H.264 License terms http://www.mpegla.com/main/programs/AVC/Documents/AVC_TermsSummary.pdf Accédé le 23 juin 2010

11 Philippe Le Hégaret, ISSUE-7: codec support and the <video> element. *W3C*, 14 mai2010 http://www.w3.org/QA/2010/05/html5_video.html Accédé le 23 juin 2010

12 W3C Patent Policy, February 2004 http://www.w3.org/Consortium/Patent-Policy-20040205/ Accédé le 22 juin 2010

13 J. Kincaid, Apple Gives Adobe The Finger With Its New iPhone SDK Agreement, *TechCrunch*, 8 avril 2010 http://techcrunch.com/2010/04/08/adobe-flash-apple-sdk/ Accédé le 17 juin 2010

14 David Feugey, Apple assouplit les règles d'accès à l'App Store : Flash et Mono de nouveau acceptés, 10 septembre 2010, *Silicon.fr* http://www.silicon.fr/apple-assouplit-les-regles-d%E2%80%99acces-a-l%E2%80%99app-store-flash-et-mono-de-nouveau-acceptes-41846.html Accédé le 16 septembre 2010.

15 K. Auletta, Publish or Perish : Can the iPad topple the Kindle, and save the book business?, *The New Yorker*, 26 avril 2010 http://www.newyorker.com/reporting/2010/04/26/100426fa_fact_auletta Accédé le 21 juin 2010

16 J. Bosman, Joyce Found Too Graphic, This Time by Apple, *The New York Times*, 13 juin 2010 http://www.nytimes.com/2010/06/14/technology/14ulysses.html Accédé le 15 juin 2010

17 Le Crosnier, Hervé Google et la neutralité du réseau, Les Puces Savantes, 9 août 2010. http://blog.mondediplo.net/2010-08-09-Google-et-la-neutralite-du-reseau Accédé le 16 septembre 2010

# La guerre des étoiles ou nouvel ordre documentaire

**Auteur : Equipe SID[1] / Laboratoire GERiiCO**

*Université Charles de Gaulle - Lille 3*

## 1 Introduction

La guerre des étoiles fait rage… sur Amazon et prochainement sur les objets communicants ! Le système de « rating » (attribution de notes ou d'appréciations) aux mains des internautes contribue à inaugurer un nouvel ordre documentaire qui se dessine selon une double injonction, technique et sociale. Les étoiles apposées sur les livres inscrits au catalogue du vendeur en ligne inscrivent le livre traditionnel dans un processus d'hybridation à la fois technique, papier/numérique, mais aussi organisationnelle où le « tagging - rating » social s'associe aux métadonnées traditionnelles des experts.

Ainsi, la modification des régimes de production et de circulation documentaire ne repose pas sur une simple logique de rupture du monde analogique vers le monde numérique mais sur un processus d'hybridation qui peut s'analyser tant au niveau des dispositifs techniques qu'à celui des pratiques informationnelles, individuelles et collectives. En effet, dans l'économie générale de l'espace documentaire (technique, social, politique, marchand, symbolique…), se pose la question de l'émergence de (nouveaux ?) objets documentaires dans le cadre du processus de « redocumentarisation du monde », au sens où l'entend le collectif Roger T. Pédauque qui définit « la redocumentarisation [comme] une nouvelle forme de documentarisation qui reflète ou tente de refléter une organisation post-moderne de notre rapport au monde, repérable aussi bien dans les sphères privées, collectives et publiques » (Pédauque, 2007).

Face à ce processus, se pose alors la question de la permanence et de la pertinence des méthodes d'analyse appliquées à ces (nouveaux ?) objets. L'objet de la communication est d'interroger le caractère opératoire des différentes approches d'analyse du document face à l'émergence d'objets documentaires innovants. Dans cette optique, nous allons tout d'abord présenter un exemple qui nous semble significatif d'un tel objet documentaire, puis nous étudierons la manière dont les approches développées au sein de GERiiCO (équipe Savoirs, Information, Document) sont en mesure d'en rendre compte. Les différents regards portés sur cet objet témoignent à la fois de la complémentarité et des points de convergence des différentes approches adoptées en sciences de l'information et de la

[1] Ce texte a été écrit collectivement par plusieurs membres de l'équipe SID (par ordre alphabétique) : Laure Bolka, Stéphane Chaudiron, Perrine Cheval, Jean Debaecker, Muhammad Ijaz Mairaj, Susan Kovacs, Yolande Maury, Widad Mustafa El Hadi, Ismaïl Timimi, Shafiq Ur Rehman.

communication. Ainsi, nous verrons comment l'approche historique-diachronique, celle de la réception et de la médiation, l'approche sémio-pragmatique et l'approche classificatoire contribuent à définir un objet hybride et transitionnel.

## 2 Présentation de l'objet

L'objet que nous nous proposons d'étudier est le produit d'une nouvelle technologie développée par Pranav Mistry, chercheur au Media Lab du Massachusetts Institute of Technology, et intitulée « Sixth Sense ». Préalablement à tout développement, il importe ici de préciser que notre étude ne porte pas sur cette nouvelle technologie mais concerne le nouvel ordre documentaire potentiellement induit par ce dispositif et ses impacts sur le processus de « redocumentarisation » du monde.

Le dispositif construit à partir de cette technologie est totalement mobile. Bien qu'il n'en soit actuellement encore qu'au stade de prototype, il est composé d'une caméra, d'un mini-projecteur, de capteurs de mouvements situés au bout des doigts ainsi que d'un miroir reliés à un téléphone portable permettant l'accès à Internet. Jusqu'ici, les informations pouvaient être lues sur des supports papiers ou numériques mais nécessitaient la présence d'un écran pour lire l'information (dès lors qu'elle était issue du numérique). La technologie Sixth Sense présente la particularité de relier les informations numériques au monde physique, en transformant n'importe quel objet ou surface en interface tactile, afin de permettre aux personnes d'interagir numériquement avec leur environnement et d'obtenir des informations complémentaires, dont ils auraient besoin pour des tâches ou situations de la vie courante, telles que lire un journal, acheter un livre dans une librairie, réserver un billet d'avion, faire des courses ou obtenir des informations sur une personne que l'on rencontre.

Dans le cadre de notre réflexion, notre attention s'est portée sur l'une des applications de la technologie Sixth Sense qui a pour objet le livre, et plus particulièrement un livre lors de son achat dans une librairie. L'exemple ci-dessous correspond à la description d'une situation choisie par les inventeurs de la technologie dans laquelle un acheteur souhaitant obtenir des informations à propos d'un livre est mis en scène.

***Imaginons donc un acheteur, venu dans une librairie pour choisir un livre…***

Dans une configuration classique, il se rendrait dans les rayons, lirait, si elles sont présentes, les critiques rédigées par le libraire ou un critique littéraire et posées sur le livre, ou pour obtenir plus d'informations, il demanderait au libraire de le conseiller. Cet acheteur pourrait également utiliser son téléphone portable pour demander le conseil d'un ami ou se connecter à Internet afin de lire les avis des internautes sur l'ouvrage choisi. Avec Sixth Sense, le « fossé » entre ces deux manières de faire est « comblé ». La caméra, le miroir et les capteurs reconnaissent le livre. La connexion au téléphone permet d'avoir un accès potentiel à toutes les informations relatives à ce dernier. Le projecteur transmet l'information directement sur les pages du livre. Par le biais du dispositif, le livre devient donc une interface tactile, faisant, par exemple, apparaître les avis des lecteurs du site Amazon.com sous forme d'étoiles, permettant aux images du livre de déclencher des fichiers audio ; offrant en superposition des informations supplémentaires issues du web sur l'auteur, les

critiques La figure 1 ci-dessous illustre ce nouvel objet-document[2] qui se caractérise par une double strate informationnelle (celle qui correspond à l'écrit papier et celle qui est ajoutée en surimpression numérique).
Nous sommes ici en présence d'un objet physique « papier », situé dans un espace classique de classement (la librairie), avec ses éléments de repérage connus (sommaire, chapitres, résumé, présentation d'auteur…) mais qui en même temps est aussi une interface tactile, reliée au monde numérique, à ses informations et à toutes ses possibilités de fragmentation, de découpage, d'enrichissement, de portabilité et d'interopérabilité.


**Figure 1.** *Cet objet est un bon exemple de redocumentarisation du monde. Comment l'analyser ?*

## 3 Selon une approche historique-diachronique

Parmi les éclairages apportés par une étude de l'évolution des formes documentaires à long terme on peut en citer trois. En premier lieu, l'approche historique offre la possibilité de relativiser – et donc de préciser – l'aspect véritablement innovateur des nouvelles formes et technologies de (re)production des documents en les comparant notamment à des innovations du passé, dans le sens technique ou conceptuel du terme. Ici on peut montrer que la 'convergence' ou l'intégration au sein du document des formes de médiation (paratextuelles) ayant pour but de communiquer la valeur sociale d'un document, n'a rien de foncièrement novatrice dans l'histoire du document : comparons ces étoiles d'Amazon aux 'prix littéraires' matérialisés par les bandeaux rouges, et qui renvoient à des critères de sélection et de légitimation reliés à des enjeux économiques et aux champs d'exercice du pouvoir culturel (Bourdieu, 1992).
Comparons encore ces étoiles aux fiches ou aux étiquettes superposées aux ouvrages et présentant le 'coup de cœur du libraire' ou le 'coup de cœur de nos lecteurs', ou bien à la traditionnelle fiche de prêt, en bibliothèque, montrant le nombre et la date des sorties de l'ouvrage et offrant ainsi au lecteur indécis un indicateur de la 'popularité' de l'ouvrage. Pensons aussi aux choix de classement des ouvrages en librairie ou en bibliothèque en tant que formes de médiation liées non seulement à des normes mais à des valeurs sociales. Cet ensemble de repères

[2] Une vidéo présentant différents exemples d'application de cette technologie est disponible à : http://www.ted.com/talks/lang/eng/pattie_maes_demos_the_sixth_sense.html

contextuels qui servent à situer un ouvrage dans l'univers symbolique du lecteur et à préparer sa réception, s'est développé historiquement par rapport à des critères d'assignation et d'identification des discours qui composent notre « ordre des livres » : la mise en valeur de l'auteur, du genre, des thèmes, de la nouveauté, et des indices variables de pertinence ou de recevabilité des textes (Chartier, 1996).
Dans ce sens, l'affichage des étoiles d'Amazon sur la couverture d'un ouvrage consulté en librairie constituerait un nouveau contrepoids au discours éditorial et aux formes de médiation incitatives proposées en librairie. L'évaluation codée par les lecteurs s'intègrant ainsi à l'appareil paratextuel fait participer un indicateur de réception sociale à la désignation-même de l'ouvrage ; ce bouleversement de « l'ordre des livres », s'il est facilité par l'appareillage technique du lecteur, s'opère non pas par la technique en elle-même mais par le statut inédit donné au résultat du sondage Amazon, devenu critère d'assignation du document. De même, cette nouvelle « convergence » conçue dans une optique d'économie cognitive, pour éviter des démarches trop coûteuses en temps de la part d'un acheteur potentiel pressé, ne renvoie pas moins à un souci fondamental, présent de longue date dans l'histoire des technologies, de créer des dispositifs informationnels qui seraient l'extension ou des facultés du cerveau à traiter simultanément plusieurs types d'informations. Une approche historique aux documents et à leurs technologies afférentes, permet d'interroger les liens entre technique et culture, selon une perspective anthropologique.
Pourtant, et on en vient au deuxième apport potentiel d'un regard historique, le besoin imputé au lecteur ici correspond-il véritablement à sa pratique ? Ce dispositif permettant un accès facilité à des données provenant du réseau Internet, aura-t-il des incidences sur les pratiques sociales du document ? Une relativisation historique de l'innovation montre combien, à la différence d'une définition des médias par leur « potentialité » plutôt que par leurs effets (Bautier, 1994), les appropriations sociales se construisent le plus souvent en différé ou en décalage par rapport à la vision du concepteur.
En dernier lieu, et en conclusion à cette réflexion sur la pertinence d'un regard diachronique porté sur le document, « l'histoire du livre » en tant que discipline offre un protocole d'analyse d'un intérêt particulier pour les SIC, qui tente de décrypter dans sa complexité l'imbrication entre techniques, producteurs et réception (Jeanneret, 2007). Situer les innovations dans une perspective historique, c'est reconnaître – et ce n'est pas un point négligeable – la variabilité et la mouvance des conceptions sociales du document, ainsi que les permanences.

## 4 Selon une approche de la médiation et de la réception

Envisagés sous l'angle de l'appropriation, les nouveaux objets documentaires (ici, le livre hybride qui nous est présenté) offrent aux usagers, via la technologie, de nouvelles formes de médiation, sous forme d'informations qui viennent « augmenter » le livre (dans son sens traditionnel), ouvrant des perspectives d'usage renouvelées. Nous appelons « signes médiateurs » ces informations qui, intervenant dans la mise en forme, la circulation, la communication des documents, « font signe » à l'usager, constituant des modalités de lecture et d'interprétation, donc d'appropriation[3] du document ; signes médiateurs également parce qu'ils participent

[3] Par appropriation, nous entendons, selon une définition partiellement empruntée à Serge Proulx (et transposée au document) « la maîtrise cognitive et technique d'un minimum de savoirs et de savoir-faire permettant éventuellement une intégration significative et créatrice

à la transformation de l'usager, transformation comprise comme déplacement rendu possible grâce à l'exposition au document, avec travail de subjectivation pour construire son propre point de vue sur le document et sur le monde via les documents (dimension socio-cognitive et anthropologique)

C'est à une nouvelle logique d'exposition du livre que correspond le système d'étoiles présent sur la page de couverture. Ce signe médiateur est à la fois d'ordre documentaire (métadonnée) et relationnel (métacommunication). Métadonnée, il est d'abord une forme de signalement pour l'usager, il joue le rôle d'un marqueur, vecteur d'attention ; en situation de surinformation, il fournit un indice de recommandation au futur lecteur qui peut décider ou non de l'adopter. Instrument de métacommunication, il a pour visée la satisfaction de l'usager : il lui permet de « se situer » dans un espace commun, dans une logique d'échange et de partage. Mais si ce système de marquage situe le lecteur dans une perspective dynamique, puisqu'en retour il peut exprimer ses préférences et attribuer aussi une valeur au document, la logique du lien social semble cependant avoir autant sinon plus d'importance que la valeur intrinsèque du document.

Ce qui ne peut qu'interroger sur le processus de légitimation ainsi développé et sur la (re)définition de la valeur et de la pertinence du document qui lui est sous-jacente. Car si ce principe d'intelligibilité distribuée peut paraître séduisant pour l'usager-lecteur qui se sent pris en considération en tant qu'acteur individuel, la personnalisation dans l'approche du document via ce signe médiateur relève aussi d'une catégorisation des singularités, régularisant les pratiques existantes, et visant à rendre calculable le désir (Collins et al., 2005). Une logique de transaction est alors présente sous la dimension communicationnelle, en apparence ouverte à de nouveaux usages. Pris dans ce mélange de considérations informationnelle et documentaire, communicationnelle, sociale, marchande, un risque pour l'usager est de se laisser porter par ce mouvement de préférences partagées, se documenter revenant alors pour lui à accepter de se voir proposer, « par propagation réticulaire », ce que d'autres ont plébiscité (Merzeau, 2009). Dépasser une logique d'adaptation (présente derrière l'idée d'adoption ; le signe médiateur étant alors essentiellement un « connecteur ») pour une logique d'appropriation, orientée vers des usages choisis et réfléchis, fidèles ou non à l'esprit du document, passe par une nécessaire prise de distance et un recul critique.

Avec la zone sensible de la page, signe médiateur qui permet d'ouvrir une nouvelle fenêtre apportant des informations complémentaires, l'usager-lecteur se trouve placé dans une posture dynamique, son geste entraînant une modification du document originel via la technologie. Pour autant si au premier abord la manipulation apparaît comme une manière d'augmenter la valeur ajoutée du document et d'en favoriser l'appropriation, dans un processus dialogique, la réponse à la sollicitation de l'usager est déjà inscrite dans la technologie, elle est davantage de l'ordre de la réactivité que de l'interactivité. Le caractère dynamique du document, son enrichissement sur le mode de l'agrégation, l'hybridation des contenus, sont inhérents à sa structure.

L'échange n'est pas symétrique, mais la carte de la convivialité et de la découverte

---

de cette technologie [pour nous, du document via la technologie notamment] dans la vie quotidienne de l'individu ou de la collectivité ». La démarche d'appropriation vise pour le sujet à acquérir les clés d'accès au document, au service de ses propres objectifs, elle peut être individuelle (acquisition individuelle de connaissances et compétences) ou collective (appropriation sociale).

ainsi offerte peut constituer un levier pour favoriser l'attention et la construction du sens par le lecteur. Car la lecture ne se fait pas sur le mode linéaire, l'appropriation suppose une part d'invention, bien au delà de la simple réception. La dimension interventionniste du lecteur, amené à repérer les signes médiateurs qui lui permettront d'animer le document, s'en trouve majorée ainsi que le remarque Jean-Louis Weissberg (Weissberg, 2001) : il doit comprendre la scénarisation du document, interpréter les indices exprimant des propositions d'action, animer le document en mettant en œuvre les programmes correspondants. Dans l'exemple donné, derrière l'éventail des possibles que permet le nouvel objet documentaire, la dynamique des savoirs se décline sur différents registres, articulant les dimensions sensori-motrice, affective, cognitive, esthétique. L'appropriation passe à la fois par un travail de construction physique, via le geste et le regard (pensée de l'action), et par un travail de re-construction faisant le lien entre la pensée et les signes externes de la culture (médiation sémio-techno-socio-cognitive).
Dans leurs évolutions successives, les modes d'entrée dans les documents se sont diversifiés, les conventions du lire enrichies, favorisant de nouvelles modalités de construction du sens ; pour autant, l'information ne s'est pas affranchie de ses dimensions sociales et matérielles, il y a permanence des supports derrière les changements. La question des modes d'appropriation des nouveaux objets documentaires – et de leur élargissement – semble ainsi toute entière contenue dans la sémiotique de ce que nous avons appelé les signes médiateurs (Jeanneret, 2007).

## 5 Selon une approche sémio-pragmatique

L'approche sémio-pragmatique se distingue d'une sémiologie structurelle dans le sens où ce ne sont pas les signes dans leur existence intrinsèque qui l'intéressent (n'est-ce pas ?) mais la manière dont les significations se construisent en fonction du contexte dans lequel elles apparaissent, phénomène que Peirce a désigné sous le terme de « sémiose » (Peirce, 1978). Le contexte de lecture d'un document, par exemple, comprend à la fois le contexte physique, environnemental dans lequel le document est lu, mais également le co-texte - les éléments textuels entourant le document - ainsi que les dispositions diverses du lecteur, parmi lesquelles notamment son bagage culturel et les attentes qu'il nourrit vis-à-vis de la lecture. Cet ensemble d'éléments contextuels conditionne l'interprétation des signes fournis par le document et donc la lecture qu'il va en faire.
L'analyse du document est aujourd'hui profondément modifiée par l'interdépendance des médias et les contaminations formelles des documents engendrées par cette interdépendance : les signes autrefois caractéristiques de l'internet, tel le curseur de la souris, sont par exemple repris à la télévision, de même l'Ipad cherche à créer une proximité avec l'expérience de la lecture du papier en couplant un visuel de document faisant référence à l'aspect d'un document « papier » à l'action du feuilletage imité par le doigt glissant sur l'écran tactile.
Dans le cas de la « réalité augmentée », l'analyse sémio-pragmatique se trouve confrontée à une étape supplémentaire de la redocumentarisation : gardant sa qualité de livre « papier », par son volume et son caractère manuscrit, l'objet-livre, à l'aide d'un appareillage technique, devient un document personnalisé : ce sont non seulement les avis des internautes qui apparaissent en couverture, mais les annotations de notre « réseau social » qui scandent la lecture, ou plutôt le feuilletage dans le contexte de la librairie que nous donne à voir la simulation de « Sixth Sense ». L'interprétation du document est conditionnée par l'appareillage technique,

qui lui reste annexé et non intégré de manière inconditionnelle. Mais qu'y-a-t-il de nouveau au niveau sémiotique dans cette projection de l'achat du livre dans le contexte de la réalité augmentée par rapport au document numérique actuel ? Le document numérique peut déjà être annoté par divers lecteurs, comme peut l'être également manuellement le document manuscrit. N'y aurait-il alors que la question de la matérialité du document et du contexte spatial de la lecture qui se pose ?
L'analyse sémio-pragmatique nous aide à identifier les marques de l'hybridité documentaire et à poser des hypothèses sur leur interprétation. Ici, l'objet-livre, dans son existence « augmentée » porte des traces de l'Internet : la figure de la notation par le système des étoiles de couleur jaune renvoie d'ores et déjà culturellement aux avis des internautes sur les sites commerciaux (tels Amazon ou la Fnac) et la simple vue des cinq étoiles permet à l'internaute averti, par inférence, de reconnaître que ce sont les avis des usagers des librairies en ligne qui sont mobilisés.
Le papier n'est pas seulement ici un support qui remplace l'écran, il est déjà lui-même un document, auquel viennent se superposer des informations de nature électronique pour créer un document hybride, mi-livre, mi-écran. Il ne s'agit plus seulement d'étudier l'effet systémique des signes dans un contexte d'hybridation documentaire mais d'étudier la superposition de deux strates documentaires, dont la première est indépendante mais conditionne l'existence de la seconde. Ainsi, le livre garde une existence autonome, la réalité augmentée venant y ajouter par exemple des traces de l'expertise communautaire : celles des usagers des librairies en ligne et des contacts des réseaux sociaux. La forme du livre ne change pas mais la superposition de ces deux strates modifie l'appréhension visuelle du livre par le lecteur : dans le feuilletage « augmenté », l'épitexte devient péritexte. Le texte s'en retrouve modifié.
L'approche SIC de la « réalité augmentée » trouve dans la sémiotique pragmatique les outils d'analyse des formes documentaires, de leurs hybridations et de leurs interdépendances ainsi qu'une réflexion ajoutant à l'approche formelle celle de l'appropriation et donc du contexte d'usage du document.

## 6 Selon une approche d'organisation et de classification

Si L'objet communicant est le produit, voire le processus, d'une technologie parue depuis plus d'une décennie, les (r)évolutions techniques actuelles lui confèrent un fort regain d'intérêt en diversifiant les usages et les modes d'appropriation. Le livre – et après une parution timide en versant électronique – se retrouve, lui aussi, assujetti à cette mouvance technologique et s'inscrit dans les réflexions de R.-T. Pédauque «les bouleversements induits par les nouveaux usages du web affectent autant la valeur attribuée aux contenus (crédit, autorité, représentativité) que les modes de médiation eux-mêmes (conditions spatio-temporelles de l'interaction, brouillage des rôles et des sphères « public/privé », camouflage des identités, rupture dans les genres, les discours et les usages, etc.) ». (Pédauque, 2006).
Reposée sur les dynamiques développées dans le web social (collaboration, partage, communauté, réseau…), la technologie de l'information ubiquitaire, largement manifestée dans le projet Sixth Sense, permet, ici dans notre étude, aux usagers d'émettre ou de réceptionner des avis sur des livres présentés sur des rayons du commerce traditionnel et non sur des sites de l'e-commerce. Ces avis sont exprimés au moyen d'indicateurs quantitatifs (nombre d'étoiles) et qualitatifs (commentaires). Il s'agit d'une pratique sociale décentralisée et spontanée, qu'on peut éventuellement désigner par le concept de la folkview, en raison de la proximité de ses aspects avec

ceux la folksonomie.
Les folksonomies, apparues avec l'essor du Web 2.0, forment un système non-traditionnel de classification, mettant en avant la participation de l'usager dans le processus d'inscription de métadonnées. La terminologie employée n'est pas définie, laissant ainsi toute liberté dans la création de l'indexation.
Si l'étude de la folksonomie (classification collaborative basée sur l'indexation), a suscité plusieurs débats entre promoteurs et opposants, le cas de l'évaluation par avis des usagers (classement collaboratif basé sur l'opinion) reste sensible et problématique compte tenu des enjeux socio-économiques, scientifiques et politiques sous-tendant. Les derniers débats sur les modèles d'évaluation de l'activité scientifique au niveau national et international ne sont pas sans rapport.
Comme pour les autres objets, la technologie Sixth Sense confère de nouvelles dimensions spatio-temporelles au livre communicant. Elle accélère le processus de la « redocumentarisation » qui voit l'usager attribuer un jugement de valeur au document, s'approprier des droits sur l'auteur, l'éditeur et le distributeur. On assiste de nouveau à un glissement dans le rôle, le pouvoir et les droits du lecteur. Avec son système d'annotation et de notation, commodité par les outils et les interfaces intuitives développés aujourd'hui par cette technologie, le lecteur devient acteur et force de proposition. Il peut directement remplir des fonctions auparavant dévolues aux autorités expertes. Le livre se retrouve ainsi basculé vers une nouvelle logique de notoriété loin des modèles traditionnels. La question du rôle joué par cette fonction de médiation devient préoccupante et soulève de nombreuses interrogations (Broudoux et Charton, 2009) : Quels rapports entretiennent sciences et médias ? Quelles sont les dérives de la communication médiatisée par la technique ? Pour quelles raisons certaines communautés s'approprient les outils plus rapidement que d'autres ?
Au-delà de son caractère informel et subjectif, cette forme d'évaluation devient un outil de médiation et de mise en espace, un outil de veille et d'aide à la décision. Si l'avis des clients n'a pas eu suffisamment de partisans dans le cas des livres dispensés dans l'e-commerce (Amazon, Ciao), la technologie Sixth Sense propose ici la médiation des internautes mais dans le commerce traditionnel (en présentiel). L'avis exprimé en temps réel sans aucun consensus peut être une valeur ajoutée.
Cependant, cette pratique d'évaluation suscite certaines remarques :

- Si la notation pour son propre besoin n'est pas contestable en soi, le partage de ces avis susceptibles et informels (portant sur des arguments disparates) risque de les convertir, en raison du poids des réseaux sociaux, en indices de notoriété.
- Pour une question de neutralité, le partage d'une évaluation en amont est ipso-facto un biais d'évaluation, non sans influence sur les futurs usagers évaluateurs.
- Si l'évaluation de certains produits de consommation de grand public peut s'avérer exacte sur la base de satisfaction et de témoignages (avis de consommateurs), l'évaluation d'un livre et moins de sa qualité scientifique ne peut être exercée que par des paires. La consommation du livre ne doit pas être conçue de la même sorte que la consommation des autres produits.
- Par des détournements possibles, le dispositif mis pour cette évaluation ne garantit pas la fiabilité de l'information et peut devenir un lieu de promotion de livres et d'auteurs et le lecteur-évaluateur devient acteur clé de la distribution.

Malgré ses limites, cette procédure d'évaluation et de notation sociales a le mérite de co-exister avec les autres projets du web 2.0 et de l'articuler en son sein (Le deuff,

2006). Dans ce contexte, on peut envisager sur l'objet communicant des systèmes de notations hybrides, mutualisant les deux types d'évaluation où l'évaluation sociale est perçue comme complément et non alternative de l'évaluation institutionnelle. Les deux versants d'évaluation ne répondent pas toujours aux mêmes objectifs ; le positionnement n'est pas synonyme de pertinence et la popularité n'est pas l'équivalent de notoriété.
Enfin, l'évolution technique des dispositifs Sixth Sense ne doit compromettre l'évaluation intellectuelle du livre. Certes, le système est pensé aux usagers par les usagers, plusieurs travaux sur l'évaluation ont montré l'intérêt pour les approches orientées usage (Timimi et Chaudiron, 2008). Nous observerons si cette pratique d'évaluation atteint ses engagements et parvient à résister aux contournements techniques, aux stratégies marchandes et aux pouvoirs des spécialistes du marketing viral très présents dans les réseaux sociaux, si elle reste fidèle à sa fonction de réguler de manière experte l'intelligence collective que représentent les futurs utilisateurs du livre communicant.

## 7 Conclusion

Au terme de cet essai de réflexion sur le dispositif étudié selon une approche plurielle et représentative des nos approches théoriques et méthodologiques au sein de l'équipe SID, nous nous interrogeons sur l'apport et la complémentarité de ces approches par rapport aux travaux de Pédauque, mis en rapport avec le contexte de l'information ubiquitaire. L'auteur collectif dégage une méthodologie d'analyse en trois dimensions : anthropologique (le document/forme comme objet à voir), cognitif (le document/texte comme objet à penser) et social (le document/relation comme objet à transmettre). Nous pensons que les approches croisées qui caractérisent notre travail apportent des éléments complémentaires qui contribuent à préciser les outils d'analyse du collectif.

## 8 Bibliographie


1 Bautier R. *De la rhétorique à la communication*. PUG, 1994.

2 Béguin-Verbrugge A. *Images en texte, images du texte. Dispositifs graphiques et communication écrite*. Presses universitaires du septentrion, 2006, collection «communication».

3 Bourdieu P. *Les règles de l'art : genèse et structure du champ littéraire*. Seuil, 1992.

4 Broudoux E., Chartron G. *La communication scientifique face au Web2.0 : Premiers constats et analyse* [en ligne]. Septembre 2009. Disponible sur : <http://hal.archives-ouvertes.fr/sic_00424826/>. (consulté le 01/06/2010).

5 Chartier R., *Culture écrite et société. L'ordre des livres (XIVe -XVIIIe siècle)*. Albin Michel, 1996.

6 Collins G., Crépon M., Perret C., Stiegler B. et Stiegler C. *Ars industrialis association internationale pour une politique industrielle des technologies de l'esprit : manifeste*. Ars industrialis, 2005.

7 Cotte D. « Ecrits de réseaux, écrits en strates. Sens, technique, logique ». In *Hermès* n°39, 2004.

8 Jeanneret Y. *Y a-t-il (vraiment) des technologies de l'information ?* Septentrion, 2000.

9 Jeanneret Y. (dir.). Métamorphoses médiatiques, pratiques d'écriture et médiation des Savoirs. Rapport final de recherche – février 2005 [ACI cognitique - programme société de l'information, écriture, nouvelles technologies, communication et cognition.]

10 Jeanneret Y. « Usages de l'usage, figures de la médiatisation ». In *Communication & Langages* 151 mars 2007, p. 3-19.

11 Le Deuff O., « Folksonomies, les usagers indexent le web », **BBF**, 2006, n° 4, p. 66-70.

12 Leroi-Gourhan A. *Le geste et la parole.* Albin Michel, 1964-65.

13 Merzeau L., « Présence numérique : les médiations d l'identité ». In *Les enjeux de l'information et de la communication*, 2009.

14 Pédauque R.-T. *Document et modernités* [en ligne]. Mars 2006. Disponible sur : <http://halshs.archives-ouvertes.fr/docs/00/06/28/26/PDF/Pedauque3-V4.pdf> (consulté le 03/06/2010).

15 Pédauque , R.T. *La redocumentarisation du monde.* Editions Cépadués, 2007.

16 Peirce, C.-S. *Ecrits sur le signe : textes choisis.* Seuil, 1978.

17 Proulx S. « Usages de l'Internet : la « pensée-réseaux » et l'appropriation d'une culture numérique ». In Guichard, Eric (dir.). *Comprendre les usages de l'Internet.* Presses de l'Ecole Normale Supérieure, 2001, p. 139-145.

18 Souchier E., Jeanneret Y., Le Marec J. (dir.). *Lire, écrire, récrire. Objets, signes et pratiques des médias informatisés.* Paris : BPI, 2003, Collection Etudes et recherche, 349 pages.

19 Timimi I., Chaudiron S. « Information Filtering as a Knowledge Organization process: techniques and evaluation ». In the *International Society for Knowledge Organization* (*ISKO'08*), Montreal, Canada, August 5-8 2008.

20 Weissberg J.-L. « Figures de la lectature. Le document hypermédia comme acteur ». In *Communication et langages*, 2001, n° 130, p. 59-69

# Le document électronique dans le cadre juridique : une précieuse occasion pour la compréhension du droit[1]

**Marina Pietrangelo (1) and Maria Angela Biasiotti (2)**

(1) Institut de Théorie et des Techniques de l'Information Juridique , Italie
(2) Conseil National de Recherches (CNR) - Institut de Théorie et des Techniques de l'Information Juridique (ITTIG), Italie

**Résumé** : Cet essai représente une contribution au débat scientifique sur l'utilisation de documents électroniques dans le contexte juridique, notamment en ce qui concerne les termes de connaissance du droit. En particulier, une description synthétique du cadre juridique général sera fournie pour ce qui en est de l'utilisation des documents électroniques pour la diffusion et la publication d'informations juridiques. Un examen plus approfondi portera sur les outils informatiques et juridiques employés de nos jours pour rédiger, publier et répertorier des documents électroniques au sein de systèmes juridiques avancés pour la connaissance du droit. Enfin, un système juridique fondé sur l'utilisation avancée de documents électroniques sera fourni en exemple, pour terminer dans la dernière partie avec des propositions pour une approche intégrée à la gestion des documents électronique dans le domaine juridique.

**Mots-clés :** systèmes avancés d'information juridique, accès au droit, structuration sémantique du domaine juridique

## 1 Introduction

Cet essai représente une contribution au débat scientifique sur l'utilisation de documents électroniques dans le contexte juridique, notamment en ce qui concerne les termes de connaissance du droit.
L'on considère, en effet, que la connaissance des informations juridiques provenant de différentes sources (règles, jurisprudence, doctrine) soit une valeur fondamentale de tout système démocratique, dans la mesure où il reflète le principe de certitude juridique. Une des conditions matérielles essentielles pour la certitude juridique - entendu ici comme la prévisibilité des conséquences juridiques d'une action – est justement la possibilité pour un agent de connaître les règles (ou la jurisprudence pertinente dans une

[1] Ce document est le résultat d'une réflexion commune des auteurs. Toutefois, l'introduction et la première partie doit être attribué à Marina Pietrangelo, les deuxième et troisième parties à Maria Angela Biasiotti, tandis que les conclusions ont été élaborées conjointement par les auteurs.

juridiction de *droit commun*) en vertu de laquelle une action peut être qualifiée par la loi.
Grâce à l'utilisation de technologies informatiques, au cours des dernières années, de nouvelles formes et méthodes ont été conçues et mises en œuvre pour la divulgation des lois et des informations juridiques qui s'y rapportent (débats parlementaires, arrêts des Cours, décisions des tribunaux de grande instance au niveau national, européen et international, doctrine), ce qui a permis un accès plus conscient au patrimoine juridique public.
Dans ce contexte, le rôle du document électronique, et notamment des documents électroniques rédigés conformément à certains standards (spécifiquement exposés dans la deuxième partie de cette contribution) a été sans doute décisif dans la construction de systèmes juridiques avancés pour la connaissance du droit. Il s'agit de recherches développées dans le cadre de l'informatique juridique, une discipline dans laquelle ont récemment été développées de nombreuses applications spécialisées de façon à soutenir l'évolution imprévisible du droit dans la société contemporaine.
Dans la première partie de ce travail, une description synthétique du cadre juridique général sera fournie pour ce qui en est de l'utilisation des documents électroniques pour la diffusion et la publication d'informations juridiques, en référant en particulier de l'état de l'art au niveau communautaire.
La deuxième partie portera sur un examen plus approfondi des outils informatiques et juridiques employés de nos jours pour rédiger, publier et répertorier des documents électroniques au sein de systèmes juridiques avancés pour la connaissance du droit.
Dans la troisième partie de ce document sera présenté l'exemple d'un système juridique fondé sur l'utilisation avancée de documents électroniques, conçu et établi sous forme de prototype au sein d'un projet du Conseil National des Recherches italien.
Enfin, dans la dernière partie, nous allons faire des propositions pour une approche intégrée de la gestion électronique des documents dans le domaine juridique.

## 2 Première partie. Le document électronique pour la compréhension du droit.

Dans de nombreux pays de l'"Union européenne ainsi que d'autres pays du monde, les expériences d'informatisation publique commencées au début des années quatre-vingt-dix ont porté à la prédisposition de systèmes informatifs aux caractéristiques techniques sophistiquées, qui permettent actuellement aux citoyens de récupérer facilement les textes juridiques publiés sur l'Internet par une entité publique. Ceci permet de réaliser, dans la plupart des cas, la publication électronique des normes juridiques aux fins de divulgation, ainsi que les arrêts des principales Cours nationales et supranationales.
En ce qui concerne notamment les normes, leur consultation s'effectue par un accès unique aux ressources publiées volontairement sur les sites institutionnels des différentes administrations publiques. Il s'agit généralement d'un méta-moteur de recherche dans lequel les documents contenant des informations juridiques sont mis à disposition par chaque entité gouvernementale sur leur respectif *site Web*. Dans de nombreux cas, les documents électroniques relatifs aux différents textes législatifs ont été rédigés et publiés de façon à permettre l'apport de modifications et amendements de manière semi-automatique, ainsi que de permettre à l'utilisateur de

consulter le contenu d'un texte législatif au cours de toute sa période de validité (la soi-disant "multi-validité").
Il s'agit, en fait, d'une modalité avancée pour la consultation des normes qui permettrait de faciliter la connaissance du droit, en particulier dans les systèmes de *droit civil* qui sont généralement composés d'un grand nombre de normes. Certaines des premières expériences à cet égard sont le projet italien «Norme in Rete » (désormais remplacé par le projet «Normattiva », consultable à l'addresse *www.normattiva.it* [2], et le portail français « Legifrance » (www.legifrance.gouv.fr).
Dans d'autres jurisdictions, en particulier celles *de droit commun,* cependant, la législation primaire et secondaire est accessible librement *en ligne,* mais la consultation ne peut être effectuée qu'après avoir accédé à une base de données centralisée spécialement conçue pour archiver les documents électroniques relatifs aux singles actes normatifs, ainsi que toute modification ou amendement qui leur ont été apportés par la suite. Il s'agit de documents pour la plupart publiés en format PDF, et qui sont donc statiques, non modifiable dans leur contenu, sinon par une substitution intégrale du document électronique original. Ainsi est le cas du Royaume-Uni avec le « UK Statute Law Database » *(www.statutelaw.gov.uk)* et de l'Irlande, avec l'« *Electronic Irish Statute Book database* ( http:/ / www.irishstatutebook.ie ).
A côté de ces services, qui visent à la divulgation maximale des normes juridiques, s'ajoutent aussi les procédures de numérisation des publications officielles, qui dans la plupart des Etats membres sont également publiées sur les sites *web* institutionnels, sur lesquels la consultation et l'extraction des textes est le plus souvent gratuite (comme dans le cas de la Bulgarie, de Chypre, du Danemark, de l'Allemagne, de la Grèce, de la Lettonie, de la Lituanie, du Luxembourg, de Malte, des Pays-Bas, de la Pologne, du Portugal, de Monaco, de la République Tchèque, de l'Espagne, de la Suède, et de l'Hongrie ; pour l'Italie - malheureusement - la gratuité n'est que temporaire *: www.gazzettaufficiale.it).* Il est entendu, évidemment, que cette représentation du droit en format numérique est à titre exclusivement indicatif.[2]
Le droit communautaire est de même actuellement accessible que par édition imprimée, étant donné que la publication des documents communautaires en format numérique dans le système EUR-Lex ne constitue qu'un outil de diffusion. L'objectif, cependant, semble être un passage définitif à la publication numérique: : "*It is certain that the future of European Union law is eLaw – eLaw meaning law which id electronic, efficient, ergonomic and European law. The Official Journal of the Union in 2016 is probably an authentic electronic journal, with some paper copies distributed to the Member States*"[3].
Si un recours aux documents électroniques a permis de réaliser des systèmes informatiques pour la publication de normes, tout d'abord à titre uniquement d'information, et successivement à valeur légale, l'emploi de ces documents pourrait permettre de concevoir des systèmes dans lesquels la publication des normes serait accompagnée d'autre informations ou de méta-informations

---

[2] Pour une reconstruction complète du cadre communautaire en ce qui concerne la publication du droit sur le Web, voir *Access to legislation in Europe. Guide to legal gazettes and other official information sources in the European Union and European Free Trade Association,* Publications Office of the European Union, 2009.

[3] K. Rissanen, *Access to EU law and eLaw – visions and challenges,* in *Séance académique « 25 années de droit européen en ligne »,* Publications Office of the European Union, 2006.

juridiques, qui conduiraient à l'établissement de systèmes juridiques plus raffinés et spécialisés.
Ces systèmes permettraient à l'utilisateur de repérer, dans un domaine d'intérêt sélectionné, tout type d'information juridique qui lui est associé: un arrêt, le commentaire d'une norme ou d'un arrêt, la législation connexe, ainsi que toute information contextuelle, pas nécessairement juridique, mais qui pourrait toutefois se révéler utile pour comprendre le cadre juridique d'intérêt.
Dans la société contemporaine, en effet, différentes catégories d'utilisateurs, plus ou moins spécialisés (les professionnels, l'administration publique, les operateurs du droit, les citoyens) doivent pouvoir accéder chaque jour à une multitude de documents de contenu juridique ; c'est à eux que sont principalement adressés les systèmes mentionnés ci-dessus, étant donné qu'ils visent à faciliter l'accès à l'information à l'aide outils de recherche et de navigation spécialisés. La facilité d'accès, la capacité de récupérer rapidement tout les renseignements nécessaires, la gestion et la facilité d'utilisation représentent donc des paramètres clés.

## 3 Deuxième Partie. Outils informatiques et juridiques pour la compréhension de l'information juridique.

Actuellement, un utilisateur dispose de différents types d'outils ou de canaux de recherche pour repérer une particulière information[4]:

1. Recherche basée sur les détails d'une mesure
2. Recherche basée sur les références normatives
3. Recherche à l'aide de certains outils de soutien pour la recherche
4. Recherche textuelle ou au travers d'un mot clé

En fait, peut-être par souci de simplicité, que l'utilisateur soit avec ou sans expérience en matière juridique, et bien qu'il connaisse certains détails de la mesure qu'il recherche, il utilise généralement une recherche par mots-clés ou une recherche plus souvent définie en tant que "recherche textuelle". Cette option oblige l'utilisateur à préciser dans la requête les *mots* exacts qu'il considère pourraient se trouver dans le document, avec le résultat que si le choix de l'utilisateur (qui doit être particulièrement prudent à ne pas utiliser des mots trop généraux ou trop spécifiques) ne correspond pas au choix effectué, par exemple, par le juge ou les législateurs, l'utilisateur se trouvera incapable de trouver l'information qu'il recherche, bien que cette information existe. Si, au contraire, il a sélectionné le mot correct, l'utilisateur va certainement faire face à un large nombre de documents provenant de domaines différents et contenant le même mot, et devra donc effectuer lui-même une sélection pour obtenir l'information qu'il recherche. L'emploi du terme 'loyer' au lieu du terme technique «location» peut comporter que si le législateur n'a jamais utilisé le premier mais s'est limité au terme technique, l'utilisateur ne trouvera pas l'information qu'il recherche. Le domaine juridique est donc un domaine où la recherche rencontre des limitations que nous pouvons caractériser comme naturelles, étant donné qu'elles inhérent à la nature même de la matière qui exige que le langage commun utilisé au cours d'une requête soit successivement traduit en jargon technique de façon à ce que la recherche effectuée par des modèles de *Information Retrieval* soit efficace.
Le mot assume donc une valeur fondamentale dans la recherche d'information juridique et en particulier au sein de la recherche spécialisée, démontrant par conséquent le besoin pressant de réduire la naturelle ambiguïté du langage. Disposer

[4] G.Pascuzzi, C*ercare il diritto. Come reperire la legislazione, la giurisprudenza e la dottrina consultando libri e periodici specializzazione*, Bologna, Zanichelli, 1998.

d'outils capables de mieux gérer les informations électroniques en général et les informations juridiques en particulier est devenu un besoin de plus en plus évident.
Ainsi, une telle facilité pour les utilisateurs peut dériver de l'adoption de systèmes pour la gestion et la recherche d'informations juridiques : des outils et des ressources sémantiques d'une part, des techniques de références croisées d'autre part, de façon à guider l'utilisateur en orientant la recherche vers un meilleur résultat. Les premiers sont des outils qui, au travers d'un système de métadonnées, ajoutent aux documents des informations de type sémantique relatives à leur signification ou à leur contenu. Grace à ce procédé, se rendent souvent explicites, et donc reconnaissables et capable d'être professées par une machines, des informations qui étaient auparavant implicites au document, bien que extrêmement significatives et représentatives. Les secondes sont des techniques de connexion et de référencement qui relient différents types de données conformément à une logique de domaine, créant ainsi un ensemble d'informations complètes et intégrées (par exemple, une loi peut être accompagne par des références aux décisions des tribunaux qui s'y rattachent, ainsi que les contributions doctrinales sur le même thème, etc ..).

## 4 Outils d'indexation sémantique

Les outils sémantiques (métadonnées sémantiques) sont des ressources qui s'appliquent au delà des limites de la recherche syntaxique, c'est à dire la recherche qui se base sur l'identification des textes d'une base de données qui contiennent les termes recherchés [5]
Les plus communs outils d'indexation sémantique sont les rèpertoires des taxonomies, les schémas de classification, les thésaurus, les ontologies [7]. En particulier:
- Les rèpertoires des taxonomies sont des listes de termes qui permettent d'identifier les questions pertinentes à un certain domaine disciplinaire;
- Les schémas de classification comportent une structuration hiérarchique des termes qui y figurent: les termes de portée plus restreinte sont subordonnés aux termes de plus grande envergure: la matière indiquée par un terme général englobe toute les matières indiquées par les termes plus précis;
- Un thésaurus est une liste structurée d'expressions qui vise à représenter sans ambiguïté, dans un système de documentation, les concepts contenus dans un ou plusieurs documents. Un thesaurus comporte (i) des descripteurs, à savoir des mots ou des expressions qui désignent de manière non ambigüe les concepts constitutifs du domaine couvert par le thesaurus, (ii) des non-descripteurs, à savoir des mots ou des expressions qui désignent, dans un langage naturel, un concept identique ou équivalent, ou même un concept considéré, dans la langue du thésaurus, comme étant équivalent aux concepts représentés par les descripteurs, (iii) des relations sémantiques, à savoir des relations relatives à la signification des termes, soit parmi les différents descripteurs que entre des descripteurs et des non-descripteurs. Les relations représentées dans un thesaurus sont des relations d'équivalence (entre descripteurs et non-descripteurs), ou des relations hiérarchiques et associatives (entre les descripteurs);
- Une ontologie définit les termes employés pour décrire et représenter un domaine de connaissance. Une ontologie peut être utilisés par des personnes, des

[5] R.Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley,1999; P. Bertolotti, *Semantic Web: analisi e utilizzo di strumenti per la sua realizzazione*, Université de Turin, 2002.

applications et des bases de données qui ont besoin d'échanger des informations d'un certain domaine de connaissance (c'est-à-dire un secteur spécifique de la connaissance, comme la médecine, la musique, la gestion d'une entité publique, etc.) Une ontologie comporte les définitions des concepts de base du domaine et les relations qu'ils entretiennent, en utilisant un langage compréhensible et utilisable par un ordinateur.[6] Chaque ontologie encode la connaissance d'un domaine spécifique mais est également capable d'encoder la connaissance de différents domaines, c'est pour cela que l'on peut dire que les ontologies rendent la connaissance réutilisable. A l'aide d'une ontologie, tout objet peut être regroupé dans une certaine catégorie de concepts, dans laquelle chaque concept est défini en termes de propriétés distinctives, chaque élément appartenant à une catégorie partage avec les autres éléments une certaine propriété (nécessaire), de telle sorte que, s'il en était dépourvu, ce concept ne pourrait pas faire partie de cette classe ; il y a donc des propriétés prototypiques (la dimension spatio-temporelle, la masse sont des propriétés communes à tout objet physique) et d'autres propriétés qui servent à distinguer les sous-catégories.

Les outils d'indexation sémantique mentionnés[7] sont des ressources qui se composent d'un certain nombre de termes / concepts / descripteurs / éléments dont le but est celui de décrire un certain domaine auquel ils se réfèrent, aussi bien à un niveau général que d'un point de vue plus spécifique. Ce qui les différencie est, d'une part, le degré de dépendance du langage, et d'autre part, le type de relation qui existe entre les différents concepts et la formalité des contraintes. Il s'agit de ressources caractérisées par une croissante expressivité des liaisons qui s'interposent entre les entités qui les constituent: en partant par les schémas de classification plus basiques dans lesquels les éléments ne sont liés que par une relation hiérarchique, l'on passe aux thésaurus, dans lesquels, en plus d'une relation hiérarchique, les termes sont également liés par une relation d'équivalence ou une relation associative, pour arriver ensuite aux ontologies dans lesquelles les relations sont une expression des liens sémantiques entre les différent concepts. Ces dernières comportent différents niveaux de complexité et exigent des techniques de construction différentes. Généralement, elles peuvent être réalisées, d'une part, sur la base de la connaissance du domaine auquel elles se réfèrent, et d'autre part, sur l'analyse du contenu concret des documents recueillis dans le système documentaire. Dans le cadre d'un projet visant à élaborer une base de données, il est possible de mettre à la disposition des utilisateurs différents outils d'indexation sémantique à des temps différents, dans un processus de raffinement progressif, en expérimentant éventuellement avec des versions multilingues (en ce qui concerne la documentation en une autre langue que l'italien).

Utiliser une ressource plutôt qu'une autre dans la recherche d'une information juridique peut avoir un impact considérable sur le succès de l'opération. Les relations qui existent parmi les concepts d'une ontologie sont expression des liens sémantiques qui existent parmi les concepts d'un domaine spécifique.

Les récents développements des technologies de l'information en matière juridique suggèrent de nos jours un outil qui s'inspire aux modèles ontologiques, capable d'offrir à l'utilisateur la possibilité de rechercher un document à partir d'un mot clé

[6] Sur ce point, voir, entre autres: D. Tiscornia, *Il diritto nei modelli dell'intelligenza artificiale*, Bologna, Clueb, 1996; B. Smith, *Ontologia e Sistemi informativi*, Networks, 2006; Hirst G., *Ontology and the Lexicon,* in S. Staab and R. Studer (eds), *Handbook on ontologies*, Springer, Berlin-Heidelberg, 2004.

[7] D.A. Cruse, *Lexical Semantics*, Cambridge University Press, 1986.

significatif, qui ne doit pas pour autant faire nécessairement partie du contenu textuel du document.
L'analyse linguistique constitue la saisie de données pour les modules d'extraction, d'acquisition et da structuration des connaissances. La connaissance qui en résulte représente une ressource pour l'utilisateur final, qui lui permet de remplir et d'élargir les répertoires de vocabulaire linguistique et terminologique utilisés pour l'analyse des documents. Ainsi se réalise un cycle vertueux entre les outils. Les ressources linguistiques, lexicales et textuelles permettent de construire, développer, d'exploiter, et d'évaluer les modèles, les algorithmes, les composants, des systèmes qui sont, à leur tour, des instruments nécessaires à alimenter et à développer de manière dynamique ces ressources.

### 4.1 La référence croisée

En outre des possibilités offertes par l'emploi de métadonnées sémantiques, il ya aussi les possibilités offertes par les techniques de référence croisée. Cette expression se réfère à tous les cas où le texte d'une loi renvoie à une autre loi (références externes) ou à une autre section de la même loi (référence interne). Bien que les références internes se basent généralement sur une logique purement textuelle plutôt qu'une logique de domaine, les références externes comportent des liens à d'autres documents qui se basent souvent une logique de domaine. Cette technique se prête donc à être utilisé de façons différentes à selon du résultat que l'on souhaite obtenir: si l'on veut, par exemple, permettre à l'utilisateur d'acquérir une image complète de la manière dont une institution juridique est régie par la loi, interprétée par les juges, conçue et comprise par la doctrine, différentes liaisons seront crées parmi les différentes sources qui concerne cette institution spécifique.

## 5 Un projet de système informatique juridique avancé: l'Observatoire sur les règles du droit agro-alimentaire (ORAAL)

Au sein des questions abordées dans cet essai est d'un intérêt particulier, en tant que *best practice*, la création d'un observatoire sur les règles de l'agriculture et l'alimentation[8] Un cas unique en Italie pour sa taille et sa modalité d'action, survenu à l'initiative de l'Université de Pise et du Département d'identité culturelle du CNR, l'ORAAL a pour objectif de coordonner, développer et diffuser les connaissances dans le domaine du droit agro-alimentaire, ainsi que de promouvoir la diffusion et la formation sur les questions de sécurité et de qualité des aliments, de santé alimentaire et de protection des identités.
Dans une industrie caractérisée par une multitude de normes qui régissent la production, la commercialisation et la distribution alimentaire, l'Observatoire a pour objectif de créer une base de données et d'établir un réseau à disposition des différents acteurs qui interviennent dans le cycle économique de la production alimentaire, des citoyens consommateurs, ainsi que des institutions et des praticiens du droit, à fin de rendre leurs actions plus efficaces et mieux coordonnées. L'utilisateur est donc fourni avec un aperçu global des connaissances, ce qui comporte une globalité des sources et une globalité des contextes reliés entre eux dans une logique de domaine (des arrêts qui offrent une interprétation

[8] ORAAL surgit à l'initiative de l'Université de Pise et du Département d'identité culturelle du CNR. Il profite de la collaboration de deux instituts du CNR, celui de Théorie et Iechniques de l'Information Juridique (Ittig) - situé à Florence - et celui de Linguistique Computationnelle ( ILC) – situé à Pise – qui s'occupent de l'analyse de l'information et du langage juridique.

spécifique des lois, des universitaires qui écrivent des notes sur le même sujet, la Cour de justice qui propose des idées particulières sur le même sujet, etc.). L'information devient donc un nœud duquel peuvent se dénouer de nombreux parcours connectés entre eux de différentes manières.
Dans ce contexte, la recherche peut donc fournir des services complets, intégrés et facilement accessibles et disponibles à tous indépendamment de leur origine. L'outil informatique combinée à l'expertise du secteur alimentaires (dans ce cas, bien que cela s'applique aussi à d'autres domaines de compétence) permet de traiter de manière complexe, uniforme et harmonisée un contexte tel que celui du droit agro-alimentaire qui semble être inconstant, incohérent et internement déconnecté. Une fenêtre privilégiée sur les règles et les dynamiques qui régissent la discipline de ce secteur, ainsi qu'un instrument de surveillance continue par rapport aux changements qui interviennent dans l'industrie alimentaire dans d'autres contextes. Cela favorise la fertilisation croisée de la réglementation provenant de différents contextes.
Tout matériel pertinent en matière de droit agro-alimentaire est recueilli dans sa version intégrale et officielle, lorsque celle-ci est disponible. La documentation juridique qui provient de sources différentes est donc rassemblée et organisée selon les méthodes et techniques propre des bases de données. En particulier, l'objectif est celui de créer un système d'information à travers duquel il est possible d'effectuer une recherche soit par matière que par source, en laissant à l'utilisateur la possibilité de combiner deux options différentes : restreindre la recherche à seulement l'une ou l'autre des possibilités ou l'étendre à toutes les deux. Le traitement des documents mis à disposition des utilisateurs prévoit l'établissement de systèmes de repérage des informations qui reposent sur une organisation uniforme des connaissances contenues dans les documents et sur l'utilisation d'outils de langage universel , en s'appuyant sur une analyse sémantique des textes qui en augmente la capacité de traitement et de recherche.
Il s'agit d'une méthodologie de travail organisée et de solutions technologiques accolées, qui conduisent à un choix d'outils spécialisés pour l'interopérabilité, avec une attention particulière pour ce qui concerne les standards. Tout ceci compte tenu de la croissance globale et pour le plus partagée des connaissances et des technologies, et en vue d'une représentation des connaissances juridiques, comme précédemment exposé, capable de rendre extrêmement efficace les systèmes de recherche d'information (pour ce qui en est de l'organisation et la récupération d'information et de la représentation sémantique des documents) utilisés soit par les spécialistes, que par les simples citoyens et entreprises.
À ces fins, trois éléments peuvent se regarder essentiel: (i) le recueil et l'organisation de textes intégraux, (ii) la fourniture et l'application de schémas de métadonnées appropriées, (iii) l'utilisation d'outils pour l'indexation sémantique.

### 5.1 Le recueil et l'organisation de textes intégraux

Le premier objectif du système informatif est celui de rendre disponible aux utilisateurs, dans la mesure du possible, tous les textes intégraux pertinents à un document spécifique.
Afin d'obtenir un aperçu utile et approprié des normes relatives à l'objet en question, les textes normatifs doivent être identifiés, reconstruits dans leur cycle de vie (amendements, abrogations, etc.), interprétées et mis en relation entre eux. Une opération qui n'est pas sans critiques en raison de la multiplicité des documents concernés (par nombre et par type) et en vue de la division des responsabilités entre le gouvernement central et les entités périphériques (sans qu'il n'y ait eu

nécessairement un transfert direct de compétences du niveau central au niveau local) au bénéfice des citoyens et des entreprises, ainsi que des operateurs de l'industrie et du droit en général. Les documents normatifs seront donc successivement marqués en langage XML en conformité avec les normes établies par le projet «Norme in Rete» (NIR) et incorporées dans les Circulaires AIPA du 6 Novembre 2001, n AIPA/CR/35 "Assignement de noms uniformes pour les documents juridiques » et du 22 avril 2002 N ° AIPA/CR/40 "Format pour la représentation électronique des mesures législatives par le langage de balisage XML ». Telle représentation permettra de profiter des fonctionnalités avancées pour la consultation, la recherche et le marquage des documents au plus haut niveau de détail, ainsi que de la gestion des schémas de classification spécifiques au secteur agro-alimentaire, qui bénéficient d'une façon plus détaillée de marquer les textes normatifs et qui permettent ainsi de produire un patrimoine à partager avec la communauté de professionnels du secteur, les références automatiques ... et le mode d'affichage en modalité multi-valide (à savoir la version du texte en vigueur à tout moment de modification).

## 6 Métadonnées

Afin de systématiser l'information recueillie, ORAAL a adopté un schéma de métadonnées qui peut être associé aux textes intégraux de manière à pouvoir représenter soit les éléments formels (auteur du document, titre, année et numéro d'identification. ...) que les contenus (par exemple au moyen de sommaires ou de notes et commentaires). D'une part, il a fallut tenir compte des meilleures expériences dans le contexte scientifique de référence (l'ensemble de métadonnées développées dans le cadre de l'initiative de Dublin Core), d'autre part, l'expérience acquise par les institutions participantes a été réutilisé pour ce qui en est de la préparation des bases de données en matière d'informations juridiques.
Notamment, les documents législatifs les plus importants sont accompagnés par des notes de coordination et des commentaires pour la plupart orientées aux utilisateurs non-spécialistes. Ces notes contiennent généralement une brève synthèse de la loi, les détails relatifs à toute modification de la législation précédente, ainsi qu'une référence aux normes corrélées les plus importantes (en absence de références formelles).
Dans le contexte d'un accès multilingue aux ressources normatives, telles que celui des ressources communautaires, les potentialités offertes par ces deux différents types de ressources deviennent beaucoup plus évidentes. Surtout dans un domaine spécifique comme celui du vocabulaire juridique, un accès lexico-conceptuel ne suffit pas à lui seul à garantir à l'utilisateur la fiabilité d'un système de recherche. Dans le contexte européen, en effet, la traduction des concepts en différentes langues a été réalisé *re ipsa*, c'est à dire que ce sont les concepts qui ont été transposé en différentes versions linguistiques à servir de pivot - en fait les deux types d'accès ont nécessairement besoin de cohabiter et de fonctionner, bien que d'une manière diamétralement opposée par rapport à une recherche mono-linguistique: c'est l'accès ontologique qui permet de contextualiser la requête de l'utilisateur en transformant une notion juridique, que l'utilisateur ne connaît que dans sa propre langue, en un terme qui identifie le même concept dans une autre langue pour lui permettre d'effectuer une recherche des ressources juridiques étrangères en utilisant les expressions lexicales appropriées pour identifier ce concept spécifique.

## 7 Conclusions

À la lumière des expériences analysées dans le présent document et en vue des instruments informatiques réalisés jusqu'à présent dans le domaine de l'information juridique, il apparaît évident que l'énorme quantité de données juridiques contenues dans les documents électroniques en ligne exige d'une organisation appropriée pour pouvoir être effectivement consultés et reconnus par les citoyens. À ces fins, la tendance la plus désirable est celle de créer des outils qui facilitent une approche simplifiée à la complexité juridique et qui rendent les systèmes de recherche juridique, qui devront être utilisés soit par des utilisateurs spécialistes que par des citoyens ordinaires, beaucoup plus efficaces.
En effet, autant plus grande est la complexité du domaine, autant plus puissant devront être les instruments informatiques pour la récupération et pour la représentation sémantique des documents juridiques, de façon à rendre la recherche du système plus efficace. Des systèmes informatifs de ce genre doivent par conséquent garantir un accès intégré, complet et global au domaine juridique d'intérêt (normes, jurisprudence et doctrine) en fournissant à l'utilisateur non seulement les informations qu'il recherche, mais aussi tout ce qui résulte implicitement contenu dans le document de référence, ainsi que tout ce qui lui est logiquement associé mais que l'utilisateur ignore. Les instruments d'information juridique doivent donc être au service du droit fondamental du citoyen à la connaissance et à la compréhension des normes qui régissent ses actions.
La difficulté principale à l'utilisation du document électronique dans un contexte juridique est sans doute celle de créer des ressources normalisées qui soit reconnues par une certaine communauté et utilisées de manière uniforme afin d'assurer un désirable niveau d'interopérabilité.

## 8 Bibliographie


1 R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval,* Addison-Wesley, 1999

2 Benjamins, Casanovas, Breuker, Gangemi (eds.), *Law and the Semantic Web,* Berlin, Springer, 2004

3 C. Ciampi, *Legal Information on the Web. The NIR Portal for the Citizen,* in *Proceedings of LEFIS Workshop on "Legal Aspects of E-Government" (Vilnius, 20th September 2003),* pp. 40 ss.

4 D.A. Cruse, *Lexical Semantics*, Cambridge University Press, 1986

5 European Union, *Access to legislation in Europe. Guide to legal gazettes and other official information sources in the European Union and European Free Trade Association,* Publications Office of the European Union, 2009

6 G. Hirst, *Ontology and the Lexicon,* in S. Staab and R. Studer (eds), *Handbook on ontologies*, Springer, Berlin-Heidelberg 2004, pp. 210 ss.

7 K. Rissanen, *Access to EU law and eLaw – visions and challenges,* in *Séance académique «25 années de droit européen en ligne»*, Publications Office of the European Union, 2006.

8 B. Smith B., *Ontologia e Sistemi informativi*, Networks, 2006

9 D. Tiscornia, *Il diritto nei modelli dell'intelligenza artificiale,* Bologna, Clueb, 1996

# Lire numérique : nouveaux dispositifs - nouvelles pratiques ?

**Animateur.** Alexandra Saemmer, Paragraphe-Université Paris 8

**Participants.** Joël Gardes, Orange Labs / INSA Lyon-LIRIS ; Claire Bélsile, Lire-CNRS ; Emmanuël Souchier et Etienne Candel, CELSA-Université Paris-Sorbonne ; Caroline Courbières, LERASS-Université Toulouse 3.

Sont abordées dans cette table ronde les possibilités ouvertes pour la lecture numérique par les dispositifs numériques (de l'écran d'ordinateur aux « liseuses »). Au-delà des questionnements suscités par les caractéristiques des dispositifs et des interrogations concernant les potentialités hypermédiatiques au sein des créations (pensons aux relations intersémiotiques entre texte, image, mouvement et son couplées à l'interaction), il nous semble d'une part important de poser la question de l'intérêt de la lecture : les nouveaux dispositifs vont-ils susciter de nouvelles motivations, de nouvelles émotions, bref, un « plaisir du texte » renouvelé ? D'autre part, nous examinons les imaginaires et les représentations visuelles du lire comme une activité sociale située. Comment les dispositifs sont-ils pris dans une « mise en scène » de la lecture comme pratique médiatique ?
Les attentes des lecteurs évoluent avec la confrontation aux dispositifs, leurs contraintes et compositions, et en fonction des contenus présentés sur ces dispositifs : nous observons par exemple l'émergence de gestes spécifiques qui ne conditionnent pas seulement l'accès à une création numérique, mais prennent sens en fonction des mots ou images manipulables. Les attentes sont également forgées par les apprentissages, les discours ambiants (promotionnels ou critiques) dans un contexte socio-culturel. Se croisent, s'entrelacent des pratiques de lecture encore et toujours apprises sur papier, des pratiques forgées par les écritures numériques (qu'elles soient informationnelles, commerciales ou artistiques, qu'elles s'exercent lors de la lecture de blogs, de profils ou d'états facebook, de twits ou de résultats fournis par un moteur de recherche), des pratiques de manipulation de la matière numérique issues du jeu vidéo… Est-ce que ces pratiques de lecture multiples s'entrecroisent pourtant toujours sans heurt ?

Dans des productions comme *Moving tales*[1] ou *Alice in Wonderland*[2] pour i-pad par exemple, qui présentent sur une même « page » du texte fixe et animé, de l'image fixe et manipulable, l'expérience d'une lecture textuelle intensive peut éventuellement disparaître au profit d'un jeu avec la matière des mots et des images, s'épuiser dans l'exploration (certes jouissive) des potentialités offertes du dispositif.

[1] <http://www.moving-tales.com/>
[2] <http://itunes.apple.com/us/app/alice-for-the-ipad/id354537426?mt=8>

Beaucoup d'études montrent déjà que le lecteur, lorsqu'il consulte les résultats hyperliés fournis par un moteur de recherche, n'épluche pas tous les résultats mot à mot ; à la recherche d'une information précise sur un portail, il parcourt les pages web affichées sans se concentrer sur tous les détails - même au cas où il a envie de se laisser surprendre par des informations dépassant ses attentes. R. Simanowski constate ainsi le web fournit certes un accès illimité et immédiat à l'information, induit la ramification et relativisation constante des données, et provoque la mise en place de nouvelles formes d'archivage et d'indexation ; une rhétorique « de l'arrivée et du départ », et une structuration de l'information « par petite bouchées » est pourtant également caractéristique du web[3]. Poussé par le désir de découvrir encore et encore, le lecteur ne prend peut-être plus suffisamment le temps de s'arrêter. Encore une fois, le but du lecteur ne semble plus être la compréhension, mais le mouvement lui-même. Nous posons donc entre autres la question de savoir si ce mouvement empêche de fait la pratique documentaire et/ou remet en cause le rapport au savoir.

R. Simanowski, pour sa part, parle d'une mort du lecteur au profit de l'interacteur[4]. Doit-on craindre effectivement que ces pratiques de « lecture post-alphabétique »[5] empêchent la réflexion, induisent l'acceptation léthargique du statu quo ? Ou sont-elles au contraire positivement ludiques, insouciantes et irrespectueuses, comme le proclament certains artistes du numérique et certains critiques lorsqu'ils affirment : « Dans l'expérimentation qu'offre l'art numérique, rien à gagner sinon l'expérimentation pour elle-même, plus exactement la découverte intime du plaisir de l'expérimentation »[6], ou : « P*layfulness is a defining quality of this new medium [...] No matter how competitive, the experience of reading in the electronic medium remains a game* »[7] ?

Nous devons certainement élargir la définition de la « lecture » dès qu'elle s'exerce sur un dispositif numérique, et sortir de la comparaison encore trop largement convoquée avec le livre ou le journal papier ; nous devons néanmoins nous demander ce que le lecteur « gagne » dans ces nouvelles pratiques de la lecture numérique, et ce qu'il y perd peut-être.

Ces dernières années, c'est notamment la discussion autour des prodiges et vertiges procurés par l'hypertexte qui a occupé le devant de la scène. Investi d'abord d'espoirs parfois démesurés (comme support d'une interconnexion quasi neuronale des savoirs), l'hypertexte a ensuite été soupçonné d'induire des dissonances cognitives. D'un côté, un mot hyperlié reste en effet un signe linguistique ; d'un autre côté, il est affecté d'une signalétique qui rapproche sa fonction de celle d'un bouton. Qu'est-ce qui prime alors dans la perception et dans la compréhension de ce signe : l'incitation au déchiffrage ou la tentation d'une manipulation immédiate ? Avant de cliquer, lorsqu'il perçoit un texte parsemé de liens pour la première fois, le lecteur est rarement renseigné sur les contenus à découvrir. Certes, la relation entre les médias activables et activés par le geste de manipulation peut ensuite devenir signifiante, et permettre au lecteur d'approfondir ses connaissances, de faire de nouvelles découvertes, de satisfaire sa curiosité. Néanmoins, l'incitation permanente à l'interaction, qui repose d'une part

[3] R. Simanowski, Digitale Medien in der Erlebnisgesellschaft. Kultur – Kunst – Utopien, Reinbek bei Hamburg, Rowohlt, 2008, p. 216.

[4] Ibid., p. 49.

[5] Ibid., p. 119.

[6] J.-P. Balpe, « Quelques concepts de l'art numérique », <http://transitoireobs.free.fr/to/article.php3?id_article=42>.

[7] J. D. Bolter, Writing Space : The Computer, Hypertext, and the History of Writing, Hillsdale, L. Erlbaum, 1991, p. 130.

sur les liens hypertextes internes, et d'autre part sur les barres interactives, les signets, la fenêtre de saisie du moteur de recherche, les icônes de la messagerie, du chat etc. entourant la fenêtre du navigateur, distraient en permanence le sujet de sa lecture d'un texte précis. Faut-il alors mettre en avant, comme le fait Paul Virilio, le risque majeur de l'interactivité ? Les dernières fonctionnalités du navigateur *Firefox*, qui permettent de lancer à partir de n'importe quel mot d'un texte une requête *Google* concernant ce mot (il suffit de sélectionner le mot et d'effectuer un clic droit), renforcent encore cette tendance à la « distraction » permanente : chaque mot paraît désormais virtuellement hyperlié. Par ailleurs, bien que beaucoup de critiques s'accordent pour affirmer que le support numérique demande un lecteur plus « actif », il faut regarder de près en quoi consiste exactement l'activité du lecteur, et ce qu'elle lui apporte : A partir du moment où il peut seulement cliquer sans produire lui-même du texte, à partir du moment où son geste ne sert qu'à faire fonctionner la machine ou à feuilleter les contenus proposés, cette activité n'est-elle pas en même temps incitante *et* frustrante ?

Comme le font remarquer Y. Jeanneret et al., face à la machine, le lecteur est « placé dans une situation paradoxale de distanciation et d'engagement. La distance de l'homme à la machine est plus grande que celle de l'homme au livre, car le texte semble avoir disparu 'derrière' l'écran, laissant prise à l'espace du secret et du sacré. En revanche, l'engagement physique s'accroît, car le lecteur devient manipulateur et doit 'agir' la machine à des fins purement fonctionnelles »[8].

En résulte un sentiment d'inconfort, de malaise relevé dans beaucoup d'études sur la lecture numérique[9]. Face au texte numérique, le lecteur semble par exemple à la fois jouir et souffrir d'une étrange impatience. Cette urgence est-elle due à la force de séduction des liens hypertexte (réels et virtuels), au désir de profiter immédiatement de la possibilité de non seulement lire le texte, mais de le manipuler - comme si la main ne pouvait plus rester immobile à partir du moment où elle a pris l'habitude des touches de la souris et du *touchpad* ? Effectivement, la main s'impatiente lorsqu'il y a trop de choses à faire, car elle ne sait pas où cliquer d'abord ; elle s'impatiente pourtant également lorsqu'il n'y a rien à faire. Y. Jeanneret et al.[10] ont pu observer que le premier geste effectué par beaucoup de sujets devant un texte numérique, peu importe sa nature et son contenu, consiste à faire avec la souris un balayage global sur l'écran. Ce geste exploratoire s'explique par le désir de tester les potentialités de l'interface, et semble ainsi constituer une préparation à l'activité de consultation « sérieuse » des liens. D'autres raisons se rajoutent à cet argument principal pour expliquer l'impatience du lecteur devant le texte numérique. Même s'il ne contient ni hypertextes internes ni animations, le texte numérique est entouré d'un contexte interfacique manipulable qui envoie en permanence des signaux au lecteur. Le curseur lui-même indique la présence du lecteur dans le texte. Certains sujets essaient de ne pas perdre de fil de la lecture numérique en suivant les lignes du texte avec le curseur. Mais contrairement au

---

[8] Y. Jeanneret, A. Béguin, D. Cotte, S. Labelle, V. Perrier, P. Quinton, E. Souchier, « Formes observables, représentations et appropriations du texte de réseau ». Lire, écrire, récrire. Objets, signes et pratiques des médias informatisés, éd. Emmanuël Souchier, Yves Jeanneret et Joëlle Le Marec, Paris, BPI Centre Pompidou, 2003 ; p. 98-99.

[9] Voir par exemple l'enquête dirigée par C. Bélisle, mise en place en 2006 auprès d'étudiants universitaires pour étudier dans un premier temps leur usage des encyclopédies en ligne. Plus de six cent quarante étudiants ont répondu à un questionnaire en ligne sur le bureau virtuel des étudiants de l'université. Synthèse et analyses des résultats consultables à l'adresse : <http://lire.ish-lyon.cnrs.fr/spip.php?rubrique88>.

[10] Y. Jeanneret et al., op. cit., p. 120.

doigt qui, lors d'une lecture papier, a toujours constitué un repère fixe pour l'écolier timoré, le curseur se transforme en butant sur les liens hypertexte, et incite au « divertissement ». En outre, comme le font remarquer Y. Jeanneret et al.[11], beaucoup d'indices avertissent le lecteur du temps qui passe : sur la plupart des ordinateurs, une horloge tourne en haut de l'écran, rappelant au lecteur que les secondes lui filent sous la main et sous les yeux.

Face à ces caractéristiques et difficultés, il nous paraît ainsi incontestable que ces nouvelles pratiques du « lire numérique » nécessitent de nouveaux apprentissages : savoir trouver l'information, savoir filtrer, savoir utiliser les outils, maîtriser de nouveaux gestes ; mais aussi mettre en perspective des contenus hyperliés sans se trouver découragé ou dispersé par le flux permanent de nouvelles informations ; savoir s'arrêter pour se concentrer sur un texte ou une vidéo, malgré les sollicitations permanentes envoyées par les différents cadres du dispositif. Quelles sont les compétences attendues en la matière notamment chez les jeunes, par exemple dans le cadre du « B2i – Brevet informatique et internet » (®) mis en place par l'Éducation nationale depuis quelques années afin de permettre aux futurs citoyens de faire « une utilisation raisonnée des TIC » ?

Dans cette table ronde, nous avons fait l'inventaire des caractéristiques principales de la lecture numérique. Nous avons abordé les potentialités induites par les nouveaux dispositifs dans leur contexte social, sans pour éviter les questionnements critiques. Car décidément, la lecture numérique constitue un défi : pour les créateurs et les éditeurs, qui sont amenés à inventer des formes et figures nouvelles de l'écrit numérique ; pour les lecteurs, dont les attentes et habitudes se retrouvent régulièrement remises en question.

Enfin, nous avons aussi posé la question de savoir si nous avons encore la bonne approche du document dans ce contexte actuel où l'on ne peut plus envisager de services sans multimodalité et sans multimédia ; ou si la distinction conceptuelle entre donnée et information ne se trouve pas au contraire réellement réaffirmée par le numérique. Quelles sont les nouvelles « contraintes » apportées par l'accessibilité aux contenus ?

[11] Ibid., p. 124.