Le "Document" à l'ère de la différenciation numérique
14e colloque international sur le document électronique
Auteurs / Editors :
Mostafa Bellafkih, Joël Gardes, Mohamed Ramdani, Khaldoun Zreik

Edité par / Published by :
europia 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

ISBN13 : 979-10-90094-07-9

Le “Document” à l’ère de la différenciation numérique
14e colloque international sur le document électronique

Mostafa BELLAFKIH, Joël GARDES,
Mohamed RAMDANI, Khaldoun ZREIK

europia

Présidents du colloque CiDE.14
Mostafa BELLAFKIH, INPT, Maroc
Joël GARDES, Orange, France
Mohamed RAMDANI, Université Mohamadia, Maroc
Khaldoun ZREIK, Université Paris 8, France

Comité d'organisation local
Mostafa BELLAFKIH, INPT, Maroc
Mohamed ERRAIS, INPT et Université Mohamedia, Maroc
Ferdaous LAHMAR, CITU-Paragraphe, Université Paris 8, France
Mohamed RAMDANI, Université Mohamedia, Maroc
Brahim RAOUYANE, INPT et Université Mohamedia, Maroc
Karima TOUNSI, INPT, Rabat, Maroc

Comité Permanent des colloques CiDE
Ghislaine AZEMARD, Université Paris8, France
Mostafa BELLAFKIH, INPT, Maroc
Jean CAELEN, CLIPS-IMAG, Grenoble, France
Jacques DUCLOY, DRRT Lorraine, France
Patrice ENJALBERT, Université de Caen, France
Mauro GAIO, Université de Pau, France
Joël GARDES, Orange, France
Jean-Luc HAINAUT, Belgique
Maryvonne HOLZEM , Université de Rouen, France
Madjid IHADJADENE, Université Paris 8, France
Peter KING, Université de Manitoba, Canada
Jacques LABICHE, Université de Rouen, France
Abdelkarim MEZIANE, CERIST, Algérie
Mustapha MOJAHID, Université de Toulouse Le Mirail, France
Ghassan MOURAD, Université Libanaise, Liban
Giovanni DE PAOLI, Université de Montréal, Canada
Jean-Pierre RAYSZ, Jouve, France
Jean Marc ROBERT, Ecole Polytechniques – Université de Montréal, Canada
Zaidi SAHNOUN, Université de Constantine, Algérie
Maurice SZMURLO, Orange, France
Loïc THOMAZO, Orange, France
Eric TRUPIN, Université de Rouen, France
Christophe TURBOUT, Université de Caen, France
Jacques VIRBEL, Université de Toulouse Le Mirail, France
Jean VIVIER, Université de Caen, France
Christine VANOIRBEEK, EPFL, Suisse
Manuel ZACKLAD, CNAM-Paris, France
Khaldoun ZREIK, Université Paris8, France (Coordinateur)

TABLE DES MATIERES

Introduction

Le document à l'ère de la différenciation numérique

La tenue de CIDE à Rabat cette année revêt une valeur symbolique particulière : la première édition de notre colloque s'est tenue au même endroit à la fin du XXème siècle (1998) et, depuis, 13 ans ce sont écoulés au cours desquels le concept de document, sa perception ainsi que ses pratiques ont évolué, voire, subit des mutations.

Les bibliothèques numériques sont désormais une « réalité » en progression permanente et non plus des projets. Le succès des dispositifs collaboratifs de type Wiki confirme la nécessité et la volonté croissantes de partager des connaissances, même si la valeur éditoriale des contenus reste encore sujette à questions.

CIDE a vécu ces treize ans que l'on pourrait dorénavant qualifier les années du « big bang » des télécommunications mobiles et, en particulier, de terminaux que l'on considérait encore d'intelligents en 1998, quand on décrivait à quoi ils allaient ressembler et qui, aujourd'hui, s'appellent « smartphones » ou « tablettes ».

Concomitamment ces terminaux ont redonné, et ceci peu paraître de prime abord paradoxal de l'écrire, une place de choix au geste. Les interfaces sont largement tactiles, non seulement pour saisir de l'information, mais aussi, pour commander le système de l'on tient au creux de sa main et pour manipuler les données affichées à l'écran.

N'y a-t'il pas ici une meilleure illustration d'une des thématiques récurrente de CIDE, à savoir, le concept de document dynamique et interactif ? En y réfléchissant, ce concept n'est rien d'autre que l'interface utilisateur de nos terminaux mobiles ainsi que des nouvelles générations d'ordinateurs personnels qui tendent à mettre à mal progressivement le concept de clavier et de souris, en offrant également la possibilité de saisir et de manipuler directement des objets graphiques ou symbolique représentant de l'information.

Bref, tout ceci avait été pressenti comme « toile de fond » du document numérique lors des différentes éditions de CIDE. Mais un aspect de plus a surgi du fait de cette banalisation de l'information numérique et de l'énorme puissance de calcul et de stockage des terminaux : l'utilisateur destinataire de l'information désire s'approprier totalement le contenu restitué et accepte de moins en moins de se plier aux contraintes de l'émetteur.

Qu'est-ce que cela signifie dans notre concept de document ?

La personnalisation de la présentation de l'information devient une problématique centrale, en ce sens que le contenu restitué doit être non seulement adapté à la nature du terminal sur lequel le document s'affiche, mais aussi, aux capacités de lecture et de manipulation du contenu de l'utilisateur en fonction de sa situation de déficience permanente

(handicap) ou temporaire (communication « bruitée »). Tout ce travail d'adaptation fait que l'on ne communique plus l'information au travers de messages et de signaux préétabli, mais au travers d'une médiation qui se traduit par autant d'instanciation de la présentation de l'information que de situations et de contextes d'usages.

Il devient indispensable de préserver des notions d'authenticité et d'intangibilité de l'information, puisque la présentation de celle-ci n'est plus fixée a priori lors de la composition du document, à l'instar de la chose imprimée ; mais est reconstruite dès que l'utilisateur y accède par l'intermédiaire de ses propres moyens techniques.

L'extension « naturelle » de cette personnalisation est, toujours, la multimodalité. Ici, on touche aux mécanismes intimes de l'interaction homme machine en posant clairement la question de l'efficacité du « canal de communication » qu'est le document numérique ou l'interface homme machine et du maintien de la conformité de l'information transmise par ce canal. Prenons simplement la vocalisation d'un extrait de livre. La première condition est de disposer d'un texte cohérent et conforme à la source. La deuxième condition est d'offrir à l'utilisateur une sorte de « balisage » du document audio lui permettant de se repérer et ce balisage remet en avant le récit du texte. Ce dernier, s'il est monocorde, égare l'utilisateur, alors que s'il contient une prosodie bien adaptée, permet à notre utilisateur de s'approprier la « lecture » audio du document.

Au départ : une information à transmettre en la codant. A l'arrivée, une information restituée selon de profonds critères de personnalisation tout en préservant conformité et authenticité. Ainsi, ce n'est pas l'information elle-même qui se différencie en fonction des contextes de situation et d'usage du numérique, mais sa présentation : une même information revêt désormais une large variété de présentations dont nous devons tenir compte. Cette variété dépend du codage que l'on adopte. Historiquement, ce codage, pour le document, correspondait à tout l'art de l'éditeur associé à celui de l'imprimeur. L'éditeur garantissait la conformité à l'œuvre et l'imprimeur fixait ces marques de conformité sur un support intangible. Avec le numérique, le contenu se réédite autant de fois que l'on y accède : que deviennent les marques de conformité et d'authenticité ? Par quels mécanismes peut-on préserver la conformité ? Quelle est « l'autorité » permettant de garantir l'authenticité ?

L'utilisateur « ouvrant » un document, ouvre en fait, un fichier contenant des instructions codées. Le document est devenu une sorte de « kit d'assemblage » de l'information, mais est-on certain que le « mode d'emploi », on devrait dire les « modes d'emplois » de montage de ce kit a été également fourni ?

Ainsi peut-on résumer le « cas » du document numérique. Il a conservé sa vocation historique de communication de connaissances (document venant de doceo, docere : communiquer pour enseigner). Mais

désormais, le document, devenu multi[media, modal, lingue, culturel, …] (pardon pour cette pseudo écriture), n'est plus qu'une instance dépendant du contexte de la communication. C'est la raison pour laquelle nous avons proposé de parler d'ère de la différenciation numérique pour cette édition de CIDE. L'information se comporte comme une cellule souche en biologie, évolue en autant de présentations fonctionnelles qu'il y a d'usages et de contextes. Ceci ouvre la voie à de vastes champs de recherche qui, on peut le parier, nous mènent aux frontières des théories de l'information actuelles en nous faisant découvrir les limites de tous nos modèles de documents. En effet, la présentation de l'information devient relativisée aux contextes d'utilisation et de personnalisation du contenu.

Partie 1 - Indexation sémantique

Un modèle sémantique pour l'indexation de documents arabes et anglais

Taher ZAKI
Laboratoire IRF-SIC, Université Ibn Zohr Agadir, Maroc
LITIS EA 4108, Université de Rouen, France

Abdellatif ENNAJI
LITIS EA 4108, Université de Rouen, France

Stéphane NICOLAS
LITIS EA 4108, Université de Rouen, France

Driss MAMMASS
Laboratoire IRF-SIC, Université Ibn Zohr Agadir, Maroc

Résumé : Nous présentons ici un système pour l'indexation contextuelle et sémantique de documents en langue arabe et anglais, en se basant sur le voisinage sémantique des termes et l'utilisation d'une modélisation à base radiale. L'usage des graphes et les dictionnaires sémantiques améliore considérablement le processus de l'indexation. Dans ce travail, nous avons proposé une nouvelle mesure TFIDF-Okappi-ABR qui tient en compte la notion de voisinage sémantique à l'aide d'un calcul de similarité entre termes en combinant le calcul du TF-IDF-Okappi avec une fonction noyau à base radiale afin d'identifier les concepts pertinents qui représentent le mieux un document. Des résultats préliminaires et prometteurs sont données sur 2 bases de textes de presse en langue Arabe et Anglaise qui montrent de très bonnes performances par rapport à la littérature.

Mots-clés : Dictionnaire, fonction noyau, formule d'Okappi, graphe sémantique, indexation, TF-IDF, voisinage sémantique.

1. Introduction

La grande masse d'informations textuelles publiées sur le réseau mondial exige la mise en œuvre de techniques efficaces pour l'extraction d'informations pertinentes contenues dans de grand corpus de textes. Le but de l'indexation est de créer une représentation permettant de repérer et retrouver facilement l'information dans un ensemble de documents.

On utilise cette indexation le plus souvent dans les systèmes de recherche d'informations, mais cette indexation peut également servir à comparer et classifier des documents, proposer des mots-clés, faire une synthèse automatique de documents, calculer des co-occurrences de termes… Dans ce papier, nous allons définir un formalisme statistique pour le traitement de documents textuels en arabe et en anglais, et montrer comment ce formalisme peut servir pour le traitement de différentes problématiques telles que l'indexation ou la classification. Notre travail se positionne dans le cadre de la recherche d'information à savoir l'apprentissage statistique qui permet le développement de méthodes génériques utilisables facilement sur différents corpus. Ce formalisme permet d'exploiter à la fois la structure et le contenu textuel de ces corpus.

2. Phase d'indexation

2.1 Problématique

L'indexation est définie comme l'opération qui décrit et caractérise des données résultant de l'analyse du contenu d'un document ou un fragment de document, par des éléments d'un langage documentaire ou naturel en repérant les thèmes présents dans ce document (AFNOR, 1993). L'objectif est de trouver les termes qui caractérisent le mieux le contenu d'un document. Nous nous intéressons donc à la prise en compte des informations explicites autour du texte, à savoir la structure et la répartition des termes, ainsi qu'aux informations implicites, à savoir la sémantique.

2.2 Les étapes du processus

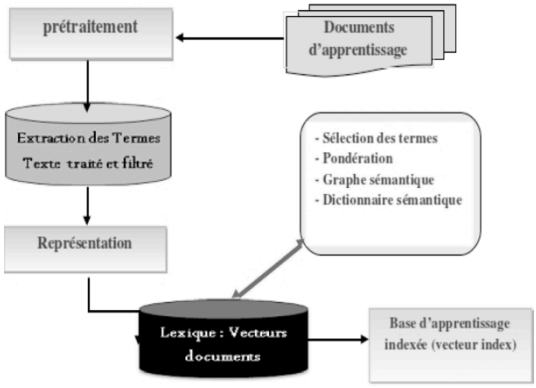


Figure 1 : le Processus de l'indexation

Le système mis au point consiste en 5 étapes fondamentales tel que illustré sur la figure 1 :

A. Base documentaire (Apprentissage et test)
Cette base est un corpus de documents de presse (the Associated Press (AP)) collectée à partir d'internet.

B. Prétraitements (Extraction des termes)
Cette phase consiste à appliquer à l'ensemble du texte une analyse morphologique (lemmatisation, stemming) en premier lieu et un filtrage des termes extraits en deuxième lieu. Ce traitement est nécessaire en raison des variations dans la façon dont le texte peut être représenté en arabe.
La préparation du texte comprend les étapes suivantes :
Les fichiers texte sont converti en codage UTF-16.
Les signes de ponctuation, les signes diacritiques, les non-lettres et les mots vides sont éliminés.
La racinisation des termes restants est opérée à l'aide du stemmer de Khoja (Khoja, 1999) pour les documents arabes, et le stemmer de Porter (porter, 1980) pour les documents anglais.

C. Espace de Représentation
Cette étape permet d'adopter une représentation vectorielle statistique du document à partir des termes retenus pour le représenter. Pour cela, nous avons étendu le modèle vectoriel de Salton en adaptant le calcul du TF-IDF par une combinaison du TFIDF et la formule d'Okappi avec une fonction noyau. Ensuite, pour éviter les problèmes combinatoires liés à la dimension de cet espace de représentation (Sebastiani, 2000) (Deerwester 1990), (Blei, 2003), nous avons adopté une approche de seuillage de fréquence (Document Frequency Thresholding) pour réduire cette dimension.

D. Classification
Pour la phase de classification, nous avons dans cette version préliminaire de notre prototype adopté l'algorithme simple des K plus proches voisins (kppv) pour sa simplicité et pour pouvoir évaluer la pertinence de nos choix de représentation. Nous avons dû également faire le choix d'une métrique adaptée à ce contexte qui est l'operateur de Dice en l'occurrence, dont l'expression est :

$$Dice\ (P_I, P_J) = \frac{2\,|P_I \wedge P_J|}{||P_I| + |P_J||} \quad (1)$$

Où $|P_i|$ est le nombre de termes dans le profile P_i

$|P_i \Lambda P_j|$ est le nombre de termes d'intersection entre les deux profiles P_i et P_j

E. Validations

Pour la validation du prototype, nous avons utilisé une base d'apprentissage très réduite comportant trois thèmes différents (sport, politique, économie et finances). Pour la phase de test, nous avons travaillé sur une base de 400 documents de presse (*Associated Press*) collectés à partir d'internet.

2.3 Pondération des unités index

La manière la plus simple pour calculer le poids d'un terme est de calculer sa fréquence d'apparition car un terme qui apparaît souvent dans un document peut être pertinent pour caractériser son contenu. Plusieurs fonctions de pondération de termes ont été proposées. Nous nous intéressons au classique TF-IDF (term frequency - inverse document frequency) utilisé dans le modèle vectoriel que nous adaptons dans notre travail. Il existe un certain nombre de variantes de TFIDF (Seydoux, 2006). Les critères retenues pour calculer le poids d'un terme sont :

- Une pondération locale qui détermine l'importance d'un terme dans un document. Elle est généralement représentée par sa fréquence (tf).

- Une pondération globale qui détermine la distribution du terme dans la base documentaire. Elle est généralement représentée par l'inverse de la fréquence des documents qui contiennent le terme (idf).

$$a_{ij} = tf(i,j).idf(i) = tf(i,j)\log\left(\frac{N}{N_i}\right) \quad (2)$$

où ***tf(i,j)*** est le *term frequency*, c'est-à-dire le nombre de fois que le terme $\mathbf{t_i}$ apparaît dans le document dj , et idf(i) est l'inverse document frequency, c'est-à-dire le logarithme du rapport entre le nombre N de documents dans le corpus et le nombre Ni de documents qui contiennent le terme ti. Ce schéma d'indexation donne plus de poids aux termes qui apparaissent avec une haute fréquence dans peu de documents.

L'idée sous-jacente est que de tels mots aident à discriminer entre textes de différents thèmes. Le tfidf a deux limites fondamentales : la première est que la dépendance du *term frequency* est trop importante. Si un mot apparaît deux fois dans un document dj, ça ne veut pas nécessairement dire qu'il a deux fois plus d'importance que dans un document d_k où il n'apparaît qu'une seule fois. La deuxième est que les documents plus longs ont typiquement des poids plut forts parce qu'ils contiennent plus de mots, donc les *term frequencies* tendent à être plus élevés. Pour éviter

ces problèmes, nous avons adopté une nouvelle technique d'indexation connue comme la formule d'Okapi (Robertson, 2000) :

$$a_{ij} = \frac{tf(i,j) \cdot idf(i)}{[(1-b) + b \cdot \mathrm{NDL}(d_j)] + f(i,j)} \quad (3)$$

Où $NDL(d_j)$ est la longueur normalisée de $\mathbf{d_j}$, c'est-à-dire sa longueur (le nombre de mots qu'il contient) divisée par la longueur moyenne des documents dans le corpus.

- La mesure N-Gramme

La notion de n-grammes a été introduite par (Shannon, 1948) et est souvent utilisée pour la prédiction d'apparition de certains caractères en fonction d'autres caractères. Les N-Gram sont des séquences de termes dont la longueur est N. Par exemple, l'utilisation des N-Gramm sur le mot « TEXT » est :

bi-grams _T, TE, EX, XT, T_
tri-grams_TE, TEX, EXT, XT_, T__
quad-grams _TEX, TEXT, EXT_, XT__, T___

Les tri-grams pour le mot الممودعيين sont : ا لم, لمو, مود , ودع , دعي, عين

La méthode des N-gramme offre l'avantage d'être une technique indépendante de la langue et permet ainsi une recherche basée sur un segment de mot.

Les systèmes basés sur les n-grammes n'ont pas besoin des prétraitements qui consistent à l'élimination ni des mots vides, ni au Stemming, ni à la lemmatisation, qui sont indispensables pour avoir des performances correctes dans les systèmes à base de recherche de mots (key matching). Pour les systèmes n-grammes, de nombreux travaux ont montré que les performances ne s'améliorent pas en procédant à des traitements d'élimination des "mots vides", de "Stemming" ou de lemmatisation. Nous avons donc mis au point une version à base de N-grammes de notre système pour comparaison.

3. Ressources sémantiques

3.1. Dictionnaire sémantique auxiliaire

Nous avons mis au point un dictionnaire sémantique auxiliaire qui est un dictionnaire hiérarchisé contenant un vocabulaire normalisé sur la base de termes génériques et de termes spécifiques à un domaine. Il ne fournit qu'accessoirement des définitions, les relations entre termes et leur choix l'emportant sur les significations. Les relations communément exprimées dans un tel dictionnaire sont :

les relations taxonomiques (de hiérarchie).

les relations d'équivalence (synonymie).
les relations d'association (relations de proximité sémantique, proche-de, relié-à, etc.).

3.2. Construction du dictionnaire

Le dictionnaire est initialement construit manuellement sur la base des termes retrouvés dans la base d'apprentissage. Mais ce dictionnaire peut être enrichi au fur et à mesure durant la phase d'apprentissage et la classification pour donner plus de flexibilité à notre modèle.
Prenons par exemple le thème finances et économie, le dictionnaire construit est comme suit :

economy, finances, enterprise, industrialism, market, capitalism, socialism, system, brevity, conservation, downsizing, financial status, productive power **…**
finances, budget, account, bill, financing, money, reckoning, score, banking, business, commerce, economic science, economics, political economy, investment **…**
budget, account, bill, calculate, estimate, finance, money, matters, reckon, reckoning, score, assortment, bunch, balanced, cheap,operating budget **…**
….
….

Figure 2 : Dictionnaire de finances et économie

3.3. Les réseaux sémantiques

Les réseaux sémantiques (Quillian, 1968) ont été conçus à l'origine comme un modèle de la mémoire humaine. Un réseau sémantique est un graphe étiqueté (un multigraphe plus précisément). Un arc lie (au moins) un noeud de départ à (au moins) un noeud d'arrivée. Les relations vont des relations de proximité sémantique aux relations partie-de, cause-effet, parent- enfant, etc.
Les concepts sont représentés sous forme de noeuds et les relations sous forme d'arcs. Les liens de différentes natures peuvent être mélangés ainsi que les concepts et instances.
Dans notre système, nous avons utilisé la notion de réseau sémantique comme outils de renforcement du graphe sémantique issu des termes extraits des documents d'apprentissage pour améliorer la qualité et la représentation des connaissances liées à chaque thème de la base documentaire.

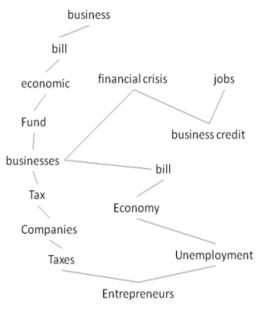
3.4. Construction du graphe

Il est important de noter que l'extraction des termes index se fait dans l'ordre de leur apparition dans le document. Les figures 3 et 4 illustrent ce processus pour un exemple de document du thème finances et économie

WASHINGTON (Reuters) – President Barack Obama signed a $30 billion small **business** lending **bill** into law on Monday, claiming a victory on **economic** policy for his fellow Democrats ahead of November congressional elections. The law sets up a lending **fund** for small **businesses** and includes an additional $12 billion in **tax** breaks for small **companies**."It was critical that we cut **taxes** and make more loans available to **entrepreneurs**," Obama said in remarks at the White House. "So today after a long and tough fight, I am signing a small **business jobs bill** that does exactly that." Obama is trying to show voters, who are unhappy about 9.6 percent **unemployment**, that he and his party are doing everything they can to boost the tepid U.S. **economy**. Democrats said they backed the **bill** because small **businesses** had trouble getting **loans** after the **financial crisis** that began in December 2007. They estimate the **incentives** could provide up to $300 billion in new small **business credit** in the coming years and create 500,000 new **jobs**.	**Business** Bill Economic Fund Businesses Tax Companies Taxes Entrepreneurs Business Jobs Bill Unemployment Economy Bill Businesses loans financial crisis incentives business credit jobs

Figure 3 : texte brute *Figure 4 : Texte après prétraitement et filtrage*

La construction du graphe sémantique tient en compte l'ordre de l'extraction et la distribution des termes dans le document. Chaque terme est associé à une fonction à base radiale qui fixe la proximité à un certain voisinage (zone d'influence sémantique du terme). Ce graphe est ensuite enrichie via le dictionnaire sémantique auxiliaire par l'adjonction de connexions. La correspondance requête- document se fait par une projection des termes de la requête sur le graphe sémantique. Si ces termes sont dans une zone d'influence sémantique forte, alors ce document est pertinent à cette requête. Dans ce qui suit nous allons définir notre fonction à base radiale et nous verrons l'utilité du graphe sémantique pour le calcul de la proximité sémantique entre la requête et le document.



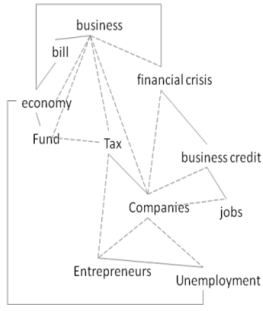


Figure 5 : Graphe Sémantique extrait à partir du document
Figure 6 : Renforcement du Graphe par les connexions sémantiques à partir du dictionnaire auxiliaire

4. Indexation sémantique à fonction à base radiale

Plusieurs travaux ont adapté le modèle vectoriel en indexant directement les concepts à la place des termes. Ces approches traitent essentiellement la synonymie en remplaçant les termes par leurs concepts. Nous traitons des liens plus riches entre les termes en prenant en considération tout les types de relations sémantiques (dans l'idée de construire une ontologie informelle du domaine au sens de conceptualisation). Ceci peut résoudre le problème de la synonymie mais aussi évite les complications causées par les autres relations de spécialisation et de généralisation par exemple.

4.1. Notre contribution pour l'indexation et la classification

Contrairement aux méthodes existantes, nous ne nous restreignons pas à l'utilisation des concepts. En effet, les termes sont enrichis s'ils sont fortement reliés aux concepts voisins et s'ils assurent une bonne connectivité sémantique. Il est important de noter que lors de la recherche, nous pouvons aussi retrouver les termes qui ne sont pas reliés au sein du réseau sémantique.

Pour calculer la similarité entre termes, nous définissons $\phi(d)$ une fonction à base radiale qui associe à chaque terme une zone d'influence caractérisée par le degré de similarité sémantique et la relation entre le terme noyau et ses voisins. (Rada & al., 1989) ont été les premiers à suggérer que la similarité dans un réseau sémantique peut être calculée en se basant sur les liens taxonomiques «is-a». Un moyen des plus évidents pour évaluer la similarité sémantique dans une taxonomie est de calculer la distance entre les noeuds comme le chemin le plus court.

Nous sommes conscients que le calcul de la mesure de similarité par restriction sur le lien «is-a» n'est pas toujours bien adapté car, dans la réalité, les taxonomies ne sont pas toujours au même niveau de granularité, des parties peuvent aussi être plus denses que d'autres. Ces problèmes peuvent être résolus en associant des poids aux liens. Ainsi nous avons choisie de prendre en considération tous les types de relations (problématique conceptuelle) et la répartition des mots dans les documents (problématique structurale).

Nous avons adapté notre système pour qu'il supporte toute sorte de relation sémantique telle que la synonymie, méronymie, hyponimie, taxonimie, antonomie, etc… et nous affectons initialement un poids unité pour les liens sémantiques.

Un réseau sémantique est construit à chaque phase pour modéliser les relations sémantiques entre les termes. Afin d'éviter les problèmes de connexité, nous avons choisi de construire un dictionnaire auxiliaire de telle sorte à avoir une connexité forte du réseau ainsi construit et d'augmenter le poids sémantique des termes descripteurs par la suite.

Dans la section suivante, nous définissons notre mesure TFIDF à base radiale et nous allons voir par la suite comment les poids des termes de l'indexation sont enrichis à partir des sorties de cette mesure.

4.2. Le TF-IDF à base radiale

Les TFIDF à fonction à base radiale (RBF pour Radial Basis Function) s'appuient sur la détermination de supports dans l'espace de représentation E. Cependant, à la différence des TFIDF traditionnels, ceux-ci peuvent correspondre à des formes fictives qui sont une combinaison des valeurs de TFIDF traditionnels, nous les appellerons donc prototypes. Ils sont associés à une zone d'influence définie par une distance (euclidienne, Mahalanobis...) et une fonction à base radiale (Gaussienne, exponentielle...). La fonction discriminante g d'un TFIDF RBF à une sortie est définie à partir de la distance de la forme en entrée à chacun des prototypes et de la combinaison linéaire des fonctions à base radiale correspondantes :

$$g(X) = w_0 + \sum_{i=1}^{N} w_i \, \phi(d(X, \sup_i)) \quad (4)$$

Où $d(x, \sup_i)$ est la distance entre l'entrée x et le support sup_i, $\{w_0, ...,w_N\}$ sont les poids de la combinaison et ϕ la fonction à base radiale.
L'apprentissage de ce type de modèle peut se faire en une ou deux étapes. Dans le premier cas, une méthode de type gradient est utilisée pour ajuster l'ensemble des paramètres en minimisant une fonction objective basée sur un critère comme les moindres carrés. Dans le deuxième cas, une première étape consiste à déterminer les paramètres liés aux fonctions à base radiale (position des prototypes et zones d'influence). Pour déterminer les centres, des méthodes de classification non supervisée sont souvent utilisées. Les poids de la couche de sortie peuvent, dans une seconde étape, être appris par différentes méthodes comme la pseudo-inverse ou une descente de gradient. Dans le cas d'un apprentissage en deux étapes, les TFIDF RBF possèdent alors plusieurs avantages. Par exemple l'apprentissage séparé des fonctions à base radiale et de leur combinaison permet un apprentissage rapide, simple et évite les problèmes de minima locaux (pertinence locale et globale). Les prototypes des TFIDF- RBF représentent la répartition des exemples dans l'espace E de représentation (termes). De plus la gestion des problèmes multi-classes est plus simple dans les TFIDF-RBF. Nous verrons dans la section suivante que les TFIDF RBF sont très semblables sous certaines conditions aux Systèmes d'Inférence Floue.
La modélisation des TFIDF RBF est à la fois discriminante et intrinsèque. En effet la couche de fonctions à base radiale correspond à une description intrinsèque des données d'apprentissage et la couche de

combinaison en sortie cherche ensuite à discriminer les différentes classes.
Dans notre système, nous utilisons des TFIDF RBF avec un apprentissage en deux étapes. La fonction à base radiale est du type fonction de Cauchy de la forme :

$$\phi(d) = \frac{1}{1+d} \quad (5)$$

Et nous définissons deux nouveaux opérateurs :

$$PoidRel(c) = \frac{degré(c)}{nombre\ total\ de\ concepts} \quad (6)$$

C'est le poids relationnel du concept (terme ou vecteur) c et *degré(c)* est le nombre des arrêtes entrantes et sortantes du sommet c. Il représente donc la densité de connexion du concept c au sein du réseau sémantique.

$$DensitéSem(c_1, c_2) = \frac{CoutMin(c_1, c_2)}{Arbre\ recouvrant\ de\ cout\ minimal} \quad (7)$$

DensitéSem(c_1, c_2) est la densité sémantique de la liaison (c_1, c_2). C'est le rapport de la distance sémantique minimale CoutMin(c_1, c_2) entre c_1 et c_2, calculée par l'algorithme de Dijkstra (Cormen et al., 2001). Cette distance est calculée à partir du réseau sémantique ainsi construit à partir de document sur la base du coût minimal de l'arbre recouvrant (c'est l'arbre de coût minimal en suivant tous les chemins minimaux de c_1 vers c_2 et les autres sommets du réseau sémantique). Cette mesure reflète l'importance de la liaison (c_1, c_2) par rapport à l'ensemble des chemins minimaux existants. Par la suite nous calculons la distance sémantique en terme conceptuel comme suit :

$$DistSem(c_1, c_2) = PoidRel(c_1) * PoidRel(c_2) * DensiteSem(c_1, c_2) \quad (8)$$

La mesure de proximité est alors une fonction de Cauchy :

$$Proximité(c_1, c_2) = \frac{1}{1 + DistSem(c_1, c_2)} \quad (9)$$

L'apport de ces opérateurs ainsi définis est qu'ils donnent plus d'importance aux concepts qui ont un voisinage sémantique dense où s'ils ont une bonne connectivité au sein du réseau. Cela a par ailleurs été vérifié durant la validation du prototype.
Nous avons également remarqué que la pondération TFIDF-OKAPPI traditionnelle de quelques termes qui sont considérés comme significatifs pour l'indexation d'un document se trouvent en bas du classement. Après le calcul de la pondération TFIDF-OKAPPI-ABR combinée par

notre fonction à base radiale, ces mêmes termes se retrouvent en haut du classement.
Pour la phase d'indexation, nous allons voir dans la partie qui suit comment les poids des descripteurs index sont générés par la nouvelle mesure à base radiale sur la base de la distance sémantique comme paramètre.

5. Nouvelle pondération des descripteurs index

Les documents sont représentés par des ensembles de vecteurs de termes. Les poids des termes sont calculés en fonction de leur distribution dans les documents. Le poids d'un terme est enrichi par les similarités conceptuelles des termes co-occurrents dans le même thème.
Nous procédons au calcul du TFIDF des termes pour l'ensemble des thèmes de la base d'apprentissage pour en déduire la pertinence globale. On calcule ensuite leur pertinence locale par l'intermédiaire de notre fonction à base radiale définie précédemment en la combinant avec le TFIDF traditionnel et en n'acceptant que les termes situés dans la zone d'influence. Ce poids noté TFIDF-ABR (t) est calculé de la manière suivante :

$$\text{TFIDF-ABR}(t,\text{theme}) = \text{TFIDF}(t,\text{theme}) + \sum_{i=1}^{n} \text{TFIDF}(t_i,\text{theme}) * \varphi(\text{Proximité}(t,t_i)) \quad (10)$$

Avec $\varphi(\text{Proximité}(t, t_i)) < \text{seuil}$

t_i ensemble des n termes dans le thème.

seuil : une valeur qui fixe la proximité à un certain voisinage (zone d'influence sémantique du terme t), nous la fixons dans un premier temps à la proximité entre le concept de t et le **concept contexte** (concept qui représente le thème).

5.1 Okapi à base radiale

Vu les limites de la mesure TFIDF évoquées précédemment, nous avons opté pour un modèle d'Okapi proposé par (Robertson, 2000) en y introduisant une extension sémantique.

Pour ce faire, la fonction $\phi(d)$ calcule le degré de pertinence pour chaque terme au niveau de son voisinage sémantique (zone d'influence). La nouvelle formule devient :

$$a_{ij} = \frac{tf(i,j).idf(i)}{[(1-b) + b \cdot \text{NDL}(d_j)] + f(i,j)} \cdot \phi(d_j) \quad (11)$$

Nous indiquons par $\phi(d_j)$ l'ensemble des termes proches sémantiquement de t_i. Un seuil de similarité est nécessaire pour caractériser l'ensemble de ses éléments. Nous fixons un seuil de similarité pour la valeur de Proximité (t_i,t) qui correspond au degré de similarité

entre t et le concept du thème où il apparaît (le terme est accepté s'il se trouve dans la zone d'influence de terme noyau définie par la fonction à base radiale φ). La relation devient donc :

$$\text{OKAPPI-ABR}(t,\text{theme}) = \text{Okappi}(t,\text{theme}) + \sum_{i=1}^{n} \text{Okappi}(t_i,\text{theme}) * \varphi(\text{Proximité}(t,t_i)) \quad (12)$$

5.2 N-Gramme à base radiale

L'utilisation de la méthode N-gramme (avec N=3 nombre de caractères) dans la recherche des documents arabes est plus efficace que celle du « keyword matching ».

Pour l'indexation et la classification de documents arabes, le choix des mesures statistiques comme les trigrammes et le poids TF*IDF semble pertinent.

L'utilisation de la méthode N-gramme pour l'indexation et la classification des documents reste insuffisante pour obtenir de bons résultats dans la recherche d'information en langue arabe. Pour cela nous avons pensé à ajouter de la pertinence sémantique à cette mesure en tenant compte de la notion du voisinage sémantique des termes extraits par une combinaison N-gramme avec une fonction à base radiale, la formule générale devient :

$$\text{NGRAM-ABR}(t,\text{theme}) = \text{NGRAM}(t,\text{theme}) + \sum_{i=1}^{n} \text{NGRAM}(t_i,\text{theme}) * \varphi(\text{Proximité}(t,t_i)) \quad (13)$$

6. Résultats

Pour la phase d'apprentissage nous avons travaillé sur une base (corpus initial) très réduite de documents étiquetés représentatifs des classes (sport, politique, économie & finance) que l'on cherche à discriminer ou à apprendre et c'est le point fort de notre mesure. Plus cette base est discriminante et représentative plus notre méthode est performante avec de meilleurs résultats.

Pour la phase de test nous avons travaillé sur deux corpus de presse (the Associated Press (AP)) de 400 documents chacun, l'un en langue arabe et l'autre en anglais. Le corpus anglais est une partie extraite d'un corpus plus large de 2246 documents (http://www.cs.princeton.edu/~blei/lda-c/ap.tgz). Pour le corpus arabe, c'est une collection de documents extraite de (www.aljazeera.net).

Le tableau 1 montre les résultats préliminaires obtenus. Ce résultats sont exprimés à travers les critères Rappel, Précision et performances en classification. Ce tableau montre en particulier la pertinence de l'utilisation de notre approche en comparaison avec l'approche N-Grammes.

Corpus	Méthode	Rappel	Précision	Performance en classification (%)
Anglais	TFIDF	0.80	0.83	80.3
	NGRAM	0.56	0.78	65.95
	TFIDF-ABR	0.89	0.92	90.5
	NGRAM-ABR	0.6701	0.8463	74.79
Arabes	TFIDF	0.81	0.81	81.0
	NGRAM	0.45	0.81	57.85
	NGRAM-ABR	0.6341	0.8762	73.57
	Okappi-TFIDF-ABR	0.98	0.98	98.79

Tableau 1 : *Tableau des résultats de l'expérimentation*

7. Conclusion

L'intégration de la notion de voisinage sémantique et de fonctions à base radiale a permis d'améliorer d'une manière très significative les performances de notre système d'indexation indépendamment de la langue manipulée. Ces résultats restent à confirmer sur des corpus plus conséquents, même si de tels corpus sont difficiles à se procurer pour la langue Arabe, qui reste notre objectif primordial.
Nous avons remarqué que les résultats de l'indexation contiennent exactement les mots-clés recherchés triés selon leur pertinence. Nous avons également fixé un seuil pour l'enrichissement sémantique, ce qui peut conduire à retourner quelques termes indésirables assez éloignés de ceux recherchés.
Nous avons aussi constaté que l'hybridation de deux méthodes statistiques améliore considérablement les performances.
Un autre point à prendre en compte et qui peut dégrader la précision des méthodes statistiques traditionnelles, est la présence de concepts complexes. Ce point peut s'avérer une piste intéressante à explorer puisque les concepts longs sont en principe moins sujets à ambiguïté.
Pour réponde à ces différentes situations, nous envisageons l'utilisation d'un algorithme de désambiguïsation et l'hybridation entre différentes mesures en les combinant avec des fonctions noyaux.

Remerciements
Ce travail est soutenu par le Programme Hubert Curien Franco-marocain Volubilis n° MA/10/233 et le projet AIDA du programme Euro méditerranéen 3+3 n ° M/09/05.

Bibliographie

AFNOR (1993). Information et documentation. Principes généraux pour l'indexation des documents. NFZ 47-102.

PORTER M. F. (1980). An algorithm for suffix stripping. *Program*, 14 :130–137, 1980. 15.

SEYDOUX F., RAJMAN M. and CHAPPELIER J.C. (2006**)**. *Exploitation de connaissances sémantiques externes dans les représentations vectorielles en recherche documentaire*. Ph.D. thesis.

BLEI D., NG A., and JORDAN M. (2003). Latent dirichlet allocation.

SEBASTIANI F., SPERDUTI A., and VALDAMBRINI N. (2000). An improved boosting algorithm and its application to automated text categorization. Technical report, Paris, France.

ROBERTSON S., WALKER S., BEAULIEU M.,(2000). Experimentation as a way of life : Okapi at TREC, *InformationProcessing and Management*, vol. 36, no 1,2000,pp. 95-108.

DEERWESTER S., DUMAIS S., FURNAS G., LANDAUER T., and Harshman R (1990). Indexing by latent semantic analysis.

Quillian M.R. (1968). Semantic memory. Semantic information processing, 1968. 65.

RADA R., MILI H., BICKNELL E., BLETTNER M. (1989). « Development and application of a metric on semantic nets », *IEEE Transaction on Systems, Man, and Cybernetics*, vol. 19, no 1, 1989, p. 17–30.

KHOJA S. and GARSIDe S. (1999). Stemming Arabic Text. Computing Department, Lancaster University, Lancaster, U.K. http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps, September 22, 1999.

CORMEN T. H., LEISERSON C. E., RIVEST R. L. and STEIN C. (2001). Introduction à l'algorithmique, (version (en) (ISBN 0-262-03293-7) deuxième édition, 2001, MIT Press and McGraw-Hill, section 24.3, Dijkstra's algorithm, pages 595–601, 2001.

SHANNON, C. (1948). The Mathematical Theory of Communication. Bell System Technical Journal, 27 :379–423 and 623–656.

Approche d'indexation automatique d'informations pédagogiques à partir de documents

Boutheina SMINE
LaLIC, Université Paris-Sorbonne
LaRODEC, IHEC de Carthage, 2016 Carthage Présidence, Tunisie.

Rim FAIZ
LaRODEC, IHEC de Carthage, 2016 Carthage Présidence, Tunisie.

Jean-Pierre DESCLES
LaLIC, Université Paris-Sorbonne

Résumé : Il y a besoin sans cesse croissant en informations pédagogiques pour les intégrer dans des ressources ou dans un processus d'apprentissage. Une indexation de ces informations s'avère donc utile en vue d'une extraction des informations pédagogiques pertinentes en réponse à une requête utilisateur. La méthode d'indexation proposée par la plupart des systèmes d'extraction d'informations pédagogiques est basée sur une annotation manuelle ou semi-automatique des informations pédagogiques, tâche qui n'est pas préférée par les utilisateurs. Dans cet article, nous proposons une approche d'indexation d'objets pédagogiques (Définition, Exemple, Exercice, etc.) basée sur une annotation sémantique par Exploration Contextuelle des documents. L'index généré servira à une extraction des objets pertinents répondant à une requête utilisateur sémantique. Nous procédons, ensuite, à un classement des objets extraits selon leur pertinence en utilisant l'algorithme Rocchio.

Mots-clés : Informations pédagogiques, carte sémantique, exploration contextuelle, Rocchio.

1. Introduction

La quantité d'informations pédagogiques disponible en ligne est en perpétuelle croissance. Dans leur processus de recherche d'informations ou d'apprentissage, les apprenants peuvent être soutenus par les moteurs de recherche. Toutefois, ces systèmes de recherche d'information sont

basés sur l'indexation des termes sans tenir compte de la sémantique du contenu pédagogique (Dehors et al., 2005), (Buffa et al., 2005). Une meilleure alternative est de proposer une approche d'indexation basée sur l'annotation sémantique des informations pédagogiques qui sont attestées dans les documents. Par une telle indexation, les informations pédagogiques présentées par l'auteur d'un document sont capturées et le processus d'apprentissage ou d'enseignement pour l'élève ou l'enseignant respectivement est facilité.
Nous proposons, dans cet article, une approche d'indexation automatique d'informations pédagogiques à partir de documents. Notre travail consiste d'abord à annoter les segments textuels (objets) reflétant un contenu pédagogique (Définition, Exemple, Exercice, etc.). Ensuite, nous procédons à une indexation de ces objets annotés pour extraire ceux qui sont pertinents par rapport à une requête utilisateur. Enfin, nous procédons à un classement de ces objets en utilisant l'algorithme de classification Rocchio.
Dans la section 2, nous positionnons cette contribution par rapport aux travaux existants. Nous consacrons la section 3 à la définition de la notion d'objet pédagogique. Une description détaillée de notre approche d'indexation d'informations pédagogiques est le sujet de la quatrième section. Avant de conclure, nous illustrons les résultats des expérimentations de notre approche dans la cinquième section.

2. Indexation des informations pédagogiques : Etat des lieux

Nous détaillons ici divers points de l'état de l'art liés à notre approche d'indexation d'objets pédagogiques, à savoir l'annotation, l'indexation, et l'extraction d'informations pédagogiques à partir de documents textuels.
L'annotation comme technique d'indexation est appliquée dans plusieurs systèmes comme le système QBLS (Dehors et al., 2005) qui est une partie de la plateforme TRIAL SOLUTION (Buffa et al., 2005). Dans cette dernière, les utilisateurs annotent les livres manuellement selon le rôle pédagogique de leur contenu, les sujets abordés dans leur contenu (mots clés) et leurs relations avec d'autres ressources (référence, prérequis, etc.). Le système QBLS a pour but de structurer le cours en se référant à une ontologie pédagogique constituée de fiches (définition, exemple, énoncé, procédure, solution, etc.) et de ressources pédagogiques abstraites (cours, thème, notion, question). Il existe aussi le système SYFAX (Smei et al., 2005) qui annote semi-automatiquement le document pédagogique selon plusieurs critères (type du document, point de vue de l'utilisateur sur le document, etc.).
En vue d'indexer les documents, les annotations proposées par les différents systèmes cités ci-dessus sont stockées dans un entrepôt de connaissances pédagogiques. Par la suite, les réponses aux requêtes sont

extraites à partir de cet entrepôt grâce à un moteur de recherche (Corese pour le système QBLS). Le système SYFAX applique un processus de raffinement de la requête basé sur une ontologie des types de documents pédagogiques et une autre ontologie des domaines des documents informatiques. Ceci permet d'extraire les documents pertinents par rapport à la requête.

Pour tous les systèmes présentés ci-dessus, une intervention humaine est requise afin d'enrichir les documents par des métadonnées. Cependant, la plupart des producteurs de ressources pédagogiques ne s'intéressent probablement pas au retour aux documents pour annoter leurs propres travaux. Notre travail se place dans cette perspective tout en procédant à l'automatisation du processus d'annotation.

D'autres travaux se sont intéressés à la recherche de ressources pédagogiques à partir du web (Thomson et al., 2003). Toutefois, le but de leur travail est limité à une extraction de métadonnées (Travaux Dirigés, Programme, Travaux Pratiques) relatives au document en entier en vue de les annoter et de les classifier. Toujours dans la même perspective, (Hassen et al., 2009) comparent l'efficacité des algorithmes Naïve Bayes et SVM dans la classification des ressources pédagogiques basée sur un ensemble de propriétés (catégorie du contenu, titre du cours, année, auteur, etc.).

A notre connaissance, ces travaux de recherche portant sur l'indexation de documents pédagogiques se sont intéressés à une indexation du document en l'annotant par un ensemble de métadonnées relatives à la totalité du document. D'autres approches basées sur des patrons linguistiques ont été appliquées dans plusieurs travaux pour extraire les définitions à partir de ressources pédagogiques afin de constituer un glossaire (Westerhout et al., 2008) ou encore pour répondre à divers types de questions (Greenwood et al., 2003). Cependant, les patrons sont appliqués la plupart du temps à extraire des objets pédagogiques de type "Définition" en raison de l'accessibilité des structures langagières relatives à ce type que ce soit sur le web (wikipédia, dictionnaires, etc.) ou dans d'autres sources comme les rapports, les manuels d'utilisation, etc.

Dans cet article, nous proposons une annotation automatique des informations pédagogiques avec des métadonnées sémantiques (Définition, Exemple, Exercice, etc.). Ce qui nous permettrait d'indexer ces informations en vue d'une extraction des informations pertinentes par rapport à une requête utilisateur.

3. Présentation des objets pédagogiques

Un utilisateur "extracteur" d'informations pédagogiques pertinentes est guidé dans sa lecture par certains passages (des segments textuels comme des phrases ou des paragraphes). L'hypothèse générale utilisée ici est

d'essayer de reproduire ce que fait un humain, en particulier l'apprenant, en soulignant certains segments textuels reflétant un contenu pédagogique. Ces segments de type pédagogique, appelés objets pédagogiques, existent, généralement, dans les documents pédagogiques sous forme de définitions, exemples, exercices, plan, questions et réponses, etc. Un objet pédagogique peut être défini comme une entité numérique ou non (Flory, 2004) qui peut être utilisée ou citée dans un apprentissage. Dans notre cas, un objet pédagogique correspond à un segment textuel reflétant un contenu pédagogique.

Un apprenant pourrait être intéressé par une définition en formulant une requête, par exemple: trouver les documents qui contiennent "La définition du langage SQL". Un autre utilisateur recherche, en explorant de nombreux textes (encyclopédies spécialisées, manuels, articles), des exemples sur un concept (par exemple, «l'inflation» dans l'économie, «polysémique» en linguistique, ..) pour l'intégrer à ses ressources pédagogiques. Un autre utilisateur peut être intéressé, à la pratique des exercices sur un concept. L'objectif de ces types d'objets pédagogiques (Définition, Exemple, Exercice) est une annotation possible des segments textuels pédagogiques qui correspondent à une recherche guidée afin d'en extraire des objets pédagogiques à partir de textes. Chaque type pédagogique, comme nous l'avons mentionné ci-dessus, est explicitement indiqué par les marqueurs linguistiques identifiables dans les textes. Notre hypothèse est que chaque type d'objet pédagogique laisse des traces discursives dans le document texte. Les types d'objets pédagogiques sont décrits comme suit :

(1) D'une part, une relation complexe entre les concepts dans une structure «carte sémantique» (Figure 1) et d'autre part un ensemble de classes et sous-classes d'unités linguistiques (indicateurs et indices).

(2) Un ensemble de règles communautaires où chaque règle concerne une classe d'indicateurs avec des indices différents.

La carte sémantique (Figure 1) est une organisation des types d'objets pédagogiques. Elle peut être conçue aussi comme une ontologie des types d'objets pédagogiques indépendamment des différents domaines d'application. En effet, les expressions de la carte sémantique pour un type d'objet sont les mêmes dans différents domaines comme l'informatique, mathématiques, gestion, ... car ces expressions sont utilisées par l'auteur pour exprimer une information pédagogique. Dans certains types de textes (textes narratifs, articles de presse,) les expressions pédagogiques ne sont pas présentes mais dans d'autres (support de cours, devoirs, travaux dirigés, ..), ces expressions organisent le texte et donnent des informations sur l'intention de l'auteur.

Le premier niveau de la carte sémantique (Figure 1) présente 6 types d'objets pédagogiques : (i) Cours, (ii) Plan, (iii) Exercice, (iv) Exemple, (v) Définition, (vi) Caractéristique. Par exemple, les règles du type d'objet "Définition" sont déclenchées par la présence de noms ou de verbes

définitoires (par exemple: "est défini", et l'annotation sémantique est attribuée si des indices linguistiques, comme les prépositions (l'indice de l'exemple précédent est "par"), sont trouvés dans le contexte de l'indicateur.

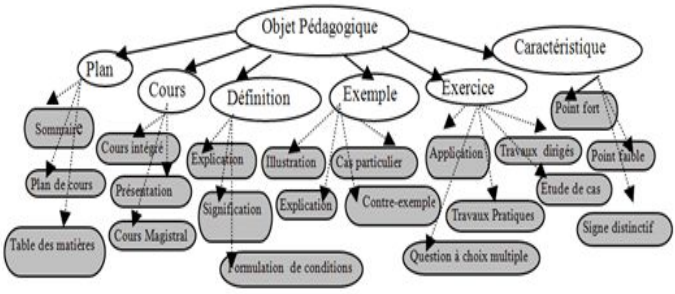


Figure 1 : Carte sémantique des types d'objets pédagogiques

4. Approche proposée pour la recherche d'informations pédagogiques à partir de documents

L'approche que nous proposons se décompose en deux principales parties: dans la première partie, nous procédons à une annotation sémantique des segments textuels représentant des objets pédagogiques (Smine et al., 2010). La deuxième partie exploite les annotations générées par la première partie pour créer un index qui est capable de localiser les segments textuels pertinents par rapport à des requêtes associées aux types pédagogiques (Définition, Exemple, Exercice, etc.). Pour classer les réponses selon leurs pertinences, nous appliquons l'algorithme de classification Rocchio sur les objets pédagogiques extraits.

4 .1. Annotation des objets pédagogiques

4.1.1. Segmentation

Mourad (Mourad, 2002) propose de segmenter le texte en se basant sur une étude systématique des marques de ponctuation. Nous avons effectué la segmentation de nos documents en intégrant les règles linguistiques développées par Mourad. Pour chaque document segmenté, le résultat obtenu est un fichier XML balisé par des balises <Section>, <Paragraphe>, <Phrase>.

4.1.2. Annotation des objets pédagogiques

Pour annoter les objets, nous explorons la technique d'Exploration Contextuelle 'EC' (Desclés, 1997). C'est une technique de traitement linguistique et sémantique du langage, qui fait appel à des marqueurs

discursifs explicites (morphèmes, mot, expression, etc.) caractéristiques d'une intention pragmatique de l'auteur. 'EC' consiste à appliquer des règles dans un contexte déterminé par des indices. Elle a l'avantage d'être indépendante d'un domaine particulier, car les règles décrivant les structures linguistiques sont indépendantes d'un domaine particulier. C'est une méthode qui a été validée par les travaux de (Djioua et al., 2006) et (Elkhlifi et al., 2010). En plus, 'EC' ne nécessite pas une analyse morphosyntaxique du texte, ce qui réduit considérablement le temps d'exécution pendant l'implémentation de la méthode.
Par l'exploration contextuelle du contenu des documents, nous pouvons repérer et annoter les objets pédagogiques contenu dans ces documents, par exemple, « des exemples de requêtes SQL», « des exercices sur le langage UML », « les définitions d'une ou de plusieurs notions », etc. Ces objets sont exprimés par des structures langagières comme « …se définit par… », « est défini par… » pour le type Définition ou « Exercices sur… », « Travaux dirigés » pour le type Exercice. Ils sont explicitement indiqués par des indicateurs linguistiques identifiables dans les textes (verbes, noms, adjectifs). Ces indicateurs sont parfois polysémiques, ils ont besoin d'indices linguistiques pour clarifier l'indétermination. Les relations reliant les indicateurs aux indices sont définis dans le cadre des règles. Une règle (IdR) se déclenche au moment de l'identification de l'un de ses indicateurs (Indicateur) ensuite elle essaye de localiser des indices linguistiques dans le contexte gauche (CL1, CL2) et/ou droite (CR1, CR2) de l'indicateur ce qui confirme ou non la valeur sémantique exprimée par l'indicateur (Figure 2).

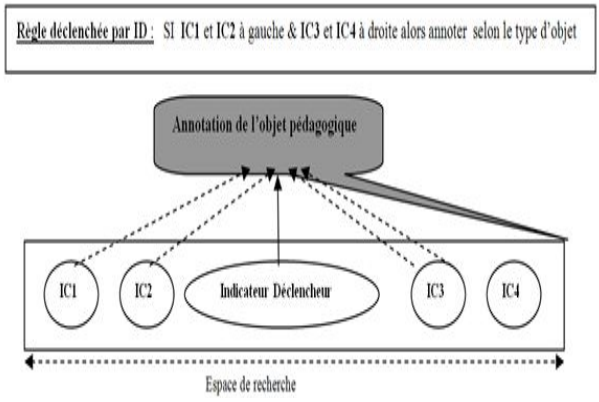


Figure 2 : Principe de fonctionnement d'une règle d'exploration contextuelle

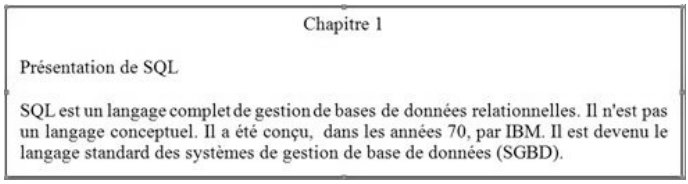
A chaque type d'objet pédagogique correspond un ensemble de règles. Des exemples de règles sont présentés dans le tableau suivant (Tableau 1).

IdR	CL_1	CL_2	Indicateur	CR_1	CR_2	Type|Sous-type de l'objet pédagogique
RD1	est|sont		défini|définis|définis	par		Définition|Explication
RD2			est | sont	le|la|un une|des|les		Définition|Explication
RC1	La|Les Des|Une		caractéristique|caractéristiques	du|de|des	est|sont	Caractéristique|Signes distinctifs
RE1	Voici	un|l' les|des	exemple|exemples	du|de|des		Exemple|Illustration

Tableau 1 : Des exemples de règles

Nous avons ajouté un composant à chaque règle qui représente l'emplacement du terme de la requête à rechercher dans le cadre du segment exprimant l'objet pédagogique. Le besoin d'ajouter ce composant est né de la variation de l'emplacement du terme à rechercher avec la variation des structures langagières exprimant les objets pédagogiques. Ceci permet d'identifier les segments textuels exprimant le type d'objet pédagogique ainsi que le concept demandés par l'utilisateur. Par exemple, pour le même type d'objet pédagogique "Définition" : le terme à rechercher "Maintenance" peut exister au début du segment "La maintenance est définie comme l'ensemble des activités destinées à maintenir ou à rétablir un bien dans un état de sûreté de fonctionnement" ou au milieu du segment pour le cas "L'AFNOR a défini la maintenance comme étant l'ensemble des activités de remise en état de fonctionnement d'un système". Sans la considération de ce paramètre, le système peut ne pas extraire l'objet demandé par l'utilisateur comme par exemple, pour le type Cours, la plupart de ses règles d'EC exigent un emplacement du terme de la requête au niveau du Titre du document. Au cas où le terme est recherché ailleurs que dans le titre, le résultat de la recherche sera erroné.
De ce fait, l'emplacement du terme est un paramètre qui diffère d'une règle à une autre selon la structure langagière exprimée par cette dernière. Nous avons désigné cet emplacement par une étiquette, qui prendra une valeur parmi un ensemble fini de valeurs désignant l'emplacement du terme par rapport aux indicateurs et indices. Par exemple, GIND indique le terme se place à gauche de l'indicateur ou TITRE indique que l'emplacement du terme est au niveau du titre du document. En fait, dans plusieurs cas, le titre peut nous révéler des connaissances sur le contenu du document.
Pour chaque type d'objet de la carte sémantique (cf. Figure 1), nous avons défini un ensemble de règles qui couvrent toutes les formes

linguistiques possibles de l'objet pédagogique. Nous avons commencé par un exemple textuel relatif à chaque type pour généraliser toutes les structures langagières. Cette méthode permet de définir de manière incrémentale une base solide de règles. Nous avons développé en totalité environ 200 règles. L'ensemble des règles développées, ainsi que la carte sémantique forment les ressources linguistiques utilisées dans notre approche.
Nous prenons un extrait de texte à partir d'un document pédagogique


Chapitre 1

Présentation de SQL

SQL est un langage complet de gestion de bases de données relationnelles. Il n'est pas un langage conceptuel. Il a été conçu, dans les années 70, par IBM. Il est devenu le langage standard des systèmes de gestion de base de données (SGBD).

Figure 3 : Un extrait d'un document pédagogique

Pour le type d'objet pédagogique "Définition", la règle RD2 (cf. Tableau 1), appliquée à l'exemple ci-dessus, permet d'annoter la phrase " SQL est un langage complet de gestion de bases de données relationnelles". Le type d'objet pédagogique est détecté grâce à l'expression "est " qui est une occurrence Ii de l'indicateur du type "Définition" et l'indice droit CR1 "un".
Pour le type "Cours", le repérage de l'occurrence Ii au niveau du titre est suffisant pour annoter le document comme un cours. L'indicateur nominal de l'objet pédagogique est le mot "Cours", et d'autres noms comme "Chapitre, "Notes de cours". A part le titre, l'existence de l'indicateur "Cours" n'implique pas l'annotation du document comme un cours.
Notons que la phrase "Il n'est pas un langage conceptuel" illustre le cas des indices négatifs. En effet, la présence de l'expression "n'….pas" annule l'annotation du segment comme Définition, malgré la présence de l'indicateur "est" et l'indice droit CR1 "un".
Afin d'annoter le segment " Il a été conçu, dans les années 70, par IBM" comme une "Caractéristique", nous détectons en premier lieu l'expression "a été conçu" ensuite nous cherchons, dans le contexte droit de l'indicateur, le CR1 "par". En cas où les deux éléments (Ii et CR1) sont présents, alors le système annote le segment comme une caractéristique.
Concernant le type d'objet "Exercice", l'indicateur peut être verbal (a) ou nominal (b), par exemple :
(3) (a) "Formulez une clause SQL……" a comme indicateur verbal "Formulez"
(4) (b) "Exercices sur requêtes SQL", son indicateur est le nom "Exercices"

4.2. Génération de l'index

Notre objectif, par l'annotation, est de générer un index sémantique contenant à la fois des objets pédagogiques annotés selon leur type, en utilisant la méthode d'annotation détaillée ci-dessus, et l'emplacement du terme de la requête spécifié par la règle appliquée pour annoter l'objet. Cet index servira à extraire les objets répondant à la requête utilisateur. Les métadonnées générées par les annotations des différents objets sont stockés dans une base de données. Pour chaque objet pédagogique annoté, les métadonnées suivantes sont introduites dans l'index : (1) L'objet pédagogique annoté (OBJECT), (2) Chemin du document analysé (PATH), (3) Type de l'objet annoté (TYPE), (4) Identifiant de la règle appliquée pour annoter le segment (IDRule) et (5) L'emplacement du terme de la requête (TERMEMP). La figure suivante (Figure 3) montre deux exemples d'objets annotés.

EDIT	OBJECT	PATH	TYPE	IDRULE	TERMEMP
	La production est une transformation de ressources appartenant à un système productif et conduisant à la création de biens ou de services.	C:\Documents and Settings\Boutheina SMINE\Mes documents\Evaluation\Gestion de Production.txt	Définition:Explication	RD2	GND
	2) Exprimer en algèbre relationnel les requêtes suivantes et donner ses résultats : checkbld Nom des immeubles ayant plus de 10 étages. checkbld Qui habite le « Koudalou » ? checkbld Nom et Profession des personnes ayant emménagé avant 1994. checkbld Gérant des immeubles ayant un appartement de plus de 150 m². checkbld Dans quel immeuble habite un acteur ? checkbld Age et profession des occupants de l'immeuble géré par « Ross » ? checkbld Qui n'habite pas un appartement géré par « Ross » ?	C:\Documents and Settings\Boutheina SMINE\Mes documents\Evaluation\Base de donnée.txt	Exercice:Travaux Dirigés	RES	TITRE

Figure 4 : Deux exemples d'objets annotés et indexés

Afin de pouvoir extraire les objets pédagogiques qui contiennent des termes de la requête, nous avons utilisé la base de synonymes WOLF (qui représente la partie traduite en Français du dictionnaire WordNet) permettant d'enrichir la requête en prenant en compte tous les termes équivalents au terme de la requête. Ce dernier est remplacé par la liste de ses synonymes. Ceci permet d'étendre le champ de la recherche. La requête est ainsi composée des termes à rechercher (par exemple "Langage SQL") et du type d'objets pédagogiques requis par l'utilisateur (par exemple : Exercice).

Grâce à un moteur de recherche (implémenté sous la plateforme Lucene), le système se connecte à l'index généré et retient les documents contenant des objets pédagogiques de même type que celui énoncé dans la requête (Exercice). Ensuite, le moteur procède à une recherche des termes de la requête (Langage SQL ainsi que ses synonymes) à partir des objets annotés et indexés. Cette recherche s'effectue dans l'emplacement désigné par la règle avec laquelle est annoté l'objet pédagogique. Par exemple, si l'emplacement du terme spécifié par la règle est DIND, le terme de la requête est recherché à droite de l'indicateur de la règle appliquée (Dans ce cas règle de type Exercice). Dans le cas où la requête

est composée du type pédagogique "Exercice" et le terme «Langage SQL", le moteur de recherche procède comme suit :
(5) Il extrait tous les objets pédagogiques trouvés dans l'index associé à l'annotation « Exercice ».
(6) Pour chaque objet extrait, il recherche le terme "langage SQL» et ses synonymes dans l'emplacement spécifié par la règle d'annotation.
(7) Sélection, à partir des objets pédagogiques extraits, les objets comportant une occurrence du terme «langage SQL» ou ses synonymes dans le bon emplacement.
(8) Afficher toutes les informations présentes dans l'index associé à chaque objet pédagogique sélectionné.

4.3. Classement des objets pédagogiques

Après l'extraction des objets pédagogiques répondant à la requête utilisateur, une autre étape suit pour classer les réponses dans un ordre croissant selon leur similarité avec la requête. Pour classer ces objets, nous avons utilisé l'algorithme de Rocchio (Rocchio, 1971), adapté à la classification des textes (Ittner et al., 1995). L'utilisateur choisit un concept pour le correspondre au terme de sa requête, parmi une liste de 15 concepts appartenant à différents domaines (domaine de l'informatique, économie, génie mécanique, biologie, etc.). Ce sont des concepts auxquels appartient l'ensemble des documents du corpus d'annotation et d'indexation. Le concept choisi représente la classe Cuser par rapport à laquelle les objets seront classés selon leur pertinence. Rappelons que nous considérons un objet pédagogique comme un segment textuel ayant différentes tailles (Phrase, paragraphe, document, etc.) selon le type de l'objet.
Nous représentons les données (les objets d'apprentissage et de test) par des vecteurs de poids numériques. Le vecteur de poids pour le m ième objet pédagogique est Vm= (p1m, p2m,……..,plm), où l est le nombre de termes index utilisés. Nous utilisons comme termes des mots singuliers et composés. Nous adoptons la mesure de poids TF-IDF (Salton, 1991) et nous définissons le poids pkm comme suit :

$$p_k^m = \frac{f_k^m \log(N/n_k)}{\sum_{j=1}^{l} f_j^m \log(N/n_j)}$$

Avec N est le nombre total d'objets, n k est le nombre d'objets dans lesquels le terme index k apparaît, et f_k^m est :

$$f_k^m = \begin{cases} 0 & q = 0 \\ \log(q) + 1 & Sinon \end{cases}$$

Avec q est le nombre d'occurrences du terme index k dans l'objet m. Dans l'algorithme de Rocchio, un prototype est produit pour chaque classe C. Ce prototype est représenté par un vecteur singulier $\vec{c}_j$ de

même dimension que le vecteur de poids original v1, ……, vN. Pour chaque classe C, the k ième terme dans son prototype est défini comme

$$\vec{c}_j = \frac{\alpha}{|C_j|} \sum_{m \in C_j} p_k^m - \frac{\beta}{|N - C_j|} \sum_{m \notin C_j} p_k^m$$

Avec Cj est l'ensemble de documents appartenant à la classe C. Les paramètres α et β contrôlent la contribution des exemples positifs et négatifs par rapport au vecteur prototype. Nous utilisons les valeurs standards $\alpha = 4$ et $\beta = 16$ (Buckley et al., 1994).

Une fois l'apprentissage achevé, nous classons les nouveaux objets fournis comme réponses à la requête utilisateur. Ce classement se fait selon leur pertinence par rapport à la classe Cuser choisie par l'utilisateur. Les objets à classer sont tout d'abord convertis en vecteurs de poids, et puis comparés aux vecteurs de poids prototypes des différentes classes en utilisant la mesure de similarité cosinus.

La mesure de similarité entre l'objet de vecteur $\vec{O}$ et la classe Cuser de vecteur $\vec{C}_{user}$ est définie comme : $\cos(\vec{C}_{user}, \vec{O}) = \frac{\vec{C}_{user} * \vec{O}}{|\vec{C}_{userl}||\vec{O}|}$

Les objets ayant une valeur de similarité avec la classe Cuser supérieure à un seuil θ sont sélectionnés, ensuite classés dans un ordre croissant selon la valeur de leurs similarités par rapport à la classe Cuser. La valeur du seuil θ varie selon le type d'objet pédagogique. Par exemple, un objet annoté par le type "Cours" contient plus de termes significatifs qu'un objet annoté par le type "Exercice" (θCourse < θExercice). Nous ne prenons en compte que les valeurs positives de la mesure de similarité. Les objets sélectionnés sont alors affichés pour constituer la fiche pédagogique demandée par l'utilisateur. Une fiche pédagogique rassemble les objets pédagogiques de type celui exprimée par l'utilisateur dans sa requête et correspondant au même concept que celui recherché par l'utilisateur. Cette fiche permet une accessibilité aux objets directement sans avoir accès au document en entier.

5. Expérimentations et Résultats

L'objectif de cette étape est d'évaluer les performances des différents modules. Un des indicateurs importants est donc le nombre des réponses pertinentes par rapport au nombre de documents indexés. Pour valider notre approche d'indexation d'objets pédagogiques, nous avons développé le système SRIDoP (Système de Recherche d'Informations à partir de Documents Pédagogiques) en utilisant le langage Java sous l'environnement Eclipse et le système de gestion de base de données Oracle. SRIDoP comporte les trois modules suivants : Annotation et

indexation des objets pédagogiques selon leurs types, Appariement entre la requête utilisateur et les objets pédagogiques indexés, et Classement des objets pédagogiques.
Notre corpus d'apprentissage ainsi que celui du test est le même pour toutes les étapes d'annotation, d'indexation et de classification. Pour le corpus d'apprentissage, nous avons collecté un ensemble de documents couvrant 15 concepts (ceux utilisés dans la génération de fiches pédagogiques). En fait, pour chacun de ces concepts, une requête a était formulée et exécutée sur le moteur de recherche Google. Les 20 premiers résultats sont collectés. Notons que le sens de quelques termes peut être ambigu, par exemple "Base" ou "Enregistrement". Pour désambiguïser la requête, nous ajoutons le terme "Données". Pour faire disparaitre l'ambiguïté, nous misons sur le type pédagogique des documents retournés en réponse. Les documents collectés sont constitués de 60 supports de cours, 65 Travaux Dirigés, 83 Présentation PowwerPoint, 30 Travaux Pratiques, et quelques documents de différentes natures (articles de Presse, articles scientifiques, etc.). La longueur moyenne de ces documents constituant le corpus d'apprentissage est 23 pages.
Notre corpus de test est composé de 1000 documents, principalement de nature pédagogique : des Supports de cours, des Travaux Dirigés, des présentations PowerPoint, des Travaux Pratiques, des manuels d'utilisation, et d'autres documents de différentes natures. La longueur moyenne des documents est 53.6 pages. Les documents ont différents formats (DOC, PDF, HTML, PPT, etc.).

5.1. Première étape : Annotation des objets pédagogiques

Pour évaluer le processus d'annotation, le corpus de test a été annoté par deux experts : pour chaque objet pédagogique repéré, ils précisent son type. Les résultats du processus d'annotation effectué par notre système SRIDoP sont illustrés dans le Tableau 2.

Type de l'objet pédagogique	NOA	NOAC	NOMAC	Précision (%)	Rappel (%)	F-Mesure (%)
Plan	88	85	98	96,59	86,73	91,40
Cours	72	60	85	83,33	70,59	76,43
Définition	228	140	266	61,40	52,63	56,68
Caractéristique	139	124	156	89,21	79,49	84,07
Exemple	357	349	376	97,76	92,82	95,23
Exercice	760	705	776	92,76	90,85	91,80

Tableau 2 : Les résultats de l'étape Annotation

$$Précision = \frac{NOAC}{NOA} \qquad Rappel = \frac{NOAC}{NOMAC} \qquad F-Mesure = 2 * \frac{Précision * Rappel}{Précision + Rappel}$$

Avec : NOA : Nombre d'objets annotés, NOAC : Nombre d'objets annotés correctement, NOMAC: Nombre d'objets annotés par les experts.
Nous remarquons que la précision de l'annotation dépasse les 85% pour la plupart des types d'objets (Exemple, Exercice, Plan, etc.). Notons que, pour le type « Définition », cette précision est moyenne. Ceci dérive du fait que certaines règles peuvent annoter à la fois des énoncés définitoires ou non. Tel le cas de la règle « R2 » ayant comme indicateur l'occurrence «est un». Cet indicateur peut identifier un segment de nature définitoire (exemple : « UML est un langage de modélisation conceptuelle orienté objet ») ou un autre segment de nature non définitoire (exemple : « Le facteur temps est un des plus importants dans la réalisation d'un projet »).
Pendant la phase d'expérimentation, nous avons pu constater aussi que la qualité de l'annotation est étroitement liée à la qualité de la segmentation du document.

5.2. Deuxième étape : Indexation des objets pédagogiques

A travers une interface de recherche d'informations, l'utilisateur saisit les termes à rechercher, et choisit le type (et sous-type) de l'objet pédagogique relatif au terme à rechercher. Les réponses aux requêtes sont affichées sous forme de liens permettant d'accéder à l'objet pédagogique répondant au besoin de l'utilisateur.
Pour tester ce module de recherche d'objets pédagogiques, nous avons formulé les mêmes 25 requêtes pour chacun des types d'objets pédagogiques. Ces requêtes appartiennent aux différents domaines du corpus. Pour chaque type d'objet, nous avons illustré le nombre de réponses ramenées et le nombre de réponses jugées pertinentes compte tenu de l'ensemble des requêtes formulées. Les résultats sont résumés dans le tableau suivant (Tableau 3).

Type de l'objet pédagogique exprimé par la requête	NR	NRP	NRRU	Précision (%)	Rappel (%)	F-Mesure (%)
Plan	72	66	77	91,67	85,71	88,59
Cours	43	35	54	81,40	64,81	72,16
Définition	156	112	193	71,79	58,03	64,18
Caracteristique	94	86	112	91,49	76,79	83,50
Exemple	213	198	230	92,96	86,09	89,39
Exercice	517	465	520	89,94	89,42	89,68

Tableau 3 : Les résultats de l'étape d'appariement Documents-Requête

$$\text{Précision} = \frac{NRP}{NR} \qquad Rappel = \frac{NRP}{NRRU} \qquad F - Mesure = 2 * \frac{\text{Précision} * Rappel}{\text{Précision} + Rappel}$$

Avec : NR : Nombre d'objets (réponses) retournés à l'utilisateur, NRP : Nombre d'objets (réponses) pertinents retournées à l'utilisateur, NRRU: Nombre d'objets pertinents.
A l'issue de ces expérimentations, nous remarquons que les résultats de l'indexation d'informations pédagogiques sont étroitement liés aux résultats de l'annotation (cf. Figure 5). La valeur de "F-Mesure" de l'extraction évolue avec la valeur de "F-Mesure" de l'annotation. Ceci s'explique par le fait, que l'extraction est effectuée à partir d'objets pédagogiques annotés et indexés. La qualité de la recherche s'améliore en améliorant celle de l'annotation. Cette dernière est elle-même dépendante de la qualité de segmentation comme nous l'avons déjà mentionné.

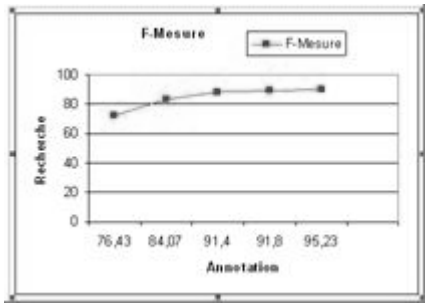


Figure 5 : Evolution des résultats de la recherche par rapport à celles de l'annotation

5.3. Troisième étape : Classement des objets pertinents

Après une extraction des objets pédagogiques, nous classons ces objets selon leur similarité avec la classe Cuser. Suite à plusieurs expérimentations, nous avons fixé la valeur du seuil θ :
0.1 pour les types "Cours" et "Définition",
0.3 pour les types "Plan" et "Exemple",
0.45 pour les types "Caractéristique" et "Exercice".
Notons que d'un côté, diminuer la valeur de θ réduit l'ensemble des objets pertinents retournés à l'utilisateur. D'un autre côté, augmenter la valeur de θ amène à une sélection des objets non pertinents.
Nous avons assigné chaque objet à l'une de ces trois catégories : A (objets classés comme pertinents), B (objets classés correctement comme pertinents), C (objets pertinents). Les valeurs de précision, de rappel et de F-Mesure sont calculées pour chaque type d'objet pédagogique comme suit :

$$\text{Précision} = \frac{B}{A} \qquad Rappel = \frac{B}{C} \qquad F - Mesure = 2 * \frac{\text{Précision} * Rappel}{\text{Précision} + Rappel}$$

Nous illustrons ces valeurs relatives à chacun des types d'objets dans la Figure 5.

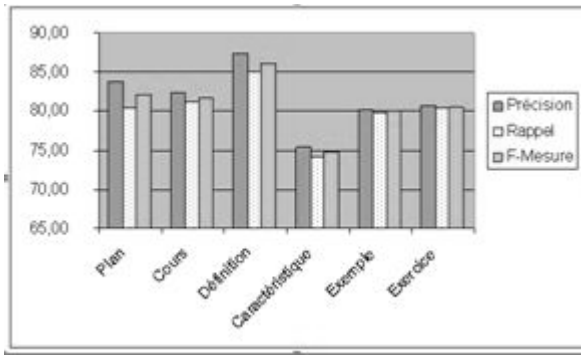


Figure 6 : Précision, Rappel et F-Mesure de l'étape de classement des objets

La figure ci-dessus présente, pour chaque type d'objet (représenté sur l'axe des abscisses), sa valeur de précision représentée en bleu, sa valeur de rappel en pointillé et sa valeur de F-Mesure représentée en rayures. Nous constatons que les valeurs de précision sont comprises entre 75% et 87% et que celles du rappel entre 74% et 85%. Notons que l'étape de classement ne dépend pas strictement de celles de l'annotation et d'appariement mais plutôt d'autres paramètres comme le corpus d'apprentissage, le choix des termes index, etc.

6. Conclusion et Perspectives

Dans cet article, nous avons proposé une approche d'indexation d'objets pédagogiques basée sur une annotation sémantique du texte par exploration contextuelle en vue d'une extraction des objets pédagogiques pertinents. Actuellement, notre travail présente un intérêt important dans plusieurs domaines d'application comme l'apprentissage en ligne, l'enseignement à distance (e-learning), l'éducation, etc. Pour évaluer notre approche, nous avons développé le système SRIDoP qui comprend les modules d'annotation, d'indexation, et de classement des objets selon leur pertinence. Nous remarquons, à travers les résultats d'évaluation, que notre approche permet d'avoir accès aux connaissances qui sont exprimées dans les textes selon un type donné, et de ramener des énoncés qu'un système de recherche d'informations classique ne parvient à capter par son approche d'indexation par mots clés.
L'un des travaux futurs que nous envisageons est l'extension de la carte sémantique des types d'objets pédagogiques par d'autres types comme Méthode, Auteur, Date, etc. Nous pensons aussi à la proposition d'une

fonction score qui fusionne les résultats des deux modules d'annotation et de classement en vue d'améliorer la pertinence des résultas.

Bibliographie

BUCKLEY C., SALTON G., ALLAN J. (1994). The effect of adding relevance information in a relevance feedback environment. Actes de International ACM SIGIR Conference, 292-300.

BUFFA M., DEHORS S., FARON-ZUCKER C., SANDER P. (2005). Vers une approche Web Sémantique dans la conception d'un système d'apprentissage. Revue du projet TRIAL SOLUTION, AFIA.

DEHORS S., FARON-ZUCKER C., STROMBONI J.P., GIBOIN A. (2005). Des annotations Sémantiques pour apprendre : l'Expérimentation QBLS. WebLearn.

DESCLES J.P. (1997). Système d'exploration Contextuelle. Co-texte et calcul du sens, Caen, 215-232.

DJIOUA B., FLORES J.G, BLAIS A., DESCLES J.P., GUIBERT G., JACKIEWIEZ A., LE PRIOL F., NAIT Baha L., SAUZAY B. (2006) Excom: an automatic annotation engine for semantic information. Dans Proc. FLAIRS, AAAI Press, Florida, 285-290.

ELKHLIFI A., Faiz R. (2009). Automatic Annotation Approach of Events in News Articles. International Journal of Computing & Information Sciences, 19-28.

ELKHLIFI A., Faiz, R. (2010). French-Written Event Extraction Based on Contextual Exploration. Dans Proc. FLAIRS, AAAI Press, Florida.

FLORY L. (2004). Les caractéristiques d'une ressource pédagogique et les besoins d'indexation qui en résultent. Journée d'étude sur l'Indexation des ressources pédagogiques numériques, Ennsib, Villeurbanne.

GREENWOOD M.A., SAGGION H. (2004). A Pattern Based Approach to Answering Factoid, List and Definition Questions. Dans Proc. RIAO 2004, Avignon, France.

HASSAN S., MIHALCEA R. (2009). Learning to identify educational materials. Dans Proc. RANLP, Bulgaria.

ITTNER D.J., Lewis D.D., Ahn D. D. (1995). Text categorization of low quality images. Actes de SDAIR, Las Vegas, US, 301-315.

MOURAD G. (2002). La segmentation de textes par Explration Contextuelle automatiques, présentation du module SegATex. Dans Inscription Spatiale du Langage : structure et processus ISLsp, Toulouse.

ROCCHIO J. (1971). Relevance feedback information retrieval. In Gerard Salton editor, The Smart retrieval system experiments in automatic document processing, Prentice-Hall, Englewood Cliffs, NJ, 313-323.

SALTON G. (1991). Developments in automatic text retrieval. Science, 253 (5023), 974-980.

SMEI H., BEN HAMADOU A. (2005). Un système à base de métadonnées pour la création d'un cache communautaire-Cas de la communauté pédagogique. Dans Proc. IEBC, Hammamet, Tunisie.

SMINE B., FAIZ R., DESCLES J.P. (2010). Analyse de documents pédagogiques en vue de leur annotation. Revue des Nouvelles Technologies de l'Information (RNTI), E-19, Ed. Cépaduès, 429-434.
THOMPSON C., SMARR J., NGUYEN H., MANNING C. (2003). Finding educational resources on the web : Exploiting automatic extraction of metadata. Proc. ECML, Workshop on Adaptive Text Extraction and Mining.
WESTERHOUT E., MONACHESI P. (2008). Creating glossaries using pattern-based and machine learning techniques. Dans Proceedings of Map of Language Resources, Technologies and Evaluation.

Indexation sémantique de documents textuels

Fatiha BOUBEKEUR
Université Mouloud Mammeri, Algérie

Wassila AZZOUG
Université M'hamed Bouguerra, Algérie

Sarah CHIOUT
Université Mouloud Mammeri, Algérie

Mohand BOUGHANEM
IRIT-SIG, Université Paul Sabatier de Toulouse

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 trois aspects: nous proposons (1) une approche d'identification des concepts en utilisant la base lexicographique WordNet, (2) une approche de désambiguïsation à deux niveaux, basée sur l'utilisation conjointe de WordNetDomains et de WordNet, et (3) une approche de pondération des concepts basée sur une nouvelle notion d'importance.

Mots-clés : Recherche d'information, indexation sémantique, désambiguïsation, WordNet, WordNetDomains.

Abstract : This paper describes a document semantic indexing approach. We propose to use WordNet as linguistic resource for retrieving the representative concepts of a document. Our contribution in this paper is threefold: we propose (1) an approach for identifying concepts using WordNet lexical database, (2) a disambiguation approach based on the joint use of WordNet and WordNetDomains, and (3) a concept weighting approach based on a novel definition of concept importance.

Keywords : Information retrieval, semantic indexing, disambiguation, WordNet, WordNetDomains.

Introduction et problématique

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. Une étape clé dans ce processus de RI, est l'indexation. L'indexation consiste à représenter requêtes et documents par un ensemble de termes (généralement des mots simples) pondérés, sensés définir 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. De la qualité de l'indexation dépend en grande partie la qualité de la recherche.
Un facteur clé impactant la qualité de l'indexation concerne la capacité du système à traiter avec l'ambiguïté de la langue naturelle et à comprendre les sens des mots dans les documents. Il ne s'agit plus alors de représenter le document par de simples chaines de caractères (entités lexicales), mais bien par des entités véhiculant des sens (entités sémantiques): les concepts. L'indexation sémantique, se base sur les concepts plutôt que sur les mots pour indexer les documents. Pour ce faire, les approches d'indexation sémantiques se basent globalement sur trois étapes : (1) une première étape d'identification des termes à l'issue de laquelle les mots (simples ou composés) contenus dans le document sont identifiés. Cette étape se base sur des techniques linguistiques classiques (tokénisation, lemmatisation, élimination de mots vides) et sur quelques techniques plus avancées d'identification des collocations de mots, (2) une seconde étape de désambiguïsation des sens des mots qui a pour objet de retrouver le sens correct d'un mot dans un contexte donné. Pour ce faire, les approches de désambiguïsation s'appuient sur des ressources linguistiques telles que les corpus d'apprentissage [5], [8], [12], dictionnaires automatisés [6], [9], [18], [20], ou encore les ontologies [14], [16], et autres Wikipédia [11], qui constituent des sources d'évidence pour les définitions et sens d'un mot. Le principe de la désambiguïsation consiste en général à associer un score de désambiguïsation aux différents sens possibles d'un mot (fournis par les dictionnaires et autres ressources …). La précision de la désambiguïsation dépend non seulement de la ressource linguistique utilisée mais aussi en grande partie du score de désambiguïsation établi. (3) Dans la troisième étape, il s'agit de pondérer les concepts identifiés à l'issue de l'étape précédente. La pondération a pour objet d'associer à chaque concept un poids numérique représentant son degré 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.
Ce papier présente la formalisation d'une approche d'indexation sémantique de documents. Dans cette approche, nous proposons d'utiliser WordNet [13] et son extension WordNetDomains comme

source d'évidence pour l'identification des sens des mots et pour leur pondération. Les mots sont alors désambiguïsés par rapport à leurs domaines associés dans WordNetDomains. La pondération d'un concept s'appuie sur une notion revisitée de l'importance d'un concept.
Le papier est structuré comme suit : Après une introduction, nous présentons en section 1 une synthèse des travaux dans le domaine, puis nous situons notre contribution. En section 2, nous donnons quelques notions préliminaires sur WordNet et WordNetDomains, puis des définitions utilisées dans la suite du papier. Notre approche d'indexation sémantique est détaillée en section 3. La section 4 présente une illustration. La section 5 conclut le papier.

1. Etat des lieux de 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 : l'identification des concepts et leur pondération. Le processus clé dans l'étape d'identification des concepts concerne la désambiguïsation des sens des mots. De nombreuses approches existantes se basent sur WordNet comme source d'évidence pour la désambiguïsation. C'est ainsi que pour désambiguïser un mot ambigu, Voorhees [19] classe chaque synset (sens correspondant à une entrée de WordNet) 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. Les concepts sont ensuite pondérés en utilisant un schéma de pondération classique *tf*idf* normalisé. Dans une approche différente, Katz et al [17] 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. Ce principe est repris dans l'approche d'indexation de Baziz et al. [1]. Les auteurs considèrent ainsi que parmi les différents sens possibles (concepts candidats) d'un terme donné, le plus adéquat est celui qui a le plus de liens sémantiques [9], [10], [15] 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. Les concepts sont ensuite pondérés sur la base d'un schéma de pondération dit *Cf*idf*, qui étend la pondération *tf*idf* pour tenir compte des termes composés. L'approche de Baziz et al. a été reprise dans Boughanem et al. [4], avec une nouvelle définition de la pondération. En effet, dans [4], les auteurs introduisent les notions de centralité et de spécificité d'un concept. La centralité définit le nombre de relations de ce concept avec les autres concepts du document. Sa spécificité définit son degré de « spécialité ». Le schéma de pondération utilisé est basé sur la combinaison de ces deux mesures. Dans notre approche d'indexation sémantique proposée dans [2], [3], le choix du concept correct dans un contexte s'appuie sur un score basé sur la somme des valeurs de similarité que le concept cible a avec les concepts les plus fréquents dans le document. Les concepts sont alors pondérés sur la base d'une mesure de leur importance dans le document, quantifiée au travers de leur proximité sémantique aux autres concepts du document. Dans une approche plus récente [7] les auteurs proposent une approche intéressante de désambiguïsation à deux niveaux : d'abord retrouver le domaine correct d'un mot dans le document, puis désambiguïser ce mot dans le domaine ainsi identifié. Le domaine correct d'un mot est celui qui maximise ses occurrences dans le contexte local du mot cible. Les auteurs utilisent WordNetDomains, qui permet de classifier les différentes entrées de WordNet dans des domaines prédéfinis.

Positionnement de notre proposition

Notre approche proposée dans ce papier tente de combiner notre approche d'indexation conceptuelle dans [2], [3] et l'approche de désambiguïsation par les domaines au sein d'un paradigme unifié. L'objectif est de représenter de manière précise le document par un noyau sémantique composé de concepts pondérés. Dans notre proposition, les termes d'indexation sont d'abord extraits en se basant sur des étapes d'indexation classiques. Cette étape inclut en outre une nouvelle proposition pour la détection des collocations à partir d'une liste pré-établie des collocations de WordNet; A l'issue de cette étape, trois listes sont construites : la liste des collocations, la liste des mots simples ayant des entrées correspondantes dans WordNet, et la liste des mots simples n'ayant pas d'entrée dans WordNet (ces mots seront dits des mots orphelins). Puis chaque mot non vide identifié dans WordNet est désambiguïsé dans son contexte global dans le document. La désambiguïsation d'un mot se base d'abord sur sa désambiguïsation de

domaine (ie. trouver le domaine correcte du mot dans le document), puis sa désambiguïsation sémantique dans le domaine choisi.
- La désambiguïsation du domaine se base sur une mesure de probabilité d'appartenance du mot cible et des mots du contexte à un domaine donné. Le domaine qui maximise cette mesure est retenu comme domaine correct du terme dans le document.
- La désambiguïsation des concepts est basée sur l'intuition que les concepts appartenant à un même domaine sont fortement (sémantiquement) liés et doivent renforcer la désambiguïsation sémantique du concept cible par rapport à des concepts qui ne sont pas dans le même domaine que lui.
L'approche de pondération des concepts est basée sur une nouvelle définition de la notion de centralité d'un concept. La centralité dépend de la pertinence d'un concept dans le document et sur sa fréquence. Ces notions seront définies en section suivante.
Dans ce qui suit, nous décrivons les différentes étapes de notre approche d'indexation sémantique. Mais d'abord, nous présentons quelques notions préliminaires sur WordNet, ainsi que des définitions et notations utilisées dans la suit de notre papier.

2. Préliminaires

2.1. WordNet

WordNet est un réseau lexical électronique qui couvre la majorité des noms, verbes, adjectifs et adverbes de la langue Anglaise, qu'il 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.
- Un terme peut être un mot simple ou une collocation de mots
* 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 nom « bank » est donné en Table 1.

The noun bank has 10 senses (first 9 from tagged texts)

1. (883) depository financial institution, bank, banking concern, banking company -- (a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home")
2. (99) bank -- (sloping land (especially the slope beside a body of water); "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents")
3. (76) bank -- (a supply or stock held in reserve for future use (especially in emergencies))
4. (54) bank, bank building -- (a building in which the business of banking transacted; "the bank is on the corner of Nassau and Witherspoon")
5. (7) bank -- (an arrangement of similar objects in a row or in tiers; "he operated a bank of switches")
6. (6) savings bank, coin bank, money box, bank -- (a container (usually with a slot in the top) for keeping money at home; "the coin bank was empty")
7. (3) bank -- (a long ridge or pile; "a huge bank of earth")
8. (1) bank -- (the funds held by a gambling house or the dealer in some gambling games; "he tried to break the bank at Monte Carlo")
9. (1) bank, cant, camber -- (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
10. bank -- (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning); "the plane went into a steep bank")

Table 1 : Les concepts de WordNet correspondants au mot bank

WordNetDomains est une extension de WordNet dans laquelle les synsets ont été étiquetés par des labels de domaines. Un exemple de domaines associés au synsets du mot *bank* est donné en Table 2.
Ces domaines sont organisés selon une hiérarchie définissant la relation de spécialisation/généralisation entre les domaines. Par exemple, le domaine *Tennis* est plus spécifique que le domaine *Sport*, et le domaine *Architecture* est plus général que le domaine *Buildings*. Une partie de la hiérarchie de WordNetDomains est donnée en Table3. Le domaine *Top-Level* représente la racine de cette hiérarchie. Le domaine *Factotum* de WordNetDomains est un domaine fonctionnel (par opposition à sémantique) qui regroupe tous les sens des mots qui n'appartiennent à aucun domaine particulier mais qui peuvent apparaître avec des termes associés à d'autres domaines. *Factotum* constitue un domaine particulier, indépendant du domaine *Top-Level* et de sa hiérarchie.

Les senses	Domaines
1. (883) depository financial institution, bank, banking concern, banking company -- (a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home")	ECONOMY,
2. (99) bank -- (sloping land (especially the slope beside a body of water); "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents")	GEOGRAPHY, GEOLOGY
3. (76) bank -- (a supply or stock held in reserve for future use (especially in emergencies))	ECONOMY
4. (54) bank, bank building -- (a building in which the business of banking transacted; "the bank is on the corner of Nassau and Witherspoon")	FACTOTUM, ECONOMY
5. (7) bank -- (an arrangement of similar objects in a row or in tiers; "he operated a bank of switches")	FACTOTUM
6. (6) savings bank, coin bank, money box, bank -- (a container (usually with a slot in the top) for keeping money at home; "the coin bank was empty")	ECONOMY
7. (3) bank -- (a long ridge or pile; "a huge bank of earth")	GEOGRAPHY, GEOLOGY
8. (1) bank -- (the funds held by a gambling house or the dealer in some gambling games; "he tried to break the bank at Monte Carlo")	ECONOMY, PLAY
9. (1) bank, cant, camber -- (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)	ARCHITECT-URE
10. bank -- (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning); "the plane went into a steep bank")	TRANSPORT

Table 2 : Les domaines associés dans WordNetDomains, aux synsets du mot bank

2.2. Définitions et notations

Soit m_i un mot d'un texte à analyser.

a. On appelle occurrence de m_i, toute instance de m_i dans le texte.

b. Une instance de m_i apparaît dans une seule phrase. L'ensemble des mots de cette phrase constitue son contexte local. On note le contexte local de l'instance m_i par ζ_i

c. On appelle contexte local droit de l'instance m_i, l'ensemble des mots à droite de m_i jusqu'à la prochaine ponctuation. Le contexte local droit de l'instance m_i sera noté : ζ_{di}

d. On appelle expression locale de l'instance m_i, la chaîne de caractères concaténant, par le biais du souligné (_), le mot m_i avec les mots de contexte local droit successivement. La taille d'une expression locale est le nombre de mots qui la composent.

e. On appelle contexte global du mot m_i, l'union de tous ses contextes locaux dans le texte du document. Le contexte global du mot sera noté par : ζ_{Gi}

Top_level	Humanities	History		
		Linguistics	Grammar	
		Literature	Philology	
		Philosophy	Psychoanalysis	
		Art	Music	
			Plastic_Arts	Jewellery
				Sculpture
			Theatre	
			Cinema	
		Paranormal		
	Free_Time	Play		
		Sport	Tennis	
		...		
	Applied_Science	Agriculture		
		Architecture	Buildings	
	Pure_Science			
		Biology	Anatomy	
		Animals		
		Earth	Geology	
			Geography	
		Mathematics		
		Physics	Acoustics	
		...		
	Social_Science	Economy	Finance	Money
		Politics		
		Fashion		
		Military		
Factotum	Quality			
	Number			
	...			

Table 3 : Extrait de la hiérarchie de WordNetDomains

3. Indexation sémantique des documents

L'indexation sémantique vise à représenter un document par un ensemble 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 :

1. l'ensemble ξ_{Expres} des expressions du document, correspondant aux collocations de WordNet.

2. l'ensemble $\xi_{Simples}$ des mots simples ayant une entrée dans WordNet,

3. l'ensemble ξ_{Orphel} des mots orphelins (mots simples n'ayant pas d'entrée dans WordNet).

Cette étape débute par l'identification des expressions. Pour cela, nous avons d'abord construit la liste φ_{Coloc} de toutes les collocations

existantes dans WordNet. Puis, pour une occurrence de mot à analyser, on extrait de φ_{Coloc} l'ensemble ς_i de toutes les collocations qui commencent par m_i. On ordonne ς_i par tailles décroissantes de ses éléments, puis on projette chaque élément de ς_i sur des expressions locales E_i de m_i. Si une expression locale s'apparie avec une collocation, elle est retenue comme expression et insérée dans l'ensemble ξ_{Expres}. Si aucune collocation de ς_i ne s'apparie avec une expression locale de m_i, alors m_i est un mot simple. Si m_i possède une entrée dans WordNet, il sera inséré dans l'ensemble $\xi_{Simples}$; Sinon il sera mis dans l'ensemble des orphelins ξ_{Orphel}.

Le principe de l'identification des termes est décrit à travers l'algorithme de la Table 4.

Algorithme de détection des termes
Entrée : document d.
Sortie : ξ_{Expres}, $\xi_{Simples}$, ξ_{Orphel}
Procédure : Soit m_i le prochain mot à analyser dans d.
Début
1. Calculer $\varsigma_i = \{C^i_1, C^i_2, \ldots C^i_n\}$ l'ensemble des collocations commençant par le mot m_i
2. Ordonner ς_i comme suit : $\varsigma_i = \{C^i_{(1)}, C^i_{(2)}, \ldots, C^i_{(n)}\}$ où $(j)_{1..n}$ est une permutation d'indices telle que $|C^i_{(1)}| \geq |C^i_{(2)}| \geq \ldots \geq |C^i_{(n)}|$, où $|C^i_{(j)}|$ est la taille de la collocation $C^i_{(j)}$
3. Bool<- faux ;
4. Pour chaque $C^i_{(j)}$ dans ς_i et bool=faux, faire :
5. Calculer l'expression locale E_i de taille $|C^i_{(j)}|$
6. Si $E_i = C_{(j)}$ alors bool <- vrai ; fait ;
7. Si (bool) mettre E_i dans ξ_{Expres}
   Sinon Si m_i mot non vide alors
   Début Si m_i possède une entrée dans WordNet alors mettre m_i dans $\xi_{Simples}$
   Sinon mettre m_i dans ξ_{Orphel} . fin ;

Fin.

Table 4: Algorithme de détection des termes d'index

3.2. Désambiguïsation des termes

Les collocations étant des expressions quasiment désambiguïsées, l'étape d'indexation concernera uniquement les mots simples ayant des entrées dans WordNet, soit donc l'ensemble des termes de $\xi_{Simples}$.

Chaque terme de $\xi_{Simples}$ peut avoir plusieurs sens possibles. Le but de cette étape est de sélectionner le meilleur sens du terme dans le document. L'approche de désambiguïsation proposée est une approche à trois niveaux :
(1) dans le premier niveau, il s'agit de déterminer la forme grammaticale (nom, verbe, …) du mot m_i dans le document, en utilisant le Stanford POS Tagger.
(2) le second niveau, permet d'identifier le domaine d'usage du mot dans le document. L'identification des domaines s'appuie sur l'utilisation de *WordNetDomains*. Ce niveau de désambiguïsation permettra de limiter le nombre de sens du terme qui seront examinés dans le niveau suivant de désambiguïsation. (3) le troisième niveau de désambiguïsation consiste alors à sélectionner parmi les sens possibles du terme dans le domaine sélectionné, celui qui est sensé le définir au mieux dans le document.

3.2.1. Identification de la forme grammaticale des mots simples
Les sens d'un mot m_i dans WordNet sont classés selon ses différentes catégories grammaticales possibles. Ainsi, nous utilisons le Stanford POS Tagger pour identifier la catégorie grammaticale du mot m_i dans le document afin de déterminer les sens appartenant à cette forme grammaticale. Cette étape permet de limiter le nombre de sens du terme qui seront utilisés dans la désambiguïsation et de récupérer les domaines qui correspondent à ses sens dans WordNetDomains.

3.2.2. Désambiguïsation au niveau des domaines
Chaque mot dans $\xi_{Simples}$ possède plusieurs sens dans WordNet. Les sens de WordNet sont étiquetés dans *WordNetDomains* par des labels de domaines. Ainsi, un sens peut appartenir à un ou plusieurs domaines.
On note :

S_i l'ensemble de tous les synsets associés au mot m_i,

D l'ensemble, non redondant, de tous les domaines associés aux éléments de S_i,

$S_{i(j)}$ est l'ensemble des synsets de S_i appartenant au domaine D_j,

$S_{i(j)}[k]$ le kième élément de l'ensemble $S_{i(j)}$

Partant de l'hypothèse que le domaine probable d'un mot est celui qui maximise sa similarité avec les autres domaines des autres mots du même contexte, nous attribuons à chaque domaine D_j associé à un sens du mot m_i, un score basé sur la somme de ses similarités avec les différents

domaines associés aux sens des autres termes t_k ($\left(t_k \in \left\{ \xi_{Simples} \cup \xi_{Expres} \right\} \right)$) d'index. Le domaine D_j ayant le plus grand score est sélectionné comme domaine adéquat pour le mot m_i dans le document.

Formellement : $Score(D_j) = \arg\max_j \left(\sum_{t_k \in \xi_{Gi}} \sum_{k \in [1..n]} Sim(Dj, D_k) \right)$

Où :

$Sim(Dj, D_k)$ désigne la similarité entre les domaines D_j et D_k.

Pour mesurer la similarité entre les domaines D_j et D_k, nous utilisons et adaptons la formule de Wu-Palmer [21] à la hiérarchie Top-Level de WordNetDomains, ce qui donne :

$$Sim(D_j, D_k) = \frac{2 * prof\ ondeur(D^*)}{prof\ ondeur(D_j) + prof\ ondeur(D_k)}$$

Où :

D^* : est le domaine le plus spécifique qui subsume D_j et D_k dans la hiérarchie de WordNetDomains.

$prof\ ondeur(D^*)$: est le nombre d'arcs entre la racine de WordNetDomains et le domaine D^*.

$prof\ ondeur(D_j)$: est le nombre d'arcs entre la racine de WordNetDomains et le domaine D_j en passant par le domaine D^*.

Remarque :
La formule de similarité est appliquée aux seuls domaines de la hiérarchie Top-Level. Le domaine factotum, indépendant de cette hiérarchie est un domaine fonctionnel (non sémantiquement informatif). Il ne sera pas considéré dans cette désambiguïsation.

3.2.3. Désambiguïsation des sens des mots

A l'issue de l'étape précédente, tout mot m_i de $\xi_{Simples}$ est associé à un seul domaine D_j dans le document. Deux cas peuvent se présenter :

- soit m_i possède un seul sens dans D_j , dans ce cas il est désambiguïsé.

- soit m_i possède plusieurs sens dans D_j, dans ce cas il est ambigu. Il faut le désambiguïser.
Nous proposons une désambiguïsation sur les seuls sens appartenant à D_j, soit donc aux seuls éléments $S_{i(j)}[k]$ de l'ensemble $S_{i(j)}$. L'objectif est alors de sélectionner parmi ces sens le sens correct pour le mot m_i dans le document.

Pour désambiguïser le mot m_i dans son domaine, on associe à chacun de ses sens $S_{i(j)}[k]$ de l'ensemble $S_{i(j)}$, un score basé sur sa proximité sémantique avec les autres sens associés aux mots de son contexte global dans leurs domaines respectifs. Le concept $S_{i(j)}[k]$ ayant le plus grand score est alors retenu comme sens adéquat pour le mot m_i dans *d*.
Formellement :

$$S_{i(j)}[k] = Arg\max\left(\sum_{\substack{l / m_l \in \subset_{G_i} \\ l \neq i}} \sum_{1 \leq n \leq \left|S_{l(m)}\right|} sim\left(S_{i(j)}[k], S_{l(m)}[n]\right)\right)$$

Où $sim\left(S_{i(j)}[k], S_{l(m)}[n]\right)$ est la similarité sémantique entre les concepts $S_{i(j)}[k]$ et $S_{l(m)}[n]$.
L'ensemble des concepts retenus constituera le noyau sémantique *N(d)* du document *d*. Pour des raisons de simplification, on utilisera la notation C^i pour désigner le ième élément de *N(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 fréquent et pertinent dans ce document, nous proposons de pondérer un concept avec un poids basé sur :
Sa pertinence, que nous définissons sa proximité sémantique aux autres concepts du document,
Sa fréquence dans le document.
Formellement, le poids $W\left(C^i\right)$ du concept C^i est défini par :

$$W\left(C^i\right) = \alpha * tf\left(C^i\right) + (1-\alpha)\sum_{i \neq l} Dist\left(C^i, C^l\right)$$

Où α est un facteur de pondération qui permet de balancer la fréquence par rapport à la pertinence. Ce facteur pourra être fixé expérimentalement.
Le schéma de pondération que nous proposons permet outre la pondération des concepts, la pondération des collocations et des termes

orphelins. Dans ce dernier cas, seule la fréquence est considérée, les proximités sémantiques inexistantes, sont initialisées à zéro.
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.

4. Illustration

Dans le paragraphe suivant, nous montrons la faisabilité de notre approche d'indexation sémantique en l'appliquant sur un exemple. Nous focalisons en particulier sur la désambiguïsation puisque de sa précision dépend en grande partie la précision de l'indexation.
Etant donné le texte suivant (extrait du document *Arthroskopie.00130003.eng.abstr* de la collection Muchmore[1])
"The posterior cruciate ligament (PCL) is the strongest ligament of the human knee joint. Its origin is at the lateral wall of the medial femoral condyle and the insertion is located in the posterior part of the intercondylar area. The posterior cruciate ligament consists of multiple small fiber bundles."

4.1. Détection des concepts

Notre algorithme de détection des concepts retourne les trois ensembles : ξ_{Expres}, $\xi_{Simples}$, ξ_{Orphel} suivants :

$$\xi_{Expres} = \{\text{human\_knee}, \text{knee\_joint}, \text{fiber\_bundle}\}$$

$$\xi_{Orphel} = \{\text{PCL}, \text{intercondylar}\}$$

$$\xi_{Simples} = \begin{cases}\text{posterior}, \text{cruciate}, \text{ligament}, \text{strong}, \text{origin}, \text{lateral}, \text{wall}, \text{medial}, \text{fem}\\ \text{condyle}, \text{insertion}, \text{locate}, \text{part}, \text{area}, \text{consist}, \text{multiple}, \text{small}\end{cases}$$

4.2. Désambiguïsation des termes simples

4.2.1. Identification de la forme grammaticale
En utilisant le Stanford POS Tagger, on obtient la forme grammaticale de chaque mot simple dans le document comme suit :

$$\xi_{Simples} = \begin{cases}\text{posterior/JJ}, \text{cruciate/NN}, \text{ligament/NN}, \text{strong/JJ}, \text{origin/NN}, \text{lateral/JJ}, \text{wall/NN}, \text{medial/JJ}, \text{femoral/JJ}\\ \text{condyle/NN}, \text{insertion/NN}, \text{locate/VB}, \text{part/NN}, \text{area/NN}, \text{consist/VB}, \text{multiple/JJ}, \text{small/JJ}\end{cases}$$

4.2.2. Désambiguïsation au niveau des domaines
Nous retrouvons pour chaque terme d'index l'ensemble des domaines associés à ses différents sens. Puis nous désambiguïsons au niveau des

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

domaines. La désambiguïsation au niveau des domaines permet d'associer les domaines adéquats aux termes d'index. Les numéros des sens et les domaines associés aux termes d'index sont présentés dans les tableaux de la figure 1 suivante. Les résultats de la désambiguïsation des domaines sont récapitulés dans les tableaux de la figure 2.

Posterior	00136931 : animals		04369872 : buildings		02574255 : factotum
cruciate	02290409 : factotum		03899176 : military	Consist	02670036 : factotum
ligament	04990644 : anatomy		08876684 : factotum		02578919 :factotum
	03528343 : factotum		04370607 :factotum		02554853 : factotum
strong	02239657 : quality	Wall	05284169 :anatomy	multiple	02140712 : factotum
	02238035 : factotum		08876934 :geology	Small	01343705 :factotum
	01766728 : factotum		04370836 :factotum		01366545 : factotum
	01764165 : quality		137331862 :factotum		02259250 : factotum
	02436955 : factotum		13731862 : factotum		01597253 : factotum
	02193170 : factotum	Femoral	02614670 : anatomy		01419050 : factotum
	01901958 : Grammar	Human knee	05254826 :anatomy		01481659 : factotum
	01118836 : factotum	Insertion	06308340 :factotum		01406192 : acoustics
	01023736 : factotum		00306609 :factotum		01503651 : factotum
	00807924 : factotum	Locate	02220173 : factotum		00845424 : factotum
Origin	07989929 :factotum		02613751 :factotum		01416987 : factotum
	04675241 :biology		02266215 :factotum		02155033 : factotum
	06874664 : factotum		00401714 :politics	condyle	05157880 : anatomy
	05654262: mathematics	area	07980485 :geography	Knee joint	05254826 : anatomy
	07610417 : factotum		13687487 :factotum	Fiber bundle	05161310 : anatomy
lateral	02353094 : factotum		05646624 : factotum		
	00746517 :factotum		02641332 : factotum		
medial	00746822 :factotum		04921220 :anatomy		
	00326753 :factotum		04842376 : factotum		

Figure 1 *: Sens et domaines associés aux termes d'index*

4.2.3. Désambiguïsation des sens des mots

A l'issue de l'étape précédente, chaque terme d'index est associé aux seuls sens liés au domaine sélectionné. Seuls ces sens sont désambiguïsés. Pour le calcul du score de désambiguïsation des sens des mots, nous nous basons sur la mesure de similarité de Lesk. Dans les tableaux de la figure 2 suivante, nous représentons pour chaque terme d'index, son domaine sélectionné ainsi que ses sens associés dans ce domaine. Le sens grisé dans le tableau représente le sens désambiguïsé (sens adéquat du terme dans le document). Un examen rapide des résultats, appuyé par une désambiguïsation manuelle, nous permet de voir que :

- La désambiguïsation au niveau des domaines a permis d'associer les domaines adéquats aux termes d'index. C'est ainsi par exemple que les termes d'index wall, area et ligament se sont vus assigner le domaine anatomy qui est le domaine le plus probable du texte indexé.
- La désambiguïsation au niveau des sens donne aussi des sens corrects dans le document. Pour vérifier cela, nous avons comparé nos résultats avec ceux obtenus par l'approche de Baziz et al. [1] pour le même texte. Les résultats obtenus montrent que nous retrouvons plus de sens

corrects que dans [1]. A titre d'exemple, le mot wall est désambiguïsé par wall#a#1 (définit dans WordNet par : an architectural partition …) dans [1], alors que notre approche nous retourne wall#a#5 (défini par : (anatomy) a layer (a lining or membrane)…) plus proche sémantiquement de la thématique (médicale) du document. Les résultats de notre approche sont encourageants mais doivent néanmoins être vérifiés sur une collection de taille réelle.

Mot simple	Domaine désambiguïsé	Les sens du mot dans le domaine désambiguïsé	Mot simple	Domaine désambiguïsé	Les sens du mot dans le domaine désambiguïsé
posterior	Animals	00136931 : posterior#a#1	locate	Factotum	02220173 : locate#v#1
cruciate	Factotum	022904409 : cruciate#a#1			02613751 : locate#v#2
ligament	Anatomy	04990644 : ligament#n#1			02266215 : locate#v#3
Strong	Factotum Grammar Quality	02238035 : strong#a#1			00401714 : locate#v#4
		02238035: strong#a#2	part	Factotum Theatre	13028617 : part#n#1
		01462766: strong#a#3			08103697 : part#n#2
		01766728: strong#a#4			05344775 : part#n#3
		01764165: strong#a#5			03746082 : part#n#4
		02436955: strong#a#6			05526011 : part#n#5
		02193170 : strong#a#7			00678112 : part#n#6
		01901958: strong#a#8			08797461 : part#n#7
		01118836 : strong#a#9			12529266 : part#n#9
		01023736 : strong#a#10			00738005 : part#n#10
Origin	Factotum Biology	00807924 : strong#a#10	consist	Factotum	02574255 : consist#v#1
		07989929 : origin#n#1			02670036 : consist#v#2
		04675241: origin#n#2			02578919 : consist#v#3
		06874664 : origin#n#3			02554853 : consist#v#4
		07610417 : origin#a#5			013433705 : small#a#1
lateral	anatomy	02353094 : lateral#a#1	small	Factotum Acoustics	00746517 : lateral#a#2
wall	Factotum	05284169: wall#n#5			01366545: small#a#2
medial	Factotum	00746822 : medial#a#1			0229250 : small#a#3
femoral	Anatomy	00326753 : medial#a#2			01597253 : small#a#4
condyle	Anatomy	02614670 : femoral#a#1			01419050: small#a#5
insertion	Factotum	05157880 : condyle#1			01481659: small#a#6
		06308340 : insertion#n#1			01503651 : small#a#8
		00306609 : insertion#n#2			00845424: small#a#9
multiple	Factotum	02140712 : multiple#a#1			01416987 : small#a#10
area	Anatomy	04921220 : area#n#5			02155033 : small#a#11

Figure 2 : Présentation des domaines et des sens désambiguïsés

5. Conclusion

Nous avons présenté dans ce papier, les fondements théoriques d'une nouvelle approche d'indexation sémantique basée sur l'utilisation conjointe de WordNet et de WordNetDomains. Notre contribution porte sur les trois aspects de l'indexation sémantique : la détection des termes d'index, la désambiguïsation des termes et la pondération des concepts. En particulier, nous avons proposé une nouvelle approche de détection des concepts incluant la détection des collocations, une approche de désambiguïsation par les domaines et dans les domaines, et enfin un nouveau schéma de pondération des concepts. Nous avons montré la faisabilité de notre approche en la déroulant sur un exemple. Des travaux sont en cours en vue de sa validation expérimentale.

Bibliographie

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.

F. BOUBEKEUR, M. BOUGHANEM, L.TAMINE, M. DAOUD. De l'utilisation de WordNet pour l'indexation conceptuelle des documents. 13 ème Colloque International sur le Document Electronique.16-17 Décembre 2010, INHA, Paris.

F. BOUBEKEUR, M. BOUGHANEM, L.TAMINE, M. DAOUD. Using WordNet for Concept-based document indexing in information retrieval. Dans: Fourth International Conference on Semantic Processing (SEMAPRO 2010), Florence, Italy, Octobre 2010.

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.

M. CUADROS, JM., ATSERIAS, J., M. CASTILLO, M., & G. RIGAU, (2004). Automatic acquisition of sense examples using exretriever. In IBERAMIA Workshop on Lexical Resources and The Web for Word Sense Disambiguation. Puebla, Mexico.

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.

S. G. KOLTE, S. G. BHIRUD. Word Sense Disambiguation using WordNetDomains. In First International Conference on Emerging Trends in Engineering and Technology. 2008 IEEE DOI 10.1109/ICETET.2008.231

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.

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.

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

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

R. MIHALCEA and D. MOLDOVAN. Semantic indexing using WordNet senses. In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October 2000

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

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).

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.

O. UZUNER, B. Katz, D. Yuret. Word Sense Disambiguation for Information Retrieval. AAAI/IAAI 1999 : 985

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.

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).

Y. WILKS & M. STEVENSON. Combining independent knowledge source for word sense disambiguation. Conference « Recent Advances in Natural Language Processing », 1–7.

Z. WU & M. PALMER. Verb semantics and Lexical selection. Proceedings of the 32th Annual Meetings of the Association for Computational Linguistics, pp. 133-138. 1994.

Partie 2 - Document interactif

Extension d'un algorithme de Diff & Merge au Merge Interactif

Xuan TRUONG VU
UMR-CNRS 6599, Heudiasyc, Université de Technologie de Compiègne, France

Pierre MORIZET-MAHOUDEAUX
UMR-CNRS 6599, Heudiasyc, Université de Technologie de Compiègne, France

Joost GEURTS
UMR-CNRS 6599, Heudiasyc, Université de Technologie de Compiègne, France

Stéphane CROZAT
Unité Ingénierie des Contenus et Savoirs, Université de Technologie de Compiègne, France

Résumé : De nombreuses recherches visant à gérer automatiquement la fusion de différentes versions de documents textuels structurés ont été menées et communément regroupées dans le théme "*Diff* & *Merge* de documents XML". Nous proposons dans cet article, une alternative appelée *merge interactif*. Cette approche consiste à ne pas appliquer systématiquement la fusion automatique mais à rendre la transformation séquentielle et interactive. L'objectif est de proposer à l'utilisateur une liste d'opérations associées au document original, qu'il/elle pourra confirmer ou non, selon l'objectif de la fusion et la détection de conflits et d'incohérences.

Mots-clés : Edition collaborative, documents structurés, documents fragmentés, *Diff* & *Merge*

Abstract : Numerous works to manage automatically the *merge* of the various versions of structured textual documents have been developed and correspond to the domain of "XML documents *Diff* & *Merge*". We propose in this paper, an alternative named *interactive merge*. This approach consists in not applying systematically the automatic merge but in keeping the transformation sequential and interactive. The objective is to propose to the user a list of operations associated with the original document, which he/she can confirm or not, according to the objective of the fusion and the detection of conflicts and inconsistencies.

1. Introduction

L'écriture collaborative est une forme d'écriture renouvelée par le numérique. Elle est devenue une pratique importante dans le monde académique, les organisations, les entreprises, et au sein des communautés en ligne. Aujourd'hui, de nombreux documents sont issus de travaux collaboratifs : journaux, manuels techniques, présentations, articles scientifiques, cours, etc. Dans un environnement collaboratif, chaque contributeur peut ajouter, modifier et supprimer des contenus.
Si cela facilite la réalisation de documents qui exploitent au mieux les compétences de chacun, cela impose une charge supplémentaire pour maintenir la cohérence de l'ensemble et coordonner les efforts individuels. La nature de la collaboration est très variée selon la dimension et la stratégie du groupe et différentes plateformes d'édition documentaire collaborative existent sur le marché pour y répondre, dont MediaWiki ou Google Docs sont des exemples populaires. Le travail présenté dans cet article s'inscrit dans le cadre du projet ANR C2M[2] dont l'objectif est de répondre aux besoins d'écriture collaborative dans le cas de documents structurés et fragmentés : On appelle document structuré un document dont la structure logique est décrite plutôt que la mise en forme physique (André et al. 1989) ; et document fragmenté un document composé par intégration de plusieurs fragments, chacun pouvant être mobilisé pour plusieurs usages au sein de différents documents (Crozat, 2007).
En mode collaboratif et fragmenté, chaque modification d'un fragment appelle des décisions délicates en terme de répercussion pour chaque utilisateur (auteurs, lecteurs, relecteurs, co-auteurs, etc.) et pour chaque document (version avancée, version simplifiée, version papier, version écran, version de relecture, version officielle, version adaptée, etc.). La question est alors de savoir quelles décisions prendre automatiquement et/ou comment aider les utilisateurs à les prendre. Une partie de la réponse porte sur la capacité à rendre intelligibles toutes les modifications aux utilisateurs. Dans cet article nous nous penchons sur les solutions permettant de comparer a posteriori deux (ou plusieurs) versions d'un même document pour en évaluer les proximités et différences.
Dans la seconde section, nous exposerons brièvement les différentes méthodes et outils existants de *differencing* et de *merging* pour les documents XML. Dans la section suivante, nous présenterons une nouvelle approche appelée "*merge* interactif" permettant à un utilisateur de visualiser, sans ambiguïté, les différences entre deux versions d'un document et de faire des choix pour procéder à leur fusion (par l'acceptation de certaines modifications et le refus d'autres). Nous

[2] http://scenari.utc.fr/c2m

donnerons en conclusion les premiers résultats obtenus, les fonctionnalités à étudier et implémenter et enfin, les perspectives associées à ce travail.

2. Méthodes et outils de différentiel de documents XML

Lorsqu'un document numérique textuel est partagé par plusieurs auteurs, il est nécessaire de pouvoir identifier les différences qui peuvent exister entre les sources pour pouvoir les synchroniser et fusionner correctement. Les travaux relatifs à ces problèmes forment le domaine du *diff* & *merge* pour lequel il existe actuellement de multiples solutions, libres et commerciales, pour divers cas d'usage, chacune apportant sa propre approche et donc ses propres propriétés et optimisations. Certaines d'entre elles sont orientées documents textuels (e.g le format MS Doc), d'autres orientées documents structurés (e.g XML). Nous donnons ci-dessous un bref panorama de ce domaine.

2.1. Edits history

Edits history est une technique consistant à capturer toutes les actions (*edit*) de l'utilisateur sur l'éditeur et les mémoriser dans un fichier appelé *edit log*. Chaque *edit log* est donc dupliqué et transféré à d'autres utilisateurs afin de le comparer avec leurs propres *edit logs*. Comparer des documents revient à comparer des *edit logs*. Pour synchroniser des documents, il suffit de rejouer sur un document les actions transférées depuis d'autres documents.
Cette méthode doit résoudre deux principaux de problèmes : premièrement, il faut capturer toutes les actions de l'utilisateur : deuxièmement, il faut s'assurer de la cohérence et la complétude du fichier. En effet, chaque *edit* va changer la position des *edits* dépendants. Un exemple d'un tel outil est Microsoft Office Groove.

2.2. Change detection

Au contraire de l'*Edit history*, *Change detection* ne requière aucune connaissance de l'histoire de l'édition du fichier. Elle cherche à déterminer, à partir des seuls fichiers courants, les changements qui ont été réalisés dans chacun d'eux. Différents algorithmes existent actuellement.

2.2.1 Les algorithmes Line oriented traitent tous les documents comme une série linéaire de lignes. UNIX diff[3] est un exemple typique et le plus connu.. Il cherche la séquence la plus longue de lignes communes entre deux fichiers. Les lignes uniques dans l'un des deux documents seront

[3] http://www.gnu.org/software/diffutils/diffutils.html

supprimées ou insérées pour passer d'un fichier à l'autre. Une variante de diff est diff3 implémentant le three-way merge, qui examine dynamiquement des mots et même des caractères au lieu de lignes (par exemple, google-diff-match-patch[4]). Ces outils sont très adaptés et efficaces pour traiter des documents textuels mais ils ne sont pas directement applicables à des documents structurés tel que XML ou XHTML, car diff n'est pas en mesure de distinguer les informations structurelles.

2.2.2 Tree oriented

Les méthodes "orientées arbre" (*tree oriented)* prennent en considération la structure d'arborescence du document. Les nœuds et les sous-arbres seront comparés et mis en correspondance à la place des lignes. Les nœuds et les sous-arbres qui ne se correspondent pas, forment les différences entre les documents.

On trouvera des études comparatives et détaillées des algorithmes différentiels orientés arbre dans (Cobéna et al. 2002 ; Coneba et al., 2002 ; La Fontaine, 2003 ; Marian et al. 2001 ; Peters, 2005 ; Rönnau, 2008 ; Wang et al. 2003). Ces algorithmes sont majoritairement généralistes ou orientés données XML.

Ils sont optimisés en temps d'exécution et utilisation de la mémoire. Certains algorithmes sont spécialisés pour *diff* et *merge* à la fois tandis que d'autres ne traitent que *diff*.. Après une étude exhaustive des algorithmes les plus utilisés (Vu, 2011) nous avons retenu 3DM de Tancred Lindholm (Lindholm, 2003 & 2004), qui est le plus efficace en termes de qualité et clarté des résultats obtenus et dont les sources sont directement accessibles.

2.2.3. Unique ID oriented

Tous les algorithmes ou outils mentionnés ci-dessus, sont essentiellement basés sur une valeur de *hash* et le contenu de chacun des nœuds pour les mettre en correspondance par un calcul de similarité (ou dis-similarité). Thao (Thao et al., 2010) a proposé une alternative au three-way *merging* consistant en *l'utilisation d'identifiants uniques*. Si chaque élément XML possède un identifiant unique, la mise en correspondance devient triviale.

2.2.4. Tree-Based textual documents

XML est utilisé non seulement pour transporter des données mais aussi pour encoder des documents textuels. Selon Angelo Di Iorio et al. (Di Iorio et al., 2009), il y a une différence entre le *diffing* d'un XML orienté document littéraire et le *diffing* d'un XML orienté données.

[4] http://code.google.com/p/google-diff-match-patch

Ils ont introduit un nouvel indicateur, *naturalness* qui reflète la capacité de l'algorithme à identifier automatiquement les changements qui pourraient être identifiés par une approche manuelle. Cependant aucune expérimentation complète ne semble avoir été réalisée avec cet algorithme.

2.3. Visualisation

Un moteur différentiel détecte des changements entre deux documents et les enregistre dans une sortie. Quel que soit le format de la sortie, il est toujours difficile pour l'utilisateur d'interpréter un changement dans le document. Il a donc besoin d'une interface de visualisation qui va permettre de mettre en évidence les changements dans leur contexte et faciliter leur manipulation.
En général, il y a deux modes d'affichage : *Side by Side* et *All In One*. Le premier mode consiste à ouvrir deux fichiers dans deux éditeurs identiques, l'un à côté de l'autre.
Les différences seront surlignées respectivement dans le premier et le deuxième éditeur. Le second mode ouvre une seule vue mais y représente tous les changements. Nous avons étudié douze outils de visualisation (Vu, 2011), qui nous ont permis de proposer un outil adapté à notre approche en prenant certaines des meilleures caractéristiques dans chacun d'eux.

2.4. Approche retenue

Notre corpus documentaire est encodé en XML et valide des modèles dédiés. Nous avons donc besoin d'un outil de *diff* & *merge* orienté XML document. L'outil 3DM semble être le meilleur candidat, car son *tree-matcher* est efficient pour le XML généraliste et peut être encore amélioré. Il exploite toutes les opérations d'édition (e.g *update*, *insert*, *delete*, *move* et *copy*) et propose une représentation de bonne qualité des différences entre les documents.
Enfin, le résultat du *three-way merge* par 3DM est en général meilleur que d'autres outils équivalents. De plus son code source est librement accessible, permettant de modifier ses modules pour réaliser nos propres optimisations.
Nous avons donc utilisé 3DM comme un framework de travail auquel nous avons ajouté des extensions spécialisées et adaptées à nos documents de façon à obtenir :
- Un *merge* interactif pour choisir, éditer des différences et résoudre des conflits.
- Une amélioration du *matching* heuristique de 3DM
- L'utilisation d'un algorithme différentiel basé sur le texte pour avoir des différences au niveau du contenu des nœuds XML.
- La visualisation des différences

3. Merge interactif

La plupart des outils de *differencing* (*diff*) et de *merging* (*merge*) fonctionnent en deux temps. L'outil de *differencing* sert à montrer en quoi deux versions sont différentes alors que l'outil de *merging* utilise ce résultat pour fusionner automatiquement les changements afin de créer une nouvelle version. Nous proposons, ici, une alternative appelée "*merge* interactif". Cette approche consiste à ne pas appliquer systématiquement la fusion automatique mais à rendre la transformation séquentielle et interactive. L'objectif est de proposer à l'utilisateur une liste d'opérations associées au document original, qu'il pourra confirmer ou non, les unes après les autres. Il doit pouvoir pré-visualiser le résultat d'une opération sur le document avant de décider de l'appliquer réellement. Le document sera modifié après chaque confirmation.
Le *merge* interactif répond à deux motivations principales : la première est qu'il permet à l'utilisateur de ne sélectionner que les changements jugés utiles pour sa propre version ; la seconde est qu'il permet de résoudre manuellement et convenablement les conflits d'un *three-way merging*.
Une raison supplémentaire concerne la visualisation des différences. Actuellement, les outils de *diff* affichent toutes les différences identifiées entre deux versions du document en même temps, ce qui permet d'avoir une vision globale de celles-ci, mais reste limitée aux trois types d'opération basiques *insert*, *delete* et *update*. D'autre part, plus le document a été changé, plus il y a des différences et plus il est difficile d'en donner une image lisible. Le *merge* interactif permet de rejouer en séquence toutes les opérations, ce qui fait perdre la vue globale mais est plus avantageux en termes d'opérations possibles (e.g. *move*, *copy*) et en termes de surcharge visuelle.
Le *merge* interactif ne produit pas lui-même la liste des opérations mais utilise celles fournies par un outil spécialisé. Selon les outils, deux sortes de listes sont disponibles : un ensemble d'opérations non-ordonnées expliquant ce qu'il se passe pendant la fusion mais qui n'est pas destiné à être exécuté ; un script d'opérations ordonnées permettant d'effectuer la fusion automatique. Cependant aucun de ces scripts ne peut être exploité tel quel, car, soit ils ne sont pas destinés à être manipulés, soit l'ordre des opérations est imposé. Avec un ensemble d'opérations non-ordonnées, il est possible de générer une séquence personnalisée d'opérations à condition de pouvoir prendre en considération le fait que certaines d'entre elles sont dépendantes de l'exécution préalable d'autres opérations.
Nous allons présenter dans la section suivante les principes d'élaboration du *merge* interactif, puis nous décrirons son implémentation dans notre prototype.

3.1. Génération des séquences d'opération

Cette section présente les aspects principaux du *merge* interactif. En particulier, elle démontre qu'il existe des relations d'ordre entre certaines opérations et qu'il est possible de recombiner dynamiquement une séquence des opérations exécutables et correctes en respectant ces relations.

3.1.1. Définitions des opérations

Nous donnons ci-dessous les définitions des opérations appliquées à un document XML qui seront utiles pour la suite. Un document XML est une structure arborescente dont les nœuds (éléments, texte, ...) sont ordonnés : modifier un document XML revient à modifier un arbre ordonné. Soit un document XML dont la structure est représentée par l'arbre *T* ordonné, dont les nœuds sont notés *m, n,* …, nous définissons les opérations :

insert(m,k,n) insère le nouveau nœud *n* en tant que k-ème enfant du nœud *m* ($m \in T$).

delete(m) supprime totalement le sous-arbre enraciné au nœud *m* ($m \in T$).

update(m,n) change la valeur initiale du nœud *m* par la nouvelle valeur *n*.

move(m,k,n) enlève tout le sous-arbre enraciné au nœud *m* de sa place initiale et le déplace au dessous du nœud *n* en tant que k-ème enfant de ce dernier ($n \in T$).

Il est à noter que ces opérations ne se comportent pas toujours de la même façon. En effet, elles dépendent du type de l'objet sur lequel elles portent : un nœud de texte ou un nœud d'élément. La valeur d'un nœud d'élément est le nom de la balise et ses attributs alors que celle d'un nœud de texte est une chaine de caractères. De plus, un nœud de texte n'a pas d'enfant. L'opération *move* peut être spécialisée par une combinaison d'*insert* et *delete*, mais l'objet sur lequel elle porte n'est ni supprimé ni inséré. On pourrait ajouter *copy* à cette liste l'opération mais l'avons exclue car elle est rarement présente dans les outils, génératrice d'erreurs, difficile à gérer lorsque les nœuds ont des identifiants, et s'applique mal à certains types de documents textuels.

3.1.2. Opérations delta

Les quatre opérations insert, delete, update et move permettent d'exprimer toutes les différences entre deux versions du document. Cependant l'expression de ces différences n'est pas unique. Dans l'exemple de la Figure 1, pour passer d'un arbre T0 à un arbre T1, il est possible que des versions intermédiaires aient donné un arbre tel que T01 . Les opérations qui font passer de l'arbre T0 à l'arbre T01 sont la modification du texte « a » en « aa » et l'insertion des nœuds c et d avec les textes « c » et « d ». Ensuite le nœud c est déplacé entre a et b, et le nœud d est supprimé. L'ensemble des opérations nécessaires pour passer de T0 à T1 forment la suite insert(R,3,c), insert(c,1, « c »), insert(R,4,d),

insert(d,1, « d»), update(a, « aa »), delete(d), move(R,2c). Supposons que seuls T0 et T1 soient enregistrés, pour passer de T0 à T1 il suffit d’insérer un nœud c avec le contenu « c » entre a et b, ce qui se résume aux trois opérations insert(R,2,c), insert(c,1, « a ») et update(a, « aa »), toutes les autres étant devenues inutiles. Ces trois opérations n'ont rien à voir avec les vraies opérations. Cela s'explique par le fait que certaines opérations de la première phase T0 à T01 et certaines autres de la deuxième phase T01 à T1, s'appliquent aux mêmes objets. Ainsi, les résultats des premières sont annulés ou altérés par les résultats des dernières. Le résultat final est donc exprimé par d'autres opérateurs. Par exemple, ici, insert(R,3,c), insert(c,1, « c »), sont remplacées par insert(R,2,c), insert(c,1, « a ») et insert(R,4,d), insert(d,1, « d»), delete(d) s’annulent et ne donnent rien. L’ensemble des cas que l’on peut rencontrer sont présentés dans le tableau 1.

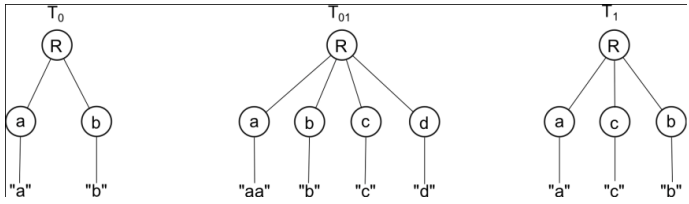


Figure 1. Etapes intermédiaires de transformation

Opérations intermédiaires annulées	Opération identifiée à la fin
update(m,v) puis update(m,v')	update(m,v')
update(m,v) puis delete(n), n est m ou un ancêtre de m	delete(n)
insert(n,k,m) puis delete(o), o est un ancêtre de n	delete(o)
insert(n,k,m) puis delete(m)	aucune
insert(n,k,m) puis update(m,v)	insert(n,k,m), m vaut v
insert(n,k,m) puis move(o,l,m)	insert(o,l,m)
delete(m) puis insert(n,k,m)	delete tous les enfants de n
delete(m) puis delete(o), o est un ancêtre de m	delete(o)
move(n,k,m) puis delete(m)	delete(m)
move(n,k,m) puis delete(o), o est n ou un ancêtre de n	delete(o) et delete(m)
move(n,k,m) puis move(o,l,m)	move(o,l,m)

Tableau 1. Opérations intermédiaires annulées par d'autres opérations

Les opérations *insert*(R,*2*,*c*), *insert*(*c*,*1*, « a ») et *update*(*a*, « aa ») ci-dessus sont appelées *opérations deltas*. Les opérations deltas ne sont pas forcément les opérations qui ont été réellement effectuées, elles utilisent des positions référencées dans l'arbre original et/ou l'arbre final pour exprimer les différences entre deux arbres. Leurs résultats sont visibles

dans au moins un arbre. D'une façon générale on définit les opérations delta de la façon suivante :
delete(*m*) est une opération delta si on trouve *m* dans T_0 mais non dans T_1.
insert(*m*,*k*,*n*) est une opération delta si on trouve *n* en tant que k-ème enfant de *m* dans T_1 mais non dans T_0.
update(*m*,v) est une opération delta si on trouve *m* dans T_0 et dans T_1 mais avec différentes valeurs (v dans T_1)
move(*m*,*k*,*n*) est une opération delta si on trouve *m* dans T_0 et T_1 mais à des positions différentes.
Un *delta* Δ de T_0 à T_1 est un ensemble d'opérations *delete*, *insert*, update et *move* satisfaisant les conditions ci-dessus. Δ ne précise aucun ordre entre les opérations, elles sont suffisantes pour passer de T_0 à T_1, cependant, il faut les appliquer une par une et dans un certain ordre pour obtenir le résultat voulu. On dira que Δ de T_0 à T_1 est optimal s'il n'existe aucun Δ', sous-ensemble de Δ permettant d'aller de T_0 à T_1.

3.1.3. Relation d'ordre
La présentation des opérations delta permettant le passage d'une version à une autre n'étant que la mise en évidence de ce qui les différentie, rien n'est dit sur la possibilité de les exécuter dans un ordre quelconque pour passer effectivement d'une version à l'autre. Prenons par exemple (Figure 2.) le cas d'un article intitulé "this is the source file" qui possède initialement deux chapitres. Chacun des chapitres contient son propre titre et des blocs constitués de paragraphes possédant éventuellement un sous-titre. On a supprimé le deuxième chapitre (*delete*) tout en conservant le seul bloc qu'il contient. Ce bloc est donc déplacé (*move*) à l'intérieur du premier chapitre en tant que 3ème enfant. Ensuite, on a changé (*update*) le titre du papier qui est maintenant "this is the cible file".

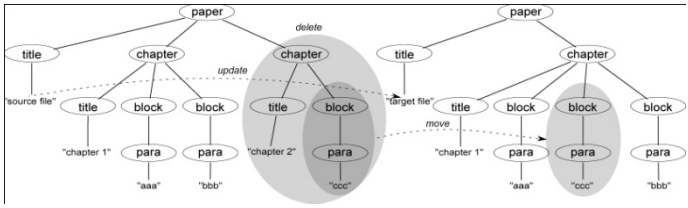


Figure 2. Modifications de "source file" à "target file"

Les opérations qui ont été effectivement réalisées sont des opérations deltas car elles sont toutes repérables sur l'arbre original et l'arbre final. Ces opérations étant a priori indépendantes, il est possible de les exécuter dans n'importe quel ordre. Cependant, en examinant le contexte de l'opération *move*, on s'aperçoit que le bloc à déplacer se trouve dans le chapitre censé être supprimé totalement. Si l'opération *delete* est exécutée avant l'opération *move*, alors tout le chapitre est supprimé, y compris le

bloc. En l'absence de son objet, l'opération *move* devient non-exécutable. Ce problème ne se présente pas si l'ordre d'exécution est inversé (*move* avant *delete*). Cet exemple montre l'existence d'une relation d'ordre sur l'exécution des opérations que nous définissons ainsi :

Définition : Deux opérations sont liées par une relation d'ordre, notée ">", lorsque l'exécution de l'une nécessite l'exécution préalable de l'autre pour assurer la faisabilité et l'exactitude des deux.

Soit ω_1 et ω_2 deux opérations, alors $\omega_1 > \omega_2$ signifie que ω_1 est dépendante de ω_2 et que ω_2 est précédente de ω_1. Dans une telle relation, l'opération précédente doit s'exécuter avant la dépendante. Ceci est nécessaire mais non suffisant pour que l'opération dépendante devienne exécutable. En effet, une opération peut dépendre de plusieurs opérations précédentes. Elle n'est exécutable qu'une fois que toutes ses précédentes ont été effectuées. Il faut aussi préciser qu'une opération est susceptible d'être à la fois précédente et dépendante d'autres opérations. Par exemple, soient ω_1, ω_2, ω_3, ω_4, ω_5, des opérations telles que : $\omega_1 > \omega_2$; $\omega_1 > \omega_3$; $\omega_2 > \omega_4$; $\omega_3 > \omega_4$; $\omega_3 > \omega_5$: ω_1 est précédente de ω_2 et ω_3 , ω_2 est précédente de ω_4, ω_3 est précédente de ω_4 et ω_5, ω_4 est directement dépendant de ω_2 et ω_3, ω_4 et ω_5 sont dépendantes par transitivité de ω_1. Ainsi, ω_1 est une précédente indirecte de ω_4 et ω_5, ω_2 et ω_3 ne sont pas en relation, de même que ω_4 et ω_5. Pour pouvoir exécuter toutes ces opérations, il faut les exécuter dans un ordre valide. Cet ordre n'est pas unique et doit prendre en compte les trois conditions suivantes :

ω_2 et ω_3 doivent être exécutées après ω_1;
ω_4 doit être exécutée après ω_2 et ω_3 ;
ω_5 doit être exécutée après ω_3.

Les opérations ω_1, ω_2, ω_3, ω_4, ω_5, forment un ensemble appelé *hiérarchie* d'opérations.

Définition : Une hiérarchie est un ensemble d'opérations dans lequel chaque opération doit être en relation d'ordre avec au moins une autre opération de cette hiérarchie. Si une opération appartient à une hiérarchie, toutes ses précédentes et ses dépendantes y appartiennent également.

Une hiérarchie peut être représentée par un graphe orienté. Une opération est représentée par un nœud qui peut avoir plusieurs prédécesseurs et plusieurs successeurs. Un arc correspond à une relation d'ordre dont le nœud sortant est la précédente et le nœud entrant est la dépendante. Une hiérarchie devient une arborescence à condition que chaque opération ait une seule précédente directe ou n'en ait pas.

Une opération qui n'est ni dépendante ni précédente d'autres opération, est appelée *indépendante*, elle n'appartient à aucune hiérarchie.
Un *delta* Δ constitué des opérations $\{\omega_1, \omega_2, \dots, \omega_n\}$ peut être réécrit sous la forme $\{H_1, H_2, \dots, \omega_1, \omega_2, \dots\}$ dans laquelle H_1, H_2, sont des hiérarchies et ω_1, ω_2 sont des opérations indépendantes.
Il faut alors répondre à deux questions : une hiérarchie peut-elle posséder un cycle ? deux hiérarchies sont-elles disjointes ?
La réponse à la seconde question est triviale par la définition d'une hiérarchie. En effet, si une opération appartient à la fois à H_m et H_n, alors toutes ses précédentes et ses dépendantes aussi. Par conséquent, H_m n'est rien autre que H_n.
Pour répondre à la première question, une solution consiste à explorer toutes les relations possibles entre les opérations, puis chercher à les enchainer afin de détecter l'existence de cycles.
Une étude exhaustive nous a permis de trouver les relations de dépendance suivantes :

insert(.,.,n) > insert(n,k,m)
insert(n,l,.) > insert(n,k,m) si l < k
move(n,l,.) > insert(n,k,m) si l < k
delete(o) > insert(n,k,m) si parent(o)= n : et : position(o) <= k
move(p,l,q) > insert(n,k,m) si parent(q)= n : et : position(q) <= k
insert(.,.,n) > move(n,k,m)
insert(n,l,.) > move(n,k,m) si l < k
move(n,l,.) > move(n,k,m) si l > k
delete(o) > move(n,k,m) si parent(o)= n : et : position(o) <= k
move(p,l,q) > move(n,k,m) si parent(q)= n : et : position(q) <= k
move(n,k,o) > delete(m) si o in T(m)

Nous pouvons distinguer deux groupes : le premier comprend les relations 1, 6 et 11. Ces relations sont des conditions consistantes sans lesquelles l'exécution des opérations concernées n'est pas possible. Les relations du deuxième groupe ne conditionnent pas l'exécution des opérations mais assurent leur exactitude en termes de résultat final, c'est donc dans ce deuxième groupe que nous cherchons à annuler l'existence de cycles. En fait, ils apparaissent dans les relations 4, 5, 9 et 10 et cela est dû au fait d'utiliser la position exacte, dans l'arbre final, du nœud inséré ou déplacé.
Une première solution consiste à "relativiser" le paramètre *k* dans la définition des opérations *insert* et *move*. La valeur de *k* n'indique pas la position dans l'arbre final, du nœud inséré (ou déplacé) mais indique la position dans l'arbre actuel où le nœud est inséré (ou déplacé). Cela permet d'exécuter correctement les *inserts* et *moves* sans dépendre des autres opérations. Pour chaque *insert* et *move*, il faut donc recalculer *k* en fonction du nombre de *deletes* et *moves* sortants ainsi que du nombre

d'*inserts* et *moves* entrants pour les nœuds du même parent et situés à gauche du nœud *m*. Chaque *insert* ou *delete* ou *move* effectué change la liste des enfants. Il faut donc également changer la valeur du *k* des *insert* ou *move* entrant. Concrètement, après l'insertion ou le déplacement d'un nœud dans la liste des enfants, les positions des nœuds les plus à droite à insérer ou à déplacer, doivent être incrémentées de 1; après la suppression et le déplacement d'un nœud en dehors de la liste des enfants, les mêmes positions précédentes doivent être décrémentées de 1 (Figure 3).
Cette première solution permettra d'annuler toutes les relations d'ordre du deuxième groupe. Il nous reste donc les relations du premier groupe 1, 6 et 11. En enchaînant ces trois relations, on peut

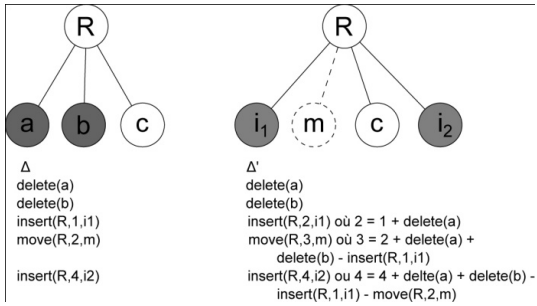


Figure 3. k est recalculé pour assurer l'exécution correcte des opérations

uniquement obtenir une séquence telle que : *insert* supérieur > *insert* inférieur > *move* ou encore inférieur > *delete* (contenant l'objet du *move*) qui ne forment jamais de cycle.
La deuxième solution, plus radicale, consiste à redéfinir *insert* et *move* : au lieu d'insérer ou déplacer un nœud à une position précise *k*, on peut l'insérer ou le déplacer après un nœud *left* :
insertAfter(n,left,m) insère le nœud *m* après le nœud *left* qui est un enfant du nœud *n*
moveAfter(n,left,m) déplace le nœud *m* après le nœud *left* qui est un enfant du nœud *n*
Le nœud *n* est nécessaire car si *left* est égal à null, le nœud *m* est inséré ou déplacé au dessous du nœud *n* en tant que premier enfant. L'*insertAfter(n,null,m)* et le *moveAfter(n,null,m)* sont donc possibles si et seulement si le nœud *n* est présent au moment de l'exécution. Le nœud *n* peut être lui-même l'objet d'un *insertAfter* supérieur. Il faut donc que ce dernier soit préalablement exécuté.
insertAfter(.,.,n) > insertAfter(n,null,m)
insertAfter(.,.,n) > moveAfter(n,null,m)

Dans le cas contraire, l'*insertAfter(n,left,m)* et le *moveAfter(n,left,m)* nécessitent que le nœud *left* soit présent au moment de l'exécution. Il se peut que ce dernier soit également l'objet d'une autre opération *insertAfter(n,left.left,left)* ou *moveAfter(n,left.left,left)* sachant que le nœud *left.left* peut être égal à *null*. Il faut donc exécuter les *insertAfter* et *moveAfter* l'un après l'autre en commençant par les nœuds plus à gauche.
insertAfter(n,left.left,left) > *insertAfter(n,left,m), left* ≠ *null*
moveAfter(n,left.left,left) > *insertAfter(n,left,m), left* ≠ *null*
insertAfter(n,left.left,left) > *moveAfter(n,left,m), left* ≠ *null*
moveAfter(n,left.left,left) > *moveAfter(n,left,m), left* ≠ *null*
Il reste à démontrer l'inexistence de cycles. Parmi les trois opérations, seules *insertAfter* et *moveAfter* peuvent être à la fois dépendantes et précédentes d'autres opérations alors que le *delete* est uniquement dépendante, ce qui implique qu'un cycle, s'il existe, contiendrait seulement des opérations *insertAfter* et *moveAfter*. L'insertAfter et le moveAfter sont pratiquement identiques en termes de relations avec d'autres opérations, il suffit donc d'examiner l'une des deux. L'insertAfter

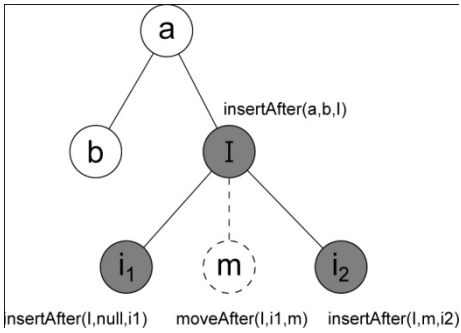


Figure 4. InsertAfer dépend d'InsertAfter à gauche et d'InsertAfter supérieur :

insertAfter(a,b,I) > insertAfter(I,null,i1) > moveAfter(I,i1,m) > insertAfter(I,m,i2) peut dépendre de l'insertAfter (ou moveAfter) à gauche qui peut également dépendre de l'insertAfter (ou moveAfter) encore plus à gauche. Si le nœud inséré (ou déplacé) de ce dernier est le premier enfant de son parent, l'insertAfter (ou moveAfter) dépend de l'insertAfter supérieur qui à son tour peut dépendre d'un insertAfter encore supérieur ou d'un insertAfter (ou moveAfter) à gauche. Le processus peut continuer ainsi de suite, mais dans une seule direction, vers le haut de l'arborescence (Figure 4). Il n'y aura jamais d'opération insertAfter (ou moveAfter) qui dépende du tout premier insertAfter, il n'y aura donc pas de cycle.
Ces deux solutions permettront d'éliminer les cycles et d'assurer l'exécution correcte des opérations. La première solution ne change pas

la définition des opérations mais elle est compliquée à cause des calculs qu'elle génère. La deuxième solution n'exige pas de calcul, mais elle nécessite d'enregistrer aussi l'information relative au nœud à gauche pour toutes les opérations insertAfter et moveAfter.

3.1.4. Principe d'acceptation et de refus

Une opération est exécutable quand toutes ses conditions d'exécution sont satisfaites et ne l'est pas si au moins une condition n'est pas satisfaite. Accepter une opération ne rend pas tout de suite exécutable ses opérations directement dépendantes mais permet de satisfaire l'une des conditions d'exécution de celles-ci. Au contraire, le fait de refuser une seule opération aura éventuellement un effet en cascade sur plusieurs autres opérations. Par simplicité, on peut dire que refuser une opération signifie refuser toutes ses opérations dépendantes (directement ou par transitivité).

Ce n'est pas tout à fait correct car il faut prendre en considération la nature de chaque condition. Comme indiqué ci-dessus, il y a deux groupes de relations. Le premier est lié à la possibilité de l'opération (est-elle possible ?), alors que le deuxième concerne l'exactitude de l'opération (est-elle correcte ?). Si la condition insatisfaite est du premier groupe, le refus en cascade est inévitable. Par contre, s'il s'agit d'une condition du deuxième groupe, l'opération reste exécutable mais sera incorrecte. Cela demande donc de modifier les paramètres des opérations dépendantes :

Refuser un *insert* ou *move* d'un nœud implique dés-incrémenter la position des *inserts* et *moves* plus à droite,

Refuser un *insertAfter* ou *moveAfter* d'un nœud implique changer le paramètre 'nœud à gauche' de l'*insertAfter* ou *moveAfter* à droite par le nœud à gauche de l'opération refusée.

3.1.5. Opération réversible

Durant l'exécution des opérations, il est parfois nécessaire d'inverser (ou annuler) des opérations acceptées. Les opérations, telles que définies, ne sont pas inversibles parce que dans leur définition, il manque des données complémentaires. Par exemple, l'opération *delete*(*m*) supprime un sous-arbre enraciné au nœud *m*. L'inverser nécessite une opération *insert*, mais on ne sait pas où exactement réinsérer le sous-arbre supprimé. Marian et al., (Marian et al., 2001) ont défini des *opérations complètes* (*completed delta*) permettant non seulement d'exécuter mais aussi d'inverser l'opération. Il est également à noter qu'inverser une opération implique rend insatisfaites des conditions relatives dans lesquelles l'opération joue le rôle de précédent.

3.1.6. Algorithme de mise en ordre des opérations

La remise en un ordre valide des opérations pour un *merge* interactif n'est pas strictement nécessaire. En effet, durant le *merge*, l'utilisateur peut ne

sélectionner que les opérations exécutables quel que soit leur ordre. En revanche, si l'opération fait partie d'une hiérarchie, il est pratique d'en percevoir immédiatement les opérations dépendantes et d'enchaîner toute la hiérarchie. Ces actions ne sont plus triviales quand les opérations relatives sont dispersées, il faut donc pouvoir les mettre ensemble. En particulier, si nous voulons être en mesure de les rejouer en séquence, il faut trouver un ordre valide pour en assurer le bon fonctionnement. Un ensemble d'opérations non-ordonnées contient éventuellement plusieurs hiérarchies distinctes et mélangées avec des opérations indépendantes. Notre algorithme de mise en ordre des opérations cherche itérativement à remettre celles appartenant à une même hiérarchie ensemble puis à placer une opération après ses opérations précédentes. Dans un premier temps, l'opération précédente est remontée devant l'opération dépendante si ce n'était pas déjà le cas. Une fois que toutes les opérations précédentes sont placées devant les opérations dépendantes, il met cette dernière juste après sa précédente dans la liste. La terminaison de l'algorithme est assurée grâce à la caractéristique acyclique des hiérarchies.

3.2. Implémentation

Le *merge* interactif ne produit pas lui-même les opérations permettant la transformation entre les versions du document. Il récupère la liste des opérations enregistrées par 3DM mais ne les exploite pas immédiatement car il faut examiner préalablement les relations éventuelles entre les opérations telles que présentées dans la section précédente. La première implémentation mobilise également un algorithme de comparaison de texte, pour compléter 3DM. Elle a été testée sur les documents au format XML issus de notre base documentaire.

3.2.1. Algorithmes utilisés

En *merge* interactif, nous voulons être en mesure de retracer et visualiser des changements à la fois au niveau de la structure de l'arborescence du document et dans son contenu textuel. Pour ce faire, nous avons eu recours à deux algorithmes complémentaires : 3DM de Lindohlm, destiné à examiner la structure d'arbre de XML et Google-diff-match-patch, pour différencier des textes. Ce dernier va mettre en évidence les mots et même les caractères du contenu d'un nœud texte qui ont été insérés ou supprimés. 3DM enregistre toutes les opérations dans un fichier de format XML appelé *edit log*. L'ordre d'enregistrement des opérations est celui de l'insertion des nœuds en vue de construire l'arbre final. Chaque enregistrement correspond à une opération précise, les propriétés de l'opération sont décrites par des paires "attribut-valeur", qui indiquent la position des nœuds impliqués par l'opérateur dans l'arbre original et final, du parent adoptif d'un nœud déplacé ou inséré. Les informations nécessaires au parcours des arbres pour retrouver les

nœuds impliqués par les opérations sont encodées. La version actuelle du *merge* interactif n'implémente que les relations du premier groupe, elle ne traite pas encore des cycles. Le positionnement de certains nœuds insérés ou déplacés peut donc être inexact. Dans la prochaine implémentation, pour éliminer les cycles, l'une des deux solutions mentionnées ci-dessus sera utilisée, la deuxième solution étant préférable, l'information sur le nœud à gauche d'un nœud inséré (ou déplacé) pouvant être enregistrée par 3DM.

3.2.2. Vue globale de l'implémentation

Le premier prototype du *merge* interactif a été réalisé en Java. La classe principale est un JFrame, qui contient un panel de la classe *InteractiveMergePanel*. Ce dernier joue à la fois le rôle de vue et de contrôleur. Elle présente les données (des opérations, la structure et le contenu textuel du document XML). Elle reçoit les actions de l'utilisateur et les traite. Les données sont réellement modifiées par la classe *Merge*.

L'*edit log* est modélisé par la classe *EditLog*, qui contient la méthode *sort*, qui implémente l'algorithme de remise en ordre des opérations. L'opération est stockée dans l'objet *Operation*, qui possède la méthode *isBelongTo* afin d'examiner si une opération est dépendante d'une autre opération. La classe *Path* est utile pour manipuler les paths des nœuds de 3DM.

Les fichiers XML sont parsés et traités par un parser de type DOM qui crée pour chacun des fichiers, un objet d'arbre interne facile à accéder et à modifier. Pour assigner cet objet d'arbre à un Swing JTree destiné à s'afficher sur l'interface, nous avons utilisé les classes *XMLTreeNode* et *XMLTreeModel* de Rob Lybarger[5] La classe *TreeCellCustomRenderer* permet de changer l'affichage de l'arbre.

3.2.3. Interface graphique du merge interactif

L'interface principale du *merge* interactif (Figure 5) est constituée de trois panneaux :

Le premier panneau affiche la liste des opérations regroupées en hiérarchies. Une opération est représentée par son type, le nœud concerné et d'autres paramètres. Les opérations activées sont susceptibles de s'exécuter immédiatement. Les opérations désactivées sont dépendantes. Elles doivent attendre l'exécution de leurs opérations précédentes afin d'être activées et exécutables. L'utilisateur peut choisir les opérations activées pour les exécuter l'une après l'autre dans n'importe quel ordre.

Le deuxième panneau représente la structure d'arborescence interactive du document XML. Cet arbre s'étend à tous les nœuds internes et non

[5] http://www.developer.com/xml/article.php/3731356/Displaying-XML-in-a-Swing-JTree.htm

aux feuilles textuelles. En cliquant sur un nœud, le contenu textuel de ce nœud est affiché dans le troisième panneau.
Le troisième panneau affiche le contenu textuel du document XML dans un format purement textuel. Les titres (du papier, du chapitre, de la session, ...) sont en gras et les paragraphes sont espacés.
Ce premier prototype a été utilisé sur des documents académiques réels de notre base documentaire et a permis d'en tester les différentes fonctionnalités (parcours d'arborescence, sélection d'opérations ou de séquences d'opération, validation de séquences dépendantes ou annulation, retour en arrière, prévisualisation de l'effet d'un opérateur, visualisation de modifications de textes, …) et d'évaluer la qualité de leur usage, tant du point de vue opérationnel que de la convivialité de l'interface. Il a été testé sur des documents dont la structure pouvait atteindre jusqu'à 17 niveaux d'éléments, contenant des balises iline de mise en forme, de référence ou de l'hyperlien et sur lesquels 3DM avait enregistré plus de 60 opérations de modifications

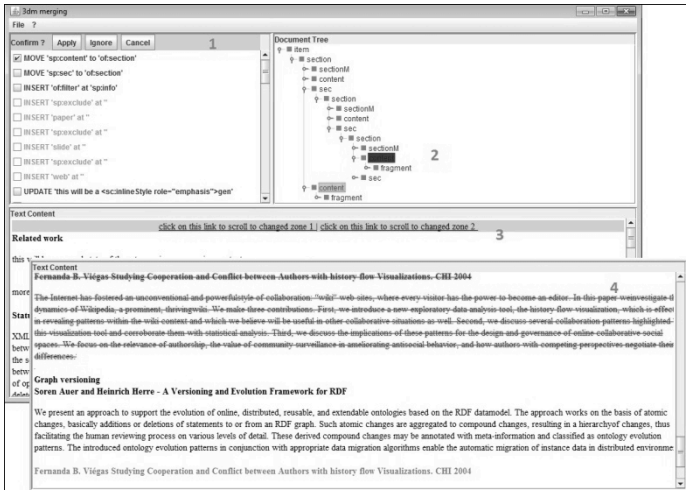


Figure 5. Interface principale du merge interactif : 1 - Liste d'opérations, avec pop-up de prévisualisation ; 2 - Structure d'arbre du document XML et mise en couleur des modifications ; 3 - Contenu textuel du document XML et mise en couleur des modifications.

4. Conclusions et perspectives

Nous avons présenté dans cet article une extension des algorithmes de *diff* & *merge* sur des documents structurés qui permet de réaliser les opérations de *merge* de façon interactive. Après avoir redéfini les opérateurs qui permettent de caractériser les différences entre deux

versions d'un document, tant sur leur structure que sur leur contenu, nous avons défini une relation d'ordre qui permet de proposer des séquences d'exécution cohérentes et exécutables de ces opérateurs. Nous avons pu en particulier montrer qu'il était possible, grâce à ces nouvelles définitions d'opérateurs, de s'affranchir en grande partie du risque de boucles et d'incohérence dans l'exécution des opérations. Nous avons enfin proposé une implémentation de l'algorithme et une interface d'exploitation qui permet à l'utilisateur de sélectionner les opérations valides ou à rejeter et de les exécuter en visualisant leur effet sur les documents. Les choix heuristiques et la gestion de conflits lors de la fusion de plusieurs éditions d'un même document s'en trouvent alors améliorés.

L'implémentation actuelle du prototype est encore limitée dans son applicabilité. Les fonctionnalités à étudier et implémenter incluent :
Gérer les conflits lors du three-way *merge* : la version actuelle est limitée à un fichier et sa modification, alors qu'il faudrait pouvoir traiter deux versions différentes d'une même source. 3DM fournit les informations nécessaires pour réaliser ce traitement en interactif.
Exploiter le schéma du document : seules les structures XML ont été prises en compte, or on pourrait intégrer la sémantique issue des modèles documentaires des chaînes éditoriales Scenari (Crozat, 2007) mobilisées dans le projet C2M.
Comparer des réseaux de fragments : un fragment contient des contenus et des références à d'autres fragments. Le fragment racine est le fragment qui n'a pas de parent et inclut par transitivité tous les fragments du document. Comparer deux documents revient à comparer deux réseaux de fragments, ce qui n'est pas possible directement avec 3DM. Cependant, on peut inclure les contenus de tous les fragments dans le fragment racine en vue de créer un seul fichier, puis comparer ces fichiers et appliquer le *merge* interactif. Lors de l'enregistrement du fichier, il faudra re-fragmenter le document résultant.

Bibliographie

ANDRE J., FURUTA R., QUINT V., Structured documents, Cambridge University Press, 1989.
CONEBA G., ADBESSALEM T. , HINNACH Y. A comparative study for XML change detection, Research Report, INRIA, 2002
COBÉNA G., ABITEBOUL S., MARIAN A., DETECTING Changes in XML Documents, Proceedings of the 18th International Conference on Data Engineering, 41-52. Feb. 2002.
CROZAT S., Scenari, la chaîne éditoriale libre, Eyrolles, 2007.

DI IORIO A., SCHIRINZI M., VITALI F., MARCHETTI C., A Natural and Multi-layered Approach to Detect Changes in Tree-Based Textual Documents, In Proceedings of ICEIS'2009. pp.90-101.

LA FONTAINE R., DeltaXML, Change Control for XML : Do It Right XML Europe, May 2003.

LINDHOLM T., XML three-way merge as a reconciliation engine for mobile data, Proceedings of the 3rd ACM international workshop on Data engineering for wireless and mobile access, 93-97, Sept. 2003.

LINDHOLM T., A three-way merge for XML documents, Proceedings of the 2004 ACM symposium on Document engineering, 1-10, Oct. 2004.

MARIAN A., ABITEBOUL S., COBÉNA G., MIGNET L. Change-Centric Management of Versions in an XML Warehouse Proceedings of the 27th VLDB Conference, Roma, Italy, 2001

PETERS L. Change Detection in XML Trees : a Survey In : third Twente Student Conference on IT ; June 2005

RÖNNAU S., PAULI C., BORGHOFF U.M., Merging changes in XML documents using reliable context fingerprints, Proceeding of the eighth ACM symposium on Document engineering, September 16-19, 2008, Sao Paulo, Brazil.

THAO C., ETHAN V., MUNSON E.V., Using Versioned Tree Data Structure, Change Detection and Node Identity for Three-Way XML Merging, DocEng2010, September 21-24, 2010, Manchester, United Kingdom.

VU X.T., Merging Interactif de Documents XML, rapport de Master mention Science et technologies de l'Information et de la Communication, Université de Technologie de Compiègne, juin 2011.

WANG, Y., DeWitt D.J., Cai, J. : X-Diff, An Effective Change Detection Algorithm for XML Documents, 19th International Conference on Data Engineering, 519-530. Mar. 2003.

La métaphore dans les relations intermédiatiques : quelles remédiatisations interactives ?

Pergia GKOUSKOU-GIANNAKOU
Laboratoire LUTIN, Cité des Sciences et de l'Industrie

Introduction

Dans cette contribution, nous examinons les particularités de la figure de la métaphore par rapport aux relations intermédiatiques qui se développent entre les fonctions et les rôles sociaux du web et ceux d'autres médias. Nous considérons que la métaphore constitue l'explicitation des cadres de perception et d'action qui émergent par les différents univers médiatiques et s'hybrident entre eux pour produire de nouvelles façons de percevoir et de communiquer. Dans notre analyse, nous mobilisons le concept de *remédiatisation*[6] et nous étudions les phénomènes métaphoriques du point de sociolinguistique ((Klinkenberg, 1973), (Lakoff & Johnson, 1986) et (Prandi, 1992)) et de la sémiologie du document numérique (Jeanneret, 2007), (Stockinger, 2005).
Plus précisément, nous considérons que les médias numériques constituent un assemblage complexe de médias anciens plongés dans le milieu numérique et agencés par un processus de *remédiatisation*s qui permet de remobiliser des pratiques d'écriture et d'usage déjà acquises et partagées par les publics. Or, nous estimons que le processus de la métaphorisation sort des limites de la simple évocation pour intervenir de façon déterminante dans la construction des cadres d'action et de raisonnement. Dans ce processus, les rapports intermédiatiques circulent entre concepteurs, usagers et supports pour encadrer le processus communicationnel.

Terrain et Approche méthodologique

Notre terrain est constitué par des sites web appartenant à des institutions de la culture scientifique et technique (Gkouskou-Giannakou, 2007). Dans ce texte, nous illustrons nos propos à travers l'exemple du site de web de l'ONERA (www.onera.fr)

[6] Plus précisément, nous reprenons le concept de remediation de (Bolter & Grusin, 2001).

Notre recherche interroge les stratégies rhétoriques de producteurs des sites web, l'organisation visuelle des pages-écrans et les pratiques interprétatives des internautes en cours d'une consultation. Cette approche induit une double grille d'analyse

– *sémiotechnique* qui consiste en :

- une *analyse morpho-dispositionnelle* (zonage[7] thématique) appliquée sur les pages des sites web étudiés et
- une *analyse cartographique* effectuée par un recensement manuel des liens « intérieurs » entre le pages du même site web ;

– *sémiopragmatique,* concernant la façon dont les acteurs (concepteurs, usagers) traitent l'objet dans leur discours. Plus particulièrement, nous avons effectué :

- un recueil des données auprès des concepteurs des sites web. Le plus souvent, il s'agit des entretiens avec les concepteurs des sites web. Les questions posées pendant ces entretiens concernaient les objectifs, les thématiques ainsi que l'histoire du site ;
- des entretiens avec des jurys des internautes. Ces entretiens avaient le style des discussions semi-dirigées et les questions concernaient les catégories thématiques des sites, leur iconicité et leurs objectifs.

Processus métaphoriques et rémédiatisations : des jeux interactifs entre acteurs et médias.

La métaphore est basée sur la perception subjective d'un rapport d'analogie entre deux objets ou unités thématiques à comprendre ou à exprimer[8]. Le processus métaphorique est donc toujours inhérent à l'existence d'un cadre de raisonnement préexistant qui a été formé dans un univers socioculturel. Il s'agit d'un processus d'inscription de la perception et de l'activité dans un cadre d'expérience déjà vécu.

De ce point de vue, nous considérons que la métaphore dans le discours constitue l'expression linguistique de la procédure sociocognitive du cadrage. Selon Erving Goffman (Goffman, 1991, p. 242), le cadrage concerne la mobilisation des prémisses organisationnelles qui aident l'acteur à interpréter les nouvelles données et à agir dans son environnement[9].

---

[7] Sur la notion de zone, voir : (Stockinger, 2005) et (Bertin, 1967).

[8] Selon Aristote, «*La métaphore est le transport à une chose d'un nom qui en désigne une autre, transport ou du genre à l'espèce, ou de l'espèce au genre ou de l'espèce à l'espèce ou d'après le rapport d'analogie*» (Aristote, *Περὶ ποιητικῆς* – *De la poétique*, traduction Hardy, 1985, page 1457 a-b).

[9] *« A partir du moment où nous comprenons ce qui se passe, nous y conformons nos actions et nous pouvons constater en général que le cours des choses confirme cette conformité. Ce sont ces prémisses organisationnelles - que nous confirmons en même temps mentalement et par notre activité - que j'appelle le cadre de l'activité ».*

Une séquence d'activité mobilise des règles et des conventions appréhendées par des expériences précédentes dans des environnements et des situations évocatrices. Ces règles et conventions constituent un cadre d'ancrage mais de transformation également pour la nouvelle expérience. Il s'agit de schémas qui encadrent toute nouvelle mise en scène[10] interactionnelle et se modifient pendant le déroulement de l'interaction pour aboutir à leur tour à la construction des nouvelles prémisses d'organisation de l'expérience. La perception du nouveau résulte de la difficulté à appliquer parfaitement les cadres d'expérience acquis dans la situation qui émerge. Il s'agit alors d'une comparaison continuelle entre l'ancien et le nouveau, d'un jeu de continuité basée sur la rupture. Sur cette tension de « conformité », Erving Goffman (Goffman, 1991, p. 287), remarque :

« Il est difficile de parler de l'ancrage de l'action dans le monde sans du même coup accréditer l'idée que nos actes sont en partie l'expression et le produit d'un soi qui subsiste. »

De leur part, les sociolinguistes mettent en évidence l'inscription de l'interprétation métaphorique dans un contexte pragmatique dans lequel la subjectivité des acteurs a un rôle primordial. Le jeu métaphorique se trouve dans mais aussi hors du texte ou image à interpréter ((LE GUERN, 1973), (KLINKENBERG, 1973), (PRANDI, 1992). Les expériences de chaque lecteur ou spectateur ainsi que les univers socioculturels dans lesquels ils ont grandi, influencent le processus interprétatif (Lakoff & Johnson, 1986).

La diffusion des métaphores impose de nouveaux cadres d'action et de communication à travers les sites web. Nous considérons que la métaphore constitue l'explicitation des cadres d'expérience qui émergent par les différents univers médiatiques et s'hybrident entre eux pour produire de nouvelles façons de percevoir et de communiquer. De ce point de vue, elle est étroitement liée au phénomène de rémédiatisation désignant la configuration des formes et des pratiques induites par un média dans un autre média. Sur ce point, nous reprenons le concept de « remediation » de Jay-David Bolter et Richard Grusin en le transformant. Pour ces auteurs, la «remediation» désigne la représentation d'un média dans un autre média[11]. En considérant que ce terme souligne plutôt la présence visuelle de l'un média dans l'autre, nous distinguons ce concept du concept de remediatisation lequel ne concerne pas la simple visualisation des formes qui font évoquer un autre média,

[10] Sur la notion de *mise en scène* dans le cadre de l'interaction documentaire (usager/document numérique), voir (Stockinger, 2005).

[11] Les auteurs définissent la « remediation » comme suit: « we call the representation of one medium in another remediation » (BOLTER & GRUSIN, 2001, p. 45). Un exemple très caractéristique de « remediation » selon ces auteurs constituent les « fenêtres » de journaux télévisés qui évoquant l'univers des « windows » des ordinateurs (Ibid, p. 189).

mais plutôt le transfert des données génératrices de l'autre média induisant des nouvelles pratiques et cadres de perception et d'action.

Processus métaphoriques et remédiatisations : l'exemple du web

Dans le cas du web, les métaphores constituent un point d'entrée très important en ce qui concerne l'observation de la mutation des formes et des pratiques de communication à travers les représentations du public. Les acteurs ont tendance à interpréter les formes émergentes selon les schémas expérientiels déjà établis (Jeanneret, 2008).
L'expression métaphorique apparaît dans le discours des concepteurs et des usagers ainsi que sur l'interface et l'architecture d'un site web. Il s'agit de la métaphore dans la matérialité du support et les pratiques des acteurs impliqués. La métaphore verbale des acteurs reflète le déplacement et la transformation des cadres d'organisation de l'expérience. La mise en place d'une métaphore donne lieu à la création d'un environnement d'action conventionnelle avec des règles et des repères qui guident le comportement des acteurs et laissent leurs traces sur le support.
Dans le cas des sites web institutionnels, les cadres qui modélisent l'action des agents concernent les pratiques culturelles et communicationnelles qui s'expriment dans la mise en forme du contenu et les modalités de circulation de l'information (par exemple les types de documents auxquels le site web fait allusion (brochures, affiches, etc.). Ces pratiques culturelles font partie des stratégies de représentation de l'image de l'identité de l'institution.
Un exemple très caractéristique est celui du site web de l'ONERA (www.onera.fr). Dans le cas de ce site, la métaphore du document de presse est très présente tant dans le discours du webmaster que de celui des internautes. Or, tandis que le webmaster présente son site comme un simple document de présentation des activités de l'ONERA, les schémas d'interprétation des internautes sont guidés en grande partie par la métaphore du « média imprimé de vulgarisation scientifique » dès la page d'accueil du site.
De sa part, le webmaster, insiste sur les avantages du système de «colonnes» qui lui permettent la gestion flexible des «rubriques» thématiques du site.
«Ce nouveau système de colonnes me permet de centraliser les rubriques. Dans chaque colonne, j'insère des tableaux avec des lignes qui correspondent à des modules indépendants. Je peux ajouter ou supprimer des rubriques sans faire effondrer mon tableau» (Webmaster de www.onera.fr)

Figure 1. Page d'accueil du www.onera.fr

La «*colonne*» gauche comprend une liste de liens concernant l'institution et ses activités tandis que la colonne droite contient des aperçus de travaux scientifiques de l'ONERA. La partie centrale de la page écran est occupée par les actualités de nature variée (scientifique ou institutionnelle) mises en thématiques.

Même si le webmaster de www.onera.fr ne se réfère pas explicitement à cela, la *métaphore* du document de la presse imprègne la construction du site. Le terme «*Magazine*» de la partie droite de la page écran est indicateur du cadre modalisant l'activité de la mise en ligne du site, celui du média de la presse.

Or, pour les internautes, le site web www.onera.fr est plus qu'un site de simple présentation des activités de l'institution puisque leurs schémas d'interprétation sont guidés en grande partie par la métaphore du « média imprimé de vulgarisation scientifique » dès la page d'accueil du site. Cette métaphore influence le parcours des internautes même devant des pages écrans dont la structure ne fait pas vraiment allusion à la forme visuelle des documents de cette nature. C'est le cas de la page écran « Images de science », dans laquelle la métaphore de la revue de vulgarisation scientifique sert à spécifier plutôt le contenu que la structure morphodispositionnelle de la page :

Figure 2. www.onera.fr, page « Images de science »

Pour P. 36 ans, cette forme évoque la revue de vulgarisation scientifique « *Reader's Digest* » quoiqu'il reconnaisse que la structure de cette page écran rappelle à peine la forme visuelle d'un magazine. Les deux listes des liens permettant l'accès au contenu constituent un hybride entre les « rubriques thématiques » d'une revue et la structure en « boutons » du panneau de commande d'une interface numérique. La liste, forme existante dans l'imprimé mais mise en valeur dans les documents numériques fait disparaître l'élément de la temporalité en spatialisant les flux d'information.

L'effet visuel provoqué par l'organisation de la page écran en listes thématiques évoque les rubriques d'un document de presse imprimé même si les utilisateurs saisissent implicitement ou explicitement la différence entre l'inscription du contenu sur un support matériel qui caractérise le document imprimé et celui de la séparation entre les deux qui fait partie de la virtualité du numérique.

« Ca donne l'impression d'être une page de « Readers digest » Tu connais « Readers digest » ? » […] «Tout à l'heure, la plupart du site était dominé par le texte. Là, on va entrer dans l'image. » […] « il nous dit : « Sélection des derniers mois » donc, des belles images qui concernent les activités de l'ONERA de ces derniers mois et puis il y a autre manière d'accéder à l'ensemble d'images y compris celles-ci de derniers mois et cette manière est thématique cette fois là » […] »Il n y a pas vraiment de texte, il n'y a que des liens… là chaque rubrique couvre un thème, disons, alors qu'ici les liens renvoient à des ensembles d'éléments d'information, des images à l'occurrence ou des animations » (P., 36 ans).

Conclusion

Dans le cas du web, les métaphores naissent dans le discours des concepteurs, s'inscrivent dans les outils de textualisation, se cristallisent éphémèrement dans la mise en forme du site et se métamorphosent dans le discours et les pratiques des utilisateurs.
L'analyse focalisée sur le support peut être éclairante en ce qui concerne les concepts métaphoriques stabilisés dans la structure techno - sémiotique du site, ce qu'on pourrait appeler «la rhétorique» du média reflétant les stratégies des concepteurs. Par contre, elle ne rend pas possible l'accès au processus d'émergence des métaphores. Les traces de celles-ci disparaissent derrière les formes cristallisées. La fixité documentaire d'un site web impose une analyse synchronique qui cache toutes les traces des métamorphoses qui s'effectuent entre les traditions culturelles et les intentions communicationnelles des concepteurs. C'est alors sur la base de la similarité, de l'évocation de la continuation médiatique que l'analyse peut s'effectuer.
A l'opposé des grilles basées sur la similarité qui imprègne l'analyse des formes fixées dans le support, l'analyse du discours des internautes révèle des processus de différenciation, d'évolution, de mutation des formes et des pratiques. Les concepts métaphoriques induisent une mise en situation, un scénario d'action, une structure narrative qui encadre l'activité des internautes et leur façon d'interpréter l'écriture d'écran. Or, une fois que les cadres d'activité se déclenchent, le processus de leur transformation s'entame. Dans cette situation évolutive, la forme sur l'écran n'est que le stimulus pour une réaction prédicative.
Les liens hypertextuels permettent l'extériorisation des processus d'interaction avec le support. Le processus de mutation observé auprès des internautes peut être comparé aux phases de l'évolution des médias décrits par André Gaudreault et Philippe Marion dans leur article «un média naît toujours deux fois» : a) la phase intégrative-mimétique pendant la quelle le nouveau média est pris dans le faisceau de déterminations des médias ou genres antérieurs et légitimes et b) l'autonomisation identitaire avec le développement de son propre langage par l' intégration transformationnelle des anciens langages.
Or, l'observation du numérique et de ses usages permet de comprendre à quel point ces deux phases s'impliquent l'une à l'autre à travers l'interaction acteur et support.

Bibliographie

ARISTOTE, *Poétique*, (traduit en français par J. Hard)y, Paris, Les Belles Lettres, 1985.
BERTIN J., *Sémiologie graphique*, Paris, Mouton/Gauthier-Villars, 1967.

BOLTER J-D., Grusin R., *Remediation. Understanding new media*, MA, MIT, 2001.
GAUDREAULT A., MARION P., «Un média nait toujours deux fois*», Sociétés & Représentations*, n° 9, Paris, CREDHESS (Paris I – Panthéon Sorbonne), 2000.
GKOUSKOU-GIANNAKOU P., *Composition médiatique des objets sites web. Le cas des sites web de la culture scientifique et technique*, Thèse de Doctorat, Université de Technologie de Compiègne, 2007.
GOFFMAN E., *Les cadres de l'expérience*, Paris, Minuit, 1991.
JEANNERET Y., « La page à l'écran, entre filiations et filières », Limoges, *Visible*, 2008, p. 153-172.
KLINKENBERG J-M., «Le concept d'isotopie en sémantique et en sémiotique littéraire», *Le français moderne*, vol. 3, n° 41, 1973, p. 285 - 290.
LAKOFF G. et Johnson M., *Les métaphores dans la vie quotidienne*, Paris, Minuit, 1986.
LE GUERN M., Sémantique de la métaphore et de la métonymie, Paris, Larousse, 1973.
PRANDI M., *Grammaire philosophique des tropes. Mise en forme linguistique et interprétation discursive des conflits conceptuels*, Paris, Minuit, 1992.
STOCKINGER P., *Les sites Web : conception, description et évaluation*, Paris, Hermès – Lavoisier, 2005.

LaSuli : un outil pour le travail intellectuel

Aurélien BENEL
Jean-Pierre CAHIER
Matthieu TIXIER

Equipe Tech-CICO, Laboratoire STMR (UMR CNRS 6279)
Université de technologie de Troyes (UTT)

Résumé : Lasuli est un logiciel d'annotation sociale à l'usage des lecteurs-interprètes développé en vue d'accompagner et d'aider au travail d'interprétation omniprésent dans les contextes de travail intellectuel et sur les connaissances. Nous présentons en détail l'importance des pratiques d'annotations pour ces activités, notamment dans l'environnement riche en document offert par l'Internet. Plusieurs fonctionnalités de Lasuli dédiées à ces activités sont présentées ainsi que les résultats d'une expérimentation réalisée avec une centaine d'élèves ingénieurs en système d'information. Ces étudiants ont utilisé Lasuli pour analyser des entretiens conduits auprès de professionnels dans le but de modéliser le système d'information de leur organisation. Nous présentons les retours d'expérience rassemblés à l'issue de cette expérimentation, lesquels nous invite à réfléchir aux améliorations futures de notre outil.

Mots-clés : Annotation, catégorisation, interprétation, coopération.

Abstract : Lasuli is an annotation tool which is aimed at supporting reading and interpretation practices widely present in the framework of intellectual professions. We detail the key aspects of these annotation practices for knowledge workers, in particular in the context of the great amount of documents available today through the web. Several functionalities specifically designed for Lasuli are presented. Besides, we present the results of an experiment carried out with about one hundred engineering school students. These students have used Lasuli to analyze the interviews they have conducted with professionals to design UML models of their organization information system. Our observation of the use of Lasuli by future engineers and the user experience feedbacks we have gathered from this experiment are detailed and invite us to consider future improvements for Lasuli.

Keywords : Annotation, tagging, interpretation, collaboration.

Introduction : Système d'information et travail intellectuel

« Intellectuel » est un mot qui gêne : prétentieux pour certains, injurieux pour les autres. On lui préfère aujourd'hui le terme de « travailleur de la connaissance ». Plus neutre, ce dernier terme est aussi plus facile à aborder par les modélisateurs que nous sommes. Quel est, en effet, le point commun entre un médecin, un avocat, un ingénieur, un historien, un sociologue, si ce n'est d'aborder l'inconnu en inventant ses propres méthodes, si ce n'est d'être engagés dans des activités de résolution de problèmes ? Il serait illusoire cependant de penser que reconnaître des méthodes de résolution de problèmes dans ces activités rendrait possible leur automatisation, ne serait-ce que parce que le « problème » n'est en général pas défini au départ mais au fil de l'activité (Zacklad, 2003).
Contrairement aux difficultés théoriques que posent le concept de « connaissances » (Rousseaux, 2005), le mot « intellectuel » présente l'intérêt de faire référence à une activité concrète : « inter-legere », étymologiquement, « lire entre les lignes ». L'intellectuel est donc celui qui lit des documents (dossiers, rapports, lois, normes, archives, entretiens), les annote et produit de nouveaux documents. Pour certains chercheurs en sciences de l'information (Buckland, 1998 ; Lund & Skare, 2010), cette idée que le document serait une « technique intellectuelle », notamment par la « fertilité documentaire » qu'elle rendrait possible (Briet, 1951), constitue une alternative radicale à la notion même d'information.
La notion d'information serait en effet une entrave à l'étude du phénomène humain et social de la transmission (Rastier, 2007). Dans le modèle de Shannon et Weaver : l'émetteur et le récepteur n'ont aucune influence sur la compréhension du « message », cette compréhension se limite à un « décodage » et la valeur du message aux propriétés statistiques de ses codes. Il en est tout autrement du document dont le sens dépend grandement de sa situation de production (auteur, date, destinataires...), et dont la compréhension demande « une sorte de *rumination* » (Virbel, 1995) car « la lecture et l'étude d'un ouvrage sur support papier incitent à vérifier quelque chose que cette forme ne permet pas d'atteindre, ou pas aisément » (Virbel, 1995).
Ces réflexions ne sont pas sans effet, nous semble-t-il sur le domaine de l'informatique des organisations. Le passage des « systèmes d'information » à des « systèmes documentaires » permettrait peut-être d'aborder plus facilement la question de l'instrumentation des métiers intellectuels et des organisations adhocratiques. En outre, cela pourrait inspirer des méthodes d'accompagnement du développement sans précédent dans les entreprises de la gestion des contenus (ECM, CMS), documents (EDM, DMS), archives (ERM) et patrimoines électroniques (DAM). Pour gérer correctement le document numérique dans les organisations, il faudrait s'occuper autant de son archivage

(métadonnées, révisions, horodatage et signature), que de son traitement automatique (recherche de motifs, recherche d'information, transformation) et de son interprétation par des humains (indexation matière, annotation).
Dans le cadre de cet article, nous ne traiterons que de l'instrumentation d'un type d'annotation qui, comme nous l'expliquerons dans une première partie, a été relativement peu traité. Nous présenterons dans une deuxième partie les choix qui ont présidé à la conception de LaSuli, « logiciel d'annotation sociale à l'usage des lecteurs-interprètes ». Enfin, nous rendrons compte de l'usage de ce logiciel par des étudiants qui, en préparation de leur futur métier de consultant, analysent des retranscriptions d'entretiens effectués auprès de professionnels.

1. Travail intellectuel et pratiques d'annotation

Rich Gazan (Gazan, 2008) note que lorsque certains étudiants ont à choisir entre des livres neufs ou usagés, ils préfèrent souvent les plus anciens, en raison des annotations qu'ils y trouvent et qui les aident à lire et à apprendre. Pionniers du travail collaboratif assisté par ordinateur, Terry Winograd et Martin Röscheisen (Röscheisen *et al.*, 1995) furent les premiers à proposer une solution pour porter sur le Web ce type d'annotations par les lecteurs. ComMentor, leur système d'annotation sociale consiste en une architecture permettant de partager des annotations structurées sur le Web à travers un navigateur modifié ou un serveur « proxy ». Ces annotations relèvent de trois types : les commentaires (cf. Fig. 1) sont organisées en ensembles correspondant à des groupes donnés. L'utilisateur doit rejoindre un groupe pour pouvoir accéder aux annotations correspondantes. Les annotations de guidage de leur côté correspondent à un filtrage dans le but que le lecteur ne soit pas distrait par des annotations trop variées, en ne gardant que celles qui aident à la visite de documents sous un certain point de vue, par exemple « chronologique », « par points marquants », « régional », etc. de façon à n'avoir qu'une indication principale par page pour poursuivre la visite de page en page. Enfin les « tampons d'approbation » correspondent à des annotations d'évaluation apposables uniquement par un groupe autorisé, mais lisibles par tous.

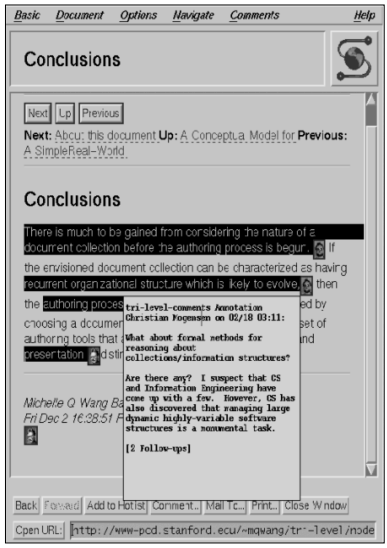


Figure 1 : Annotations de type « commentaire » (Röscheisen et al., 1995)

Par la suite, les travaux visant à étudier théoriquement l'annotation en rapport avec le Web, ont été nombreux, mais on peut regretter que peu aient donné lieu à des systèmes utilisés, en particulier par des groupes dans des situations réelles de travail, dans l'entreprise par exemple. De plus ces travaux ont été polarisés dans certaines directions. Certains proposent des solutions techniques pour une annotation individuelle, utilisant les standards du Web (Denoué & Vignollet, 2011). D'autres travaux, très nombreux, visent la structuration sémantique des annotations sur le Web, à partir de ressources terminologiques et ontologiques (Demontils & Jacquin, 2002) (Bringay et al., 2006) (Kahan et al., 2001), de Topic maps (Park & Hunting, 2002) ou d'ontologies sémiotiques et d'approches de « documents pour l'action » (Zacklad, 2005) (Lortal et al., 2006), pour ne citer que quelques exemples. Tous ces travaux ont en commun un aspect bien particulier que (Virbel, 1995) identifie comme la fonctionnalité de structuration (ou d'organisation), supposant la constitution de collection et mettant en jeu la composition d'entités textuelles « reposant sur un système de catalogage et au-delà, de représentation de connaissances encyclopédique ». Beaucoup de ces travaux sur l'annotation sémantique supposent une structuration de la représentation des connaissances selon un mode de que l'on pourrait qualifier de « descendant » (« top-down »).

Notons que, *a contrario,* les approches « ascendantes », à partir des annotations des lecteurs pour constituer (éventuellement) des structures émergentes non-préconçues, ont commencé à éveiller l'intérêt plus récemment quand le « Web 2.0 », le courant des folksonomies et du

social tagging, a fait la preuve de la possibilité de catégoriser des ressources du Web de façon collective et ascendante, tout en donnant lieu à des expériences et des travaux stimulants (Damianos *et al.*, 2006) (Millen *et al.*, 2006) (Fu *et al.*, 2010). La différence majeure avec l'annotation dynamique est que le social tagging, jusqu'à une date récente, est resté cantonné à la caractérisation de pages entières (ou de photos ou d'items globaux) repérées par des URL, comme dans le cas de FlickR ou de del.icio.us), alors que l'annotation vise par définition les fragments, à un niveau très fin. L'annotation d'un livre global ne se confond pas avec celle de ses passages. La catégorisation collective fine au niveau des fragments est un problème beaucoup plus difficile, qu'aborde justement l'annotation collaborative. Des progrès importants dans cette direction sont illustrés par certains systèmes récents de Blogs ou par le système SideWiki de Google, qui rajoute dans la barre d'outils un moyen de publier directement dans le volet latéral du navigateur des commentaires à propos de n'importe quelle page Web, de lire « en contexte » les commentaires d'autres utilisateurs de Sidewiki, et donc de discuter sur une page. Mais dans ces systèmes de commentaires collectifs, ceux-ci sont faiblement structurés (disposés l'un derrière l'autre, par exemple chronologiquement) et chaque commentaire n'est pas ancrés sur un fragment précis, désigné par exemple par « surlignage ».

Il est à noter que la structuration des annotations (unifiante ou non, descendante ou non) n'est qu'une des classes de fonctionnalités que Jacques Virbel identifie, parmi d'autres, comme utiles à la lecture dans le sens d'une « lecture expérimentale ». Toutes aussi utiles pour lui, sans qu'il soit besoin d'avoir l'obsession d'une structuration préexistant aux acteurs, sont les fonctionnalités davantage liées à la matérialité du texte et à l'acte de lecture : la fonctionnalité de marquage (balisage) permettant une structuration par l'auteur du texte pour dénoter par exemple les unités de la structure du texte, l'annotation proprement dite qui permet d'associer des caractérisations et des commentaires propres à un lecteur particulier, et la prospection, visant à exploiter les possibilités offertes pour réaliser des investigations fines dans le texte. Toutes ces fonctionnalités annotatives du modèle proposé par Jacques Virbel (appelé « MAPS », pour « Marquage, Annotation, Prospection, Structuration ») vont alors pouvoir tirer parti de l'informatisation pour permettre :

« une forme pérenne de mémorisation des travaux de lecture (y compris en distinguant diverses « campagnes » d'annotation réparties dans le temps) (…) pouvant supporter toutes sortes d'opérations (de consultation, de tri, de composition, etc.) » ;

une forme « d'opportunisme » où toute idée est immédiatement enregistrable ou effaçable

et enfin « une sorte de systématicité et d'exhaustivité, en général hors de portée dans le contexte papier, et donc la possibilité de formuler et de tester en temps quasi réel des hypothèses de toutes sortes qui restent autrement informulées ou invérifiables, et celle d'enregistrer le résultats de ces tests ».
Jacques Virbel définit ainsi une « lecture exploratoire ou expérimentale », et prévoit son instrumentabilité prochaine par un système informatisé.
Nous ne pouvons que nous inscrire dans ce modèle et faire écho à cette idée, en proposant (quinze ans plus tard) un système informatisé et opérationnel réalisant effectivement le passage de la théorie à la pratique pour un groupe, au moins pour l'essentiel de ce projet. Conformément au programme de J. Virbel, l'informatisation réalise l'identité du support du texte lu (et relu) et du texte rédigé, en correspondance avec la continuité lecture–écriture d'annotation, les deux grandes sources de nouveauté attendues de l'informatisation de l'annotation dans un cadre collectif. De ce modèle très complet, qui inventorie les nombreuse postures et figures de l'annotation dynamique pour caractériser des passages (hiérachiser, architecturer, contextualiser, programmer des actions) ou les attacher à d'autres éléments (reformuler, commenter, documenter, corréler), nous avons retenu en conformité avec ce modèle une division en deux groupes, les annotations catégorisantes et les annotations de commentaire. Notons que cette division recoupe des pratiques fort anciennes où l'annotation apparait comme une technique empirique de mémorisation et de capitalisation de résultats de lecture, au cœur du travail intellectuel déjà au Moyen Âge (Yates, 1966) (Carruthers, 2002). Ces deux grandes figures recoupent en effet deux mouvements opposés :
l'un allant de l'intérieur du texte vers l'extérieur (les « scholies », reprises dans la marge d'un terme pour en expliciter ou en discuter la signification),
l'autre de l'extérieur vers l'intérieur du texte (les « rubriques », étymologiquement « ce qui est en rouge », afin de faire ressortir visuellement des fragments du texte).
Ces techniques basiques de mémorisation, de capitalisation et de « rumination » des questions en jeu derrière le texte sont donc déjà là, et depuis longtemps ! Ce que le système informatisé va juste apporter en plus – mais ce n'est pas négligeable –, c'est la possibilité d'augmenter le potentiel de capitalisation, d'expression et de travail intellectuel des participants dans des groupes importants, en surmontant les limites mémorielles et spatiales (là encore, analysées par Virbel) concernant le support papier. Avec le papier, la mémorisation à long terme achoppe en effet sur la difficulté à capitaliser « temporellement » le très grand nombre d'annotations non systématisées et changeantes de trop nombreux participants, faute d'un système « d'additivité » des annotations ne détruisant pas la singularité de chacune. Tandis que du

point de vue spatial, les marges des ouvrages sont à l'évidence inaptes au travail collectif alors que les avancées actuelles des IHM permettent des dispositifs d'extension de la taille de la page tels que les volets additionnels, les ascenseurs, etc. libérant autant d'espace que souhaité, rendant visible les liens entre annotations, et manipulant facilement par des jeux de couleurs cohérents les « rubriques » ressortant du texte.

2. Lire entre les lignes avec LaSuli

Pour mettre en œuvre notre logiciel d'annotation sociale catégorisante, un certain nombre de choix de conception ont dû être effectués.
Tout d'abord, concernant la forme visuelle des annotations dans le texte, nous avons souhaité nous inscrire dans la tradition des « rubriques » et avons donc eu recours à l'usage de la couleur. Pour étendre le modèle à des catégories multiples et des fragments superposés, nous avons choisi la métaphore des « surligneurs ». Ce choix présentait cependant une petite difficulté afin de présenter ces fragments superposables sous forme de balises HTML (non-superposables). Un document pouvant présenter plusieurs centaines (voire milliers) de fragments, nous avons rapidement dû remplacer l'algorithme naïf défini au départ par l'algorithme suivant, plus efficace :
Supposons un ensemble de « surlignages »,

[{begin: 200, end:350, color:magenta},
{begin: 100, end:200, color:yellow},
{begin: 500, end:550, color:cyan},
{begin: 300, end:400, color:grey}]

Nous les indexons par position de début et de fin,
{ 100: {begin: [yellow]},
200: {begin: [magenta], end: [yellow]},
300: {begin: [grey]},
350: {end: [magenta]},
400: {end: [grey]},
500: {begin: [cyan]},
550: {end: [cyan]} }

Puis énumérons séquentiellement les positions en indiquant pour chacune la combinaison des couleurs actives.
{ 100: yellow,
200: magenta,
300: magenta+grey,
350: grey,
400: NONE
500: cyan,
550: NONE }

Le deuxième choix d'importance que nous avons dû faire concerne la manière de gérer visuellement le caractère « social » des annotations. En effet, comme l'illustre la figure 1, malgré la volonté de ses auteurs, il serait illusoire de montrer les détails de toutes les annotations réalisées par les lecteurs. Pour autant, nous ne souhaitions pas perdre la « sagesse des foules » chère au « Web 2.0 » (O'Reilly, 2005). Nous avons donc concilié dans la même interface :

- un onglet principal (cf. Fig. 3, section 4) correspondant à cette agrégation des points de vue portés sur le document par l'ensemble des lecteurs,
- des onglets, ouverts à la demande, permettant de ne visualiser qu'une seul point de vue à la fois.

Lorsque la vue correspondant à la « sagesse des foules » est active, les annotations de tous les lecteurs sont affichées de manière indifférenciée : un « nuage » agrège par nom les catégories mobilisées sur cette page et les passages analysés sont juste signalés dans le texte (en jaune). Au contraire, lorsque l'on choisit d'entrer dans « l'intelligence » d'une analyse, outre le filtre appliqué aux catégories et aux fragments, des couleurs différentes sont associées à chaque catégorie et aux fragments correspondants.

La liste des analyses pouvant être ouvertes posait une petite difficulté supplémentaire en termes d'interface homme-machine. En effet, le nom de l'analyse, choisi par le lecteur, n'était souvent discriminant que par rapport à ses propres analyses et non à celles des autres lecteurs. Nous avons donc préféré permettre l'ouverture des analyses à partir de la liste de leurs auteurs et du nuage des catégories qu'elles contiennent.

Un dernier point concerne l'affichage des fragments dans la marge. Nous avons préféré offrir, à l'extérieur du texte, une autre vue des fragments surlignés. Leur sélection (en vue de les supprimer ou de changer leur couleur) en est ainsi grandement facilitée. Mais, cela permet également, à condition de classer les fragments par catégorie, de faire émerger visuellement leurs points communs et leurs différences (cf. Fig. 2).

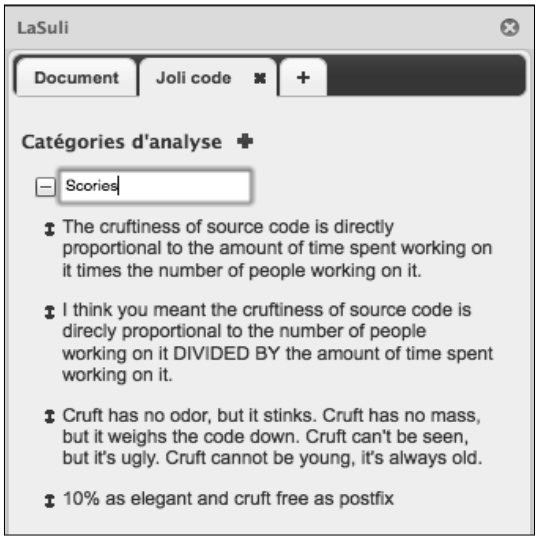


Figure 2 : Fragments classés par catégories (copie d'écran de LaSuli).

3. Illustration par les usages

Afin d'illustrer la manière dont Lasuli a été conçu pour accompagner le travail intellectuel et les pratiques d'annotation, nous proposons de présenter deux expérimentations réalisées avec des élèves ingénieurs en systèmes d'information (SI). L'évaluation d'outils de support aux activités coopératives est une question complexe, qu'il s'agisse d'intégrer la dimension coopérative à l'évaluation de l'utilisabilité (Greenberg *et al.*, 2000), ou de définir une démarche d'évaluation en relation avec l'activité ou la démarche de conception (Gauducheau *et al.*, 2005). D'autant plus dans le contexte du travail intellectuel que cherche à accompagner Lasuli où les critères d'efficacité et d'efficience classiques en évaluation de l'utilisabilité (ISO 9241-11) ne sont pas évidents à définir. Laissant l'évaluation de l'utilisabilité de Lasuli a des perspectives de recherche futures, les expérimentations que nous avons conduites s'inscrivent dans une perspective d'évaluation de l'expérience utilisateur (UX) (Hassenzahla et al., 2006). Le paradigme de l'expérience utilisateur propose d'aller au delà du caractère instrumental des outils (comme moyens efficaces ou pratiques de réaliser une tâche spécifiée) pour s'intéresser aux finalités des acteurs dans leur activité instrumentée, au ressenti et à la perception subjective des utilisateurs ainsi qu'au caractère situé de l'usage. Par ailleurs il s'agit de s'intéresser non seulement aux limites ou défauts de l'outil, mais également de mettre l'accent sur les aspects évalués positivement par les utilisateurs. L'intérêt de cette

approche pour les développements futurs de Lasuli se situe dans le retour d'expérience d'utilisateurs en situation de travail réelle en vue d'apprendre des usages des utilisateurs.
Lasuli a été utilisé par une centaine d'élèves ingénieurs (92 pour être précis) afin de les aider dans l'interprétation d'entretiens réalisés en vue de l'élaboration de modèles UML. Ces entretiens ont été conduits auprès de représentants d’organisations diverses (entreprises, associations, institutions) dans l'optique de recueillir des informations pertinentes pour modéliser le système d'information (SI) de leur organisation. L'analyse et la modélisation demandées comme livrable portaient essentiellement sur les flux de production et/ou de document au sein de l'organisation, ses produits, services et clients ainsi que la réalisation détaillée de certaines activités. Les étudiants ont été amenés à utiliser Lasuli dans le cadre de deux situations : l'une où plusieurs utilisateurs travaillaient sur une même retranscription mais où la construction de la grille d'analyse était laissée libre à l'utilisateur, l'autre où tous les utilisateurs partageaient une même grille d'analyse mais où chacun travaillait sur son propre entretien.
Les retours d'expérience des étudiants dans leur utilisation de Lasuli ont été collectés au moment des séances de TD par les encadrants et de façon plus systématique lors d'une séance de bilan à l'issue du module de cours par retours écrits et discussions. Dans ce qui suit nous présentons plus en détail chacune de ces expérimentations avant de présenter les résultats mis en lumière suite au retour fait par les utilisateurs en situation et lors de la séance de bilan des projets.

3.1. Un texte et de multiples grilles

À titre d'apprentissage et de familiarisation avec l'outil, les élèves ingénieurs ont utilisé Lasuli afin d'analyser un entretien de leur choix parmi le corpus utilisé dans le cadre du cours. Des consignes d'installation et de démarrage de Lasuli étaient à disposition sur un wiki et les séances se sont déroulées par groupe d'une vingtaine d'utilisateurs à la fois. Un même entretien pouvait ainsi être analysé par plusieurs étudiants en même temps ou d'une séance à l'autre par d’autres étudiants (cf. Fig. 3). L'objectif était de relever des informations pertinentes pour les aider à réaliser des modèles UML de l'organisation décrite au travers de l'entretien. Les catégories d'analyses étaient laissées à l'appréciation de l'utilisateur. La figure 1 montre les différentes catégories d'analyses établies par les quatre étudiants qui ont travaillé sur cet entretien.

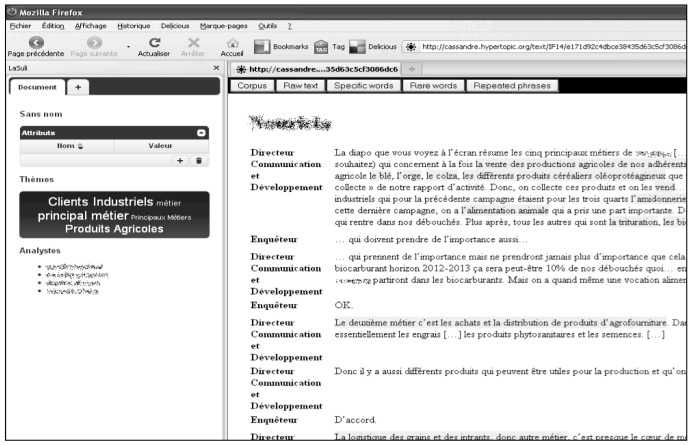


Figure 3 : Analyses concurrentes d'un même entretien (copie d'écran de LaSuli).

Plusieurs stratégies d'annotation et d'élaboration des catégories sont à disposition de l'utilisateur (cf. Fig. 4). Celui-ci peut définir ses catégories d'analyse à l'avance (approche « top-down ») dans la barre latérale pour annoter le texte. L'utilisateur peut également commencer par surligner des fragments dans le texte par un clic droit et définir le nom de la catégorie par la suite (approche « bottom-up »).

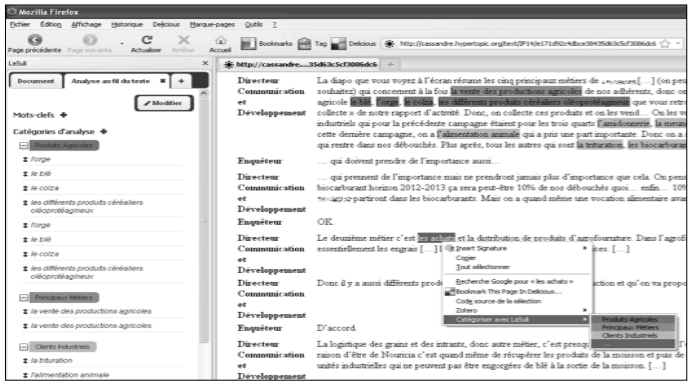


Figure 4 : « Surlignage » d'un entretien par un étudiant dans le cadre d'une analyse émergente (copie d'écran de LaSuli).

Dans l'analyse présentée (fig. 4), l'étudiant a mis en surbrillance les catégories « Produits Agricoles », les « Principaux Métiers » et les « Clients Industriels » qui montre une analyse influencée par le contexte de l'entretien portant sur une coopérative agricole. D'autres étudiants ont élaboré des catégories plus générales comme par exemple des catégories

« Services », « Résolution de problèmes », « Fonction », « Etape ». Ainsi, bien que les étudiants avaient accès aux analyses de l'ensemble des participants, les catégories proposées ont été différentes d'un étudiant à l'autre. Cette première utilisation de Lasuli a permis aux étudiants d'installer l'outil et de se familiariser avec sa manipulation avant de passer à l'analyse de leur propre entretien.

3.2. Une grille d'analyse et de multiples textes

Dans le cadre de la réalisation de leur dossier d'analyse, chaque élève ingénieur a été amené à analyser l'entretien réalisé par ses soins auprès d'une organisation. La grille d'analyse était cette fois imposée en regards des objectifs de modélisation établis pour le cours et l'analyse portait sur les « acteurs », les « flux » et les « actions » mentionnés par les interviewés (cf. Fig.5).

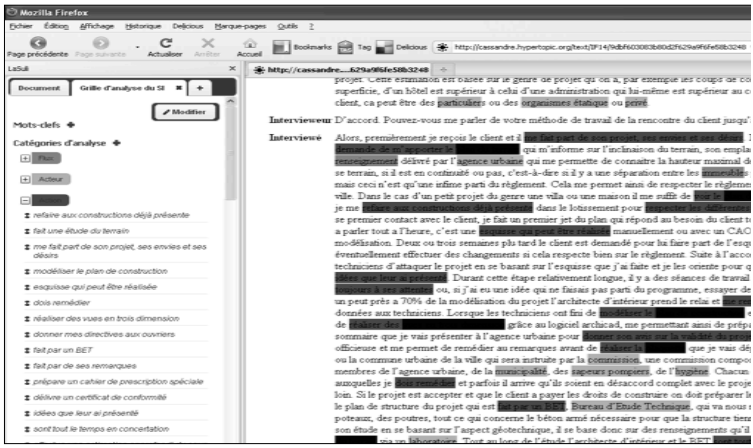


Figure 5 : Analyse réalisée par un étudiant à l’aide d’une grille imposée (copie d’écran de LaSuli)

La grille a été élaborée par les encadrants en vue de concentrer l’analyse sur les modélisations à réaliser. Les recouvrements entre des fragments faisant référence aux flux de production (ou de documents) de l’organisation avec les fragments mettant en évidence l’activité plus particulière de certains acteurs dans ce flux répond à la logique d’analyse portée par les objectifs de modélisation. Les fragments extraits permettent ainsi d’élaborer l’interprétation de l’entretien en vue de mettre en lumière différents aspects dans la densité du récit de l’activité fait par les professionnels, lequel ne s’inscrit pas de façon directe dans la logique analytique que doivent adopter les élèves ingénieurs pour leur activité de modélisation du système d’information de l’organisation.

3.3. Expérience utilisateur : quelques réactions positives et subjectives

Les retours d'expérience des étudiants au cours des deux expérimentations et plus formellement lors de la séance de bilan ont été positifs sur plusieurs points. Les participants ont trouvé que Lasuli était très adapté pour les aider à mettre en lumière les éléments mobilisables pour la modélisation dans leurs entretiens. La proximité avec leurs pratiques de lecture/analyse des entretiens sur papier où ils soulignaient et annotaient également des fragments de textes a certainement une part dans le fait que Lasuli ait été reconnu comme utile. Un étudiant insatisfait des temps de réponses de l'interface a même poussé l'analogie au point de faire une première passe de son analyse d'entretien à l'aide de surligneurs de couleurs différentes « à la façon de Lasuli », mettant bien en évidence la relation de Lasuli avec les pratiques de lecture « entre les lignes » demandée par le travail de modélisation. Autre résultat positif, plusieurs étudiants ont demandé à pouvoir utiliser Lasuli dans l'avenir notamment afin de les aider durant les analyses de documents qu'ils auront à faire en stage professionnel.

4. Conclusion

Dans cet article nous avons présenté l'importance et les enjeux des pratiques d'annotation dans les activités de travail intellectuel et les questions posées quant à l'instrumentation de telles activités, notamment dans le contexte des pratiques développées sur l'Internet. Nous avons présenté une proposition d'instrumentation de l'activité d'annotation catégorisante au travers de Lasuli en mettant en avant sa logique de conception et certaines de ses fonctionnalités caractéristiques. Les retours d'expérience positifs recueillis après l'expérimentation à grande échelle de Lasuli auprès d'élèves ingénieurs en systèmes d'information, nous incitent à continuer de perfectionner cet outil et à développer ses usages.

Les aspects coopératifs et de partage d'analyses entre utilisateurs possibles à l'aide de Lasuli on été peu mobilisés par les étudiant participant à l'expérimentation et un cadre d'observation et d'évaluation dédié reste à construire à cette fin. Les expérimentations réalisées ont par ailleurs permis de tester la maturité et la qualité technique de Lasuli et du serveur d'annotation support dans un contexte d'utilisation difficile en invitant plusieurs centaines d'utilisateurs à travailler avec nos outils. Le travail simultané de vingtaines d'utilisateurs a posé quelques problèmes de latence dans les réponses aux requêtes que nous œuvrons actuellement à résoudre. Quelques difficultés de configuration des paramètres de connexion vers le serveur d'annotation ont également été relevées au moment de l'installation. Les observations de l'utilisation de

la fonctionnalité de superposition d'annotations de différentes couleurs nous ont également interpellé quant à la nécessité de revoir la gestion des compositions de couleurs de façon à éviter que le texte soit rendu illisible par des couleurs trop sombres. Des perspectives d'amélioration sont déjà envisagées sur ce dernier point. Au final, le travail des étudiants a été mené à bien de façon très satisfaisante malgré ces désagréments mineurs. Nous projetons de renouveler l'expérience et de poursuivre l'évaluation de Lasuli au cours des semestres suivant avec d'autres groupes d'étudiants afin de continuer à perfectionner notre outil.

Au delà de la simple poursuite des expériences, l'évolution des pratiques de lecture numérique ces dernières années ouvre de nouvelles perspectives à nos outils. La généralisation de l’usage de nouveaux dispositifs de lecture tels que les tablettes et les livres électroniques pose la question de l'introduction d'un outil d'annotation sociale comme Lasuli dans ces dispositifs de lecture. Cependant le caractère propriétaire de la plupart des plateformes rend pour le moment cette intégration complexe et se présente comme une piste de développement à élaborer. Par ailleurs, la sélection de fragments sur des dispositifs de petite taille demeure un défi ergonomique.

Une autre perspective intéressante se situe dans l’extension de cette annotation catégorisante aux images, flux audios, vidéos, voire à des compositions multimédias interactives. Dans le cadre du serveur Steatite, nous savons déjà gérer des fragments d’images, cependant l’intégration avec LaSuli est assez complexe et reste largement à faire. L'extension à des contenus audio et vidéo nécessiterait, quant à elle, d’étendre la définition du fragment à un empan temporel et non plus seulement spatial. Au delà des défis technologiques posés par ces perspectives, l’enjeu est de pouvoir étendre les expérimentations à d’autres domaines d’expertise (critique de films ou d’art numérique par exemple) et tenter ainsi de mieux questionner ce qu’est l’interprétation.

Bibliographie

S. BRIET, *Qu'est-ce que la documentation ?*, Éditions documentaires et techniques, Paris.1951

S. BRINGAY, C. BARRY, J. CHARLET, Annotations: A Functionality to support Cooperation, Coordination and Awareness in the Electronic Medical Record., in *Proceedings of the 7th International Conference on the Design of Cooperative Systems COOP'06*, (Carry-le-Rouet, 9-12 Mai 2006), Hassanaly P., HERRMANN T, KUNAU G. et ZACKLAD M. (Eds.), p. 39-54

M. K. BUCKLAND, *What is a "document"?*, Journal of the American Society for Information Science, vol. 48, n°9, 1998. p. 804-809

M. CARRUTHERS, *Le livre de la Mémoire*, Coll. Argo, éditions Macula, Paris.2002

L. DAMIANOs, J. GRIFFITH, D. CUOMO, D. HIRSt, J. SMALLWOOD, Onomi: Social bookmarking on a corporate intranet, in *Collaborative Web Tagging*

Workshop at WWW2006, Collaborative Web Tagging Workshop at WWW 2006, Edinbhurg, 22th May 2006

L. DENOUE, L. VIGNOLLET, Personal Information Organization using Web annotation, In *Proceedings of WebNet 2001 - World Conference on the WWW and Internet*, Orlando, Florida, October 23-27, 2001, p. 279-283

E. DESMONTILS, C. Jacquin, Indexing a Website with a terminology Oriented Ontology, In *The Emerging Semantic Web*, IOS Press, p. 181-197

W-T FU., T. KANNAMPALLIL, R. KANG, J. He, *Semantic Imitation in Social Tagging*, ACM Transactions on Computer-Human Interaction, vol. 17, n° 3, 2010

W.-T. FU, W. DONG, From *collaborative indexing to knowledge exploration: A social learning model*, IEEE Intell. Syst., vol. 25, n°4 p. 15-23

N. GAUDUCHEAU, E. SOULIER, M. LEWKOWICZ, Design and evaluation of activity model-based groupware: methodological issues, in *Proceedings of 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Entreprises*, WETICE 2005, IEEE, 2005, p. 226-232

R. GAZAN, *Social annotations in digital library collections*, D-Lib Magazine, vol. 14, n°11/12, 2008

S. GREENBERG, G. FITZPATRICK, C. GUTWIN, S. KAPLAN, *Adapting the locales framework for heuristic evaluation of groupware*, Australian Journal of Information Systems, vol. 7, n°2, 2000, p. 102-108

M. GRUNDSTEIN, From capitalizing on Company Knowledge to Knowledge Management, in *Knowledge Management, Classic and Contemporary Works*, chapter 12, Daryl Morey, Mark Maybury, Bhavani Thuraisingham (Eds.), Cambridge (Mass.), The MIT Press, 2000, p. 261-287

M. HASSENZAHL, N. TRACTINSKY, *User experience – a research agenda*, Behaviour & Information Technology, Vol. 25, n°2, 2006, p. 91-97

J. KAHAN, M.-R. KOIVUNEn, E. PRUD'HOMMEAUX, R.R. Swick, Annotea: an open RDF Infrastructure for Shared Web Annotations, in *Proceedings of WWW10*, Hong- Kong, May 1-5, 2001, p. 623-632

M. LEWKOWICZ, G. LORTAL, A. TODIRASCU, M. ZACKLAD, M.F. Sriti, A web-based annotation system for improving cooperation in a care network, in *Engineering Advanced Web Applications, International Conference on Web Engineering*, ICWE 2004, Matera, M., Comai, S. (Eds.), Rinton Press, 2004, p. 227-239

G. LORTAL, M. LEWKOWICZ, A. TODIRASCU-COURTIER, AnT&CoW: Share, Classify and Elaborate Documents by means of Annotation, in *Proceedings of the IEEE - 1st International Conference on Digital Information Management*, ICDIM 06, Mathew T.C. and Pichappan P. (Eds.) Bangalore, India, December 6-8, 2006

N.W. LUND, R. SKARE, Document Theory, in *Encyclopedia of Library and Information Sciences*, Third Edition, vol. 1, 2010, p. 1632-1639

D. R. MILLEN, J. Feinberg, B. Kerr, Dogear: Social bookmarking in the enterprise, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, April 22-27, 2006, p 111-120

T. O'REILLY, What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, Journal personnel, 30 sept. 2005. http://oreilly.com/web2/archive/what-is-web-20.html

J. PARK, S. HUNTING, *XML Topic Maps : Creating and Using Topic Maps for the* Web., Addison-Wesley, MA, Boston.2002

F. RASTIER, *Communication, interprétation, transmission*, Semen, n°23, 2007, http://semen.revues.org/5341

M. RÖSCHEISEN, C. MORGENSEN, T. WINOGRAD*, Beyond browsing: Shared comments, SOAPs, trails, and on-line communities*, Computer Networks and ISDN Systems, vol. 27, n°6, 1995, p. 739-749

L. ROSENBLATT, The Transactional Theory of Reading and Writing, in *Theoretical Models and Processes of Reading*, 5/e edited by Robert B. Ruddell and Norman J. Unrau, international Reading Association, 1978

F. ROUSSEAUX, T. BOUAZIZ, *L'invention des Connaissances par les informaticiens : Déconstruction des Connaissances et proposition de dépassement par la notion de Collection*, in Revue Texto !, 2005. http://www.revue-texto.net/Inedits/Rousseaux/Rousseaux-Bouaziz.html

J. VIRBEL, *Annotation dynamique et lecture expérimentale : vers une nouvelle glose ?*, Littérature, n°96, 1995, p 91-105

F. A. YATES, *The Art of Memory*, Londres.1966, Trad. Française : *L'art de la Mémoire*, Paris, Gallimard.1966

M. ZACKLAD, Introduction aux ontologies sémiotiques dans le Web Socio Sémantique, in *Actes des 16èmes journées francophones d'Ingénierie des Connaissances*, Jaulent M.-C. (Ed.), Grenoble, PUG, 2005

M. ZACKLAD, Un cadre théorique pour guider la conception des collecticiels dans les situations de coopération structurellement ouvertes, *Psychologie Sociale Appliquée, Economie Médias Nouvelles Technologies*, Bonardi C., Georget P., Roland-Levy C., et Roussiau N. (Eds.), Paris, InPress, 2003. p. 135-164. http://zacklad.org/articles_communautes_action_cscw/cooperation%20structurellement%20ouverte%20livre.pdf

Partie 3 - Document participatif

Analyse exploratoire d'un wiki académique : le cas d'EFRARD

Kahina BELGAID
Equipe Index-Paragraphe
Université de Paris 8

1. Introduction

Depuis quelques années, la question de l'analyse des pratiques informationnelles suscite un intérêt croissant de la part de chercheurs en sciences de l'information et de la communication mais aussi d'autres disciplines, comme en témoignent notamment des travaux en psychologie cognitive, en informatique ou sociologie des usages. On parlera de « pratiques informationnelles » pour désigner la manière dont un ensemble de sources formelles ou non, d'outils, de compétences cognitives sont effectivement mobilisés dans les différentes situations de production, de recherche, et diffusion de l'information. (Chaudiron & Ihadjadene, 2010). Ces pratiques se diversifient avec l'apparition des dispositifs du Web 2.0 (Wiki, réseaux sociaux, CMS, logiciels de filtrage…) qui autorisent de nouvelles modalités de production et partage de l'information en fusionnant les fonctionnalités de recherche , d'édition et de communication.
Les individus, mais aussi les organisations, mettent aussi en œuvre des stratégies ou des politiques pour faciliter l'échange d'information et sa communication via des plateformes d'intermédiation (Wiki, Intranet, réseaux sociaux, etc.). Dépassant leur statut de simples récepteurs, les usagers jouent désormais un rôle actif dans l'organisation et l'évaluation de l'information. Comme le rappellent (C.Roth, Taraborelli et Gilbert, 2006), les wikis constituent sans doute un des exemples les plus saillants des systèmes de construction collective de contenus. Ces auteurs soulignent l'imbrication qui existe entre la croissance de la population et la croissance du contenu informationnel des wikis.
Les wikis sont caractérisés par la liberté d'écriture et de consultation, permettant la collaboration et le processus de partage de connaissances dans les organisations. Des études ont montré que, les wikis sont de plus en plus répandus dans les entreprises et les organisations (Danis &Singer, 2008), (Majchrzak et al, 2006). Selon ces travaux, le concept est assez bien accueilli par les professionnels en entreprise et dans le domaine de l'éducation mais le nombre d'utilisateurs actifs est moins important que le nombre d'utilisateurs passifs, les utilisateurs les plus jeunes sont les plus fidèles. Le domaine d'activité et la profession des usagers sont deux

facteurs primordiaux qui explicitent en partie le degré d'appropriation de ces dispositifs. Ainsi, la fréquence la fréquence d'utilisation des Wikis est plus important par exemple dans le domaine des technologies de l'information. La taille de l'entreprise est un autre critère distingue les différentes stratégies d'appropriation. Ainsi, les plus grandes structures utilisent plus les wikis pour plusieurs raisons, en l'occurrence, le manque de communication entre les différents employés parce qu'ils se connaissent pas ou pour la distance physique qui les sépare.
Contrairement à Wikipédia dans laquelle on trouve toute sorte d'information sur n'importe quel sujet, les wikis dans les organisations peuvent traiter un nombre restreint de thèmes. L'anonymat des utilisateurs (Danis &Singer, 200!) est un autre facteur qui différencie lew Wikis d'entreprises par rapport aux wikis grand public. Dans une organisation, les utilisateurs sont parfois réticents à l'idée qu'un des collègues critique son travail et d'autres sont au contraire très motivés car se voient améliorer leur réputation en apportant des solutions aux problèmes.
L'objectif principal de cette communication est d'analyser les pratiques informationnelles d'universitaires sur différentes plates formes communautaires sur le Web, en tenant compte des différents profils utilisateurs ainsi que les situations qui peuvent de prés ou de loin influencer leurs comportements quant à ces réseaux virtuels. Le choix s'est porté sur le réseau EFRARD qui est une plate forme communautaire francophone pluridisciplinaire, pour la recherche et le développement. Ce WIKI a aussi pour but de renforcer la coopération scientifique internationale, il compte des membres d'au moins 20 pays différents dans le monde, principalement chercheurs ou universitaires.

2. Etude empirique

2.1. Terrain et contexte

Le but de nos recherches c'est d'étudier les pratiques info-communicationnelles dans le travail collaboratif et l'utilisation des réseaux sociaux, wikis et sites de partage scientifique ou d'information au sens large. Dans ce cas présent, avec l'enquête EFRARD, on s'intéresse tout particulièrement à étudier les différentes ressources collaboratives où un public ciblé (dans notre cas des universitaires) sélectionne son information ainsi que son rapport avec de ces ressources à savoir : apports informationnels, utilisation, modes d'interrogation, facteur de choix (fiabilité, confiance, gratuité…).
Pour rappel EFRARD est un dispositif créé à l'initiative d'un groupe de chercheurs à l'université Paris 8 (Kamga&Zreik, 2009), pour renforcer la communication et le partage de connaissances au sein de la communauté scientifique francophone.

Notre enquête a été réalisée en trois étapes principales. Dans la première, on s'est concentrés sur l'étude de l'outil, sur ses fonctionnalités et la communauté, afin de mieux cerner l'échantillon d'individus sur lequel on allait travailler. L'étape suivante, on a réalisé des interviews sur une sélection de participants. Les interviews ont été réalisées soit par vidéo conférence, soit par emails interposés pour les membres résidents dans des pays avec un grand décalage horaire avec Paris. Nous avons réalisé en tout 76 interviews auprès de 36 participants.

2.2. Analyse des résultats

A- Sources d'information

Les dispositifs utilisés comme moyens d'accès à l'information sont divers. Néanmoins, la place très importante d'Internet, celle aussi de la communication avec des collègues, peuvent être retenues : il est clair que la recherche d'information est loin de passer exclusivement par des moyens d'accès formels et validés par une structure traditionnelle. Le premier moyen d'accès à l'information est Internet : 70,89 % y ont fréquemment ou toujours recours lors de leur recherche d'information. 84,61% répondent qu'ils interrogent les moteurs de recherche sur le Web, principalement Google, 38,46% interrogent des bases de données telles que la BNF ou les catalogues universitaires. Viennent ensuite les réseaux sociaux et les wikis. Près de 15% des participants citent EFRARD parmi leurs ressources principales.

L'usage de l'émail demeure le moyen de communication le plus utilisé. 82,26% de la population étudiée travaillent en collaboration, les moyens de communication qu'ils choisissent pour ce faire sont souvent les mêmes : la messagerie (eMails, messageries instantanées, Chat…), le face à face et le téléphone, avec 84,61% pour la messagerie, et 53,48% pour le face à face et le téléphone. Les participants utilisent beaucoup les forums aussi pour collaborer avec un pourcentage de 46,15, suit derrière les réseaux sociaux (Facebook, Linkedin, Viadeo…) avec juste 30,76% et sont très rares les personnes qui collaborent via des wikis.

Dans une étude antérieure, B. Evans, S. Kairam et P. Pirolli[12] (2009) ont défini trois pratiques collaboratives pour la collecte d'information *directed asking, public asking* et *researching*. Après avoir observé différents comportements d'utilisateurs, qui pour avoir l'information, envoyaient des mails ou interrogeaient directement leurs amis en utilisant la messagerie instantanée (*directed asking)*, puis s'adressaient à un groupe d'individus à la fois *(public asking*, Elle s'applique aussi sur l'utilisation des forums de discussion). Et enfin interrogent des moteurs de recherche a été catégorisée dans *Researching*. Néanmoins, Ils ont prouvé que la

[12] Brynn M.EVANS, Sanjay KAIRAM, Peter PIROLLI (2009). Exploring the Cognitive consequences of Social Search, Student Research Competition.

combinaison de ces trois activités est beaucoup plus productive pour un processus de recherche.

B- Utilisation d'EFRARD

Bien que notre échantillon d'individus soit composé de membres d'EFRARD, la fréquence de leurs utilisations de cette dernière est très réduite. Il n'y a que 7,7% des membres qui vont fréquemment sur EFRARD, 30,76% n'y vont presque jamais. Les moteurs de recherche, les bases de données et les réseaux sociaux les plus connus sont les plus fréquentés avec plus de 85% des participants qui affirment s'y rendre tous les jours.

Neuf des personnes interrogées au cours de notre étude n'ont pas réussi à définir le degré de fiabilité et le degré de facilité d'EFRARD car elles ne l'utilisent pas assez comme source d'information pour pouvoir donner un avis. 27 participants sur les 36 trouvent que c'est une ressource fiable, cependant, pas très facile d'utilisation à 88,88%. Parmi les raisons qui laissent les utilisateurs réticents envers EFRARD c'est la liberté d'accès et le manque de contrôle et de structure des contenus, la première raison est d'ailleurs le plus important car ils n'arrivent pas à faire complètement confiance au wiki et le considérer comme ressource fiable à 100%. Un des points négatifs d'EFRARD, selon les utilisateurs, c'est aussi le nombre insuffisant de membres actifs de la plate forme, la majorité des participants choisissent une plate forme collaborative par rapport à la quantité d'information qu'elle contient, et un faible pourcentage sur l'étendue de sa communauté. Ils tiennent compte également de l'information après utilisation.

C- EFRARD et le partage de connaissances

Après analyse statistique de tous les résultats, on n'a pas trouvé un lien direct entre l'utilisation d'EFRARD l'utilisation d'une autre ressource. Cependant, il en existe un avec la raison pour laquelle les individus on rejoint la communauté, ce qui est visible encore plus entre cette motivation et leur mode d'utilisation de la plate forme. 92,30% des utilisateurs affirment que ce qui les a motivé pour faire partie de la communauté EFRARD c'est de tisser des liens et rester en contact avec des collègues et la communauté scientifique francophone au niveau national et international et rester au courant des évènements scientifiques organisés, l'activité de cette communauté en l'occurrence « le forum francophone pour la recherche et le développement », contre 15,38% qui sont motivé par le partage de connaissance scientifique via le wiki. Ce qui ressort dans l'étude des usages quand on observe un pourcentage de 69,07 de membres passifs, qui ne font que consulter les pages.

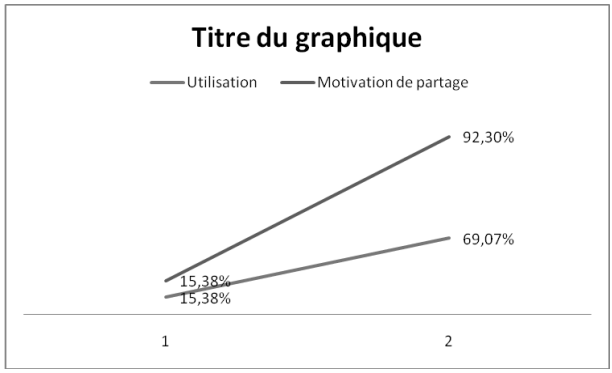


Fig 2. Le lien entre l'utilisation active d'EFRARD et le facteur partage de connaissance via le Wiki

3. Conclusion et discussion

Cette étude exploratoire nous a permis de répondre à pas mal de nos interrogations, on en déduit que le profil utilisateur n'a pas beaucoup d'influence sur l'utilisation du wiki, les enseignants chercheurs qui forment la plus grande partie de notre échantillon d'individus, travaillent en collaboration mais utilisent souvent des moyens de communication et de partage assez basiques comme le téléphone, les vidéos conférences et les eMails, ils suivent le même comportement que les internautes de façon plus large, c'est-à-dire, utilisent les plates formes les plus populaires et les plus connues sur le Web telles que les réseaux sociaux, les forums et les wikis. Cependant, ils ont une approche différente des wikis, qu'ils utilisent avec une certaine retenue, ils restent toujours fidèles aux banques de données fiables, aux récits et écrits d'auteurs non anonymes et préfèrent dans l'idéal collaborer avec des personnes qu'ils connaissent. Dans un avenir proche, nous envisageons d'explorer les usages d'une nouvelle communauté, dans le domaine professionnel de l'entreprise ou dans un environnement médical.

Bibliographie

EVANS B., KAIRAM S., PIROLLI P. (2009). Exploring the Cognitive consequences of Social Search, Student Research Competition.
PIROLLI F(. 2010). Web 2.0 et pratiques documentaires. Les Cahiers du numérique (Vol. 6), p. 81-95.
CHAUDIRON S., Ihadjadene M. (2010), « De la recherche de l'information aux pratiques informationnelles », in Études de Communication n°35, CEGES, Décembre 2010, p. 13-29.

DANIS C , Singer D. (2008). A wiki instance in the enterprise: opportunities, concerns and reality, Proceedings of the 2008 ACM conference on Computer supported cooperative work, November 08-12, 2008, San Diego, CA, USA

KAMGA R., ZREIK K. (2009) "Les nouveaux enjeux de la mise en valeur du Patrimoine scientifique et technique de la recherche dans l'espace Francophone". In Patrimoine 3.0 : Actes du douzième colloque international sur le document électronique (CIDE.12). Ed. Europia, Paris, 2009.

GIMAZANE Ret al. (2010). Nouveaux documents, nouvelles compétences. Documentaliste-Sciences de l'Information (Vol. 47), p. 56-67.

ROTH C., Taraborelli, D., Gilbert, N. (2008). Démographie des communautés en ligne: le cas des wikis. Réseaux. 205-240

SPENCE P., REDDY, M., and HALl R. (2005). A Survey of Collaborative Information Seeking Practices of Academic Researchers. In Proc. of ACM Conf. on Supporting Group Work (Group'05). Sanibel Island, Fl. Nov 6-10. pp. 85-88.

MAJCHRZAK A , WAGNER C , Yates D., 52006) Corporate wiki users: results of a survey, Proceedings of the 2006 international symposium on Wikis, August 21-23, 2006, Odense, Denmark

Les références bibliographiques dans Wikipédia

Gilles SAHUT
Doctorant, LERASS, Université Paul Sabatier, Toulouse 3.

Résumé : La crédibilité de Wikipédia repose en grande partie sur le référencement de ses contenus. Les articles de cette encyclopédie collaborative intègrent donc des références bibliographiques. Nous proposons ici une synthèse des travaux sur l'usage de ces références pour la recherche d'information et son évaluation. Nous récapitulons également les résultats des recherches portant sur le nombre et la nature des sources citées dans l'encyclopédie. Leur grande diversité suscite des interrogations quant à l'existence d'une hiérarchie documentaire sur Wikipédia. Nous proposons alors des pistes de recherche visant à mieux comprendre les spécificités de ce modèle éditorial. Comment les règles relatives au référencement des sources sur Wikipédia évoluent-elles ? De quelle manière la communauté s'approprie-t-elle ces règles et les applique-t-elle ?

Mots-clés : Wikipédia, référencement, sources d'information, jugement de crédibilité, recherche d'information.

Toute entreprise encyclopédique a pour objectif d'élaborer une synthèse des savoirs existants à des fins didactiques et suppose, par là même, la reformulation d'informations. Ce processus est à l'œuvre dans Wikipédia. Cependant, les conditions d'élaboration de cette encyclopédie collaborative sont spécifiques. Elle est souvent définie comme l'encyclopédie à laquelle *« tout le monde peut participer »*, ce qui génère, depuis le milieu des années 2000, de nombreux débats sur sa crédibilité. Face à cela, l'encyclopédie énonce le principe de vérifiabilité comme l'une des règles inhérentes à son écriture et son édition : « *Une information ne peut être mentionnée que si les lecteurs peuvent vérifier qu'elle a déjà été publiée par une source ou référence de qualité* » [1]. En ce sens, on peut considérer que les références bibliographiques[13], figurant en notes ou à la fin d'un article de

[13] Nous considérons ici qu'une référence bibliographique est *« [l']ensemble des éléments de données nécessaires pour identifier un document ou une partie de document de tout type, sur tout support (livre, article, site web, etc.)* » Boulogne, Arlette (coord.). *Vocabulaire de la documentation*. ADBS Éditions, 2004

Wikipédia, constituent des indices – voire des preuves – d'une reformulation de sources publiées. Or le paysage informationnel actuel est caractérisé par la multiplication du nombre de documents accessibles et leur extrême hétérogénéité. Le développement du web - et plus encore du web social – a favorisé la production de sources autoritatives résultant d'un processus d'auto-publication et émanant d'un auteur qui construit *« lui-même les conditions de sa reconnaissance dans l'univers électronique »* [2] ; ces documents cohabitant avec d'autres, issus de modèles éditoriaux plus traditionnels. Ce constat suscite donc une interrogation sur la nature et les caractéristiques des sources citées dans Wikipédia, l'encyclopédie étant parfois accusée de reposer sur des sources douteuses[14].
Afin de souligner l'intérêt de cette thématique, les usages des références bibliographiques faits par les lecteurs de cette encyclopédie seront tout d'abord analysés. Nous synthétiserons ensuite les recherches récentes portant sur le nombre et la nature des sources citées dans Wikipédia. Nous proposons enfin des pistes de recherche suivies actuellement dans le cadre d'un doctorat en sciences de l'information et de la communication.

Les usages des références disponibles sur Wikipédia

Les mesures d'audience sur le web indiquent que Wikipédia est actuellement l'un des dix sites les plus visités au monde[15]. Cette notoriété se retrouve chez les jeunes, population dont les usages numériques sont les plus étudiés. Toutes les enquêtes sont unanimes, Wikipédia est devenue une source de premier plan pour la grande majorité des élèves et des étudiants, que ce soit pour leurs recherches personnelles ou dans le cadre de leur travail scolaire ou universitaire [3] [4] [5]. Elle est majoritairement exploitée lors de la phase initiale de la recherche d'information, sa consultation paraissant nécessaire afin de cerner le sujet de la recherche et de comprendre les termes qui le composent [4] [5] [6]. Ainsi il est particulièrement fréquent que les jeunes activent les liens hypertextes présents dans la bibliographie ou les notes des articles de Wikipédia. Grâce à ces références, ils accèdent à d'autres documents en ligne jugés pertinents pour leur recherche [5] [7] [8] [9]. On retrouve donc chez les jeunes une stratégie particulière de recherche

[14]L'une des premières polémiques autour de la Wikipédia française est née à la fin de l'année 2006 d'un problème lié à une référence présente dans l'article consacré à l'affaire Dreyfus. Un livre violemment antidreyfusard figurait dans la bibliographie accompagné d'une recommandation « *Ouvrage fondamental à consulter en priorité* ». Plusieurs journalistes (Daniel Garcia puis Pierre Assouline notamment) s'en sont indignés sur leurs blogs, générant ainsi des débats sur la fiabilité de cette encyclopédie.
[15] Selon Alexa, Wikipédia.org figure au 7e rang des sites les plus visités au monde en juin 2011.

d'information, le « chainage » qui avait été repérée par Ellis chez les chercheurs en sciences sociales à la fin des années 1980 [10]. Ainsi Wikipédia n'est pas seulement une ressource encyclopédique, elle peut être considérée comme une passerelle vers d'autres documents. En un sens, elle fait office de répertoire de ressources sur un thème alors que la plupart des annuaires généralistes « historiques » du web ont disparu.

Les références bibliographiques ne sont pas utilisées dans le seul but de repérer de nouveaux documents. Elles sont également susceptibles d'être examinées afin de formuler un jugement de crédibilité[16]. Les études témoignent sur ce point de pratiques évaluatives qui diffèrent. Certaines indiquent que rares sont les jeunes qui les prennent en compte afin d'évaluer le degré de crédibilité des articles de Wikipédia [6] [12].

Celle de Lim et Simon souligne que le thème de la recherche d'information est susceptible d'influencer cette pratique : quand celui-ci est jugé sérieux (par exemple les questions portant sur la santé), une certaine attention serait accordée aux références afin de formuler un jugement de crédibilité. En revanche, pour des sujets relatifs aux loisirs, les références ne seraient pas prises en compte. Néanmoins, selon l'aveu même des auteurs, ce résultat mérite confirmation [9]. La prise en compte des références semble également dépendre du degré de confiance accordé à Wikipédia.

D'après une expérimentation menée en psychologie cognitive, le nombre de références présentes dans un article de Wikipédia a une incidence positive sur sa crédibilité quand les personnes accordent un haut degré de confiance à l'encyclopédie collaborative. Celles qui doutent de la fiabilité de l'encyclopédie collaborative sont davantage attentives à la qualité des références [13]. Sur un autre plan, la participation au projet Wikipédia et la connaissance de son mode de fonctionnement paraissent avoir un impact net sur les pratiques évaluatives. Ainsi les jeunes contributeurs à la version chinoise de Wikipédia accessible à Hong-Kong réussissent à formuler des jugements de crédibilité argumentés et informés même s'ils disposent de peu de connaissances sur le thème de l'article. Ils s'appuient alors sur les indices périphériques tels les références et les liens externes qui sont disponibles [14].

Les études recensées convergent donc sur un point : lors de la consultation des articles de l'encyclopédie, les usagers –notamment les jeunes- exploitent fréquemment les références citées quand elles pointent vers des documents en ligne. Il arrive aussi que celles-ci soient analysées

[16] Mobilisé dans une pluralité de discipline, le concept de crédibilité fait l'objet de définitions et d'approches multiples. Au sein du contexte des études sur la recherche d'information, il est possible de considérer le jugement de crédibilité comme une composante du jugement de pertinence. Il comporte principalement deux dimensions : la perception de l'expertise de la source et celle de sa bienveillance (ou son intention de dire la vérité). De multiples variables entrent en compte dans sa formulation : les connaissances de l'individu, son expérience, la situation dans laquelle il se trouve [9] [11].

à des fins d'évaluation de l'article. Néanmoins, la fréquence de cette pratique et les facteurs la conditionnant ne sont pas encore connus avec précision. La constitution de listes bibliographiques placées en notes ou à la fin des articles de Wikipédia permet donc d'attribuer une nouvelle visibilité aux documents cités et les rend disponibles pour de nouveaux usages, ce qui suscite alors une interrogation sur le nombre, les caractéristiques et la qualité de ces sources.

Nombre et type de sources référencées sur Wikipédia

L'étude des références dans Wikipédia s'inscrit dans un débat scientifique plus général : celui de l'évaluation de la qualité informationnelle de l'encyclopédie. Depuis les articles de Liu [15] et de Giles [16], les recherches scientifiques sur ce sujet empruntent principalement deux voies. La première méthode est celle de l'analyse d'un corpus d'articles de Wikipédia par des experts dans un domaine donné. Ceux-ci ont alors recours aux critères d'évaluation de l'information habituellement usités dans le cadre des sciences de l'information (exactitude, lisibilité, complétude…)[17]. Une approche comparative avec d'autres sources est parfois mise en œuvre. La seconde approche vise à identifier des variables (nombre de contributeurs à l'article, nombre d'éditions de l'article, profil des contributeurs …) afin de proposer des modélisations statistiques permettant de prédire la qualité d'un article[18]. L'analyse des références citées dans Wikipédia[19] s'inscrit plutôt dans la première approche, ce qui n'exclut pas le recours à un outillage statistique.

En 2007, l'étude exploratoire de Willinsky réalisée à partir d'un corpus de 100 articles choisis de manière aléatoire indique une moyenne de 1,68 références par article. Ce chiffre peu élevé pouvant en partie s'expliquer par le caractère aléatoire du choix des articles. Alors que les encyclopédies généralistes traditionnelles sont centrées sur des thèmes académiques, Wikipédia est ouverte à la culture populaire : l'échantillon de Willinsky comprend ainsi tout une gamme de sujets (personnages de bandes dessinées, émissions télévisées, albums de groupes de rock plus ou moins connus…) qui n'ont pas fait l'objet d'une consécration académique.

Il est possible que pour de tels sujets, les contributeurs se fondent uniquement sur leurs connaissances personnelles et ne jugent pas nécessaire d'indiquer leurs sources. Willinsky relève cependant que des entrées plus classiques comportent également un nombre réduit de

[17] Par exemple [16] [17] [18]

[18] Par exemple [15] [19] [20]

[19] Précisons que les études citées dans cette communication porte sur la Wikipédia en langue anglaise à l'exception de celle d'Huvila qui porte sur la Wikipédia en langue suédoise 2010 [26].

références [21]. L'étude réalisée par le même auteur l'année suivante offre un point de vue radicalement différent. Elle met en évidence que les éditeurs de l'encyclopédie collaborative ont intégré un nombre élevé de références pointant vers des écrits scientifiques numériques en accès libre (5000 liens vers arXiv.org, 1000 vers Social Science Research Network) mais aussi vers des banques de données partiellement ou totalement payantes (2400 liens vers Science Direct, 4500 vers JSTOR…). Dans le domaine philosophique, beaucoup d'articles comportent des références à la *Stanford Encyclopedia of Philosophy,* ressource gratuite en ligne élaborée par des universitaires selon un modèle scientifique. Par ailleurs, les textes de cette encyclopédie sont également mentionnés dans les pages de discussion accompagnant des entrées philosophiques de Wikipédia, nourrissant d'intenses échanges argumentatifs entre contributeurs [22].

Pour ce qui est des sciences exactes, Nielsen démontre à partir de méthodes statistiques que les articles de ce domaine de connaissance sont majoritairement dotés de références bibliographiques de travaux de recherche ; ce qui n'exclut pas la présence de sources non scientifiques tels des quotidiens d'actualité. Il relève également que le nombre de ces références scientifiques a tendance à augmenter au fur et à mesure de la croissance de l'encyclopédie et que les revues les plus citées sont celles qui, à l'instar de *Nature* ou *Science*, bénéficient d'un facteur d'impact élevé dans le *Journal Citation Reports*. D'après ces éléments, on pourrait considérer Wikipédia comme une source de bonne qualité dans le domaine scientifique [23]. En 2010, Haigh a recensé et caractérisé les références de 50 articles sur des thèmes relatifs aux sciences médicales et à la santé. En moyenne, chaque entrée comportait 50 références. Plus de la moitié de celles-ci indiquaient des sources jouissant d'une bonne réputation parmi les spécialistes du domaine.

Selon l'auteur, il est fort probable que, dans certaines disciplines telles l'anatomie et ou la physiologie, les sources mentionnées dans Wikipédia soient de nature plus scientifiques que celles utilisées pour la réalisation de certains manuels universitaires [24]. Toutefois, ces conclusions ne peuvent être généralisées. A l'issue d'une analyse d'un corpus d'articles sur l'histoire des pays, Luyt et Tan concluent que Wikipédia ne devrait pas être utilisée pour la recherche de références.

Outre le fait que leur nombre soit jugé peu élevé (en moyenne, 10,16 références par articles), les auteurs soulignent un décalage avec les pratiques de la publication scientifique dominante en histoire : alors que dans cette discipline, l'écrit de recherche est encore essentiellement diffusé sur support imprimé, les sources citées dans Wikipédia sont majoritairement issues du web. Par ailleurs, ils notent l'importance des citations de médias d'actualité en ligne (11,9% du total des références), ainsi que celle des sources relevant de sites gouvernementaux des différents pays (7,3%), de l'Etat fédéral américain ou d'institutions qui lui

sont liées[20] (9,8%) ; ceci ne correspond évidemment pas aux canons de l'édition savante. Luyt et Tan relèvent enfin que l'écrasante majorité des références est en langue anglaise (91%) alors que les articles analysés traitaient de sujets internationaux. Dès lors, des interrogations surgissent quant à la crédibilité et la neutralité des sources utilisées et, par ricochet, quant à celles des articles concernés. Le risque est ici une absence de pluralisme et l'imposition d'un point de vue nord-américain sur les histoires nationales [25].

Huvila a abordé la question de la référence au sein de Wikipédia sous un autre angle. Il s'est intéressé non pas à l'encyclopédie en tant que produit éditorial mais à ses contributeurs[21] et aux sources qu'ils utilisent afin de participer à l'écriture d'un article. Peu étonnant au regard des résultats précédemment évoqués, la majorité des informations sont issues du Web, les « Wikipédiens » faisant, comme l'ensemble des internautes, un fort usage des moteurs de recherche et en particulier de Google. Toutefois, l'éventail des sources exploitées - citées et/ou reformulées - s'avère large (livres parfois trouvés dans des bibliothèques, presse en ligne, revues scientifiques, articles de Wikipédia en d'autres langues, connaissances personnelles qui sont mobilisées sans faire l'objet de références formelles…).

Cette diversité documentaire peut se comprendre par les profils différenciés des participants. En effet, Huvila identifie cinq types de contributeurs :

- les « *investigateurs* » qui mènent des recherches approfondies sur des thèmes correspondant à leurs centres d'intérêt à partir de médias d'actualité, documents vulgarisés ou factuels ;
- les « *surfeurs* », soucieux d'économiser leur temps et leurs efforts qui fondent leurs contributions sur des sources facilement repérables sur le web ;
- les « *matérialistes prudents* » qui s'appuient prioritairement sur les connaissances issues de leurs propres expériences ;
- les « *scientifiques* », souvent des doctorants ou des jeunes chercheurs qui proposent des états de la question fondés sur des écrits de recherche ;
- les « *éditeurs* » qui ne recherchent pas d'informations pour leur contribution, se concentrant prioritairement sur les tâches administratives, les corrections grammaticales, la traduction d'articles de Wikipédia en langue étrangère [26].

[20] Le Département d'Etat, la CIA ou encore le site Country Studies réalisé par un département de la bibliothèque du Congrès et parrainé par le Ministère des armées des Etats-Unis.

[21] Suédois en l'occurrence.

Pistes de recherche

La nature et la qualité des références trouvées dans Wikipédia sont donc diverses. Elles varient selon les thématiques et les domaines de connaissance abordés mais également dans le temps, Wikipédia se caractérisant par une instabilité temporelle [27]. A l'instar d'autres types de publications [28], l'encyclopédie constitue un espace de médiations hybrides qui offre un enchevêtrement de références documentaires aux statuts différents. Ce constat peut entretenir des doutes sur la crédibilité de l'encyclopédie.

Il suscite parallèlement des interrogations sur les évolutions passées et à venir de cette entreprise éditoriale. Une hiérarchie des sources se dessine-t-elle au sein de la communauté wikipédienne? Le modèle éditorial de Wikipédia évolue t-il afin de se rapprocher des standards académiques en vigueur pour l'évaluation des sources utilisées et citées? Ou reste-t-il marqué par un principe égalitaire pouvant impliquer des appréciations indistinctes de la valeur documentaire?

Wikipédia peut être qualifiée d'espace documentaire participatif et évolutif dans lequel la communauté participante établit, selon des principes démocratiques, ses propres règles de fonctionnement et les recommandations à suivre lors de l'élaboration des articles [29]. Le dispositif technique sur lequel s'appuie l'encyclopédie comporte un système d'archivage qui garde trace des débats accompagnant les diverses prises de décision communautaires. Dès lors, il devient possible de mener une étude diachronique afin d'identifier les transformations des règles relatives à la pratique citationnelle, de repérer leurs moments d'émergence, de consolidation, voire de remise en question et de tenter de cerner les facteurs qui ont présidé à leur évolution.

De manière complémentaire, les modalités d'appropriation de ces règles par les contributeurs de l'encyclopédie méritent d'être étudiées ainsi que les éventuelles résistances à leur application, leurs détournements ou leurs contournements. Il s'agira alors d'examiner les fonctions assignées aux références bibliographiques par les participants à l'encyclopédie dans le cadre des discussions accompagnant les éditions successives des articles. Nous entendons là analyser la manière dont les différentes sources citées sont qualifiées, discutées, différenciées afin de discerner si une hiérarchie documentaire émerge au fil du temps et réussit à s'imposer au sein de la communauté «wikipédienne».

Bibliographie

[1]. Wikipédia (2011). Wikipédia:Vérifiabilité. http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:V%C3%A9rifiabilit%C3%A9. Consulté le 10/06/2011.

[2] BROUDOUX E., GRESILLAUD S., LE CROSNIER H. & LUX-POGODALLA V. , Construction de l'auteur autour de ses modes d'écriture et de publication. In Actes Conférence Hypertextes Hypermédias H2PTM'05 : Créer, jouer, échanger: expériences de réseaux. Hermès, 2005.
[3] LUCKIN R., LOGAN K., CLARK W., GRABER R., OLIVER M., & MEE A., Learners' use of Web 2.0 technologies in and out of school in Key Stages 3 and 4: reseach report. Becta, 2008.
[4] HAMPTON-REEVES S., MASHITER C., WESTAWAY J., LUMSDEN, P., DAY H. & HEWERTSON H., Students' Use of Research Content in Teaching and Learning. University of Central Lancashire, 2009.
[5] EISENBERG M. & HEAD A., How today's college students use Wikipedia for course–related research. First Monday, 2010, 15(3).
[6] SUNDIN O. & Francke, H., In search of credibility: Pupils' information practices in learning environments. Information Research, 2009, 14(4).
[7] LUYT B. et al., Youth Perception and Usage of Wikipedia. Information Research, 2008, 13(4).
[8] LIM S., How and Why Do College Students Use Wikipedia? Journal of the american society for information science and technology, 2009, 60(11).
[9] LIM S. & SIMON C., Credibility judgment and verification behavior of college students concerning Wikipedia. First Monday, 2011, 16(4).
[10] ELLIS D., A behavioural approach to information retrieval system design. Journal of documentation, 1993, 45(3).
[11] RIEH S. Y. & DANIELSON D. R., Credibility: A multidisciplinary framework. Annual review of information science and technology, 2007, 41(1).
[12] FORTE A. & BRUCKMAN A., Learning information literacy in the age of Wikipedia. In Proceedings of the International Conference of the Learning Sciences, 2008, Utrecht, Netherlands,.
[13] LUCASSEN T., NOORDZIJ M. L., & SCHRAAGEN J. M., Reference Blindness: The influence of references on trust in Wikipedia. In Proceedings of the ACM WebSci'11, June 14-17 2011, Koblenz.
[14] CHAN C., CHAN P. D. J. Y. & CHAN A. D., Young Wikipedians' perceptions of Wikipedia. In Wikimania 2011 Haïfa.
[15] LIH A., Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. In 5th International Symposium on Online Journalism 2004.
[16] GILES J., Internet encyclopaedias go head to head. Nature, 2005, 438(7070).
[17] CHESNEY T., An empirical examination of Wikipedia's credibility. First Monday, 2006, 11(11).
[18] ELVEBAKK B., Philosophy Democratized? First Monday, 2008, 13(2-4).
[19] WILKINSON, D. M., & HUBERMAN, B. A. Assessing the value of cooperation in Wikipedia. First Monday, 2007, 12(4-2).
[20] JACQUEMIN P., LAUF A., POUDAT C., HURAULT-PLANTET M. & AURAY N., La fiabilité des informations sur le web: le cas Wikipédia. In CORIA 2008, 2008, 227–235.
[21] WILLINSKY J. What open access research can do for Wikipedia. First Monday, 2007, 12(3).
[22] WILLINSKY J., Socrates Back on the Street: Wikipedia's Citing of the Stanford Encyclopedia of Philosophy. International Journal of Communication, 2008, 2.

[23] NIELSEN, F. A. Scientific citations in Wikipedia. First Monday, 2007, 12(8-6).
[24] HAIGH, C. A. Wikipedia as an evidence source for nursing and healthcare students. Nurse Education Today, 2011, 31(2).
[25] LUYT B. & TAN D., Improving Wikipedia's credibility: References and citations in a sample of history articles. Journal of the American Society for Information Science and Technology, 2010, 61(4).
[26] HUVILA I., Where does the information come from? Information Source Use Patterns of Wikipedia. Information Research, 2010, 15(3).
[27] ENDRIZZI L., Wikipédia : de la co-rédaction au co-développement de la communauté. In Chartron, G. Broudoux, E. (dir.). Actes de la conference DocSoc, 2006. ADBS, 2006.
[28] COUZINET V., Médiations hybrides: le documentaliste et le chercheur en sciences de l'information. ADBS, 2000.
[29] ZACKLAD, M. Espace documentaire participatif et gouvernance. In Congress of the European Regional Science Association (47th Congress) and ASRDLF (Association de Science Régionale de Langue Française, 44th Congress) PARIS - August 29th - September 2nd, 2007.

Enrichissement sémantique du corpus iSPEDAL

Abd El Salam AL HAJJAR
Mohammad HAJJAR
Zeinab ABDEL NABI
Georges LEBBOS
Laboratoire GRIT
Institut Universitaire de Technologie
Université Libanaise, Liban

Résumé : Dans cet article nous présentons la méthodologie utilisée pour doter iSPEDAL d'une dimension sémantique. iSPEDAL est une version améliorée de DESELA, peut être présenté sous la forme d'une base de données relationnelle facilement exploitable à l'aide des langages de requêtes appropriés. Dans notre cas trois voies sont explorées pour ajouter de certaines relations de sens entre les mots dans iSPEDAL. La première est basée sur l'exploitation des caractéristiques naturelles d'un dictionnaire classique, c'est qu'un dictionnaire propose, en générale, pour un mot donné ses synonymes, ses antonymes, etc. La deuxième voie est basée sur la traduction vers et à partir de l'anglais pour projeter les relations sémantiques entre les mots anglais aux mots arabes. La dernière consiste à exploiter le « Word-Net arabe » pour enrichir notre dictionnaire.

Mot-clé : Langage arabe, dictionnaire arabe, dimension sémantique, translation, WordNet arabe.

1. Introduction

La proposition des dictionnaires classiques arabes a constitué l'essentiel des travaux linguistiques effectués sur la langue arabe. La plupart de ces dictionnaires sont maintenant disponibles sur le web sous forme des fichiers électroniques plats. Donc, on observe actuellement, de plus en plus, une transition vers des dictionnaires électroniques [Al Hajjar et al., 2010]. Par contre, la plupart de ces dictionnaires collecte leurs données à partir des plusieurs dictionnaires classiques et offre un service de navigation et de recherche limité. Ces limites d'interrogation sont dues, principalement, à une faiblesse de structuration des entrées dictionnairiques utilisées [Habash, 2005] [Habash, 2004] [Habash and

Rambow, 2006]. Dans ce cadre, nous avons proposé un vrai dictionnaire électronique structuré et évolutif de la langue arabe iSPEDAL [Hajjar et al., 2010] qui est une version améliorée de DESELA [Al Hajjar et al., 2009a]. En effet, iSPEDAL peut être présenté sous la forme d'une base de données relationnelle facilement exploitable à l'aide des langages de requêtes appropriés. Ce nouveau dictionnaire fournit les liens d'un mot donné avec sa racine, ses affixes et son modèle éventuel. De plus, nous avons construit un système automatique qui permet d'alimenter et d'enrichir iSPEDAL à partir de plusieurs dictionnaires classiques ou à partir d'un corpus textuel arabe quelconque. Par contre, iSPEDAL souffre d'une handicape par rapport à un dictionnaire classique et qu'il n'offre pas des relations sémantique entre ses mots [Al Hajjar, 2010].
Dans cet article nous présentons la méthodologie utilisée pour doter iSPEDAL d'une dimension sémantique qui se restreint à l'ajout de certaines relations de sens entre les mots. Dans notre cas trois voies sont explorées. La première est basée sur l'exploitation des caractéristiques naturelles d'un dictionnaire classique, c'est qu'un dictionnaire propose, en générale, pour un mot donné ses synonymes, ses antonymes, etc. La deuxième voie est basée sur la traduction vers et à partir de l'anglais pour projeter les relations sémantiques entre les mots anglais aux mots arabes. La dernière consiste à exploiter le « Word-Net arabe » pour enrichir notre dictionnaire.

2. Le dictionnaire électronique arabe iSPEDAL

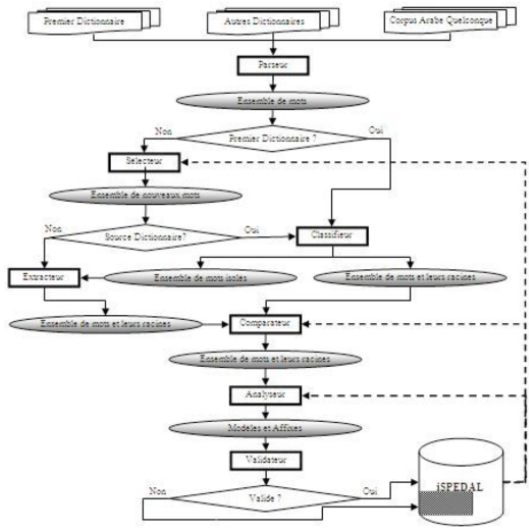


Figure 1 : Schéma général d'iSPEDAL

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.

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.

L'entrée de ce système peut être un dictionnaire arabe classique sous format plat (fichier texte, PDF,..) ou n'importe quel autre corpus arabe textuel (page web, fichier texte, …) [Al Hajjar et al., 2010]. Si le document en entrée est le premier dictionnaire, l'ensemble des mots est passé au classifieur. Ce premier dictionnaire est utilisé pour initialiser iSPEDAL [Hajjar et al., 2010]. 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 envoyés 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 pour trouver la racine d'un mot arabe [Al Hajjar et al., 2009b].

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 iSPEDAL [Hajjar et al., 2010].

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.

Ces ensembles 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 [Al Hajjar, 2010].

3. La dimension sémantique dans iSPEDAL

Dans cet article nous limitons la dimension sémantique à l'ajout de certaines relations de sens entre les mots. Ces relations permettent de situer le descripteur dans son environnement conceptuel [Walde and Zinsmeister, 2006]. Ces relations sont :

Synonymie : est un rapport de similarité sémantique entre des mots ou des expressions d'une langue. La similarité sémantique indique qu'ils ont des significations très semblables. Des termes liés par synonymie sont des synonymes [Pagin, 2000] [Walde and Zinsmeister, 2006] [Berzlánovich et al., 2008].

Antonymie : Deux mots sont en relation d'antonymie si on peut montrer une symétrie de leurs traits sémantiques par rapport à un axe. Ils sont de sens contraire [Mohammad et al., 2008] [Walde and Zinsmeister,2006] [Berzlánovich et al.,2008].

Hyponymie : est la relation sémantique d'un lexème à un autre selon laquelle l'extension du premier est incluse dans l'extension du second. Le premier terme est dit hyponyme de l'autre. C'est le contraire de l'hyperonymie [Berzlánovich et al., 2008] [Walde and Zinsmeister,2006].

Hyperonymie : est la relation sémantique hiérarchique d'un lexème à un autre selon laquelle l'extension du premier terme, plus général, englobe l'extension du second, plus spécifique. Le premier terme est dit hyperonyme de l'autre, ou super ordonné par rapport à l'autre [Berzlánovich et al., 2008] [Walde and Zinsmeister,2006].

Méronymie : est une relation partitive hiérarchisée, une relation de partie à tout. Des termes liés par méronymie sont des méronymes. Un méronyme A d'un mot B est un mot dont le signifié désigne une sous-partie du signifié de B. La relation inverse est l'holonymie. Par exemple, اليد est un méronyme de الجسم, de même que سقف est un méronyme de البيت [Berzlánovich et al., 2008] [Walde and Zinsmeister,2006].

Holonymie : est une relation partitive hiérarchisée, c'est la relation inverse de la relation méronymie. Des termes liés par holonymie sont des holonymes. Un holonyme A d'un mot B est un mot dont le signifié désigne un ensemble comprenant le signifié de B. Par exemple, الجسم est un holonyme de اليد, البيت est un holonyme de سقف [Berzlánovich et al., 2008] [Walde and Zinsmeister, 2006].

4. Méthodologie de mise en place des liens sémantiques dans iSPEDAL

Pour ajouter les relations de sens entre les mots à iSPEDAL, nous avons utilisé trois méthodes.

La première est basée sur l'exploitation des caractéristiques naturelles d'un dictionnaire classique (Fig. 2). En effet, un tel dictionnaire propose, en générale, pour un mot donné la définition, l'orthographe, les sens, les synonymes, les antonymes, les modes d'utilisation, etc. Notre procédure consiste à trouver les mots clés qui signifient qu'après ces mots on peut trouver des autres mots ou des phrases qui ont des relations sémantiques avec le mot ou la racine en question. Par exemple, dans le dictionnaire « Lesan Al Arabe » [Ibn Manzour, 2008], on peut trouver sous la racine «الكل» le mot clé «أي» qui permet de trouver le synonyme de cette racine («طعام»). Donc, il faut recenser les expressions qui expriment des relations sémantiques avec l'élément d'entrée du dictionnaire. D'autre part, on peut trouver sous chaque racine, des expressions qui expliquent les mots dérivés d'elle. De plus, on peut trouver des expressions qui donnent quelques caractéristiques de ces mots (pluriel, antonyme,…).

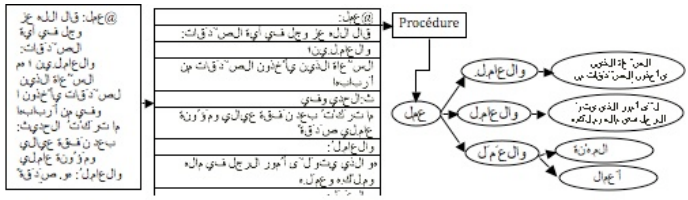


Figure 2 : Extraction des relations entre les mots à partir du dictionnaire « Lesan Al Arabe ».

La deuxième voie est basée sur la traduction vers et à partir de l'anglais pour projeter les relations sémantiques entre les mots anglais aux mots arabes (Fig. 3). Il y a beaucoup des traducteurs arabes [Sakher, 2010] [Diab, 2004], anglais ou français. Si deux mots arabes ont un même sens, probablement ils ont la même traduction en anglais, ou bien en français. Pour cela, nous prenons 2 mots arabes, on les traduit, la traduction de chaque mot peut donner un ensemble de mots, donc nous avons 2 ensembles des mots A et B, s'il y a une intersection entre ces 2 ensembles alors il y a une relation sémantique entre les 2 mots d'entrée avec un coefficient de ressemblance CR qui est définit par $A \cap B / A \cup B$. Exemple :

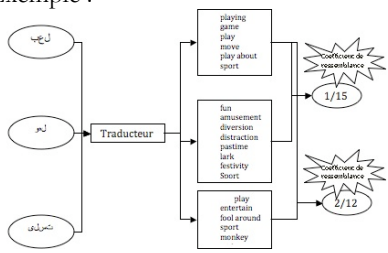


Figure 3 : Extraction des relations entre les mots en utilisant la traduction.

La dernière consiste à exploiter le « Word-Net arabe » pour enrichir notre dictionnaire. Word-Net est une grande base de données lexicale qui classe les mots arabes en noms, verbes, adjectifs, adverbes, etc. [Miller et al., 1993] , [Rennie, 2000] ,[Pedersen et al. 2004](Fig. 4). Ces éléments sont regroupés en ensembles de synonymes cognitifs (synsets) dont chacun exprime un concept distinct. Ces synsets sont reliés entre eux par des relations conceptuelles-sémantique et lexicales. La relation principale entre les mots dans Word-Net est la synonymie, par exemple : (إغلاق et إقفال) [Abouenour, 2008], [Abouenour, 2010], [Bouzoubaa, 2010]. Notre procédure consiste à extraire ces ensembles et à projeter les relations au niveau des mots de chaque ensemble.

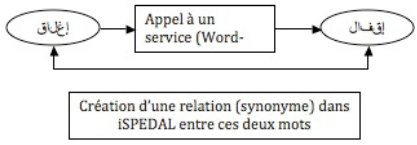


Figure 4 : Extraction des relations entre les mots à partir du Word-Net Arabe.

Bibliographie

[Abouenour et al.,, 2008] L. ABOUENOUR, K. BOUZOUBAA, P. ROSSO. Improving Q/A Using Arabic Wordnet. In: Proc. The 2008 International Arab Conference on Information Technology (ACIT'2008),Tunisia, December. 2008.

[Abouenour et al.,, 2010] L. ABOUENOUR, K. BOUZOUBAA, P. ROSSO. Using the YAGO ontology as a resource for the enrichment of Named Entities in Arabic WordNet. Workshop LR & HLT for semitic languages, LREC'10. Malta. May 2010.

[Al Hajjar et al., 2009a] A. AL HAJJAR, M. HAJJAR, K. ZREIK "Un nouveau dictionnaire électronique structuré et évolutif pour la langue arabe", CiDE.12, 12e Colloque International sur le Document Electronique, Montréal, Canada, 21 - 23 Octobre, 2009.

[Al Hajjar et al., 2009b] A. AL HAJJAR, M. HAJJAR, K. ZREIK "Classification of Arabic Information Extraction methods" MEDAR 2009 2nd International Conference on Arabic Language Resources and Tools, Le Caire, Egypte, 21-23 Avril, 2009.

[Al Hajjar et al., 2010] A. AL HAJJAR, M. HAJJAR, K. ZREIK "Structure, historique et évolution des dictionnaires arabes : le cas d'iSPEDAL", CiDE.13, 13e Colloque International sur le Document Electronique, Paris, France, 16-17 Décembre, 2010.

[Al Hajjar, 2010] A. AL HAJJAR "Extraction et gestion de connaissance à partir du Web multilingue : Spécificité de la langue arabe", Université Paris 8, France, Co-directeurs Pr. K. Zreik et Pr. M. Hajjar, thèse de doctorat soutenue le 17 décembre 2010.

[Berzlánovich et al., 2008] I. BERZLÁNOVICH, Lexical cohesion and the organization of discourse, University of Groningen, Supervisors: G. Redeker, M. Egg., 2008.
[Bouzoubaa, 2010] K. Bouzoubaa, Arabic Wordnet Use and Enrichment, Mohammadia School of Engineers, Rabat, Morocco. 2010.
[Diab, 2004] M. DIAB, Feasibility of Bootstrapping an Arabic WordNet Leveraging Parallel Corpora and an English WordNet, Linguistics Department Margaret Jacks Hall Stanford University Stanford, CA 94305, USA.2004.
[GT, 2011] Google Traduction (GT), web site : http://translate.google.fr/.2011.
[Habash and Rambow, 2006] N. HABASH and O. RAMBOW, A Morphological Analyzer and Generator for the Arabic Dialects, Center for Computational Learning Systems Columbia University New York, NY 10115, USA. july 2006.
[Habash, 2004] N. HABASH, Large Scale Lexeme Based Arabic Morphological Generation, University of Maryland Institute for Advanced Computer Studies University of Maryland College Park College Park, Maryland, 20742 USA. 2004.
[Habash, 2005] N. HABASH, Arabic Natural Language Processing: Words, Columbia University Center for Computational Learning Systems, Summer School on Human Language Technology Johns Hopkins University, Baltimore July 6th, 2005.
[Hajjar et al., 2010] M. HAJJAR, A. AL HAJJAR, K. ZREIK, P. GALLINARI "A improved structured and progressive electronic dictionary for the Arab language: iSPEDAL", IEEE ICIW 2010, The Fifth International Conference on Internet and Web Applications and Services, Barcelone, Espagne, 9 - 15 Mai, 2010.
[Ibn Manzour, 2008] IBN MANZOUR, 2008. Lisan Al-Arab, www.muhaddith.org.
[Miller et al., 1993] G. MILLER, R. BECKWITH, C. FELLBAUM, D. GROSS, K. MILLER, Introduction to WordNet: An On-line Lexical Database, 1993
[Mohammad et al., 2008] S. MOHAMMAD, B. DORR, G. HIRST, Towards Antonymy-Aware Natural Language Applications, University of Toronto, 2008.
[Pagin, 2000] P. PAGIN, A Quinean definition of synonymy, 17 May 2000.
[Pedersen et al., 2004] T. PEDERSEN, S. PATWARDHAN, J. MICHELIZZI. WordNet: : Similarity - Measuring the Relatedness of Concepts. In: AAAI (2004), p. 1024-1025. 2004.
[Rennie, 2000] RENNIE, J. WordNet::QueryData: a Perl module for accessing the WordNet database. http://search.cpan.org/dist/WordNet-QueryData. 2000.
[Sakher, 2010] Sakher Company, Al Ajeeb, http://lexicons.ajeeb.com 1998-2010.
[Walde and Zinsmeister, 2006] S. WALDE, H. ZINSMEISTER, Introduction to Corpus Resources, Annotation and Access: Semantic Annotation, Foundational Course 18th European Summer School in Logic, Language and Information Málaga, Spain July 31 - August 4, 2006.

Partie 4 - Aspect cognitif du document numérique

Terminologie hypertexte : dynamique temporelle d'une taxonomie

Nathalie PINEDE
Université de Bordeaux, MICA EA4426, MSHA, France
Université de Bordeaux, IPB, IMS (UMR5218), Bordeaux, France

David REYMOND
Université du Sud, Toulon-Var, I3M EA 3820, La Garde, France
Université de Bordeaux, IPB, IMS (UMR5218), Bordeaux , France

Benoit LE BLANC
Université de Bordeaux, IPB, IMS (UMR5218), Bordeaux, France
ISCC, CNRS, 20 rue Berbier-du-Mets, Paris, France

Véronique LESPINET-NAJIB
Université de Bordeaux, IPB, IMS (UMR5218), 146 rue Léo Saignat 33 076 Bordeaux Cedex

Résumé : Dans la lignée de travaux précédemment réalisés, l'objectif de notre étude ici est de caractériser le contenu d'un site web organisationnel par les marqueurs lexicaux associés aux liens hypertextes de la page d'accueil (unités lexicales hypertextuelles – ULH). Nous nous sommes intéressés à un ensemble de sites web issus d'un domaine organisationnel homogène (i.e le domaine universitaire français), à partir duquel nous avons généré deux corpus d'ULH relevé à deux moments différents (2009 et 2011). Ces deux corpus d'ULH ont été intégrés au sein d'une taxonomie produite à partir des niveaux hiérarchiques et catégoriels des différents modes de navigation. S'appuyant sur un premier niveau d'analyse quantitative, l'intérêt est de mettre à jour, à travers cette dynamique temporelle, les évolutions au niveau des pratiques du web tout en faisant la preuve d'une stabilité certaine des corpus d'ULH.

Mots-clés : Unités Lexicales Hypertextes, approche diachronique, taxonomie, site web organisationnel.

1. Introduction

La notion de « site web » s'impose comme une manifestation emblématique des contenus architecturés du Web, mais elle dissimule aussi une réalité multidimensionnelle qui fonde sa nature complexe. Parmi les différentes facettes qui peuvent contribuer à son approche, nous retiendrons le site web en tant que ressource relevant de l'appellation générique « document numérique » (Leleu-Merviel, 2004). Sa caractérisation en tant que document numérique permet de l'envisager notamment au plan informationnel, selon une approche originale. En effet, nombre de travaux menés autour d'une analyse des contenus de sites web opèrent soit selon une perspective qualitative (évaluation ergonomique d'interfaces, étude d'ordre sémiotique), soit selon une perspective orientée statistico-sémantique à partir du contenu intégral de documents textuels. Notre hypothèse ici est de nous appuyer sur un niveau d'organisation des connaissances déjà réalisé sur le site, à savoir les liens hypertextes des pages d'accueil des sites web, liens hypertextes considérés dès lors comme les entrées d'un sommaire donnant accès aux contenus internes du site. Un travail de recensement automatique de ces liens hypertextes et de classification des unités lexicales associées permet ensuite de disposer d'une taxonomie à partir de laquelle un certain nombre d'éléments d'analyse et d'applications potentielles peuvent être réalisés.

Nos travaux[22] ont déjà amené des résultats intéressants. Constituée à partir d'un corpus réduit, la stabilité et la fiabilité de la taxonomie ainsi obtenue ont pu être vérifiées grâce à son pouvoir de recouvrement sur d'autres sites web du même domaine (Reymond *et al.,* 2011). D'autre part, deux pistes d'exploration sont en cours : générer des profils informationnels de sites web, sur la base des occurrences dans la taxonomie, à partir des dominantes thématiques qui émergent (notamment au plan des activités organisationnelles) ; permettre la reconnaissance et le classement automatique de sites[23].

Si notre postulat de départ reste identique (i.e. caractériser le contenu d'un site web par les marqueurs lexicaux associés aux liens hypertextes de sa page d'accueil), notre positionnement ici est différent. En effet alors que jusqu'à présent, nous avions mené nos analyses dans une perspective synchronique, à un instant t, nous nous positionnons ici dans une perspective diachronique pour tenter de mettre en lumière des éléments significatifs d'évolution. Comment évoluent ces marqueurs lexicaux dans le temps ? Quels enseignements peut-on tirer des

[22] Menés dans le cadre du programme RAUDIN (Recherches aquitaines sur les usages pour le développement des dispositifs numériques) à financement Feder n°31462, Conseil Régional Aquitaine et Université Bordeaux 3

[23] Ces enjeux et perspectives sont synthétisés dans une communication présentée aux journées d'études TICIS 2010.

modifications notoires observées, au plan des pratiques du web mais aussi de la dynamique organisationnelle ?

Les unités lexicales hypertextes (ULH) des pages d'accueil peuvent être assimilées à des fragments informationnels qualifiant les contenus internes du site. A ce titre, elles représentent non seulement les activités phare mais aussi des informations d'ordre structurel ou en prise avec l'actualité et la vie de l'organisation ; elles donnent également accès à des fonctionnalités rendues possibles par le média web (outils logiciels, flux RSS, etc.). Elles donnent donc à voir une certaine image de l'organisation et incarnent de fait une mémoire intéressante, signifiante et significative, des changements d'une structure, de ses choix, de ses représentations, de l'intégration de problématiques telles celle de l'accessibilité, ainsi que des phénomènes de lissage et normalisation terminologiques qui peuvent s'effectuer dans le temps (Rouquette, 2009). Dans une perspective plus opérationnelle, les analyses extraites, aux plans qualitatif et quantitatif, peuvent générer des préconisations en matière de qualification de l'information sur le web, en s'appuyant sur des pratiques consensuelles.

Partant d'un corpus unique de sites web, nous avons donc analysé l'évolution de ces marqueurs lexicaux à deux années d'intervalle[24]. Les premiers éléments de cette étude comparée que nous présentons ici sont essentiellement d'ordre quantitatif, pour tenter de qualifier les évolutions globales au plan des corpus[25]. Dans ce texte, nous développerons tout d'abord les arguments clefs de notre problématique, à savoir l'articulation site web/document numérique, la dimension organisationnelle pour le site web, le rôle déterminant de la page d'accueil et des ULH en tant que représentations informationnelles signifiantes de l'ensemble des contenus du site. Dans un deuxième temps, nous expliciterons le principe d'organisation de nos ULH en taxonomie, en nous appuyant sur une catégorisation en fonction des types de navigation. Ensuite, nous présenterons nos deux terrains d'investigation, choisis dans le domaine organisationnel des universités. Une université en sciences sociales et une université en sciences ont été sélectionnées pour constituer notre corpus d'ULH sur deux temporalités distinctes, 2009 et 2011. A partir de ce premier niveau d'analyse diachronique, la comparaison quantitative de nos corpus d'ULH permettra de dégager des éléments génériques intéressants, qui seront à affiner ultérieurement par l'auscultation de la dynamique des formes et de la sémantique des ULH.

[24] Cet intervalle de temps permet de faire émerger des tendances, à confirmer par l'analyse du même corpus à t+2.

[25] Une approche plus fine des évolutions terminologiques est actuellement en cours.

2. Problématique

2.1. Du site web au document numérique

Le terme de « site web » renvoie naturellement à un carrefour de pages et d'hyperliens, participant à la définition de territoires numériques (Musso, 2009). Mais au-delà de cette première approche, plusieurs déclinaisons peuvent être mises en avant pour tenter de définir un site web.

Tout d'abord on retiendra au premier chef celle d'un dispositif technique s'appuyant sur un serveur informatique et identifié par un nom de domaine de type DNS (Domain Name Server), réalité technique pouvant être rapprochée des contours et limites d'une organisation[26].

Au-delà de cette définition du site web par la mobilisation de ressources informatiques et protocoles techniques d'échanges de données (image du site web peu accessible à l'usager), le site web se donne à voir à travers son contenu. Stockinger définit le site web comme « lieu de prestations, lieu de services à destination d'un certain public » (Stockinger, 2005). Les sites web organisent de façon plus ou moins sophistiquée, diversifiée, cohérente et exhaustive, des ressources de nature informationnelle à destination de certaines catégories d'usagers-internautes. Ils sont dès lors des produits d'information éditée, renvoyant à une responsabilité éditoriale identifiée.

Le site web peut également être appréhendé comme un dispositif communicationnel « qui définit des rapports énonciatifs, attribue des rôles, inscrit des marques pour l'interprétation, à ceci près qu'il renvoie à un espace de communication dit « planétaire », en fait assez indéterminé sur le plan culturel et social » (Souchier et al., 2003). Il s'inscrit dès lors dans une tension entre intentionnalité(s) et usage(s), incluant les contraintes inhérentes au dispositif technique sollicité. Dans cette articulation d'éléments, l'intention (de communication, de proposition de services) se traduit par la production d'un site web au service d'une stratégie à destination de publics supposés.

Enfin, le site achève de prendre sens dans cet espace web à travers le réseau d'hyperliens qui le jalonnent et qui guident l'action des usagers tout en leur autorisant de multiples parcours, dans un paradoxe permanent entre clôture et évasion. Le site web peut ainsi être assimilé à un document numérique dynamique inscrit dans un espace topologique hypertextuel (Leleu-Merviel, 2004). On peut aussi adhérer aux propositions du collectif R.T. Pédauque selon lequel le document numérique s'articule autour de la tridimensionnalité Forme-Texte-Medium et possède les propriétés de mémorisation, organisation, création et transmission, propriétés dont nous nous inspirons pour bâtir notre argumentation présente.

[26] Nous verrons toutefois, à partir des sites de notre corpus, que cette proximité DNS/organisation n'est pas si évidente que cela en pratique.

2.2. Site web organisationnel

Ces dimensions multiples (et non exhaustives) des sites web génèrent de nombreuses possibilités d'approche et d'analyse. Il est toutefois difficile d'appréhender ces sites dans leur globalité, tant est grande l'hétérogénéité des formes, des contextes de production, des intentionnalités de communication, ou encore des données informationnelles. Dans l'optique de traiter, notamment au plan quantitatif, voire automatique, des ensembles de sites, il paraît indispensable de repérer des cohérences et régularités parmi les différents sites web.

Là encore, plusieurs angles de lectures peuvent être choisis. Stockinger propose une typologie des sites sur la base de leurs caractéristiques les plus saillantes : sites personnels « simples », sites d'information, sites d'accès à des ressources (sites portails), sites fac-simile ou dématérialisés de services organisationnels, sites à thème (Stockinger, 2005). On peut bien entendu regrouper les sites selon d'autres perspectives : par secteurs d'activité, selon la finalité, etc.

En ce qui nous concerne et par rapport à nos objectifs de recherche, nous avons choisi de travailler sur des corpus de sites représentant une organisation-type : nous appellerons ces sites « sites web organisationnels » (SWO). Ce choix de constitution de corpus, sous couvert d'identification d'organisations relevant d'un même domaine (au plan de l'activité), garantit un facteur de cohérence et d'homogénéité en termes de missions, services et publics visés. D'une façon plus spécifique, nous nous sommes intéressés aux sites web des universités françaises.

En considérant d'une part la présence historique du milieu universitaire sur le Web et d'autre part l'activité de communication des universités via leur site, le support internet est au cœur de la communication des universités. Ainsi que le rappelle Lainé-Cruzel, « le site de l'université se doit d'être à jour : renseigner d'une manière efficace, présenter des formations et des équipes existantes, des annuaires exploitables, etc. Sa fonction est d'être utile et de rendre des services. Sa nature est d'être évolutive, pour restituer une image aussi fidèle que possible d'un univers en transformation permanente.

La qualité du site sera liée à sa capacité à évoluer en même temps que l'univers sur lequel il informe » (Lainé-Cruzel, 2004). À ce titre, elle le caractérise en tant que ressource, qui s'inscrit dans une double logique de médiation et d'usage.

Par ailleurs, l'intérêt de travailler sur ce domaine organisationnel est d'avoir une architecture du web complexe, puisque sous l'appellation « site web d'université » se dissimule une démultiplication de sous-sites web, autonomes, mais participant tous de l'image organisationnelle de l'université.

2.3. Fonction de la page d'accueil et des ULH pour une discrimination informationnelle des SWO

Vis-à-vis de cet objet web aux facettes multiples, la page d'accueil joue un rôle déterminant. Celle-ci met en effet en représentation une grande partie de l'information gérée par l'organisation ou la structure concernée, tout en traduisant des choix sélectifs et stratégiques, qui en valorisent certaines dimensions plutôt que d'autres (Nielsen & Tahir, 2002). Mais la page d'accueil d'un site web se donne aussi à voir en tant qu'écrit d'écran, pour lesquels « nous disposons de « signes outils », de « signes passeurs » qui nous donnent accès aux multiples modalités du texte » (Souchier *et al.*, 2003). La fonction d'orientation des usagers du site vers des zones informationnelles spécifiques est donc assurée par ces signes passeurs, choisis notamment pour répondre aux contraintes issues de l'écriture pour ce média. Au plan des usagers, les stratégies de navigation web varient en fonction du niveau d'expérience Web (Thatcher, 2008). Il a été montré que les novices du web s'appuient principalement sur la recherche par sommaire (i.e. les menus de navigation), a contrario de la recherche par mots-clés qui s'avère plutôt efficace seulement pour ceux qui ont déjà des connaissances dans le domaine (Dressen-Hammouda & Drot-Delange, 2009) indépendamment de leur niveau d'expertise web. Dans le cadre de notre recherche, nous considérons donc la page d'accueil en tant que « sommaire » du site (au plan lexical), sommaire qui, tout en présentant des régularités de contenus, pourra prendre différentes formes pour un ensemble de sites du domaine.

Nous restreignons également la notion générique (Jeanneret, 2007) de « signe passeur » (incluant par exemple la dimension iconique) à celle de « texte passeur ». Marqués à la fois par des contraintes d'édition et de structuration web, les objectifs stratégiques de valorisation de l'information, et les capacités rédactionnelles des auteurs, les ULH de la page d'accueil, produisent une signature textuelle et sémantique de la page d'accueil (Reymond et Pinède, 2010a) et par extension, du site web concerné. En effet, la majeure partie de ces ULH sont des mots-clés représentant les contenus sous-jacents et constituent des marqueurs thématiques des contenus du site, que l'on peut traduire en signature informationnelle. À ce titre, les ULH d'une page d'accueil de site peuvent être considérées comme représentatives de l'ensemble de l'information portée par le site et, pour certaines, discriminantes (dans leur co-présence) par rapport à des sites web organisationnels.

Partant de ces postulats, nous nous intéressons ici à la comparaison de deux corpus d'ULH issus d'un ensemble de SWO du domaine universitaire. L'exploitation de ces deux collectes s'effectue à plusieurs niveaux :

- Comparaison générique entre ces deux corpus d'ULH permettant d'analyser l'évolution des marqueurs lexicaux aux plans quantitatifs

(augmentation/diminution du volume global de termes, à mettre en perspective avec les évolutions organisationnelles et techniques).
- Comparaison diachronique des entrées de la taxonomie par classe. Il s'agit ici, en superposant les disparitions et nouvelles entrées d'ULH dans les classes de la taxonomie, en prenant également en compte l'évolution des occurrences d'ULH dans les classes, de faire émerger des dynamiques thématiques significatives dans le temps (entre t et t+1, à savoir 2009 et 2011).

3. Méthodologie

3.1. Constitution des corpus d'ULH

Nous avons collecté les ULH présentes sur les interfaces de pages d'accueil[27] de sites web relevant du domaine organisationnel de deux établissements universitaires, se distinguant par leur champ disciplinaire (Sciences et Techniques / Sciences humaines et sociale)[28].
Par rapport à une première extraction de l'ensemble des sites web des zones DNS (Domain Name Service) de Bordeaux 1 et Bordeaux 2, nous avons opéré à une sélection en ne retenant que les sites répondant aux critères d'accessibilité (en langage HTML) ainsi que ceux présents simultanément en 2011 et sur le site webarchive.org pour la collecte 2009. Dans un deuxième temps, n'ont été conservés que les sites dont :
les ULH ont une sémantique propre (les critères d'exclusion sont : hyperliens en soi (http), courriels, nombres et ULH à 1 caractère) ;
le nombre d'ULH collecté par site est supérieur à deux.
Notre corpus de site est constitué au final de 96 pages d'accueil (ou pages de menu lorsqu'il existe une page d'introduction) de sites pour moitié issus respectivement des domaines DNS de l'université de Bordeaux 1 et de Bordeaux 2.
La collecte automatique s'appuie sur un analyseur lexical qui en premier lieu requête la page d'accueil ou son éventuelle redirection. La seconde phase extrait du code source les termes passeurs (ancres) des pages ainsi que le contenu des balises <ALT>[29]. L'extraction opère dans le contenu en langage html tout ce qui se trouve entre les balises et ainsi que dans le champ des balises ALT pour les hyperliens sur images, icônes ou zone (<xxx src=..>) inscrits en tant qu'hyperliens. Selon le format éditorial choisi par les éditeurs, la collecte automatique souffre d'un biais issu de la possibilité de masquer certains éléments du

[27] Par pages d'accueil, nous entendons chaque page d'accueil de tous les sites recensés dans une zone DNS identifiée. Soit une centaine de pages d'accueil en moyenne pour chaque organisation universitaire.
[28] Respectivement universités de Bordeaux 1 et Bordeaux 2.
[29] Ces balises sont destinées à l'accessibilité de la page pour combler la déficience visuelle et sont donc à vocation de décrire textuellement un marqueur de lien ou d'image.

code source lorsque le navigateur en produit le rendu via des feuilles de style (CSS par exemple) ou par programmation (ECMAscript). Le parseur n'interprète pas ces langages et est en conséquence insensible à ces variétés de mise en forme. Inversement, le collecteur peut aussi capter les marques d'accessibilité qui sont des compléments d'information textuels des données multimédia. Nous notons que le collecteur réalisé n'est pas exempt d'erreurs d'analyse lexicale lors du traitement des données. Toutefois après notre traitement automatisé nous avons pu mesurer une erreur peu significative : moins de 5% des ULH sont erronées (par exemple, , <xt javascript>).

3.2. Principes de constitution de la taxonomie

L'objectif est d'organiser les ULH collectées au sein d'une taxonomie (Tableau 1). Cette taxonomie se découpe selon trois catégories de navigation, déclinées à partir de la segmentation de (Nielsen et Tahir, 2002), elles-mêmes scindées en différentes classes :

- Catégorie 1 « Navigation thématique » : correspond aux ULH permettant de donner accès aux contenus des zones profondes des sites. Ces contenus peuvent être directement liés aux activités de l'organisation (5 classes) ou faire référence à des contenus génériques c'est à dire transversaux à différentes organisations (9 classes).
- Catégorie 2 « Navigation fonctionnelle » : correspond aux ULH faisant référence aux liens types outils (8 classes).
- Catégorie 3 « Navigation par profil » : correspond aux ULH permettant d'avoir accès à une recomposition du site pour les usagers en fonction de spécificités linguistiques ou autres (2 classes).

Catégories		Classes	Définition	Exemples d'ULH
Navigation thématique	Activités	Recherche	Activités ayant trait à la recherche	Ecole doctorale Laboratoires
		Formation	Activités ayant trait à l'enseignement	Master 2 Nouvelle licence 2011-2012
		Ressources documentaires	Ressources documentaires générales ou spécialisées accessible en ligne	Livres numériques Thèses électroniques
		Partenariat/transfert/ valorisation	Partenariats affichés de l'université (hors international)	Nos partenaires Collaborations et contrats
		International	Relations et politique internationales de l'université	Etudes à l'étranger Relations internationales
	Génériques	Accueil/présentation	Contenus généralistes sur l'organisation	Bienvenue Accueil
		Actualités	Contenus ayant une dimension temporelle, éphémère	A la une Evènements à venir

		Composantes extérieures	Mention de composantes ou structures extérieures à l'organisation	Accueil CNRS Institut Polytechnique de Bordeaux
		Informations pratiques	Informations, contenus démarches dépendantes de l'organisation	Infos pratiques étudiants Logement Santé-social
		Services dématérialisés	Accès à des services identifiés accessible en ligne (annuaire, etc.)	Intranet Annuaire
		Recrutement	Proposition d'emploi, mode de recrutement (hors formation)	Offres d'emploi Travailler à l'université
		Culture / loisirs	Activités liées au culturel, à l'associatif	Atelier photo Ciné-club
		Composantes de l'organisation	Dimension structurelle de l'organisation (hors activités)	Conseil d'administration Bibliothèque
		Logistique / équipement	Mention d'infrastructures techniques et des modalités de gestion de celle-ci	Equipement Réservation salles
Navigation Fonctionnelle		Accès web	Navigation en termes d'action dans le site	Cliquer ici Aller au pied de page
		Contacts	Renvoi sur un contact	Nous écrire Contactez-nous
		Technologies	Renvoi sur des formats, langages, ou outils informatiques (hors outils web 2.0)	Joomla Template XHTML 1.0
		Accès géographique	Indications géographiques d'accès au site physique	Plan d'accès Localisation
		Outils de communication web	Renvoi sur des outils d'échange et de communication du web 1.0 et 2.0	Flux RSS Liste de diffusion
		Authentification	Procédure d'authentification (sans mention du service associé)	Identifiez-vous Mot de passe oublié ?
		Mentions légales	Renvoi sur des mentions, informations obligatoires et légales	Crédits et mentions légales Législation
		Fonction rechercher	Renvoi sur la fonction de recherche (sur/hors du site)	Rechercher Moteur de recherche
Navigation par profils		Profil utilisateurs	Renvoi sur un espace dédié à un profil précis d'utilisateurs	Espace Entreprise Futur étudiant Accès étudiants
		Profils linguistiques	Renvoi sur l'interface traduite dans la langue concernée	En fr

Tableau 1 : Définition des principales classes de la taxonomie

Il est à noter que certaines ULH ne peuvent être classées actuellement et ce, pour trois raisons principales :
- ULH rejetées : certaines ULH renvoient à des mentions marginales ou anecdotiques (par exemple, « âme du bâtiment ») qui ne peuvent entrer dans un recensement terminologique dont la finalité est d'aboutir à une normalisation.
- ULH de spécialité : certaines ULH renvoient à des termes relatifs à des spécialités disciplinaires (par exemple, « Nanostructures Organiques », « Harmoniques XUV et impulsions attosecondes ») qu'on pourra tenter d'intégrer ultérieurement dans la taxonomie en s'appuyant sur des référentiels dédiés.
- ULH ambiguës : des ULH comme « programme », « équipe » sont évidemment des termes pertinents qu'il serait souhaitable d'intégrer à la taxonomie mais qui, à l'heure actuelle, demeurent ambivalentes sans qualification par le contexte associé : relèvent-elles de la classe « formation », de la classe « recherche » ou d'une autre classe ? Il s'avère donc impossible actuellement de les classer sans risque d'erreur.

3.3. Méthodologie d'analyse et de comparaison des corpus

Concernant le corpus, pour chaque année (2009 et 2011) nous mesurerons les variables suivantes :
- Le nombre d'ULH totales (ie. ULHt) ainsi que le nombre moyen d'ULH par site (nous préciserons l'étendue). Cet indicateur inclut les occurrences.
- Le nombre d'ULH distinctes (i.e. ULHd). Les ULH distinctes correspondent aux unités lexicales se distinguant par la forme. Par exemple, « Présentation » / « présentation » / « présentation » correspondent à trois ULH distinctes même si c'est le même terme désigné.
- La taille moyenne des ULH en nombre de caractères (espace et ponctuation inclus) (par exemple : « contactez-nous » : ULH de taille 14 / « programme de recherche » : ULH de taille 22)

Nous effectuerons des comparaisons de moyennes sur ces différents critères. Afin d'étudier l'évolution temporelle du corpus et de la taxonomie, plusieurs analyses seront effectuées :
- au plan de la caractérisation du contenu des interfaces pour déterminer les évolutions quantitatives,
- au plan des catégories d'ULH de la taxonomie pour transposer les évolutions précédentes aux catégories informationnelles présentées par les dispositifs et, indirectement, par les organisations.

De plus, nous précisons le nombre d'ULH t et d en fonction de leur appartenance à 3 types de corpus :
- corpus exclusif 2009 : ULH n'apparaissant qu'en 2009 et ayant disparu en 2011

- corpus commun aux deux années : ULH apparaissant aussi bien en 2009 qu'en 2011
- corpus exclusif 2011 : ULH n'apparaissant qu'en 2011

4. Résultats

4.1. Analyse sur le corpus

4.1.1 Caractéristique des interfaces

Le propose une comparaison concernant les principales données sur les ULH. Une progression, entre les deux années, a été observée sur le nombre moyen d'ULHt par site avec une progression de 136%. Une diminution en 2011 a été observée sur le nombre maximum d'ULHt par site (diminution de 9%).

	Nb moyen d'ULHt par site	Etendue du nb d'ULHt par site
2009	22	min 1 / max 174
2011	30	min 1 / max 160
Différence	+ 8	Max – 14

Tableau 2 : Caractéristiques des pages d'accueil du corpus en critères d'ULH (présence, taille)

Plusieurs hypothèses peuvent être formulées pour expliquer ce phénomène d'augmentation. On peut l'expliquer par une convergence au plan technique des modes de composition de l'interface page d'accueil. La progression du nombre d'ULHt peut être corrélée aux augmentations de formats d'écran (le plus souvent contraint en 1024x768 en 2009, et à 1280x1024 en 2011), laissant plus de place pour les menus, ou encore à l'utilisation de CSS et javascript pour dispenser via la même page menus ET sous-menus. L'autre hypothèse est de l'attribuer à une augmentation en volume des contenus publiés sur les sites. Pour vérifier cela, il faudrait corréler cette observation avec l'évolution de la taille des sites web sur la période concernée.

4.1.2. Caractéristiques du corpus d'ULH

La comparaison des deux corpus montre des différences importantes. Les évolutions temporelles concernent plusieurs indicateurs et révèlent une progression en faveur de 2011. Le Tableau 3 montre les caractéristiques générales des corpus d'ULH collectés sur les 96 sites. Rapportée au nombre d'ULH par site, l'augmentation entre les deux années (+725, soit 36% d'augmentation) montre la croissance de ces ensembles lexicaux. Si l'on regarde toutefois l'augmentation des ULH distinctes au plan lexical (+418, soit 21% d'augmentation), cela pondère

l'effet d'augmentation et permet de mettre en évidence une croissance relative, en phase au plan des contenus sur l'ensemble du corpus.

	2009	*2011*	*Variation*
Nombre d'ULHt - ULHt	2096	2940	+844
Nombre d'ULHd	1589	1895	+298
Proportion des ULHd / ULHt	76 %	64 %	

Tableau 3 : Caractéristique des corpus d'ULH collectées (nombre, occurrences, taille)

Le Tableau 4 montre la dynamique globale du corpus des ULH. Les ULH communes 2009-2011 représentent l'ensemble des ULH stables sur les deux années. Les ULH exclusives 2009 sont celles qui ont disparu entre les deux périodes de collecte. Les exclusives 2011 sont celles apparues en 2011.

	Nb ULHt (% / au nb total)	Nb ULHd (% / au nb total)	Proportion ULHd / ULHt
Exclusif 2009	841 (16,7%)	779 (29%)	92,7 %
Commun 2009-2011	2762 (54,8%)	818 (30,6%)	29,6 %
Exclusif 2011	1433 (28,5%)	1077 (40,3%)	75,2 %
Corpus complet	5036	2674	53,1 %

Tableau 4 : Dynamique temporelle du corpus d'ULH : les communs aux deux années, exclusifs 2009 et 2011.

Plusieurs points sont à mentionner :
- 55 % des ULHt du corpus complet sont issus des ULHt communes 2009-2011 alors que cette proportion diminue à 29,6% lorsqu'on s'intéresse aux ULH distinctes.
- au niveau des ULH communes 2009-2011, un principe d'occurrence très important est observé : 1 ULH apparait 3 fois en moyenne (proportion ULHd/ULHt).
- Entre 2009 et 2011, nous notons une diminution notoire de la dispersion des ULH (93% en 2009, contre seulement 75,3 % en 2011)..
- Sur le corpus complet, si on s'intéresse uniquement aux ULHd alors on se rend compte que ce corpus d'ULHd est surtout composé d'ULHd issues du corpus exclusif 2011 (43 %).

Le corpus commun illustre une forme de consensus autour de ces unités lexicales, à la fois dans le temps (permanence et stabilité de ces termes), et dans le partage (nombreuses occurrences). Nous pouvons constater une augmentation des occurrences sur le corpus 2011, montrant que si de nouvelles ULH sont utilisées, elles en appellent pour leur majorité à

des ULH existantes dans le corpus, ou encore à de nouvelles ULH mais « harmonisées » puisqu'à occurrences multiples.

4.2. Taxonomie

Nous avons classé chacune des ULH dans les différentes classes de la taxonomie présentée précédemment (cf. Tableau 1). Le Tableau 5 donne les statistiques de classement des ULH pour les deux années en précisant les ULH classées des ULH non classées.

	Corpus complet Nb ULHt / ULHd	Corpus classés Nb d'ULHt / ULHd	Corpus non classés Nb d'ULHt / ULHd
Corpus 2009	2096 / 1589	1385 / 961	711 / 636
Corpus 2011	2940 / 1782	2028 / 1156	912 / 739
Variation entre les 2 années	+844 / +185	+643 / +195	+201 / +103
Progression en %	+40% /+11,5 %	+46,4% / +20,3%	+28%/ +14%

Tableau 5 : Caractéristiques des différents corpus.

Ainsi, une progression du nombre d'ULH entre les deux années s'observe aussi bien sur le corpus classé que le corpus non classé. Il est à noter que sur les 5036 ULH total du corpus complet (2009 et 2011 inclus), 32 % n'ont pas pu être classées au sein de notre taxonomie. Nous retrouvons dans cette catégorie « non classées » les ULH rejetées, de spécialité et ambiguës (cf. 3.2 pour la définition des ULH non classées). Ce qui est important de noter, c'est que l'augmentation des non classés ne suit pas l'augmentation du nombre d'ULH (moins de 24%).

Les données suivantes ne concernent que les ULH classées au sein de notre taxonomie, afin de permettre une meilleure visibilité, nous ne présenterons que les données concernant les ULHt en fonction des catégories.

Concernant les ULH classées, si on reste au niveau des 3 catégories de notre taxonomie, il y a une grande différence en terme d'utilisation des ULH, avec une surreprésentation de la catégorie « navigation thématique » qui reste stable sur les deux années (Tableau 6) : pour les deux corpus exclusifs, 80 % des ULHt appartiennent à la classe « navigation thématique » alors que pour le corpus commun 2009-2011 cette catégorie diminue à 66%. Dans le corpus commun 2009-2011, 32 % des ULHt appartiennent à la catégorie « navigation fonctionnelle ».

Lorsqu'on compare les deux corpus exclusifs catégorie par catégorie, certaines évolutions temporelles apparaissent :

15% des ULHt de la classe « thématique » appartiennent au corpus exclusif 2009 contre 28 % appartenant au corpus exclusif 2011.
5 % des ULHt de la classe « navigation par profil » appartiennent au corpus exclusif 2009 contre 35 % appartenant au corpus exclusif 2011.
Concernant les ULHt de la classe « navigation fonctionnelle » il n'y a pas de différence entre les deux corpus exclusifs (10% pour 2009 et 15% pour 2011).
La catégorie « navigation fonctionnelle » est la plus stable des trois, celle où les ULH restent majoritairement présentes sur les deux années (peu de disparition, peu de nouvelles ULH). On note par contre une prise en compte croissante des profils utilisateurs (apparition importante d'ULH dans cette catégorie en 2011). Quant à la navigation thématique, dont la présence majoritaire reste constante, elle témoigne de la dynamique la plus importante, avec des renouvellements d'ULH importants, ce qui est à rapprocher des évolutions dans le temps de l'organisation concernée.

	Type de corpus d'ULHt						
	Exclusifs 2009		Commun 2009-2011		Exclusifs 2011		
Catégorie de la taxonomie	ULHt (%/type corpus)	% par catégorie	ULHt (%/type corpus)	% par catégorie	ULHt (%/type corpus)	% par catégorie	Total
« navigation thématique »	366 (80%)	15%	1382 (66%)	57%	684 (80%)	28%	2432
« navigation fonctionnelle »	84 (18%)	9,5%	663 (32%)	75%	135 (16%)	15%	882
« navigation par profil »	5 (1%)	5%	59 (3%)	60%	35 (4%)	35%	99
Total	455		2104		854		3413

Tableau 6 : Poids respectif des 3 catégories de la taxonomie en fonction du type de corpus

Par ailleurs, 5 classes (sur les 24 classes de la taxonomie) représentent à elles seules à peu près 60 % des ULHt (Tableau 7). On a ainsi une concentration importante autour de 5 pôles d'intérêt qui concernent à la fois les missions phare de l'organisation concernée (Recherche, Formation, Ressources documentaires) et une aide à la navigation sur le média web. Néanmoins, certaines différences apparaissent en fonction du type de corpus considéré. Pour les deux corpus exclusifs, les deux classes les mieux représentées sont la recherche et la formation avec une priorité à la recherche dans le corpus exclusif 2009 (23%) et une priorité à la formation pour le corpus exclusif 2011 (20%). Alors que les deux

classes les mieux représentées au sein du corpus commun sont respectivement : accueil/présentation (13,5%) et accès web (13%), la recherche n'apparaissant qu'en 3ème position (12%) et la formation en cinquième position (8%).

	% d'ULH classées / aux ULHt		
Ordre d'importance des classes	Exclusif 2009	Commun 2009-2011	Exclusif 2011
Rang 1	Recherche (23%)	Accueil/présentation (13,36%)	Formation (21%)
Rang 2	Formation (18,7%)	Accès web (13%)	Recherche (17%)
Rang 3	Ressources documentaires (9,6%)	Recherche (12%)	Informations pratiques (9%)
Rang 4	Accès web (7,7%)	Ressources documentaires (10,6%)	Ressources documentaires (9%)
Rang 5	Accueil/présentation (5,3%)	Formation (8%)	Accès web (8%)
% de représentation des 5 classes	64 %	57 %	62 %

Tableau 7 : Poids des 5 classes les plus représentées dans la taxonomie en 2009 et 2011

Afin d'observer plus précisément la hiérarchisation de ces 5 classes dans les corpus (exclusifs 2009 et 2011, commun 2009-2011), une carte cognitive30 a été réalisée. L'intérêt de cette carte est de faire apparaître visuellement à la fois les classes les plus fortement représentées et l'évolution de leur positionnement dans le temps. Les classes Formation et Recherche restent prédominantes sur les corpus exclusifs. Par contre, dans le corpus commun, ce sont les classes Accueil / Présentation et Accès web qui sont majoritairement représentées, ce qui illustre bien la stabilité des termes présents dans ces classes. Enfin, on relèvera présence croissante dans le corpus exclusif 2011 de la classe Informations pratiques, ce qui pourrait indiquer un changement dans les pratiques de diffusion d'informations (passage du papier au numérique).

[30] Réalisées avec le logiciel XMind

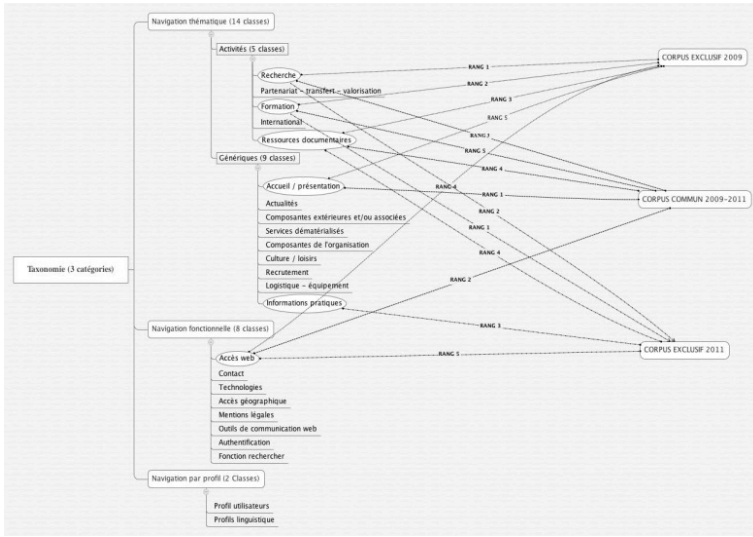


Figure 1 : Carte cognitive de la taxonomie en fonction des 3 corpus

Conclusion

Ces travaux ouvrent donc des perspectives de recherches intéressantes en terme d'analyse des contenus numériques. En l'occurrence, le fait d'observer, à travers le prisme de cette taxonomie, deux corpus de sites web appartenant au même domaine organisationnel permet tout d'abord de mettre en évidence des grands mouvements d'ensemble au plan des choix terminologiques. Ainsi, les deux images (en 2009 et 2011) de la taxonomie gardent les mêmes proportions, avec une relative stabilité des poids des classes informationnelles et des catégories sur les deux années. Les tests montrent une convergence (dans des temporalités courtes) de la gamme des ULH utilisées et de surcroît une normalisation de leur forme. Cela concerne essentiellement les classes de la navigation thématique. On voit notamment un ancrage temporel fort pour ces classes, et notamment pour la partie Activités. D'autres classes semblent moins dépendantes à cette dimension temporelle, notamment les classes Accueil / Présentation et Accès web, avec des ULH de marquage énonciatif plus pérennes. Ces tendances émergentes constatées ici seront à confirmer dans le temps, en observant le même corpus de sites web à échéance de deux ans.

Cette approche permettra aussi d'apporter des éléments d'information sur l'évolution des thématiques et des priorités, en accord avec les changements organisationnels. Cela suppose de descendre à un niveau

micro pour étudier les évolutions terminologiques à l'intérieur des classes informationnelles, tout en les situant par rapport au contexte organisationnel concerné, en l'occurrence l'université. Dans cette optique, il serait également intéressant d'intégrer la dimension sémiotique des pages d'accueil pour enrichir les dimensions d'analyse.

Au plan synchronique, d'autres perspectives sont à creuser, notamment en terme de consolidation de la taxonomie, par inclusion des termes de spécialités et traitement des ULH ambiguës. Même si le pouvoir de recouvrement de la taxonomie actuelle est satisfaisant, dans la perspective d'une alimentation automatique de la taxonomie (Rastier et al., 1994, Tellier, 2009), il est nécessaire d'améliorer la structure de la taxonomie. In fine, ces travaux nous permettront, à l'aide d'une typologie de sites web appuyée sur des secteurs d'activités (collectivités territoriales, santé), de transposer cette approche pour des applications à d'autres domaines (Reymond, Pinède, 2010b).

La démarche déployée ici, incluant la dimension temporelle, offre l'intérêt de proposer un angle d'analyse inédit du site web et de voir les permanences et renouvellements à travers l'image saisie de deux corpus d'ULH. Au prisme de la comparaison de ces deux corpus, l'évolution des marqueurs lexicaux –signes et traces-, tant au plan quantitatif que qualitatif, témoigne des changements, que ceux-ci se situent aux niveaux sémantique, informationnel ou encore strictement langagier. Ils témoignent aussi des inflexions de pratiques, en matière de représentation de l'information sur le web, ainsi que de l'évolution des contextes, d'organisation ou d'usage. C'est à l'intersection, aux interstices de ces différentes sphères que s'inscrit la production de sens pour ce document numérique particulier qu'est le site web.

Bibliographie

DRESSEN-HAMMOUDA D., DROT-DELANGE B. (2009). « Expertise et maîtrise des structures informationnelles : le cas de la documentation professionnelle en ligne pour les concepteurs web », Colloque «Evolutions technologiques et information professionnelle », organisé par le GRESEC, Université Stendhal Grenoble III, 10-11 décembre 2009.

JEANNERET, Y. (2007). *Y a-t-il (vraiment) des technologies de l'information ?* Lille, Presses universitaires du Septentrion.

LAINE-CRUZEL S. (2009). Documents, ressources, données : les avatars de l'information numérique. *Information-Interaction-Intelligence*, vol. 4, n°1, p 105-119.

LELEU-MERVIEL S. (2004). Effets de la numérisation et de la mise en réseau sur le concept de document. *Information-Interaction-Intelligence*, vol. 4, n°1, p. 121-140.

MUSSO P. (2009). « Critique de la notion de « territoire numérique ». Les dilemmes de l'économie numérique, sous la direction de Laurent Gille, Limoges, FYP Éditions (collection Innovation), p. 168-175.

NIELSEN J., TAHIR M. (2002). *L'art de la page d'accueil*. Paris, Eyrolles, 2002.
PINEDE N., Reymond D. (2010). De la diversité au lissage informationnel : création d'une taxonomie inductive pour les sites web universitaires. 17e congrès de la SFSIC: au cœur et à la lisière des SIC, Dijon 23-26 juin 2010.
RASTIER, F. CAVAZZA, M, ABEILLE, A. (1994). Sémantique pour l'analyse : de la linguistique à l'informatique, Paris, Masson.
REYMOND D., PINEDE N., LESPINET-NAJIB V., LE BLANC B. (2011). "Une étude terminologique de la communication hypertexte web. Application au domaine universitaire." 9th international conference on terminolgy and articial intelligence. Paris, 8-10 novembre.
REYMOND D., PINÈDE N. (2010a) "Using a taxonomy based fingerprint: classification and recognition of the academic webspace". *Proceedings of the Sixth International Conference on Webometrics, Informetrics and Scientometrics (WIS) & Eleventh COLLNET Meeting*, Mysore, India, 2010.
REYMOND D., PINÈDE N. (2010b) "Website and communication strategy alignment: a librarian science approach to webometrics tools". *Proceedings of the Sixth International Conference on Webometrics, Informetrics and Scientometrics (WIS) & Eleventh COLLNET Meeting*, Mysore, India, 2010.
REYMOND D., PINEDE N. (2010c). "Améliorer la "lecture" du web. Synthèse informationnelle des interfaces web". Communication aux journées d'études TICIS, Paris, 13-15 décembre.
ROUQUETTE S. (2009). L'anayse des sites internet. Une radiographe du cyberspace. Bruxelles, de Boeck , 2009.
SOUCHIER E., JEANNERET Y. et LE MAREC J. (2003). (dir.). *Lire, écrire, récrire. Objets, signes et pratiques des médias informatisés*, Paris, BPI, 2003, 335 p.
STOCKINGER P. (2005). *Les sites Web. Conception, description, évaluation*. Paris, Hermès- Lavoisier, 2005, 270 p.
TELLIER I. (2009). « Apprentissage automatique pour le TAL : Préface », *Traitement Automatique des Langues* 50, 3, p.7-21.
THATCHER A. (2008). « Web search strategies: The influence of Web experience and task type », Information Processing and Management, vol. 44, n°3, p. 1308-1329.

Un modèle d'architecture de pages web pour une accessibilité augmentée destinée aux non-voyants

Mustapha MOJAHID
Bou Issa YOUSSEF
Bernard ORIOLA
Nadine VIGOUROUX

Institut de Recherche en Informatique de Toulouse - Equipe Elipse
(Etude de L'Interaction Personne SystèmE)
Université Paul Sabatier, Toulouse

Résumé : Dans cet article, nous montrons que les non-voyants ne peuvent accéder à toutes les informations sur le web malgré le respect des normes définies par le W3C et les outils disponibles sur le marché. Un sous-ensemble de ces informations inaccessibles est constitué par les structures visuelles et les relations existantes entre différents objets visuels. Notre contribution consiste à élaborer un modèle pour intégrer cette accessibilité augmentée. Pour mener ce travail, nous nous basons (1) sur les concepts de l'architecture de texte et de la granularité des niveaux proposés par le langage notationnel des images de pages et (2) sur l'analyse de corpus et l'observation d'experts et de sujets non-voyants. Un système a été développé pour permettre la génération des pages tactiles et les premiers résultats sont avérés très positifs.

Mots-clés : Non-voyant, accessibilité augmentée, architecture textuelle, relations rhétoriques, image de page.

1. Introduction

La contribution de cet article se situe dans le cadre de l'amélioration de l'accessibilité à l'information visuelle des pages web pour les non voyants. Nous proposons un modèle qui prend en compte les informations visuelles qui restent cachées aux non-voyants telles que l'organisation globale (du « first glance » au « first touch ») de la page, certaines propriétés visuelles d'objets constituant la page web et certaines relations entre ces objets. Ce modèle se base fondamentalement sur : d'une part le concept des *Images de Pages* (IdP) (Luc, Mojahid et Virbel 2001, Mojahid 2011) et sur le *modèle d'architecture textuelle* (MAT) (Vibel

1988, Pascual 1991) où un objet textuel ou une relation peuvent être décrit par un *méta-discours* ou un ensemble de *méta-phrases* vérifiant un ensemble de règles ; et d'autre part, sur le modèle *RDF (Resource Description Framework)* (RDF-Schema, 2004) développé par le World Wide Web Consortium (W3C). Notre modèle, que nous appelons MAP-RDF (modèle d'architecture des pages web) permet de produire une image tactile de la page web à partir des *métas-données* des éléments visuels constituant la page.
Nous présentons d'abord le contexte et les problèmes qui se posent à un non voyant lorsqu'il souhaite consulter une page web (section 2) et nous dressons un état de l'art des recherches dans ce domaine (section 3). Nous présentons ensuite notre modèle MAP-RDF (section 4) et nous finirons par l'évaluation (section 5) du système GIVRA (Gestion des Informations Visuelles des pages web en vue de les Rendre Accessibles) (section 4) développé sur la base de ce modèle.

2. Contexte

L'organisation mondiale de la santé (OMS) estimait en 2003 à 180 millions le nombre de personnes atteintes d'incapacités visuelles, dont 40 à 45 millions sont non-voyants. Plus récemment, en mai 2009, les chiffres étaient de l'ordre de 314 millions de personnes souffrantes de déficiences visuelles dont 45 millions de non-voyants. Ces chiffres pourraient doubler d'ici 2020, en raison notamment de l'accroissement démographique et du vieillissement des populations.
Les interfaces classiques actuelles utilisant les écrans graphiques et la souris posent de sérieux problèmes contrairement aux outils de la génération d'avant où toute interaction s'effectuait avec un écran en mode texte et un clavier. Les solutions apportées consistent à transformer une information en mode graphique en un texte lu par une synthèse vocale ou affiché en braille, en prélevant à la source les informations en mode texte. Quant à la souris, on a compensé son utilisation par le clavier. Cela résout partiellement les problèmes posés puisqu'il arrive souvent que les consignes de rédaction formulées par le W3C ne soient pas suivies. Aussi, remplacer la souris par le clavier rend l'interaction difficile puisqu'il faut augmenter le langage de commande, ce qui provoque une grande charge cognitive vis-à-vis des traitements supplémentaires occasionnés.
Pour illustrer les problèmes de navigation rencontrés par un non-voyant, voici une liste non exhaustive :
- les graphiques sans proposition d'alternative textuelle (un drapeau indiquant le choix de la langue pour un site ou un texte en image à recopier dans une boîte d'édition pour sécuriser l'accès à une page ou l'achat en ligne),

- l'ouverture de nouvelles fenêtres lorsque l'on clique sur un lien, les pages qui se rafraîchissent régulièrement, l'utilisation de flash, les bandes sonores qui tournent en boucle sur un site et couvrant le son de la synthèse vocale,
- les claviers logiciels,
- les mises en évidences (couleur, gras, encadrement...),
- les structures spatiales des pages composées de plusieurs objets et linéarisées pour une possible oralisation pour un non voyant,
- le repérage des informations.

Les outils actuels (voir section 2) ont montré leur limite pour résoudre ces problèmes. Nous montrerons à l'aide de l'exemple suivant (site de Biarritz) que malgré les normes proposées par le W3C, la plupart de ces problèmes persistent et restent sans solution.

Pour vérifier l'accessibilité de ces sites, nous avons utilisé l'outil Wave de Webaim (Wave, 2010). Il s'agit d'un outil gratuit, permettant l'évaluation de l'accessibilité d'une page web. Wave aide à évaluer une page web, en se référant aux recommandations des WCAG et de la section 508 (Etats-unis) en indiquant s'il y a une erreur, sans toutefois pouvoir vérifier automatiquement l'accessibilité. À titre d'exemple, Wave ne peut pas dire si l'alternative textuelle est appropriée, il se contente de pointer son absence. En passant au test la page d'accueil, l'outil Wave a détecté 46 erreurs et alertes d'accessibilité. Nous avons constaté que parmi toutes ces erreurs signalées, aucune indication n'est mentionnée sur l'accessibilité de l'organisation spatiale de la page et les possibles relations entre les différents objets textuels ou non textuels de la page. L'accès à ce deuxième niveau, appelé accessibilité augmentée est validé par une expérimentation (voir section 6). Par ailleurs, l'observation des indices non captés nous a amené à étudier un corpus plus large (voir section 3) pour définir les objets visuels et les relations à modéliser et à intégrer dans notre modèle. Dans cet article, nous proposons un modèle et un outil qui fournissent des éléments de réponse aux problèmes correspondant à l'accessibilité augmentée apportée par les structures spatiales de la page. Une de nos hypothèses générales est de considérer que ces structures vont faciliter (voire permettre) le repérage des informations dans une page web.

Dans la section suivante, nous dressons un état de l'art des travaux qui ont oeuvré dans la thématique de l'accessibilité.

3. État de l'art

Dans cette section, nous présentons quelques travaux qui traitent l'accessibilité à l'information visuelle notamment liée à la structure de la page web et au contenu informationnel organisé visuellement.

3.1. Accessibilité des tableaux de données

Oogane & Asakawa (1998) ont proposé une méthode de conversion de représentations visuelles, et particulièrement les tableaux, en des représentations non visuelles.
Leur méthode traite de la représentation des tableaux à deux dimensions, mais ne convient pas pour les tableaux non cartésiens qui ont une représentation différente comportant par exemple des cellules fusionnées.
Dans le même contexte, Filepp et al. (2001) ont proposé le TTPML (Table To Prose Markup Language), un langage de balises compatible avec XML. Ce langage facilite la génération des descriptions discursives des tableaux HTML. La méthode consiste à transformer les données dans un tableau en des phrases explicitant les relations entre les lignes et les colonnes. Cette opération de transformation est réalisée en associant des règles de transformation des tableaux pour permettre aux lecteurs d'écrans de lire correctement les éléments du tableau.

3.2. Accessibilité des documents de présentation

Ishihara et al. (2006) ont travaillé sur l'accessibilité des diagrammes dans les documents de présentation. Un diagramme est formé d'objets ou de groupes d'objets qui sont visuellement reliés par des flèches. Les auteurs ont alors proposé une méthode pour créer un modèle de description concernant trois types de relations entre les objets : un objet visuellement inclus dans un autre ; plusieurs objets localisés et alignés par rapport aux autres objets de la même diapositive ; et un objet ou un groupe d'objets lié à un autre par une flèche.
En appliquant ce modèle aux différents objets et relations détectés, la méthode consiste à transformer le document en une organisation arborescente qui contient la structure visuelle du document et l'ordre d'apparition des objets. Cette méthode a été ainsi implémentée en développant DocExplorer qui permet de naviguer entre les différents objets et relations.

3.3. Accessibilité de la structure visuelle d'une page web

Asakawa et al. (2000) ont travaillé sur l'accessibilité de la structure de la page aux non voyants. Ils ont proposé un système de transformation basé sur l'annotation de la structure afin de distinguer les groupements visuels d'une page. Ces groupements ayant différents rôles (menu principal, sous-menu, publicité, liens) sont souvent identifiables par des indices visuels (couleur, forme, disposition des éléments). La séquence de balises HTML cependant lue par les lecteurs d'écrans ne reflète pas cette structure et n'explicite pas les relations liant les objets.
Dans leur système, les auteurs distinguent deux annotations : celle de la structure et celle du contenu textuel. Les annotations de la structure sont utilisées pour reconnaître les groupements visuellement fragmentés, et

par conséquent pour montrer l'importance sémantique de chaque groupe. Les annotations du contenu textuel ont pour rôle de décrire les objets multimédias inaccessibles par le non-voyant. Le système produit une page reformulée, indiquant la réorganisation des groupes suivant leur rôle et leur importance.
De son côté, le W3C par le biais de sa plate-forme ARIA (Accessible Rich Internet Applications) aborde le problème de l'accessibilité au contenu dynamique en utilisant le contenu des balises. Il fournit une méthode standard pour l'attribution de rôles et des états des éléments DHTML, et pour décrire la mise à jour de ces éléments dynamiques.
Ainsi, le non voyant dispose des descriptions textuelles concernant la composition de la page, mais il ne dispose toujours pas d'une représentation réelle bidimensionnelle de la page.
Pour sa part, IBM Accessibility (2009) a proposé une réorganisation de la page suivant les différents groupements et leurs rôles. Ceci altère la structure initiale de la page, et produit une structure modifiée qui reste toutefois non bidimensionnelle.
Comme on l'a vu, aucun des travaux pré-cités ne permet réellement à un déficient visuel d'appréhender la structure d'une page Web dans sa globalité par le concept du « premier toucher ». C'est pourquoi nous proposons donc notre modèle MAP-RDF, que nous allons décrire dans la section suivante.

4. Le modèle MAP-RDF

Pour élaborer notre modèle, nous avons commencé par analyser un corpus de pages web afin d'identifier les indices, les objets et relations qui posent problème aux non-voyants malgré le respect des directives du W3C.
Dans cette étude, nous nous intéressons particulièrement à l'accessibilité des indices visuels indispensables à la compréhension de la structure de la page.
Nous avons choisi d'analyser des sites web touristiques[31]. Ce choix est justifié d'une part, par la loi européenne qui consiste à rendre tous les sites officiels publics de l'Union Européenne accessibles par tous les internautes, sans discrimination, en suivant les recommandations. D'autre part, le choix des sites web correspondait à celui d'une étude psycholinguistique (Etcheverry, I. 2007 ; Etcheverry, I. Terrier, P. Marquie, J.C. 2007) qui a porté sur les difficultés de localisation visuelle et les exigences cognitives dans la recherche des informations dans une page web. Les critères du choix de ces sites web répondent ainsi aux

[31] Les analyses ont été effectuées courant juillet 2009.

exigences suivantes. Les pages doivent constituer un corpus homogène et comparable en termes de quantité de données textuelles et de graphiques et traiter les mêmes types de contenus d'information.

Ces sites web déclarent qu'ils respectent les règles du W3C conformément aux lois européennes. À titre d'exemple, le site web de la Grande Bretagne dédie une page à l'accessibilité[32]. Ils déclarent que le site est développé « *afin de servir un public aussi large que possible* », en apportant « *une attention toute particulière aux besoins des moins valides, en créant un site compatible aux programmes spécifiques qui leur sont destinés* ». Le site se dit conforme à la « Web Accessibility Initiative » du « World Wide Web Consortium », et dit respecter les standards définis par les « UK Government Web Accessibility Guidelines ».

Nous avons analysé un corpus de huit sites web touristiques de pays ou de villes européennes : nicetourisme.com ; biarritz.fr ; parisinfo.com ; bordeaux-tourisme.com ; allemagne-tourisme.com ; visitbritain.fr ; visitportugal.com ; holland.com/fr.

Une évaluation avec l'outil Wave a été réalisée sur les 8 sites. Nous avons rejeté le site web de Bordeaux parce que nous avons constaté que sa page d'accueil est constituée d'objets flash avec animation audio-visuelle qui n'est pas accessible.

Cette évaluation a permis de vérifier que le corpus des sites web choisi est conforme aux normes Content Accessibility Guideline (WCAG) ; cela représente le point de départ pour tester le deuxième niveau d'accessibilité : celui de l'accessibilité augmentée. Les mêmes sites touristiques ont été consultés par des non-voyants avec le lecteur d'écran Jaws[33]. Cette consultation a pour but d'identifier les objets visuels et les relations visuelles non pris en compte par Jaws.

D'une façon générale, Jaws, comme tous les lecteurs d'écran, lit linéairement le contenu textuel et ne fournit pas d'information concernant la disposition et l'emplacement des objets dans la page web. Cependant Jaws fournit à la demande, les couleurs des caractères, les couleurs de fond et le contraste entre les couleurs, les propriétés typographiques d'italique, de gras et de soulignement sous forme textuelle et par conséquent, cette restitution orale entraîne une énorme surcharge cognitive pour le non-voyant.

Dans notre observation des sites web, nous avons pu identifier les groupements d'objets visuels constitués par des « menus », « rubriques », « bandeaux », « agendas » et « formulaires ». Ces groupements visuels sont formés à partir d'objets visuels élémentaires possédant des

---

32 http ://www.visitbritain.fr/corporate/accessibility.aspx.

33 Jaws (Job Access With Speech) est un lecteur d'écran : un logiciel pour les déficients visuels. Il transforme un texte affiché sur un écran en un texte oral (par un système de synthèse vocale) ou un texte enbraille, et permet d'interagir avec le systhème d'exploitation et les logiciels.

propriétés visuelles ou discursives et permettant de les identifier isolément. Nous détaillons aux paragraphes suivants les caractéristiques des différents objets visuels et nous donnons l'exemple du groupement visuel, le menu.

4.1. Les objets visuels

Il s'agit essentiellement des :

- blocs de texte : segments de texte constitués d'un ou de plusieurs mots, formant une ou plusieurs lignes avec les mêmes propriétés typographiques homogènes (police, graisse, italique, souligné, couleur de caractère et de fond) ;
- titres : unités lexicales identifiées par des indices discursifs et des marqueurs organisationnels, par exemple la numérotation, ou « info plage, météo, culture, que faire, où se loger, et des indices dispositionnels, par exemple, les espaces avant, après à gauche et à droite (Virbel et al. 2005 ; Ho-Dac et al. 2010) ;
- liens hypertextes : éléments (textuels ou graphiques) de la page web identifiés à partir des balises HTML. Ils sont différentiés visuellement par un habillage de texte et avec un lexique spécifique, par exemple, lire la suite, +Infos, > en savoir plus ;
- images : elles peuvent avoir plusieurs rôles : images de fond, images-liens, logos, images informationnelles du contenu. Elles se caractérisent notamment par leurs dimensions, leurs formes et leurs couleurs ;
- Champs d'entrée : il peut s'agir d'un champ à remplir dans un formulaire ou une zone de recherche, ou encore un bouton. Les champs d'entrée sont identifiés à partir des balises HTML et généralement, ils sont différentiés visuellement par des cadres et des habillages visuels spécifiques selon les fonctions.

L'ensemble de ces objets visuels forment des groupements visuels. Pour nous aider à définir et caractériser ces groupements, nous avons repris d'une part, les deux principes fondamentaux du modèle d'architecture textuel (Virbel 1989, Pascual 1991) concernant le principe de ressemblance et de différence et celui des équivalences de formulations discursives développées et à réalisation réduite et des lois de la théorie de la Gestalt (Guillaume 1937). Ces groupes d'objets dénotent et sont la manifestation de relations contextuelles intra ou inter objets visuels que nous nous allons illustrer.

A l'état actuel de ce travail (Bou Issa 2010), nous avons exploré les groupements visuels constitués par les différents menus, les rubriques et les bandeaux. Dans cet article, nous focaliserons la présentation sur les menus. Nous en donnons une définition, que nous avons pu dégager par l'étude du corpus et nous illustrons quelques problèmes liés à l'accessibilité de ces groupements et leurs relations par le non-voyant.

4.2. Les Groupement visuel : l'exemple du menu

Nous définissons un menu par un ensemble d'objets élémentaires liens hypertextes, groupés visuellement par une même typographie et qui sont adjacents. Les liens hypertextes des menus sont formés à partir d'un groupe de mots ou d'images labellisées formant une structure énumérative fonctionnellement parallèle (Luc 2001 ; Virbel et al. 2005 ; Ho-Dac et al. 2010).

Il peut s'agir en particulier du menu principal de la page, d'un sous-menu ou d'un menu secondaire. Dans le cas du menu principal, il est disposé horizontalement, dans la première moitié de la page et occupant une largeur supérieure à 50% de la largeur de la page. Si deux menus sont disposés horizontalement et que chaque menu occupe plus de 50% de la largeur de la première moitié de la page alors les attributs typographiques de saillance comme la différence de gras et la taille des caractères ou bien le gras et la couleur de fond du menu permet de caractériser le menu principal mis en saillance.

Problèmes liés à l'accessibilité aux menus

L'exemple du site d'Allemagne illustre les problèmes rencontrés dans la plupart des sites du corpus. Le menu principal est formé d'items disposés horizontalement dont la couleur du fond est la même, mais le côté inférieur du lien a une couleur différente des autres. C'est cette même couleur, mais avec un contraste plus faible, qui indique que le curseur est placé sur une case donnée. Cette même couleur est reprise dans le sous-menu qui contient trois niveaux de bleu différents et l'arborescence est réalisée par des effets de contraste de couleur, d'indentation et de soulignement vertical, effets qui ne sont également pas perçus par un lecteur d'écran.

Figure 1. Menu principal et sous-menu du site de l'Allemagne

4.3. Relations entre les groupements visuels

Les groupements visuels peuvent être associés entre eux et entretenir des relations non perceptibles par un non-voyant. Ces relations complètent les éléments de l'accessibilité augmentée pour capter la structure d'une page web. Nous avons donc observé particulièrement deux types de relations : les relations de nature visuelle, et les relations contextuelles en rapport avec le contexte/les thèmes du site web.

4.3.1. Les relations visuelles

Deux types de relations visuelles ont été distinguées. Les relations géométriques ou spatiales sont définies par le voisinage entre objets (à gauche, à droite, en haut, en bas, symétrie) et les relations typographiques de ressemblance/différence fournies par les propriétés de la couleur, le style de police, ou la couleur de fond.

4.3.2. Les relations contextuelles

Ces relations concernent l'environnement, les circonstances, ou les développements qui précisent le contenu des groupements visuels. Nous distinguons deux types de relations contextuelles : les relations thématiques associées à un champ lexical commun donné par leurs titres. Il s'agit des groupements qui se rapportent à un même thème ou illustrent un même sujet. Par exemple les rubriques « Guide de voyage » et « Réservation d'hôtel » se rapportent au thème du voyage. Les relations de développement associent un objet et un groupement le détaillant, l'exemplifiant… Par exemple « où aller » du menu principal est détaillé dans la rubrique informationnelle dont le titre est « où aller ». Une relation contextuelle « détail » est donc dénotée entre les deux groupements visuels en question.

Après cette phase d'analyse du corpus, nous avons élaboré le langage de formalisation des objets visuels élémentaires des groupements visuels et des relations. L'objectif est de préparer les données pour produire ensuite une présentation dans la modalité tactile.

4.4. Le langage MAP-RDF

Plusieurs modèles et théories nous ont servi : (1) le Modèle d'Architecture Textuelle développé par Virbel (1989) et Pascual (1991) pour représenter la dimension visuelle du texte et les équivalents langagiers des phénomènes textuels (métalangage) ; (2) la théorie des structures rhétoriques (RST) développée par Mann et Thomson (1981) pour représenter les relations rhétoriques entre les différents objets de la page ; (3) le modèle des Images de Page (IdP) développé par Luc et al. (2001), pour « matérialiser » visuellement les représentations architecturale et rhétorique obtenues grâce et à la RST et au modèle MAT. L'IdP est basée sur un langage notationnel s'inspirant de la théorie de notation de Goodman (Goodman 1990).

Les principaux avantages du modèle IdP se résument par :
son expressivité : représentation facile et compréhensible des propriétés visuelles du texte ;
- sa représentation bidimensionnelle des phénomènes textuels pour permettre une vue globale de la structure du texte ;
- la possibilité de fournir des descriptions selon plusieurs niveaux de granularité.
Étant donné le volume important et hétérogène d'informations contenues dans une page web (objets visuels, groupement et relations), le langage des images de page nous donne un outil efficace pour répartir ces informations dans plusieurs niveaux d'image de page selon des critères liés aux différents processus/métaphores de lecture (en diagonale, approfondie, recherche d'information…).
Pour définir une stratégie d'affectation des objets/relations prélevés dans le corpus, nous avons réalisé une enquête auprès de 12 experts concepteurs de sites. Pour chacun des objets du corpus, les sujets devaient proposer une note de 0 à 10 évaluant selon eux, les degrés d'importance.
Les résultats de l'enquête nous a amené à diviser l'affichage pour les non-voyants en trois niveaux, nous donnerons un exemple à la fin de cette section :
- le niveau I (notes ≥ 6) contiendra les menus et leurs rôles ; les rubriques et leurs rôles ; le bandeau, son caractère visuel, formes visuelles des objets ;
- le niveau II (4 < notes < 6) contiendra en plus des informations du niveau I, le caractère visuel des menus ; le caractère descendance des menus ; le caractère visuel du bandeau ; le caractère visuel des rubriques ; les titres des rubriques ; les relations contextuelles entre ces groupements.
- le niveau III (notes ≤ 4) contiendra en plus des informations précédentes la disposition interne des groupements visuels ; les styles et les polices de caractères avec et sans effet souris ; la couleur principale des groupements.
Nous avons également fait appel (4) au modèle RDF[34] (Resource Description Framework) qui permet de décrire les propriétés des pages web. Les avantages de ce modèle se résument par : la possibilité de créer nos propres ontologies pour décrire les métadonnées de la page web ; la facilité de son implémentation ; et la compatibilité avec les navigateurs.
Dans le modèle MAP-RDF, la page web est structurée suivant un ensemble de triplets :
1. le sujet ou la ressource : qui peut être un groupement visuel ou une relation entre deux groupements ;

[34] Resource Description Framework (RDF) est un modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs méta-données, de façon à permettre le traitement automatique de telles descriptions. Développé par le W3C, RDF est le langage de base du web sémantique. Une des syntaxes de ce langage est RDF/XML.

2. le prédicat : qui représente l'ensemble des propriétés visuelles et lexicales applicables à cette ressource. Nous distinguons trois types de prédicats : de mise en forme, de caractérisation et de présentation ;
3. l'objet : qui correspond à la valeur du prédicat concerné.

4.4.1. Les ressources « groupements visuels »
Dans le cadre de cet article, nous nous focalisons sur le cas des menus. Le lecteur se rapportera à (Bou Issa 2010) pour les rubriques et les bandeaux. Plusieurs prédicats sont nécessaires pour une modélisation complète, Nous citons quelques exemples :
Prédicat de caractérisation :
Caractère : {déroulant ; fixe} ; Rôle : {menu principal ; menu secondaire}
Descendance : {menu ; sous-menu} ; Type des liens : {texte, texte imagé, image labellisé}
Prédicat de mise en forme :
Couleur originale du texte ; lors événement souris et après visite : {RGB}
Couleur du texte du fond, avec événement souris : {RGB}
Dimensions : {Dimension} ; Coordonnées : Abscisse Ordonnée
Prédicats de présentation :
Niveau d'affichage : {niveaux de granularité des IdP}.
Pertinence : pour associer un degré d'importance dans la page.
Modalité de sortie : dans cette version la seule modalité est tactile (voir perspective).
Selon le même schéma, nous avons défini les ressources « relation ».

4.4.2. Génération de la présentation tactile à partir du modèle MAP-RDF
Image de page web tactile
La perception visuelle est complètement différente de la perception tactile (Bingham et al. 2007). Il est nécessaire de retoucher les représentations visuelles afin de les reproduire en tactile. Nous nous référons aux travaux de (Socrate-Comenius, Project, 1999-2000) sur les consignes à suivre lors de la production d'un document tactile. Voici deux extraits à propos des distances minimales et des reliefs et perçues par le non-voyant.

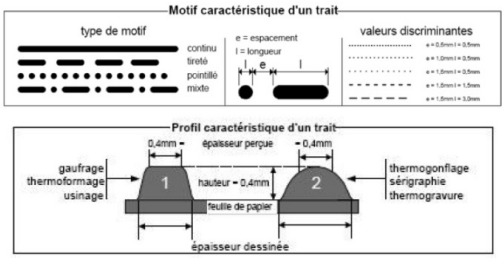


Figure 2. Motif et profil caractéristique d’un trait en relief emprunté à Socrate-Comenius, Project, 1999-2000

En tenant compte de ces consignes sur les limitations du tactile, nous avons conçu des symboles pour représenter les triplets de notre modèle MAT-RDF, par exemple (Figure 4) :

Ressource	Prédicat	Valeur	Symbole
menu	caractère	déroulant	
		fixe	
	rôle	principal	
		secondaire	

Ressource	Prédicat	Valeur	Symbole
relation contextuelle	caractère	même thème	
		père-fils	
		extrait	
		zoom	

Figure 3. Exemples de symboles associés aux groupements visuels et aux relations

Pour illustrer une instanciation du modèle MAP-RDF, voici la représentation tactile de l’exemple du niveaux 1 de l’IdP de la page d’accueil du site d’Allemagne telle qu’elle est imprimée puis passée dans un four thermogonflable de type « ZY-Fuse » pour avoir une reproduction tactile de la page.

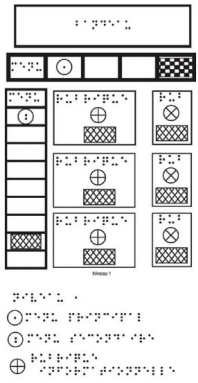


Figure 4. Premier niveau d’affichage d’une Image de Page Web et sa légende

5. Système de gestion des informations visuelles en vue de les rendre accessibles : vers un nouveau standard d'accessibilité

Le système de Gestion des Informations Visuelles des pages web en vue de les Rendre Accessibles (GIVRA) permet :
d'annoter par des concepteurs web, les pages selon le modèle MAP-RDF ;
de fournir et générer aux utilisateurs non voyants une interface de lecture des images de pages.
Tout d'abord, le système extrait automatiquement certaines propriétés des groupements telles que leurs coordonnées, les dimensions et les propriétés de typographie. Ensuite, cette extraction automatique est complétée par une annotation manuelle selon la description du modèle MAP-RDF (voir section 3.). Le système pourra ainsi fournir une reformulation de ces propriétés selon les trois niveaux d'affichage défini par le modèle MAP-RDF.

5.1. Module d'annotation

Ce module propose au concepteur web d'annoter une page web en la décomposant selon les groupements visuels, en utilisant une interface graphique (figure 5). Cette interface propose les fonctionnalités nécessaires pour ajouter des objets, des groupements d'objets ainsi que leurs propriétés et les relations qui les relient.

Figure 5. Interface graphique du module d'annotation du système GIVRA

5.2. Module de génération

Une fois annoté les groupements et les relations dans une page web, le module de génération se charge de la transformation dans la modalité tactile. Cette opération est réalisée en plusieurs étapes :

la récupération des symboles visuels de la base des symboles développés sous Visio (voir 3.4.2).
l'arrangement des symboles et la génération de l'Image de Page correspondante à la page analysée, selon le niveau de granularité souhaité, en suivant les règles que nous avons défini dans le modèle (figure 3).
l'impression : l'image de page ainsi produite sera imprimée, puis passée dans un four du type « ZY-Fuse » pour avoir une reproduction tactile de la page.

6. Résultats préliminaires

Nous avons réalisé une expérience pour tester les productions tactiles des images de page auprès de six personnes non-voyantes. Tous les sujets utilisent régulièrement Internet pour des tâches quotidiennes (courriels, achats en ligne, consultation de factures…).
La première étape a consisté en une phase d'apprentissage du langage des éléments des images de pages afin de s'approprier les représentations tactiles associées. Aucun des sujets n'avait de mal à comprendre et à mémoriser les symboles proposés.
Nous avons cherché ensuite à tester quatre hypothèses qui avancent l'intérêt des IdP pour l'accès aux groupements visuels de la page ; aux propriétés visuelles des constituants de la page ; aux relations visuelles et contextuelles entre les constituants de la page ; et enfin pour l'accès global et non seulement séquentiel à la page. Les résultats obtenus nous permettent de valider toutes nos hypothèses (Bou Issa 2011). Les utilisateurs ont globalement bien perçu le rôle, le titre quand il y en avait un, et la position de chaque groupement dans la page.
Pour finir, l'enquête de satisfaction auprès des utilisateurs nous a montré qu'ils accueilleraient très favorablement un tel dispositif s'il était disponible.

7. Conclusion et perspectives

Nous avons montré dans cet article, que même si un site web respecte les normes d'accessibilité définies par le W3C, il reste un niveau inaccessible pour les non-voyants qui comprend toutes les informations visuelles des pages web, y compris la structure globale et les relations entre les objets et les groupements visuels. Des tests préliminaires et une étude d'un corpus de pages web nous a permis de cerner les problèmes posés aux non-voyants et les éléments à prendre en compte. Nous avons ensuite proposé un modèle d'architecture de page web qui améliore l'accessibilité à ces informations et nous l'avons illustré sur

l'exemple des menus. Pour mener à bien ce travail, nous nous sommes appuyés sur les modèles d'architecture textuelle des images de pages et sur les théories notationnelles et des structures rhétoriques. Le modèle RDF a été utilisé pour formaliser le langage de représentation sous forme de triplets ressource-prédicat-valeur. Nous avons présenté l'architecture générale du système de gestion des informations visuelles qui permet aux concepteurs d'annoter les pages web et de générer automatiquement une sortie tactile. Cette production tactile est structurée sous forme d'images de pages selon trois niveaux de granularité identifiés expérimentalement et en tenant compte des règles à respecter pour les non-voyants. Nous avons pu valider notre modèle à l'aide d'une première expérience préliminaire qui a pu nous montré des résultats très prometteurs.

Nous projetons plusieurs perspectives à ce travail. Nous envisageons d'abord, de réaliser une expérience sur un plus grand nombre de sujets pour conforter nos premiers résultats. Nous examinerons un plus large corpus de sites web pour développer les concepts de groupements visuels et de relation afin de proposer un modèle encore plus représentatif de la diversité du web (formulaires, agendas, tableaux, etc.), au-delà des groupements particuliers déjà étudiés (menu, rubrique, bandeau). Nous comptons également réaliser améliorer l'outil GIVRA, notamment en ce qui concerne l'extraction automatique des propriétés visuelles et des relations dans les pages web.

Une dernière perspective à ce travail est d'étudier de nouvelles stratégies pour combiner les deux modalités tactile et orale et de réaliser l'affichage sur un écran tactile. En effet, la non-existence d'afficheur dynamique, nous a amenée à simuler une production tactile sur un papier thérmogonflable.

Bibliographie

ASAKAWA C., TAKAGI, H. (2000). Annotation-based transcoding for nonvisual web access. ASSETS'00 Proceedings of the fourth international ACM Conference on Assistive technologies (pp. 172-179). Arlington, Virginia, United States: ACM. Association Valentin Haüy. (2001). Code de Transcription en Braille des Textes Imprimés. Retrieved 9 3, 2008, from http://www.avh.asso.fr/rubriques/infos_braille/nouveau_code_braille.php

BINGHAM J., CAVENDER A., BRUDVIK J., WOBBRACK, J. (2007). A comparative analysis of blind and sighted browsing behavior. Assets'07: Proceedings of the 9th international ACM SIG ACCESS conference on Computers and Accessibility. Tempe, Arizona, USA.

BOU ISSA Y. (2011) Accessibilité aux informations visuelles des pages web pour les non-voyants, thèse de l'université Paul Sabatier, Toulouse.

BOU ISSA Y., MOJAHID M., ORIOLA, B., VIGOUROUX N. (2010). Analysis and evaluation of the accessbility to visual information in web pages. ICCHP, International Conference on Computers Helping People with Special Needs. Vienna University of Technology, Austria.

CAUCHARD, F. (2008). Empan perceptif en lecture et en recherche d'information dans un texte : influence des signaux visuels. Thèse de Docotorat . Université de Toulouse II.
COLAS, S. (2008). Outils d'amélioration de l'accessibilité du web pour les personnes visuellement handicapées. Université de Tours.
ETCHEVERRY, I., TERRIER, P., & MARQUIE, J. C. (2011). Are older adults less efficient in making attributions about the origin of memories for web interaction? European Review of Applied Psychology - Special Issue on Information Searching (J. Dinet & A. Chevalier, Eds).
Lorch, R. F., Jr, Lemarié, J., & Grant, R. A. (2011). Signaling Hierarchical and Sequential Organization in Expository Text. Scientific Studies of Reading.
HO-DAC, L. M., FABRE, C., PÉRY-WOODLEY, M. P., REBEYROLLE, J. On the signalling of macrodiscourse structures, 8th Multidiscplinary Approaches of Discourse - MAD2010, Moissac – France, 17-21 march 2010.
FILEPP, R., CHALLENGER, J., ROSU, D. (2001). Improving the accessibility of aurally rendered HTML tables. ASSETS'02: Proceedings of the fifth international ACM conference on Assistive technologies (pp. 9-16). Edinburgh, Scotland: ACM.
GUILLAUME, P. (1937). La psychologie de la forme. Paris, Flammarion, 1979.
GOODMAN. (1990). Langages de l'Art. Trad. J. Morizot, Editions Jacqueline Chambon.
Hailpern, J., Guarino-Reid, L., Boardman, R., Annam, S. (2009). Web 2.0: blind to an accessbile new world. Proceedings of the 18th international conference on world wide web (pp. 821-830). Madrid, Spain: ACM.
IBM, Human Ability and Accessibility Center. (2009). Retrieved 3 1, 2010, from Creating accessible microsoft powerpoint documents: http://www-306.ibm.com/able/guidelines/documentation/docmsppt.html.
ISHIHARA T., TAKAGI H., ITOH T., ASAKAWA C. (2006). Analyzing visual layout for a non-visual presentation-document interface. ASSETS'06: Proceedings of the 8th international ACM SIG ACCESS Conference on Computers and Accessibility (pp. 165-172). Portland, Oregon, USA: ACM.
LUC C., MOJAHID M., VIRBEL J. (2001). Système notationnel de l'architecture textuelle par image de page. Conférence Internationale sur le Document Electronique, Toulouse.
LUC, C., (2001) Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte, Traitement Automatique du Langage Naturel.
MANN, W. C., THOMPSON, S. A. (1988). Rhetorical structure theory: toward a functional theory of text organization. Text , 8 (3), 243-281.
MOJAHID, M. (2011) Fondements du langage des Images de Page et sa mise à l'épreuve. 2nd International Conference on Linguistic & Psycholinguistic Approaches to Text Structuring, Louvain, Belgique.
OOGANE, T., ASAKAWA, C. (1998). An interactive method for accessing tables in HTML. ASSETS'98: Proceedings of the third international ACM conference on Assistive technologies (pp. 126-128). Marina del Rey, California, United States: ACM.
PASCUAL, E. (1991) Représentation de l'architecture textuelle et génération de textes. Thèse de l'Université Paul Sabatier, Toulouse.

RDF-Schema. (2004). RDF Schema, language for declaring basic class and types for describing the terms used in RDF. Retrieved from http://www.w3.org/TR/rdf-schema/

SOCRATE-COMENIUS, Project. (1999-2000). Recommendations for transcribing documents. Annexe technique 1. Documentation Michel Bris S.D.A.D.V. CNEFEI Suresnes.

VIRBEL J. (1989). The contribution of linguistic knowledge to the interpretation of text structures. In Structured documents (pp. 161-180). Cambridge University Press.

VIRBEL J., GARCIA-DEBANC C., BACCINO T., CARRIO L., DOMINGUEZ C., JACQUEMIN C., LUC C., MOJAHID M., PERY-WOODLEY M. P., SCHMIDS S. Approches cognitives de la spatialisation du langage. De la modélisation de structures spatiolinguistiques des textes à l'expérimentation psycholinguistique : le cas d'un objet textuel, l'énumération. Dans : Agir dans l'espace. Catherine Thinus-Blanc, Jean Bullier (Eds.), Editions de la Maison des sciences de l'homme, p. 233-254, Cognitique, 2005.

VIRBEL J., NESPOULOUS, J. L. (2004) Handicap langagier et recherches cognitives : apports mutuels. Revue parole.

WAVE : web accessibilty evaluation tool. (2010). Retrieved 8 9, 2009, from Webaim: http://wave.webaim.org

Partie 5 - Pratique du document numérique dans l'univers de la recherche

Pratiques de lecture numérique et usages des technologies de l'écrit chez le chercheur tunisien

Abderrazak MKADMI
Institut Supérieur de Documentation, Université de la Manouba, Tunisie
Membre associé au laboratoire Paragraphe, université de Paris8, France

Besma BSIR
Institut Supérieur de Documentation, Université de la Manouba, Tunisie

Résumé : La lecture numérique est une activité très complexe qui exige la mise en exergue plusieurs acteurs à savoir l'auteur, le concepteur, le lecteur, l'annotateur, etc. ainsi que plusieurs objets comme le support de lecture, le format, la structure et le type de document, etc. Tous ces éléments influent certainement sur les pratiques de lecture. Nous tenterons dans ce travail de présenter cet acte de lecture chez les chercheurs tunisiens à travers la présentation de quelques résultats d'une enquête menée par notre équipe de recherche sur la lecture numérique. Ces résultats concernent essentiellement les types de documents les plus lus par les chercheurs, les formats de documents préférés, ainsi que l'implication des chercheurs dans des activités de collaboration pédagogiques et scientifiques. Nous mettons l'accent essentiellement sur l'indicateur genre dans le but de déceler l'évolution des tendances et des habitudes de lecture et des usages des technologies de l'écrit dans un contexte universitaire et de recherche en Tunisie.

Mots-clés : Chercheur tunisien, lecture numérique, travail collaboratif, document numérique, pratiques de lecture.

Abstract : The digital reading is a very complex activity that requires highlighted several actors like author, designer, reader, annotator, etc. and several objects such as support for reading, format, structure and type of document. All these elements have certainly an effect on practices reading. We will try in this work to present this act of reading among Tunisian researchers through the presentation of some results of a survey conducted by our research team on the digital reading. These results concern mainly the types of documents read by most researchers, formats of documents preferred, as well as the involvement of researchers in collaborative activities teaching and research. We

emphasize mainly on the gender indicator in order to detect the evolution of tendencies and reading habits and uses technology writing in academic and research environment in Tunisia.

Keywords : Tunisian researcher, digital reading, collaborative work, digital documents, reading practices.

Introduction

La lecture sur écran ou la lecture numérique est une pratique qui a profondément évolué ces dernières années vu la dématérialisation des contenus et de la généralisation de l'accès à Internet. L'écran est devenu le support privilégié quant à nos rapports à la culture, à la communication, à l'apprentissage et à la distraction. Dans le monde universitaire le rapport avec l'écran est encore plus présent, et ce dans presque toutes les activités pédagogiques, scientifiques et de recherche.
Notre travail consiste à étudier les nouvelles pratiques de lecture numérique des chercheurs tunisiens à travers une enquête menée auprès de 307 chercheurs exerçant dans les cinq universités du Grand Tunis. Ces cinq universités rassemblent la plupart des chercheurs tunisiens: enseignants et étudiants au 3ème cycle et regroupent presque toutes les disciplines scientifiques [Mkadmi, 2010][35]. Les résultats que nous allons avancer dans ce travail concernent essentiellement les types de documents les plus lus par les chercheurs, les formats de documents préférés, ainsi que l'implication des chercheurs dans des activités de collaboration pédagogiques et scientifiques. Nous mettons l'accent essentiellement sur l'indicateur genre dans le but de déceler l'évolution des tendances et des habitudes de lecture et des usages des technologies de l'écrit aussi bien chez le genre féminin que chez celui masculin dans un contexte académique de recherche, dans un pays du sud telle que la Tunisie.
En effet, en matière de pratiques de lecture numérique, très peu de travaux traitent de l'influence du genre. La majorité de littérature font référence à des études se rapportant à l'usage d'Internet et des technologies de l'information et de la communication en général. Nous pouvons néanmoins les évoquer pour comprendre la tendance générale de l'impact du facteur genre à ce niveau. Collet [2006] relève l'impact du facteur genre au bénéfice des hommes quant à l'usage d'Internet. En Afrique et dans les pays en voie de développement, cet impact en faveur des hommes renforce ce que nous appelons communément la fracture numérique entre les pays. Women's Learning Partnership [2007] présente une étude sur les utilisateurs d'Internet qui montre que le pourcentage

[35] - Voir aussi [Ben Romdhane, 2008], [Limam, 2008] et [Hachicha, 2010].

des femmes utilisant l'Internet n'atteint que 4% dans les pays en développement alors qu'il va jusqu'au 54% dans les pays développés. Cependant, cette tendance masculine de l'utilisation aisée de l'ordinateur et de l'Internet est en partie bouleversée avec l'avènement des réseaux sociaux[36].

1. Lecture numérique et recherche scientifique : equity[37] entre homme et femme

Avant d'aborder les pratiques de lecture dans le milieu universitaire chez les chercheurs tunisiens, nous essayons d'identifier le rapport que les deux genres[38] homme et femme essaient d'entretenir dans le contexte de recherche scientifique.

En effet, dans une logique de parité, hommes et femmes peuvent avoir les mêmes rapports avec la recherche scientifique, puisque tous les deux constituent un genre unique « genre humain ». Dans ce sens on essaie de nier ou même d'abolir toute différence entre la nature des deux genres et aussi fait-on face aux préjugés résiduels résultant de l'histoire. Cette dernière prouve d'un côté que l'homme a fait naître la science pour comprendre la nature, mais de l'autre, elle rappelle que la femme n'a pu entrer la sphère de l'activité scientifique qu'à partir du 19e siècle [Keller, 2003].

Mais en dépit de cette réalité, « la question de savoir si le corps d'un scientifique est mâle ou femelle n'a aucune pertinence », tant que chaque femme pourvue d'un esprit scientifique doit avoir accès à la science [Kelley, 2003]. Ainsi, expliquons-nous la parité du pourcentage des hommes-femmes ayant répondu à notre enquête, dans un contexte scientifique tunisien à l'an 2008[39], par une conscience de l'identité et

---

[36] Nous évoquons ici les réseaux sociaux dans notre enquête parce que nous estimons comme disait Claire Belisle : qu'« une des caractéristiques des réseaux sociaux est de développer la lecture en ligne, tout autant que l'écriture. Se tenir informé, c'est lire, regarder, écouter, échanger, réagir » [Belisle, 2011, p.197].

[37] - Le terme anglais *equity*, désigne la recherche conjointe de l'égalité et de la justice. [Keller, 2003]

[38] - « Le genre est une catégorie culturelle qui modèle notre développement en tant qu'hommes et femmes adultes. En ce sens, le genre représente une transformation culturelle du sexe » ; sachant que le sexe est « une catégorie biologique dans laquelle nous sommes nés comme enfants mâles ou femelles ». [Keller, 2003].

[39] - Sachant que dans le contexte occidental américain, « En 1956, près d'un siècle après avoir admis sa première étudiante, Ellen Swallow Richards, le MIT a réuni une commission spéciale pour examiner la question de savoir s'il fallait ou non continuer à accepter des femmes parmi ses étudiants ; et cette commission a recommandé qu'il soit mis fin à la mixité de la formation dispensée » [Keller, 1981] cité in : http://cedref.revues.org/509#bodyftn3. Consulté le 08 juin 2011.

l'aptitude scientifiques des femmes ayant pour effet le développement et l'amélioration des productions scientifiques.

Sexe	Nombre	Pourcentage
Masculin	150	48,9%
Féminin	157	51,1%
TOTAL.	307	100%

Tableau 1 : Nombre des enquêtés par sexe

De plus nous remarquons une tendance de féminisation des chercheurs tunisiens vu que les tranches d'âges des enquêtés allant de moins de 25 ans jusqu'au 39 sont plutôt féminisées :

Sexe/Age	Moins de 25 ans	De 25 à 29 ans	De 30 à 39 ans	De 40 à 49	De 50 ans et plus	TOTAL
Masculin	13	46	39	29	23	150
Féminin	30	57	42	13	15	157
TOTAL	43	103	81	42	38	307

Tableau 2 : tranches d'âge des chercheurs par sexe

Ainsi que, le nombre des étudiantes-chercheures dépasse celui des étudiants-chercheurs dans l'échantillon des chercheurs tunisiens faisant l'objet de notre étude :

Sexe/statut	Etudiant chercheur	Enseignant chercheur	TOTAL
Masculin	79	71	150
Féminin	94	63	157
TOTAL	173	134	307

Tableau 3 : Statuts des chercheurs par sexe

Suite à la présentation du genre faisant l'objet de notre population enquêtée, nous essayons d'évaluer leurs pratiques de lecture dans le contexte numérique en ligne.

2. Lecture numérique vs lecture papier : quelles préférences pour les femmes et les hommes chercheurs ?

Plusieurs interrogations se posent de plus en plus sur ce qu'est lire. Ces interrogations sont liées essentiellement à la consultation du web sur des

écrans qui a engendré une « étendue conceptuelle » de l'acte de lire. Il peut désigner des activités très diverses. « Pour certains, la lecture est une activité en déclin appelée à être remplacée par la communication orale et visuelle ; multimédia et virtuelle ; pour d'autres c'est un élargissement de la représentation de ce qu'est lire qui se met en place avec le texte à l'écran. »[Rosado, 2011]. Ce qui est important à signaler qu'avec la lecture numérique, nous ne parlons pas seulement des textes imprimés transposés sur support numérique, mais aussi et surtout d'autres genres de documents multimédias, vidéos, blogs, sites web, messages électroniques, etc.

Dans ce contexte, les historiens signalent deux changements majeurs dans le passage du papier à l'écran, le premier concerne l'association écriture-lecture qui n'était pas « possible » et le deuxième est lié à la distinction entre le lieu du document et le lieu du lecteur. En d'autres termes, la lecture numérique permet au lecteur de faire des annotations et des modifications sur l'objet de lecture et peut également accéder à cet objet à distance [Cavallo, 2001].

Toutes ces interrogations dans la littérature se rapportant à la lecture en ligne, sa manière, les capacités de la pratiquer, son essence par rapport à la lecture sur papier attestent selon [Valéry, 2011] « du malaise et du rejet qui surviennent lorsqu'il s'agit de parler de la lecture dans un monde numérique ».

Ces différents changements dus au passage de l'imprimé à l'écran, nous incitent à savoir jusqu'à quel point ceci a influé sur les pratiques de lecture chez les chercheurs tunisiens au niveau du temps consacré à la lecture et de leur vision au document numérique par rapport au document papier.

En réponse à la question ayant pour objectif de savoir si les chercheurs consacrent plus de temps à la lecture sur papier qu'à la lecture sur écran, la majorité a répondu « oui » à raison de 66,1% contre 32,9% du nombre total d'enquêtés. En effet, les chercheurs femmes et hommes se rejoignent pour confirmer leur rattachement au papier.

Sexe/ Lecure papier Vs lecture sur écran	Non réponse	Oui	Non	Total
Masculin	2	100	48	150
Féminin	1	103	53	157
TOTAL	3	203	101	307

Tableau 4 : Temps consacré à la lecture papier par rapport à la lecture sur écran

Ces résultats montrent que les repères de lecture chez les chercheurs tunisiens sont toujours liés au papier. La majorité des enquêtés considère que la lecture sur écran, même si elle est plus rapide et permet de créer une interactivité avec l'écrit, elle reste une activité fatigante. Ce point de

vue est approuvé aussi bien par les hommes que par les femmes ayant tendance plus à pratiquer une lecture numérique superficielle en survolant des parties des textes. [cf. tableau 5]. Ceci peut être expliqué par le fait que la lecture numérique « demande au lecteur la maîtrise de nouvelles capacités. Les technologies de l'information en elles-mêmes modifient les pratiques de lecture, car le message écrit à l'écran s'accompagne le plus souvent d'un message audio–visuel » [Mihaela-Luminita, 2005]

Sexe/ LecNum	Non répo nse	Plus attrayante	Plus fatigante	Plus rapide	Plus inter-active	Plus superficielle	plus ennu-yeuse	Tota l
Masculin	0	24	80	65	40	33	11	253
Féminin	2	17	97	84	33	33	15	281
TOTAL	2	41	177	149	73	66	26	534

Tableau 5 : lecture numérique aux yeux des chercheurs tunisiens

La lecture sur écran est considérée aussi plus épuisante que la lecture papier parce qu'elle est perturbée par des éléments externes au texte lu (boutons, informations sur la structuration du corpus). Le lecteur n'a plus de vision globale du contenu, mais seulement des portions de texte dans lesquelles il effectue une sélection. Cette sélection demande déjà un effort de déplacement dans le texte autre que les instructions traditionnelles de la lecture. Il est appelé à construire son propre texte pour pouvoir le lire.

Toutes ces entraves peuvent limiter le temps consacré par le chercheur tunisien pour la lecture sur écran. Ainsi, en réponse à la question se rapportant au temps consacré à la consultation des ressources numériques, nous pouvons dire que l'activité de la lecture numérique est plutôt masculine [cf. tableau 6].

Sexe/temps de consultation des ressources numériques	Non réponse	Moins de 2 heures	Entre 2 et 7 heures	Entre 7 et 15 heures	Plus de 15 heures	Total
Masculin	1	32	55	35	27	150
Féminin	2	35	65	32	23	157
Total	3	67	120	67	50	307

Tableau 6 : temps de consultation des ressources numériques selon le genre

D'après notre enquête, les grands lecteurs (07 heures et plus par semaine) sont de majorité hommes. Les femmes lisent entre deux et sept heures par semaine, et ceci peut être expliqué par leurs occupations et leurs charges matrimoniales. Ces résultats sont confirmés par une étude

réalisée au cours du dernier trimestre 2010 par la société Gartner et menée auprès de 1569 personnes, à travers 6 pays : les Etats-Unis, l'Angleterre, la Chine, le japon, l'Italie et l'Inde [Gartner, 2010]. Il ressort de cette étude que les hommes lisent plus facilement sur écran que les femmes.

3. Formats de documents et lecture numérique

L'histoire de lecture nous montre que l'étude des pratiques fait partie de l'éphémère [Belisle, 2001]. Ces pratiques sont généralement liées aux changements technologiques, aux supports plus particulièrement, mais aussi aux formats dans lesquels atteint le document son lecteur. Les formats les plus utilisés pour les ressources numériques sont HTML, DOC, PDF, etc. Nous avons voulu à travers notre enquête de savoir lesquels de ces formats sont les plus utilisés par les chercheurs tunisiens. Les réponses montrent que les formats PDF et word sont les plus utilisés aussi bien par les femmes chercheurs que par les hommes chercheurs [cf. tableau 7].

Sexe/Format DocNum	Non réponse	HTML	PDF	DOC	TXT	RTF	PS	TeX	Autres	Total
Masculin	0	76	122	101	44	17	13	12	11	396
Féminin	1	72	128	114	43	4	11	6	8	387
TOTAL	1	148	250	215	87	21	24	18	19	783

Tableau 7 : formats de documents numériques utilisés par les chercheurs tunisiens

En cherchant à comprendre pourquoi PDF et DOC sont les deux formats qui sont placés avant le HTML, malgré que ce dernier représente le format et le langage du web par excellence, nous avons compris que ces deux formats comportent, selon les chercheurs, des repères semblables à ceux du document papier. En plus ils offrent une meilleure présentation de documents.
Par ailleurs, si nous pouvons expliquer l'usage du PDF en tant que format adapté à la lecture sur écran par ses caractéristiques intrinsèques reproduisant en fac-similé les pages du document source, y compris ses illustrations, fixant la pagination et offrant une fonction d'agrandissement pour modifier la taille globale d'une page, nous n'avons pas les mêmes caractéristiques pour le format Doc sauf que ce dernier offre des nouvelles fonctionnalités de lecture/écriture numérique. D'autres formats s'imposent aujourd'hui comme des standards de description et de structuration des documents tels que le XML ne sont pas connus par la majorité des chercheurs tunisiens.

4. Types de documents numériques consultés par le chercheur et la chercheure[40] tunisiens

Lire pour chercher de l'information nécessite la consultation de différents types de documents. Les dictionnaires et les encyclopédies peuvent être consultés pour se faire une idée sur un sujet. Les monographies et les articles de périodiques permettent d'approfondir cette idée et aussi de l'analyser. De plus différents autres types de documents peuvent être utilisés selon les besoins des chercheurs, tels que les thèses et mémoires, les rapports, les cartographies, les manuscrits, les documents pédagogiques, documentaires ou scientifiques et autres.
Les femmes, considérées auparavant fortes lectrices du papier comme le rappelle Olivier Donnat[41] : « quel que soit l'âge ou le niveau de diplôme, les femmes se distinguent par un niveau de lecture légèrement supérieur à celui des hommes et par une préférence pour la fiction », se trouvent dans le nouveau contexte numérique en concurrence avec les hommes dans la pratique d'accès et de lecture sur le web.
En effet, les chercheurs hommes et femmes ont tendance à préférer les articles des périodiques au détriment de la lecture des monographies, tel qu'il est illustré dans le tableau ci-après [cf. tableau 8] : signalant 121 hommes et 127 femmes affirmant lire des articles de périodique. Cependant, si les hommes devancent les femmes dans la lecture de la presse et des magazines en ligne (57 hommes contre 37 femmes) pour des objectifs culturels, sociaux et politiques; les femmes semblent plus patientes dans la lecture sur écran des travaux volumineux tels que les thèses et mémoires (109 femmes contre 85 hommes) pour des objectifs plutôt scientifiques et pédagogiques. Pour les autres types de documents, une certaine parité entre femmes et hommes peut-être remarquée dans leurs lectures des œuvres de fictions, des livres à caractères documentaires, ou scientifiques, des rapports, des cartes et plans, et des manuscrits.

Sexe/Types Document numérique	Masculin	Féminin	Total
Non réponse	0	2	2
Articles de revues	121	127	248
Presse (journaux, magazines)	57	37	94

[40] Michèle Lenoble-Pinson. -In **:** Revue belge de philologie et d'histoire = Belgisch tijdschrift voor filologie en geschiedenis, ISSN 0035-0818, Vol. 84, N°. 3, 2006 , page 637-652.
[41] Donnat, Olivier. – Les français face à la culture. De l'exclusion à l'éclectisme, Paris, la Découverte, 1994. Cité par (Mauger, 95).

Livres à caractère documentaire ou scientifique	65	65	130
Œuvres de fiction (romans, nouvelles,...)	5	5	10
Thèses et mémoires	85	109	194
Rapports	60	60	120
Cartes et plans	19	20	39
Manuscrits	18	17	35
Dictionnaires, encyclopédies	62	76	138
Documents pédagogiques (cours en ligne, exercices	74	88	162
Autres (précisez svp)	4	4	8
Total	570	610	1180

Tableau 8 : Types de documents numériques préférés par les chercheurs tunisiens

Nous expliquons cette tendance vers la lecture des périodiques plutôt que les monographies par la fluidité de la lecture des presses sur écran et par la lourdeur de la lecture des livres entiers ; ce qui incite les chercheurs à survoler plutôt qu'à lire ou même à imprimer pour une lecture papier ou encore à enregistrer pour une lecture différée.

5. Pratiques de lecture dans l'environnement numérique

Selon les résultats de notre enquête, il est confirmé que nos enquêtés enseignants chercheurs pratiquent le plus souvent une lecture fragmentée dans l'environnement électronique ; ils peuvent suivre plusieurs parcours outre que le fil conducteur de l'auteur. Il s'agit plutôt d'une lecture hypertextuelle et interactive permettant de réécrire pour soi le texte.
En parité, femmes et hommes chercheurs considèrent que les liens hypertextuels constituent plutôt un « enrichissement pour la compréhension des textes ». Une minorité seulement des chercheurs considère que les liens peuvent dévier leur lecture sur écran, tel qu'il est illustré dans le tableau suivant :

Sexe / Liens hypertextuels	Masculin	Féminin	Total
Non réponse	2	9	11
Enrichissent votre compréhension du contenu	102	103	205
Vous font perdre le fil des idées	28	30	58

Vous dévient de vos centres d'intérêt initiaux	22	15	37
Autres (précisez svp)	4	7	11
TOTAL	158	164	322

Tableau 9 : Chercheurs hommes et femmes activent les liens hypertextuels

Ce résultat confirme que la linéarité de la lecture disparaît sur écran au profit de la navigation et de l'interaction, surtout que les documents les plus consultés par les chercheurs sont de type scientifique adapté plutôt à la lecture sélective. Nos chercheurs en tant qu'usagers du Web se trouvent bien à l'aise avec l'accès à l'information et l'activation des liens. Néanmoins, dans cet environnement numérique, les chercheurs tunisiens associent rarement les pratiques de l'écriture à celles de la lecture sur écran, telles que celles de prise de notes sur écran, de marquage, de soulignement ou de surlignement.

Prise de notes sur écran		
Sexe/Notes écran	Masculin	Féminin
Non réponse	33	40
Très souvent	10	11
Souvent	18	24
Parfois	48	34
Jamais	41	48
Total	150	157

Marquage (souligner, surligner, colorer)		
Sexe/Marquage	Masculin	Féminin
Non réponse	36	42
Très souvent	12	16
Souvent	16	19
Parfois	45	37
Jamais	41	43
Total	150	157

Tableaux 10-11: Prises de notes et marquage sur écran

Avec un certain degré d'égalité, aussi bien les femmes que les hommes pratiquent rarement les prises de note sur écran ainsi que le marquage des documents, ce qui les amène encore à avoir recours aux pratiques d'impression et d'enregistrement des documents. En effet, une certaine égalité paraît encore entre femmes et hommes chercheurs dans ces pratiques, tel qu'il est illustré dans les deux tableaux suivants :

Impression des documents		
Sexe/Impression	Masculin	Féminin
Non réponse	21	31
Très souvent	31	32
Souvent	38	34
Parfois	53	50
Jamais	7	10
Total	150	157

Enregistrement des documents		
Sexe/Enregistrement	Masculin	Féminin
Non réponse	18	16
Très souvent	62	85
Souvent	41	40
Parfois	26	14
Jamais	3	2
Total	150	157

Tableaux 12-13 : Impression et enregistrement des documents

Un fort attachement empêche le chercheur tunisien de quitter le papier, ayant constitué le lieu de permanence de ses repères de lecture, ce qui l'incite à imprimer et à participer à la constitution des tonnes de nouveaux documents papiers.
L'enregistrement des documents permet aussi aux chercheurs la constitution des bibliothèques personnelles, mais cette pratique les incite à lire peu sur écran et surtout à accumuler des documents enregistrés non lus.
D'autres pratiques d'échange et de travail collaboratif peuvent être instaurées dans le contexte numérique ; ainsi nous essayons dans la section suivante d'évaluer l'apport du chercheur tunisien (femme ou homme) dans le but de favoriser le développement de la culture participative caractérisant la deuxième phase du web, appelé dores et déjà Web 2.0.

6. Culture participative du chercheur tunisien

Tout chercheur est appelé à passer du stade de la simple consultation des ressources numériques au stade d'une véritable « culture participative[42] » où il peut être actif au niveau de l'échange, de la mise en ligne, de la coproduction et de la création des documents numériques.
Cette culture s'instaure de plus en plus suite au développement des différents services du web2.0 tels que les blogs, les wikis, l'étiquetage, les réseaux sociaux, les fils RSS, les mashups (notamment la géolocalisation)[43].
Selon certaines études, les fonctionnalités du web 2.0 attirent plus les femmes que les hommes. En effet, les femmes sont plus actives sur le réseau en partageant des photos, des vidéos, des jeux et en utilisant la messagerie instantanée [Abraham, 2010]. Elles passent plus de temps sur les réseaux sociaux que les hommes.
Dans notre enquête nous n'avons pas pu vérifier ce bouleversement parce que les chercheurs tunisiens étaient peu impliqués dans les services web 2.0. Les nouveaux services que nous avons évoqués sont essentiellement les blogs et les archives ouvertes. Notre étude montre que les blogs sont plus utilisés par les hommes, et les archives ouvertes sont utilisés à égalité entre hommes et femmes pour la diffusion de leurs travaux scientifiques ou pédagogiques :

---

[42] - Le concept de « culture participative » est proche de la notion de « démocratie participative », où les individus deviennent actifs. (Rieder, 2006) cité par (Bsir, 2010)
[43] - Détails : in : Tout sur le web 2.0 / Capucine Cousin. – Paris : Dunod. 2008. – (CommentCaMarche.net : l'encylopédie pratique de l'informatique). - ISBN 978-2-10-051177-8.

Sexe/ diffusion des travaux scientifiques sur le web	Masculin	Féminin
Non réponse	46	69
Site personnel	25	17
Site de l'institution universitaire	24	23
Blog personnel	24	11
Plate-forme d'enseignement à distance	5	7
Archives ouvertes	16	17
Autres	23	27
Total	163	171

Tableau14 : Diffusion des travaux scientifiques sur le web

Par contre pour l'échange et l'accès aux annotations des autres, ils sont en faveur des femmes [cf. tableau 15].

Sexe	**Non réponse**	**Accès aux annotations des autres**	**Accès aux marquages des autres**	**Diffusion de vos annotations et/ou marquages personnels**	**Autres**	**Total**
Masculin	32	71	30	27	7	167
Féminin	31	**79**	18	**32**	11	171
Total	63	150	48	59	18	338

Tableau 15 : impact du genre sur l'échange et la collaboration entre les chercheurs

Le travail collaboratif de nos chercheurs tunisiens s'est manifesté plus au niveau d'échange des documents et de leur communication via les différents services classiques d'Internet en l'occurrence la messagerie et les listes de diffusion :

Sexe/ communication	Masculin	Féminin
Non réponse	06	11
En signalant l'adresse URL	68	60
En envoyant tout le document par courrier électronique	79	83
En le signalant à travers des listes de diffusion ou des groupes de discussion	14	16
Autres	18	17
Total	185	187

Tableau 16 : Echange et communication des documents

Les hommes paraissent plus habiles à signaler les adresses URL des documents à échanger via la messagerie, par contre les femmes paraissent plus rassurées en les faisant joindre en fichier attaché. Malgré ces initiatives de mise en ligne des travaux scientifiques et pédagogiques, ainsi que leur échange via l'Internet, le chercheur tunisien n'est pas vraiment passé à la phase participative sur le web ; ceci est approuvé par sa réticence à la production collaborative. En effet, hommes et femmes se déclarent plutôt passifs dans cette action aussi importante pour être à la phase 2.0 du web, tel qu'il est illustré par le tableau suivant :

Sexe/ProdTrav	Non réponse	Oui	**Non**	Total
Masculin	0	40	110	150
Féminin	3	34	120	157
TOTAL	3	74	230	307

Tableau 17 : Participation à la production des travaux collectifs en ligne

Cette déclaration aussi frappante doit être encore vérifiée suite à l'évolution technoculturelle que connait l'environnement des chercheurs tunisiens et au développement des différents services web2.0 demandant moins de technicité d'usage, citons essentiellement les wikis et les blogs, ayant plus d'essor ces dernières années.

7. Développement d'une culture technique

La lecture numérique nécessite la maîtrise d'un espace d'action en plus de celui de l'information. En effet, les chercheurs sont en face d'un espace à la fois technique et informationnel demandant une certaine culture technique assurant le « savoir faire pour lire » et le « savoir lire pour faire ». [Ghitalla, 2003].

Ainsi, expliquons-nous que la variable genre n'a pas établi à ce stade une influence claire sur les résultats, parce que les chercheurs tunisiens hommes et femmes sont encore en phase de développement de cette culture technique. Cette entrave amène les 2/3 des enquêtés (en parité 100 hommes et 103 femmes) à consacrer encore plus de temps à la lecture papier qu'à la lecture numérique.

L'évolution des technologies de l'écrit peut à son tour drainer les chercheurs à lire sur écran en les remettant dans la situation et la scène conventionnelles de lecture. Ainsi le format de fichier pdf est apprécié par un fort taux (250 enquêtés, divisés en parité 122 masculins et 128 féminins) parce qu'il présente l'avantage de reprendre les repères du document papier.

La lecture sur le web modifie les tendances de choix des types de documents à lire : femmes et hommes lisent plus les articles de

périodiques quelque soit leur discipline scientifique. Mais, suite à la diversité des ressources numériques sur le web et à la multiplication des possibilités de téléchargement des monographies, des rapports, des vidéos et autres, nos chercheurs ont plus de possibilités d'enregistrement et d'impression pour la constitution des fonds et des bibliothèques personnels.
La lecture sur écran incite de plus en plus les chercheurs à échanger les documents pédagogiques et scientifiques essentiellement par le biais des outils classiques d'Internet en l'occurrence la messagerie, les listes de diffusion, ainsi que les archives ouvertes. Cependant, les chercheurs tunisiens sont encore moins actifs à participer à la production collective.
Les initiatives des chercheurs au niveau du travail collaboratif peuvent être soutenues s'ils maîtriseraient les nouveaux services du Web participatif en l'occurrence les blogs pour la coproduction, les wikis pour la gestion des contenus et le travail collaboratif, les réseaux sociaux pour l'échange des informations et des événements scientifiques, les fils RSS pour suivre les nouveautés des domaines de recherche, et autres.

Conclusion

La parité des résultats entre femmes et hommes chercheurs peut émerger de nouvelles tendances. Les femmes considérées auparavant « fortes lectrices du papier » se trouvent sur le web concurrencées par les hommes ayant plus de temps à lire, à naviguer et à dénicher dans les flux d'informations en ligne.
En contre partie, la monopolisation de la maîtrise des technologies et des techniques informatiques par les hommes depuis des années, est désormais conquise par les femmes attirées par les nouveaux services du web en l'occurrence les réseaux sociaux.
Ainsi, une deuxième enquête portant sur les usages des nouveaux services Web2.0 par les chercheurs tunisiens peut-elle nous confirmer ce résultat de tendance de masculinisation des pratiques de la lecture et nous permet-elle de révéler à quel point ces chercheurs maîtrisent-ils les outils du web participatif pour se collaborer et coproduire sur le web ?

Bibliographie

[Belisle, 2011] Lire dans un monde numérique / Claire BELISLE (coord.). - Lyon : Presses de l'ENSSIB, 2011.
[Ben Romdhane, 2008] « Nouveaux modes de lecture-écriture et travail collaboratif des enseignants-chercheurs tunisiens dans l'environnement numérique. »/ Ben ROMDHANE M., MKADMI A., HACHICHA S. – In : actes de la conférence « Document numérique et société », Novembre 2008,

CNAM, Paris, ADBS éditions, 2008, ISBN : 978-2-84365-116-8, ISSN : 1762-8288.

[Bsir, 2010] De la lecture en ligne via le web documentaire à l'émergence de la « culture participative » via le web2.0 / Besma BSIR. - In : actes du 2ème colloque France-Maghreb "Les médias et les mémoires de demain", Toulon, – 9-10 décembre 2010.

[Bsir, 2010] Passage au numérique: nouvelles pratiques de lecture/ Besma BSIR. - In : Journées d'études organisées par le groupe de recherche Lecture numérique, Institut supérieur de documentation, 14-15 avril 2009. Article publié dans un ouvrage collectif : Lecture numérique et usage du web, sous la coordination de Raja FENNICHE, édité par l'Institut supérieur de documentation, 2010.

[Cavallo, 2001] Histoire de la lecture dans le monde occidental/ Guglielmo CAVALLO, Roger CHARTIER (dir.). Paris : éditions du seuil, 2001, 2ème édition.

[Collet, 2006] L'informatique a-t-elle un sexe ?: hackers, mythes et réalité/ Isabelle COLLET, Paris; Budapest; Torino : l'Harmattan, 2006, 312 p., (Savoir et formation. Série Genre et éducation), ISBN : 2-296-01480-1.

[Donnat, 2009] Les pratiques culturelles des Français à l'ère numérique, enquête 2008/ Olivier DONNAT. - La Découverte, Ministère de la Culture et de la Communication, 2009.

[ENDA, 2005] Fracture numérique de genre en Afrique francophone : une inquiétante réalité/ Enda tiers-monde, Dakar, 2005. ISBN : 92 9130 055 8.

[Fee, 1983] « Women's nature and scientific objectivity »/ FEE E. – In : Hubbard R. and Lowe M. (eds), Women's Nature : Rationalisations of Inequality, New York, Pergamon Press, 1983.

[Ghitalla, 2003] L'outre – lecture : Manipuler, (s')approprier, interpréter le Web/ Franck Ghitalla, Dominique BOULLIER, Pergia GKOUSKOU-GIANNAKOU, Laurence LE DOUARIN, Aurélie NEAU. – Paris : Bibliothèque publique d'information / Centre Pompidou, 2003. – (Etudes et recherche). – ISBN 2-84246-081-2.

[Hachicha, 2010] La lecture-écriture numérique à l'ère du Web 2.0 : le cas des chercheurs tunisiens/ S. HCHICHA,A. MKADMi, M. BEN ROMDHANE. - In : Journées d'études organisées par le groupe de recherche Lecture numérique, Institut supérieur de documentation, 14-15 avril 2009. Article publié dans un ouvrage collectif : Lecture numérique et usage du web, sous la coordination de Raja Fenniche, édité par l'Institut supérieur de documentation, 2010.

[Keller, 2003] Le/La scientifique : sexe et genre dans la pratique scientifique/ Keller, Evelyn Fox. - In : Les cahiers du CEDREF (centre d'enseignement, d'études et de recherche pour les études féministes), 11/2003. En ligne : http://cedref.revues.org/509 consulté le 26 octobre 2010

[Limam, 2008] Les pratiques de la lecture numérique : cas des enseignants chercheurs tunisiens/ LIMAM, L., BSIR, B., HACHICHA, S., BEN ROMDHANE, M. MKADMI, A.- In : Interagir et transmettre, informer et communiquer : quelles valeurs, quelle valorisation ? Actes du colloque international des sciences de l'information et de la communication, ISD, IPSI, SFSIC, 17-19 avril 2008, pp. 343-356.

[Mihaela-Luminita, 2005] Du papier à l'écran : quelles compétences documentaires développer chez l'élève ? Les mutations de la lecture, mémoire I.U.F.M de l'Académie de Montpellier/ Anin–Maffre Mihaela-Luminita, dirigé

par Anne–Marie Filleron, 2005. En ligne : http://www.crdp-montpellier.fr/ressources/memoires/memoires/2005/b/0/05b0032/05b0032.pdf

[Mkadmi, 2010] Lecture numérique : impact du genre et de la discipline scientifique sur l'usage du web 2 .0/ A.Mkadmi, M. Ben Romdhane, S.Hachicha.- In : actes du 2ème colloque France-Maghreb "Les médias et les mémoires de demain", Toulon, – 9-10 décembre 2010.

[Rosado, 2011] Qu'est ce que lire ? – in : Lire dans un monde numérique / Claire Belisle (coord.). - Lyon : Presses de l'ENSSIB, 2011, pp. 67-110.

[Turkle, 1988] «Computational reticence. Why women fear the intimate machine ? »/ TURKLE S. : In : Kramarae C (ed), Technology and women's voces. Keeping in touch, London, 1988.

[Zuckerman, 1991]. - « The Wo/man Scientist : Issues of Sex and Gender in the Pursuit of Science », in Harriet ZUCKERMAN, Jonathan R. COLE and John T. BRUER (eds), the Other Circle :Woman in the Scientific Community, New York and London, W.W. Norton, 1991, p. 227-236.

Présentation de l'information comme support d'aide à des processus cognitifs

Mustapha MOJAHID
Nesrine NOUGHI
Philippe BOISSIERE

Institut de Recherche en Informatique de Toulouse - Equipe Elipse
(Etude de L'Interaction Personne SystèmE)
Université Paul Sabatier, Toulouse Cedex

Résumé : La recherche présentée dans cet article s'inscrit à la fois dans le cadre du traitement automatique du langage et plus précisément l'architecture textuelle, de l'interaction Homme/Machine et de la psycholinguistique. Nous contribuons à ces problématiques en étudiant l'apport de l'architecture textuelle et de la présentation de l'information dans des processus cognitifs et dans dans l'amélioration de l'accessibilité à des personnes en situation de handicap.
Nous proposons une approche d'analyse basée sur les modèles en cherchant à conserver les bénéfices de chacun. Nous proposons également une stratégie de transformation de la représentation obtenue en une représentation en termes de langage d'images de pages (IdPs). Enfin, nous développons un outil expérimental qui permettra à l'utilisateur l'interaction avec les différents niveaux d'IdPs et d'améliorer ainsi ses performances en terme d'accessibilité et de recherche d'information.

Mots-clés : Traitement automatique du langage, modèles de représentation de texte, interaction avec le texte, accessibilité.

Introduction

Plusieurs axes en TALN visent à traiter les documents textuels écrits (Pery-Woodley et Condamines 2007 ; Jackiewicz 2005 ; Laignelet, 2003) et plus précisément les structures visuelles (Virbel et al. 2005 ; Luc et al. 2001 ; Jacques 2010, Ho-Dac 2010). L'approche soutenue par ces derniers suggère que ces niveaux de structure sont fortement indissociables du contenu et participent au sens du texte. Il apparaît que des objets textuels de types particuliers (tels que les titres, les énumérations ou les définitions) ont des comportements spécifiques qui

appellent la mise au point de représentations locales propres : la notion classique de classes de documents recoupe ainsi celle de classes d'objets, ces classes de documents étant régies par leur logique (syntaxique, rhétorique et visuelle) propre (Luc 2001).
Par ailleurs, l'étude des objets textuels ne peut être menée à bien que si nous considérons la composante humaine. Beaucoup de travaux ont montré (Veyrac 1998 ; Virbel et Nespoulous 2004 ; Eyrolle et al. 2008 ; Léger et al. 2005 ; Etcheverry 2009 ; Alamargot et Chanquoy 2002) l'intérêt d'étudier l'impact des différents niveaux de structures de texte sur les processus cognitifs (mémorisation, recherche d'informations, compréhension, réalisation de tâches).
Veyrac (1998) a montré dans sa thèse que la non prise des structures de textes (logique et visuelle) dans la rédaction et la présentation des consignes pour les conducteurs peut poser d'énormes problèmes dans la phase de l'exécution de ces consignes.
Trois contributions sont proposées dans cet article. Dans la première section, nous proposons une méthode pour améliorer les processus d'analyse manuelle (convergence d'annotation) en combinant plusieurs méthodes d'analyses dans le domaine du traitement automatique des langues : modèle d'architecture textuelles (MAT) ; théorie des structures rhétoriques (RST) et la questionnabilité du texte (Q-R), développés initialement par, respectivement Pascual (1998), Mann et Thomson (1988) et Chali (1997). Dans la deuxième section, nous présentons une stratégie pour élaborer le contenu des images de pages qui constitueront l'interface d'interaction des utilisateurs avec le contenu du document.
Notre hypothèse générale est de montrer l'apport de notre approche de structuration du contenu de l'information, à l'aide des IdPs en rendant plus perceptible, pendant le processus de lecture et de recherche, certaines informations et relations du discours et à l'inverse en cachant d'autres. Le recours aux Images de pages favorisera le focus d'attention et améliorera la mémorisation et l'accès à l'information. Pour évaluer cette interaction avec des textes structurés en Images de Pages, nous avons développé une plateforme expérimentale, présentée à la dernière section. Nous présentons à la fin le protocole expérimental élaboré pour valider nos hypothèses. Une première expérience préliminaire nous a permis de valider notre approche (voir section 3).

1. Méthode d'analyse

Le principe général consiste à combiner les modèles (MAT/RST/Q-R) pour bénéficier des avantages de chacun. La figure 1 résume notre approche que nous explicitons dans cette section en l'illustrant sur un exemple de texte tiré du corpus d'étude (Sarda 2010).

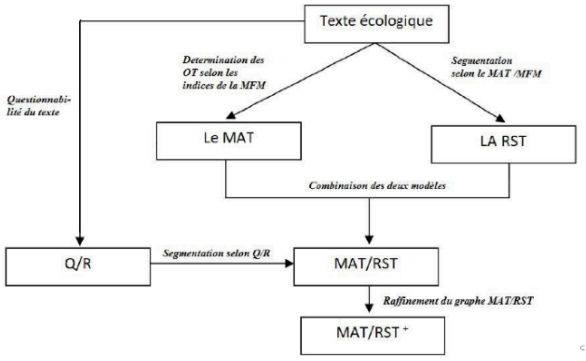


Figure 1 - Approche d'analyse

1.1. Etude du corpus

Le corpus est constitué d'un recueil de textes procéduraux composé de 31 recettes de cuisines (Sarda, 2010). Nous illustrons notre méthode d'analyse à travers l'exemple de la Tortilla (figure 2). Toutes les recettes sont rédigées en français par un seul rédacteur[44]. La rédaction s'est faite après la réalisation de ces recettes par des personnes âgées assistées par des aides à domicile. Les textes tiennent sur une seule page et comportent diverses marques lexicales et de mise en forme matérielle. Trois types de structures imbriquées existent dans ce corpus : les titres, les rubriques et les énumérations.

Tortilla

Ingrédients pour 4 personnes
½ kg de pommes de terre
4 oeufs
Huile d'olive, sel

Préparation (30 mn)
Laver et couper en lamelles les pommes de terre. Les jeter dans une poêle où l'on aura fait chauffer de l'huile (3 bonnes cuillères à soupe). Les faire dorer et les réserver.

Battre les oeufs avec le sel et y plonger les pommes de terre. Les pommes de terres doivent s'imprégner des oeufs. Jeter le tout dans la poêle en retirant au préalable, l'excès d'huile. Laisser dorer et la retourner.

Ne pas trop cuire la tortilla, elle ne doit ni être sèche ni baveuse et les pommes de terre doivent rester tendres.

Figure 2 - Une copie de la recette « Tortilla »

[44] Nous remercions l'auteur du recueil des recettes de cuisine, Noëlle Sarda.

Un problème important qui se pose à l'analyse du corpus pour la segmentation et l'étiquetage est la subjectivité. L'analyse d'un même texte par plusieurs analystes est susceptible de produire autant d'interprétations.
Notre objectif est d'apporter des éléments de réponse pour définir une méthode qui permet d'améliorer les processus d'annotation et d'analyse manuelle et proposer ainsi des pistes pour des analyses automatiques (ou semi-automatique) de texte en prenant en compte à la fois des indices visuels et discursifs. Les travaux actuels ne traitent essentiellement que les marqueurs discursifs (Annodis, 2009).

L'analyse est réalisée en deux étapes et le résultat est représenté sous forme de deux graphes MAT/RST. Les deux graphes se distinguent par le degré de granularité de leurs segments minimaux et par les nouvelles relations fournies par la phase de questionnabilité du texte.

Etape 1
La segmentation dans cette étape se base principalement sur le modèle d'architecture textuelle (MAT) et les indices de la MFM : lexico-syntaxiques, typographiques, dispositionnels, diacritiques ou une combinaison. La décomposition contient 13 segments où chacun est constitué d'une phrase élémentaire au sens de (Harris, 1937), d'une phrase composée ou d'une phrase complexe. Les parenthèses sont des indices dans la segmentation, puisqu'il s'agit d'un procédé de réalisation qui peut trouver son équivalent langagier pour traduire une fonction métatextuelle et architecturante du texte.
Les indices dispositionnels ont permis de délimiter les objets textuels (étapes, blocs, items, actions).
Pour illustrer on présente un extrait de la segmentation :
ut5 = Préparation
ut6 = (30 min)
ut7 = Laver et couper en lamelles les pommes de terre.
les jeter dans une poêle où l'on aura fait chauffer d'huile
ut8 = (3 bonnes cuillères à soupe).
ut9 = Les faire dorer et les réserver.
ut10 = Battre les oeufs avec le sel et y plonger les pommes de terre.

Nous obtenons pour l'exemple de la Tortilla, le graphe architectural et son équivalent en termes de la RST en ajoutant les relations rhétoriques suivantes :

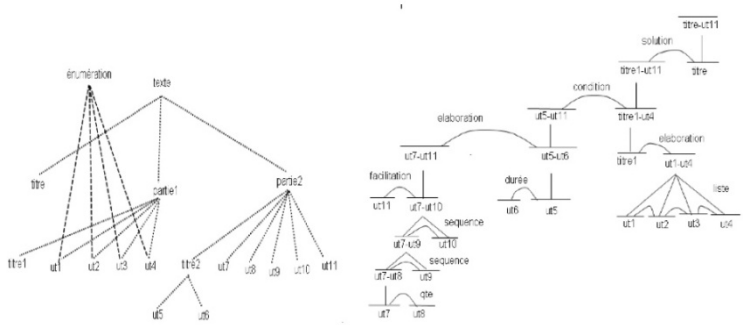


Figure 3 - graphe MAT et graphe RST de la Tortilla à l'issue de l'étape 1

Nous rappelons que le graphe MAT est une transcription schématique du métadiscrous suivant :
Le texte est composé d'un titre et de deux parties contrastées par le gras, la première identifiée par « Ingrédients pour 4 personnes » et la seconde identifée par « Préparation (30 mn) ». La première partie est composée d'une énumération qui est la liste des ingrédients et la seconde est composée de trois blocs (paragraphes) qui ont été délimités grâce aux espaces verticaux, où chaque paragraphe est composé d'une séquence d'actions (qui peuvent être dépendantes ou indépendantes entre elles). Les deux premiers paragraphes représentent à leur tour une liste d'actions constituant une séquence.
Deux intérêts majeurs pour coupler le modèle MAT et la théorie de la RST sont d'une part prendre en compte des relations rhétoriques (RST) et d'autre part considérer les structures visuelles du texte et exprimer des compositions entre segments non-adjacents (MAT). La figure 3 représente la structure résultante de la combinaison de l'arbre RST et le graphe MAT de la recette Tortilla.

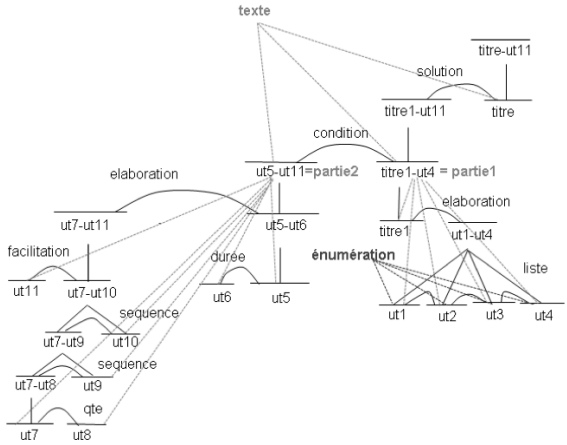


Figure 4 – Représentation MAT/RST de la Tortilla de l'étape 1

Etape 2

Dans la deuxième étape, la segmentation se base sur le système Question-réponses.

Selon l'approche de Harris un texte peut être décomposé en un ensemble de phrases élémentaires et en un ensemble d'opérations linguistiques associées (Harris, 1971 ; Daladier, 1990).

L'hypothèse que Y. Chali a proposée en se basant sur les travaux de Harris est :

si nous admettons qu'un texte répond ou contribue à répondre à une question qu'il soulève, explicitement ou implicitement, nous pouvons admettre plus analytiquement qu'une partie au moins du sens d'une unité de ce texte sera fonction des questions dont le texte peut être le support, et de la réponse qu'elle donne à d'autres questions supportées part d'autres unités.

Le choix de considérer comme unités de base du texte des phrases élémentaires constituant les phrases textuelles, permet de constituer un ensemble fini de questions telles que les réponses à ces questions figurent dans le texte.

Ceci va nous permettre pour un ensemble donné de type logico-linguistiques de questions, de définir une structure questions/réponses sur l'ensemble des phrases élémentaires d'un texte. Un mécanisme de sélection permet, pour une phrase élémentaire support d'une question, de hiérarchiser l'ensemble des phrases élémentaires qui constituent une réponse à cette question.

L'idée de la méthodologie de Y. Chali est d'identifier une question par l'ensemble de ses réponses.

Une distinction peut être introduite entre un groupe de questions dites internes et un groupe de questions externes.

Les questions internes
Une question interne est une question dont la réponse est comprise dans la phrase élémentaire qui la supporte. Ce type de question découle syntaxiquement de la structure du schéma de phrase associé à la phrase élémentaire.

Exemple : [Nom0] verbe nom1. Nous pouvons associer la question que verbe nom1 ?

Couper les pommes de terre → que doit-on couper ?

Les questions externes
Contrairement aux questions internes, une question externe ne contient pas de réponse dans son support : elle fait intervenir une autre phrase (ou plusieurs) que la phrase élémentaire qui la supporte.
C'est pourquoi l'ensemble des phrases élémentaires produites de la décomposition du texte est muni, en plus des caractérisations lexico-syntaxiques issues de l'opération de décomposition, de relations interrogatives.
Ces relations appliquées aux phrases élémentaires du texte engendrent une structure de questions/réponses. Il s'agit d'associer à toute phrase élémentaire support d'une question, l'ensemble des phrases élémentaires réponses à cette question.
Sept types de relations interrogatives (Tableaux 1 et 2) ont été ainsi distinguées dans le contexte du corpus des recettes de cuisine pour répondre aux questions (5 externes : quelle, combien, comment, où, avec quoi et 2 internes : de quoi, que). Ces questions déterminent les phrases élémentaires qu'elles peuvent supporter, et y associer les phrases élémentaires qui constituent les réponses à ces questions.

Nous avons donc appliqué ce principe à l'exemple de la Tortilla :

pe support questions	pe support de réponses				
	quelle	combien	comment	où	avec quoi
titre1(1)		titre1(2)			
ut1(1)		ut1(2)			
ut2(1)		ut2(2)			
ut5	ut6				
ut7(2)			ut7(3)		

pe support	pe support de rép.	
question	de quoi	que
ut1(1)	ut1(1)	
ut2(1)	ut2(1)	
ut3	ut3	
ut4	ut4	
ut7(1)		ut7(1)

ut7(4)				ut7(5)	
ut7(6)		ut8			
ut10(1)					ut10(2)
ut10(3)				ut10(4)	
ut10(6)				ut10(7)	
ut11(1)			ut11(2)		
ut11(1)			ut11(3)		
ut10(9)			ut11(4)		

ut7(2)		ut7(2)
ut7(4)		ut7(4)
ut7(6)		ut7(6)
ut8		ut8
ut9(1)		ut9(1)
ut9(2)		ut9(2)
ut10(1)		ut10(1)
ut10(3)		ut10(3)
ut10(5)		ut10(5)
ut10(6)		ut10(6)
ut10(8)		ut10(8)
ut10(9)		ut10(9)
ut10(10)		ut10(10)
ut11(1)		ut11(1)

Tableaux 1et 2 - Graphes des questions/réponses externes (à gauche) et internes (à droite)

Si nous prenons comme exemple la ligne 10 du graphe des questions/réponses internes, nous pourrons lire : Que doit-on faire dorer (ut9(1) = faire dorer les pommes de terre).
Le support de réponse de cette question est la même phrase élémentaire.
Cette deuxième étape d'analyse vient enrichir celle obtenue dans la première étape en affinant la granularité des segments (34 segments) et en définissant de nouvelles relations rhétoriques entre ces nouveaux segments.
Nous illustrons par un extrait cette nouvelle décomposition :
titre1(1) = Ingrédients
titre1(2) = les ingrédients sont pour 4 personnes
pour combien de personnes les ingrédients sont prévus ?
ut1(1) = on a besoin de pommes de terre
de quoi a-t-on besoin ?
ut1(2) = on a besoin de 1/2 kg
combien a-t-on besoin de pommes de terre ?
ut2(1) = on a besoin d'oeufs
de quoi a-t-on besoin ?

ut2(2) = on a besoin de 4 oeufs
combien a-t-on besoin d'oeufs ?
Nous obtenons ainsi une nouvelle représentation enrichie MAT/RST+ (fig. 5)

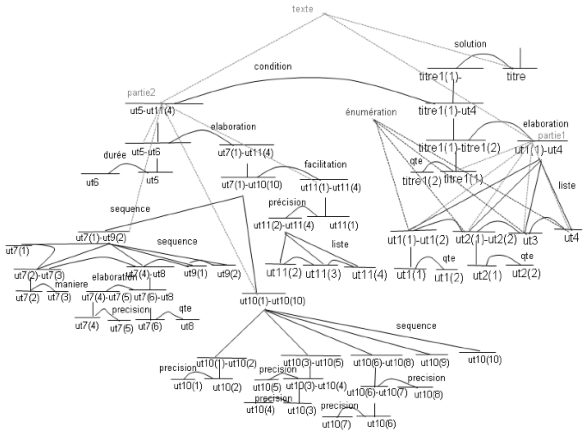


Figure 5 graphe finale MAT/RST+

2. Passage de la représentation MAT/RST+ aux Images de pages

L'objectif final de cette étude est de proposer une interface pour aider l'utilisateur[45] à améliorer ses performances de mémorisation et de recherche d'information. Nous proposons des stratégies de transformation de la représentation obtenue (MAT/RST+) vers une nouvelle présentation qui servira d'interface d'interaction avec l'utilisateur. Nous avons opté pour le langage des Images de Pages dont nous allons donner les principes généraux (Luc et al. 2001 ; Mojahid 2011).

2.1. Langage des images de page

Le langage des IdP est un langage notationnel permettant de représenter les phénomènes architecturaux de texte. Il sert de support pour la description et la visualisation de différents niveaux de structures de texte (visuelles, syntaxiques et rhétoriques).
Le principe du langage notationnel est sa construction à partir d'un ensemble d'images de pages hiérarchisées, chacune prenant en compte un certain niveau de granularité de l'ensemble (ou sous-ensemble) des

[45] Les perspectives est d'étudier l'impact des IdPs sur des sujets ayant des troubles du langage (aphasie), de la mémoire (amnésie) et déficiences mentaux légers.

propriétés d'un niveau de structuration (typo-dispositionnelle, syntaxiques et rhétoriques). Une image d'IdP peut-être vue ainsi comme un transparent que l'on peut superposer (/retirer) pour ajouter (/supprimer) des informations concernant les propriétés d'un texte.
En se référant à la typographie invisible définie par Twymann (1982), la première IdP peut montrer le « first glance » et rendre perceptible une vue globale d'un texte ou ce qui peut-être vu lorsqu'on se trouve éloigné à une certaine distance du texte ou encore les éléments pertinents à lire en premier. Certaines propriétés de nature discursive ou visuelle seront donc cachées et les IdP suivantes superposeront et exhiberont progressivement ces nouvelles propriétés.
Pour définir le contenu de chacun des niveaux des IdP, nous avons d'une part exploité les résultats de l'analyse du corpus et d'autre part, nous avons réalisé un pré-test pour cerner les difficultés majeures rencontrées par le sujet.

2.2. Pré-test

Nous avons choisi de faire un pré-test auprès d'un sujet ayant un handicap moteur. Dans la première tâche effectuée, le sujet devrait réaliser la recette. Le sujet ne pouvant pas utiliser ses mains, notre but était de favoriser la verbalisation de ses actions et de ses intentions. Il donnait ainsi des instructions à une tierce personne.
Dans la deuxième tâche, le sujet devait saisir (recopier) une deuxième recette. Il saisissait avec un clavier adapté à son handicap et recopiait à partir d'un écran.
Nous avons posé un questionnaire au sujet, après chaque tâche, sur les deux recettes pour tester ses capacités de mémorisation et de localisation des informations dans le texte.
Plusieurs problèmes rencontrés par le sujet ont été relevés :
- localisation des informations dans le texte de la recette : par exemple quand nous lui avons demandé de nous citer les actions que nous devons appliquer sur les pommes de terre, le sujet avait du mal à localiser l'ingrédient pommes de terre.
- le sujet avait du mal à reprendre (continuer) là où il s'est arrêté : lors de l'exercice de recopie le sujet a eu du mal à voir où il en était car l'image de son écran ne correspondait pas à l'image du texte qu'il avait à recopier.
- de nombreux retours en arrière vers les différentes parties du texte ont été observés pour l'ensemble des tâches effectuées.
Ces différentes difficultés ont provoqué beaucoup d'erreurs et ont fait accroître le temps de réalisation de la recette et le temps de sa saisie.

2.3. Stratégies de passage aux IdPs

L'analyse du corpus des 31 recettes de cuisine et les résultats du pré-test nous ont fourni une excellente base pour extraire ces stratégies. En

outre, étant donné le genre de textes concerné (texte à consignes de type recettes de cuisine) et les profils de sujets *handicapés* visés à long terme, nous avons considéré les deux éléments suivants : d'une part, l'interface avec les IdP doit tenir compte de l'évolution dans le temps du processus de réalisation de la recette. D'autre part, pour faciliter l'utilisation de cette interface, nous avons cherché à optimisation le nombre de niveaux.
Nous sommes parvenus à une généralisation sur l'ensemble des recettes en utilisant trois niveaux d'IdPs :
- le premier niveau d'Idp contient l'architecture globale de la recette ainsi que les « outils » (ingrédients et ustensiles) utilisés,
- le deuxième niveau d'IdP est constitué de plusieurs sous niveaux correspondant aux étapes de préparation de la recette structurés en actions,
- le troisième niveau d'IdP est facultatif correspondant à une suggestion d'accompagnement ou à une facilitation de préparation.
De plus nous avons enrichi tous les niveaux par deux types de relations qui peuvent relier des segments adjacents ou non-adjacents. La première relation fait apparaître les actions qui peuvent être réalisées en parallèle ; c'est le cas quand le sujet est au repos et peut lancer une nouvelle action. La deuxième relation sert à faire apparaître les actions qui ont un ordre inversé par rapport à l'ordre dans la formulation de la recette ; c'est l'exemple de *« Jeter le tout dans la poêle en retirant au préalable,* l'excès d'huile ». Cette relation a été détectée à travers la relation rhétorique condition et l'expression au préalable.

2.4. Langage notationnel des IdPs et interaction avec l'utilisateur

Le tableau suivant présente les différents éléments et notations utilisés pour définir les niveaux des IdP et leur interaction avec l'utilisateur.

Notation	Sémantique
Le flou	– cacher des segments
	– garder la structure globale de la recette
Les couleurs	– colorier les ingrédients
	– localiser les ingrédients
Flèche à un seul sens	– localiser les actions avec un ordre inverse
Flèche à double sens	– localiser les actions parallèles

Tableau 2 – Éléments du langage notationnel des IdP

A fin de minimiser le nombre de couleurs utilisés (certaines recettes contiennent plus de 10 ingrédients !). Nous avons choisi de regrouper les ingrédients selon leurs catégories (ingrédients de base, condiments, épices...) et d'attribuer à chaque catégorie une couleur en cumulant avec

un indice de MFM (soulignement, italique, gras). La figure 5 montre le cas où l'utilisateur a coché sur les deux étapes 1 et 2.

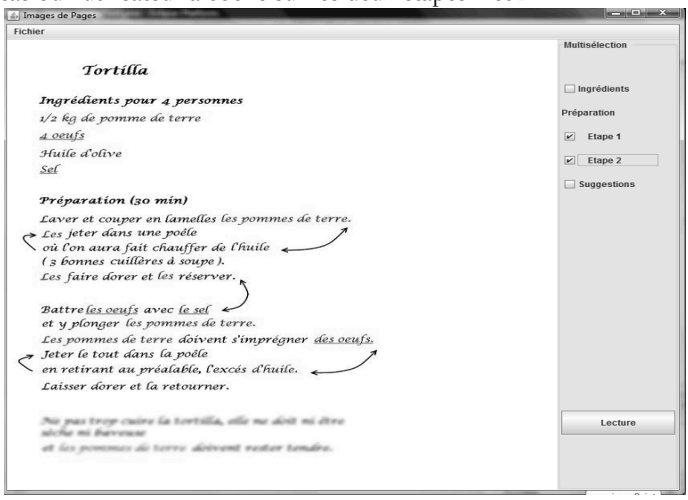


Figure 6. – Superposition des niveaux 2-1 et 2-2 de la Tortilla.

3. Outil d'expérimentation et premières évaluations

Nous avons opté pour le langage Java pour développer l'interface de notre outil à cause de son approche orientée objet et de sa portabilité. Java permet également de développer des applications autonomes et de créer facilement des interfaces graphiques grâce à la plateforme « Java Virtual Machine ». Pour notre application, nous avons utilisé l'environnement de développement d'Eclipse IDE qui est un environnement intégré, libre et extensible.

3.1. Interface

L'interface (figure 5) est constituée principalement de :
- un menu *Multi-sélection* composé des niveaux d'IdP qui sont présentées par des *cases à cocher*, où le sujet peut choisir de visualiser chaque niveau d'IdP indépendamment des autres ou bien sélectionner la combinaison de plusieurs niveaux afin de les superposer,
- un bouton *Lecture* permet d'afficher le texte de la recette en entier pour une première lecture de découverte du contenu,
- un espace de lecture, où les IdP seront affichées,
- un menu *Fichier* est un menu déroulant qui comporte l'ensemble des recettes de notre corpus.

Nous présentons dans cette dernière sous-section le protocole que nous avons fait valider par une équipe d'ergonomes[46] et les résultats préliminaires d'une première expérience.

3.2. Protocole et résultats préliminaires

Le protocole défini une évaluation comparative des processus de mémorisation et de recherche d'informations dans deux recettes de cuisine prises dans le corpus (Sarda 2010) et présentée selon deux formats de présentation différente.

Deux situations sont prévues, une première où la recette est présentée sur un écran dans son format initial (écologique) telle qu'elle a été produite par le rédacteur. Dans la deuxième situation, la recette est reformatée (selon le langage des IdP défini dans la section 2) et proposée au sujet en deux phases. D'abord le sujet pourra la découvrir dans une première lecture et ensuite, il aura à sa disposition l'interface supplémentaire à l'aide des niveaux d'IDP (figure 5). Le sujet dans cette deuxième phase a la possibilité de lire le contenu de la recette de manière sélective en choisissant le (ou les) niveau(x) d'IdP voulu(s).

Le but du protocole est de montrer l'apport de notre approche d'interaction de structuration du contenu de l'information à l'aide des IdP, en rendant plus perceptible certaines informations et relations et à l'inverse en cachant d'autres. Le recours aux IdPs favorisera le focus d'attention et améliorera la mémorisation et la recherche d'information.

Plusieurs hypothèses ont été formulées :

- Les sujets auront plus de facilité à mémoriser et à localiser les actions, les ingrédients et les ustensiles à l'aide de l'interface des IdP.
- Les sujets auront plus de facilité à mémoriser et à localiser l'ordre temporel des actions.
- L'interface sera encore plus adaptée à des sujets ayant des troubles du langage (aphasie), de la mémoire (amnésie) ou pour des déficients mentaux légers.

Les variables dépendantes pour la tâche de mémorisation sont le nombre d'informations rappelées, l'enchaînement temporel des actions respectées, les actions parallèles, l'association des actions aux étapes de la recette, la précision des mots rappelés, les temps de réponse.

Les variables dépendantes pour la tâche de recherche d'informations sont le temps de réponse, la localisation de l'information et le nombre de relectures.

[46] Nous tenons à remercier Isabelle Etcheverry, Julie Lemarié et Patrice Terrier (psychologues ergonomes) pour leurs critiques et leurs remarques pertinentes sur l'élaboration du protocole.

Conclusion et perspectives

L'étude présentée dans cet article se trouve à l'insertion de trois domaines qui sont le traitement automatique du langage naturel et plus spécifiquement le texte écrit, l'interaction Homme/Machine et le handicap.
La contribution se situe à quatre niveaux :
- La définition d'une approche d'analyse de corpus en couplant trois modèles et théories : le Modèle d'Architecture Textuelle, les théories des structures rhétoriques et de la questionnabilité.
- L'élaboration d'une stratégie pour transformer la représentation obtenue (MAT/RST+) en une représentation dans les termes du langage des Images de Pages.
- La conception d'une interface qui propose l'interaction à l'aide de ces niveaux des Images de pages.
- La spécification d'un protocole pour valider les hypothèses qui stipulent l'apport bénéfique des IdP et de son interaction dans l'amélioration de la mémorisation et dans la recherche d'informations.

De nombreuses perspectives sont possibles pour ce travail. Nous avons retenu celles-ci :
Une première perspective consiste à réaliser deux expérimentations, la première auprès de sujets valides et la deuxième auprès de sujets non valides avec des troubles langagiers et/ou de mémoire.
Valider notre méthode d'analyse sur d'autres corpus, éventuellement de genre différents (narratifs, descriptifs...), ainsi que sur des domaines spécifiques, par exemple en mathématiques où l'architecture des documents (exemples, démonstrations, définitions, formules, dessins...) est extrêmement riche d'indices de mise en forme matérielle et de discours.
Tester d'autre tâches comme la réalisation de la recette ou de la saisie de texte (Maurel et Antoine). Dans le cadre de cette dernière tâche, un travail a été initialisé pour étudier l'intérêt des IdPs dans une tâche de recopie de texte. L'hypothèse générale considère que les IdPs améliorent les performances du sujet en diminuant le temps de saisie et les erreurs. Cette collaboration se situe dans le cadre du projet Palliacom (2008-2011).

Bibliographie

ALAMARGOT, D. & CHANQUOY, L. (2002) Les modèles de rédaction de textes. In. M. Fayol (Ed.). Production du langage : Traité des Sciences Cognitives. Hermes.

ANNODIS (2009) Annotation discursive : corpus de référence pour le français et outils d'aide à l'annotation et à l'exploitation. Projet ANR (CLLE-ERSS-Toulouse ; GREYC-Caen ; IRIT-Toulouse).
CHALI Y. (1997) L'expansion de texte Une approche basée sur l'explication par questions/réponses pour la génération de versions de textes. Thèse de l'Université Paul Sabatier.
DALADIER A. (1990) Aspects constructifs des grammaires de Zellig Harris. PhD thesis, 1990.
ETCHEVERRY, I. (2009). Les exigences cognitives de la recherche d'informations sur Internet et les difficultés liées à l'âge examinées sous l'angle de la recollection. Thèse de l'Université Toulouse.
ETCHEVERRY, I., Baccino, T., & Mojahid, M. (2009). Eye movements and recollective experience in web search tasks: Is what you see what you get? Proceedings of the 5th European Conference on Eye Movements (ECEM'09). University of Southampton, 23-27 août.
EYROLLE H. CELLIER J.-M. LEMARIÉ, J. (2008) The segmented presentation of visually structured texts: effects on text comprehension. Computers in Human Behavior, 24, 888-902.
Harris Z. Structures méthématiques du langage. PhD thesis, 1971.
Ho-Dac L. M., Fabre C., Péry-Woodley M. P., Rebeyrolle J. (2010) On the signalling of macrodiscourse structures, 8th Multidiscplinary Approaches of Discourse - MAD2010, Moissac.
JACQUES M.-P (2010) Étudier des structures de discours : préoccupations pratiques et méthodologiques Cognition, Représenttion, Langage, 2, Volume 8.
JACQUES M.-P, MOJAHID M., Sarda L. (2001) Repérer les structures du texte, élement pour la construction d'un module d'analyse. CIDE 2001, Toulouse.
JACKIEWICZ A. (2005) Les séries linéaires dans le discours. Langue française, 148, 95–110. 2005.
LAIGNELET M. (2003) Les cadres de discours spatiaux et temporels dans les documents
géographiques : interactions et croisements. Mémoire de maîtrise de Sciences du Langage, Université de Toulouse – Le Mirail.
LEGER L., TIJUS C, BACCINO T. (2005) La discrimination visuelle et sémantique : Pour la conception ergonomique du contenu de sites web. Revue d'Interaction Homme-Machine, 6, 81-106.
LUC C. (2000) Représentation et composition des structures visuelles et rhétoriques du texte, Approche pour la génération de textes formatés. Thèse de l'Université Paul Sabatier, Toulouse.
LUC C. (2001) Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. Dans : TALN2001, Université de Tours.
MANN, W.C., & THOMPSON, S.A. (1988) Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8 (3). 243-281.
MAUREL D., ANTOINE J.-Y. (2007) La communication assistée, TAL 48-3.
MOJAHID, M. (2011) Fondements du langage des Images de Page et sa mise à l'épreuve. 2nd International Conference on Linguistic & Psycholinguistic Approaches to Text Structuring, Louvain, Belgique.
PALLIACOM (2008-2011) Communicateur multimodal pour la palliation et l'augmentation alphabétique et logographique de la communication écrite et orale, Projet ANR, telecom Bretagne, IRIT-Toulouse. http://recherche.telecom-bretagne.eu/palliacom/.

PASCUAL E. (1991) Rprésentation de l'architecture textuelle et génération de texte. Thèse de l'Université Paul Sabatier. Toulouse.
PÉRY-WOODLEY M.-P., CONDAMINES, A. (2007) Linguistic markers of lexical and textual relations in technical documents. in d. alamargot, p. terrier , j-m. cellier (eds.), written documents in the workplace. studies in writing. amsterdam : Elsevier, pp. 3-16.
Sarda N. (2010) Petits secrets de famille. ADPAM, Toulouse.
TWYMANN M. (1982) The Graphic Presentation of Language, Information Design Journal 3(1), 2-22.
VEYRAC MERAD-BOUDIA H. (1998) Approche ergonomique des représentations de la tâche pour l'analyse d'utilisations de consignes dans des situations de travail à risques, Thèse de l'Université du Mirail, Toulouse.
VIRBEL J. & NESPOULOUS J.-L. (2004) Handicap langagier et recherches cognitives : apports mutuels. (ii). parole.
VIRBEL J. (1998) The Contribution of linguistic knowledge to the interpretation of text Structure. In André J., Quint V. Furutra R. eds.
VIRBEL J., GARCIA-DEBANC C., BACCINO T., CARRIO L., DOMINGUEZ C., JACQUEMIN Ch. Luc Ch., MOJAHID M., PERY-WOODLEY M.-P, SCHMIDS S. (2005) Approches cognitives de la spatialisation du langage. De la modélisation de structures spatiolinguistiques des textes à l'expérimentation psycholinguistique : le cas d'un objet textuel, l'énumération. Agir dans l'espace. Catherine Thinus-Blanc, Jean Bullier (Eds.), Editions de la Maison des sciences de l'homme, p. 233-254, Cognitique.

Tendances lourdes et tensions pour les filières du document numérique

Ghislaine CHARTRON
DICEN, CNAM

François MOREAU
LIRSA, CNAM

Résumé : Sur la base des travaux réalisés dans le cadre de l'ANR Digital 3.0 PRISE[47], cette communication rend compte du bilan de l'atelier « Culture, médias et numérique », sur les tendances lourdes et les tensions majeures qui transforment actuellement les principales filières du document numérique, document étant entendu ici de façon extensive, à savoir la presse, les livres, l'audiovisuel, la musique (produits des industries culturelles). Les tendances et tensions identifiées concernent l'organisation économique, les régulations, les usages et les pratiques dominantes.

Mots-clés : Analyse socio-économique, industries du contenu, media, mutation numérique, modèle économique, prospective.

Abstract : Based on work done during the project ANR Digital 3.0 PRISE, this paper reports the results of the workshop "Culture, Media and Digital," concerning the trends and major tensions that are transforming the main sectors of the digital document, document understanding here extensively as the press, books, audiovisual, music (products of cultural industries). Trends and tensions identified concern the economic organization, regulations, customs and main practices.

Keywords : Socio-economic analysis, content industries, media, digital transformation, economic model, prospective.

47 http://digital3prise.net/ , projet ANR, thème "Quelles innovations, quelles ruptures dans la société et l'économie numériques ?". Atelier « Culture, media, numérique » co-piloté par G. CHARTRON et F. MOREAU
http://digital3prise.net/pg/groups/59051/

Introduction : méthodes et objectifs de ce travail

Cette proposition s'inscrit dans la continuité de nombreuses analyses et études prospectives déjà réalisées mais généralement limitées à un secteur particulier (Leiba, 2011)(1), (Boureau & all, 2007)(2), (Chartron, 2010)(3), (Bajon, 2008)(4), (Udecam, 2010)(5). Il s'agissait au contraire dans ce projet de confronter l'analyse d'acteurs impliqués dans différents secteurs (édition, musique, presse, audiovisuel) pour déterminer les caractéristiques communes partagées des transformations numériques. Ce travail privilégie les dimensions socio-économiques (modèles économiques, acteurs, usages) apportant ainsi une complémentarité d'analyse par rapport à des approches plus orientées vers la conception ou l'ingénierie.
La méthode utilisée est triple : à la fois l'observation participante des auteurs de cet article aux mutations de ces filières depuis plusieurs années (G. Chartron pour l'édition et en particulier l'édition scientifique, F. Moreau pour la musique), mais aussi, à l'occasion de cet atelier « Culture, médias et numérique », l'interview de 10 experts[48] de filières différentes au cours de deux journées d'étude, et le travail collaboratif des participants aux trois journées de l'atelier (chercheurs, praticiens, étudiants en thèse).
Ces différentes auditions et les travaux collectifs ont permis de mettre en évidence une série de tendances lourdes, de tensions et d'inconnues. L'objectif poursuivi est ici de rendre compte ce qui semble dominant et partagé pour le domaine de la culture et des médias, sachant par contre que des spécificités sectorielles restent plus finement à discerner.

1. Tendances lourdes partagées

1.1. Brouillage des objets, éclatement des frontières

Le numérique ne se réduit pas à la reproduction des objets précédents dans la nouvelle donne des réseaux. La délinéarisation marque l'ensemble des contenus (vente et « consommation » au chapitre, à l'article, au titre musical et à la vidéo sélectionnée via un moteur). Les frontières que l'on connaissait entre media textuels et audiovisuels sont de plus en plus poreuses, laissant place à de nouveaux objets comme le « livre augmenté », à savoir un objet à l'origine textuel et linéaire « augmenté » de contenus multimédia et interactifs, de potentialités calculatoires

[48] Denis ZWIRN (Numilog, Livre), Laurent SORBIER (MySkreen, Audiovisuel), Simon BALDEYROU et Marianne LE VAVASSEUR (Deezer, Musique), Olivier DONNAT et Alain GIFFARD (Ministère de la Culture et de la Communication) et Nicolas CURIEN (ARCEP, régulation), Françoise BENHAMOU (Paris 13), Nathalie SONNAC (Paris 2), Alain BUSSON (HEC), PJ BENGHOZI (Ecole Polytechnique).

diverses et de personnalisations variées (Pédauque, 2006) (6), les web documentaires mixant le genre documentaire avec l'écriture multimédia non linéaire associant texte, photos, vidéos, sons et animations interactives (Broudoux, 2011) (7).

Le développement du concept« transmédia » est significatif de ce débordement des frontières, il désigne un type de narration qui articule un univers narratif sur différents médias, le lecteur peut découvrir une histoire dans la presse papier, avoir des prolongements sur un site Internet, rester en contact sur les services mobiles, suivre des épisodes télévisuels. Les points d'entrée sont multiples et peuvent être compris indépendamment. Ce genre de narration induit de nouveaux contrats d'auteurs permettant des exploitations multiples et dérivées des œuvres (« franchises transmedia »). L'aspect calculatoire (algorithmique) des contenus numériques marque de façon transversale cette nouvelle donne, l'objet fini laisse place à des objets potentiellement toujours ouverts dont les contours sont décidés en grande partie par l'interactivité avec l'utilisateur. Le moteur de recherche associé à des métadonnées est le dispositif central de cette déconstruction/reconstruction des contenus.

Enfin, ce brouillage des objets pose aujourd'hui de réels problèmes de redéfinition des régulations en place, l'appareil législatif qui avait été pensé devient inadapté et pourrait freiner les initiatives de nombreux acteurs. La loi récente sur le livre numérique est emblématique de cette confusion : si le bilan de cette loi est plutôt positif dans le contexte papier ayant permis de protéger en partie le réseau des petits distributeurs sur le territoire ; sa transposition numérique, si elle vise aussi à contrer les offensives des géants du web sur le secteur du livre numérique, oublie toutefois que le livre numérique se dissout dans des formats plus proches de la base de données aujourd'hui et que les initiatives sont multiples pour offrir d'autres modalités de vente inédites que pourraient figer une loi trop interventionniste.

1.2. Les valeurs sociétales en évolution

Tous les secteurs de contenus ont construit leurs repères par des marques dominantes qui ont structuré la diffusion, que ce soit des éditeurs (Gallimard, Hachette, Elsevier, Armand Colin…), des groupes de presse (Le Monde, Ouest-France, …), des maisons de disques (Universal Music, EMI ..), des chaînes de télévision (TF1, Antenne2, CBS,…) et ont construit le paysage médiatique des « consommateurs ». Or, l'émergence de nouveaux acteurs, le développement des contenus générés par les utilisateurs, le poids de nouveaux distributeurs brouillent le paysage et redéfinissent progressivement de nouveaux repères : on parle de revues Cairn, de musique Deezer ou I-tunes, de livres Numilog, … alors qu'il ne s'agit que de distributeurs opérant pour des producteurs de contenus mais qui tendent à s'effacer au profit de la marque de distribution, les conflits potentiels sont possibles.

La redistribution de la valeur s'opère aussi au sein de consommations collectives, le groupe rassemble des individus qui partagent certains goûts et dont on suivra les recommandations. De la même façon, des leaders prescriront des lectures, des écoutes comme le font notamment des Disc Jockey pour la musique auprès des jeunes. Ces potentialités d'auto-publication portées par le web participent activement à l'évolution de la hiérarchie des valeurs sociétales en matière de références médiatiques.
Par ailleurs, la connaissance et la culture que diffusent en continu les réseaux sont fortement associées à la gratuité, à la non-exclusivité, à la notion de « biens communs ». La propension à payer a diminué fortement dans tous les secteurs des contenus.

1.3. La valeur économique bousculée

Cette faible propension à payer les contenus se traduit au final par une destruction de la valeur du contenu et la remise en question des droits de la propriété intellectuelle au profit d'abonnements à des réseaux, d'achats, de matériels, d'interfaces de plus en plus sophistiquées. La migration de la valeur s'opère en grande partie sur l'accès (Rifkin, 2000) (8), notamment les point centralisés dominés par les géants du web (Google, Apple, Amazon) même si, à tous les niveaux, des moteurs de recherche spécialisés tentent de gagner la fidélité des usagers[49]. Les procédures de référencement et d'agrégation sont devenues centrales dans tous les secteurs, renforçant constamment ces points d'entrées dont les recettes s'appuient massivement sur la publicité en ligne. La migration de la valeur s'opère aussi sur les appareils de lecture, c'est la stratégie d'Apple et d'Amazon qui verrouillent l'accès aux contenus par le support et qui, particulièrement pour Apple, peaufinent le design et l'ergonomie, misant sur les relations affectives des usagers à leurs appareils de lecture. La marchandisation des données personnelles et des relations sociales est également un déplacement de valeur convoité par les agents économiques tout en provoquant de fortes contestations, de fortes résistances pour des raisons éthiques auprès des internautes.

1.4. Servicialisation et renouvellement des modèles économiques

Les contrats de droits d'usages se substituent à l'acquisition de biens de façon très significative. Les licences d'usage individuel ou collectif se sont installées Un abonnement forfaitaire donne droit à une consultation illimitée de revues scientifiques par exemple (Springer, Elsevier), de musiques (Spotify, Deezer), d'ouvrages (Cyberlibris), l'usager ou un tiers (la bibliothèque) souscrit un abonnement annuel, mensuel, voire à la

[49] Comme par exemple l'initiative du moteur français Myskreen pour la video en ligne de toute nature, films et programmes télévisés, vidéo à la demande (VOD) et télévision de rattrapage, www.myskreen.com

journée… De façon complémentaire et à la fois opposée dans la logique de vente, les achats à l'unité sont le deuxième pilier de cette nouvelle donne. Licence globale et offre hyper segmentée sont aujourd'hui présentes selon des poids différents dans les secteurs.
La possession des biens cède du terrain à la consommation ponctuelle, répondant à un besoin exprimé, laissant de vrais dilemmes à la constitution des mémoires, aux lieux de capitalisation de savoirs que sont les bibliothèques personnelles ou partagées. Mais le numérique séduit par son agilité potentielle pour les consommateurs : accessibilité omniprésente, externalisation de la gestion et maintenance matérielle. La technologie Cloud Computing (« Informatique dans les nuages ») gagne du terrain pour la distribution des contenus[50].
Face à la gratuité dominante, les modèles économiques essaient de composer avec cette donne : coexistence de version gratuite et de version premium payante pour une qualité de service supérieure (Deezer Premium, Revues.org Freemium, Orange Premium…), développement de marchés bifaces (existence de deux clientèles : annonceurs et consommateurs) adossés principalement aux revenus publicitaires mais ce n'est pas tout à fait nouveau pour les medias… Mais ces modèles restent incertains face à la réelle propension des consommateurs à payer pour un service de plus grande qualité dans le premier cas et les risques de dépendance des produits et services soumis aux marchés bifaces dans le second.

1.5. Renouvellement de la fonction d'intermédiation

Le web a permis des relations directes entre producteurs et utilisateurs, sous des formes inédites de publication et de diffusion selon les secteurs. Mais ceci reste souvent marginal face au besoin d'un marché de masse nécessitant des agents organisant les transactions face à une demande de plus en plus exigeante en qualité de service. La fonction intermédiaire, en particulier pour la distribution et la diffusion, n'a pas disparu mais convoque de nouvelles compétences, notamment technologiques dont se saisissent de nouveaux entrants, concurrençant et mettant en difficulté les anciens intermédiaires tels que les libraires, les grandes surfaces culturelles, les salles de cinéma, les bibliothèques…
Les géants du web, par leurs compétences technologiques concentrent de plus en plus cette fonction de distribution pour des raisons techniques (complexité de la gestion de flux de données massifs) et pour des raisons commerciales (attrait pour des catalogues élargis de contenus). La distribution est soumise à de forts mouvements de concentration, ce n'est pas nouveau mais le processus est amplifié dans le contexte numérique.

[50] Voir notamment le service Google Livres.

1.6. La nécessaire évolution de la régulation

Les medias sont régis par un ensemble de régulations sectorielles avec d'éventuelles instances désignées comme par exemple le CSA (Conseil supérieur de l'audiovisuel), le CNC (Centre national du cinéma et de l'image animée). Un des problèmes majeurs qui se posent aujourd'hui est de savoir comment continuer à réguler les médias sans réguler Internet ? Une partie de l'audiovisuel sur le web est aujourd'hui régulé : une web TV, par exemple, tombe sous le coup de la loi sur l'audiovisuel de 1986, YouTube est régulé dans sa partie professionnelle (hors contenus générés par les utilisateurs), si le fournisseur est établi en France… mais le problème est aujourd'hui que seulement 2% environ de l'audiovisuel est régulé contre 98% de contenus web audiovisuels non régulés… Les principes de pluralisme, de diversité culturelle, de protection du jeune public sont-ils encore contrôlables dans le contexte du web ? Plus largement, la question est de savoir si les anciens cadres sont extensibles au numérique ou s'il faut reconsidérer tout autrement la régulation.

Certains prêchent pour une intervention modérée de l'Etat et misent sur des processus d'autorégulation, de confiance entre les acteurs. Il s'agirait de passer d'une régulation centralisée et prescriptive à une co-régulation et à une auto-régulation décentralisée qui favoriserai la confiance entre acteurs (Curien, 2011) (9).

Une autre dimension importante de la régulation sur le plan économique concerne le financement de la création qui repose à l'heure actuelle sur le prélèvement de taxes dans les différentes filières via les activités des éditeurs, des bibliothèques, des chaînes de télévision… Si le transfert de valeur se fait sur le réseau, sur les outils d'accès, sur les appareils de lecture et que les acteurs technologiques ne sont pas soumis à la réglementation nationale, le problème va devenir crucial pour le financement de la création, à moins de penser que ces financements centralisés par l'Etat seront suppléés par d'autres sources potentielles, lesquelles de façon pérenne ?

1.7. Transformations des pratiques culturelles et médiatiques

La dernière édition de l'enquête majeure (5000 personnes sondées) menée tous les 8 ans par le Ministère de la Culture (Donnat, 2008) (10), et couvrant l'ensemble des usages des médias et des activités de temps libre au-delà des pratiques culturelles, montre un retournement de tendances sur la dernière décennie, notamment chez les moins de 35 ans. Pour la 1ère fois, on constate une diminution des consommations de TV et radio et une montée en puissance d'une « culture d'écran », de plus en plus de pratiques culturelles passent par des écrans dorénavant.

L'observation que l'on peut faire, mais qui mériterait des vérifications empiriques, concerne des pratiques de consommation de plus en plus fragmentées et dispersées, à l'opposé des logiques des anciens médias. Les lectures sont délinéarisées : après une recherche par moteur de

recherche, on lit un article, un chapitre d'ouvrage, on visualise une émission en télévision de rattrapage, on écoute un titre de musique et non plus un album. Les pratiques sont sélectives, fortement induites par les outils d'accès. Les consommations sont aussi collectives, guidées par les recommandations des groupes auxquels on appartient.

2. Tensions communes aux différents secteurs des contenus numériques

Après avoir mis en évidence précédemment certaines tendances lourdes, le travail conduit dans l'atelier « Culture, médias et numérique » de l'ANR Digital 3.0 PRISE nous permet de rendre compte de quelques tensions majeures associées dans l'ensemble des filières de production des contenus, que ce soit la presse, l'édition, le secteur musical ou cinématographique.

2.1. Dématérialisation vs. biens physiques

La tension est fondatrice et centrale à la mutation numérique, elle engendre de nombreux corollaires :

- La logique de location/accès vs. possession

Pour tous les produits culturels, la consommation sur support physique acheté par le consommateur a longtemps été simultanément le mode de consommation dominant et la source principale de revenus pour les producteurs/éditeurs. C'est particulièrement vrai pour le livre ou la musique. Et même pour les films cinématographiques, les recettes vidéo dépassent en France les recettes en salles (source CNC). Avec le numérique et le développement apparent du streaming au détriment du téléchargement, la notion de « possession » est remise en cause. Pour la musique par exemple, s'abonner à Spotify ou à Deezer conduit à être locataire de sa bibliothèque musicale et non plus propriétaire. Pour l'édition, l'abonnement à des revues en ligne conduit à vider les bibliothèques physiques sans garantir un accès pérenne à ces contenus. Le numérique, tout en répondant à des besoins immédiats, installe une insécurité sur les représentations et les mémoires.

- Les tensions liées à la captation de la valeur de plus en plus difficile sur les contenus

Un contenu numérique possède les deux caractéristiques d'un bien collectif pur : non-rivalité et non-excluabilité[51]. La possibilité de reproduire à coût quasiment nul un fichier numérique fait en effet disparaître la propriété de rivalité du support physique. Quant à

[51] La rivalité traduit le fait que la consommation d'un phonogramme par un individu empêche un autre de le consommer et l'excluabilité le fait que l'accès à ce bien est réservé aux seuls individus acceptant d'en acquitter le prix

l'excluabilité, si elle peut toujours être théoriquement assurée ex post par la stricte application des droits de propriété littéraire et artistique et/ou ex ante par des moyens techniques de protection, en pratique l'ampleur du trafic sur les réseaux peer-to-peer souligne que les contenus numériques revêtent potentiellement la propriété de non-excluabilité. C'est ainsi la structure traditionnelle de la commercialisation des produits de contenus qui est battue en brèche. Il est possible de jouer sur le déplacement de la valeur des contenus vers des consommations liées, soit des biens rivaux soit la méta-information. La stratégie de déplacement de la valeur vers des biens rivaux vise à lier un contenu libre à des biens rivaux utiles, voire nécessaires, pour une consommation pleinement satisfaisante. Les marchés de l'accès à Internet haut débit, des appareils de lecture (baladeurs numériques, téléphone portables, etc.) constituent ces principaux biens rivaux[52]. La complémentarité entre ces marchés peut être mise en œuvre par les acteurs économiques eux-mêmes ou par les pouvoirs publics. Des exemples de cette première solution sont fournis par Apple, dont le site de vente de musique en ligne (iTunes), n'a pas pour vocation d'être rentable mais de booster les ventes d'iPod et maintenant d'iPhone et autres appareils qui représentent la principale source de profits de la société californienne (11).

- Tensions liées aux « business models » et aux modes de financement des contenus

La vente de supports physiques s'est toujours faite à un prix largement supérieur au coût marginal (et donc une marge unitaire élevée) souvent justifié par une politique de mutualisation des risques. Les quelques succès finançant les nombreux échecs. Avec le passage à un système d'abonnement ou de financement publicitaire, le prix apparent par « clic », écoute ou visionnage devient infime. D'où une impression de « bradage » des produits culturels. Or ce sont les sommes globales collectées via les abonnements ou la publicité qu'il convient de mettre en regard des sommes qui étaient perçues des ventes de supports physiques. Par ailleurs, se surajoute à cette équation, comme nous l'avons dit précédemment, la faible propension de l'utilisateur à payer sur Internet, sauf pour l'accès au réseau et ses appareils de lecture. La donne est différente dans les secteurs où un tiers paie pour l'utilisateur et souscrit notamment des licences d'usage annuel. La bonne santé financière des gros éditeurs scientifiques tels qu'Elsevier dans le domaine académique s'explique en grande partie pour cette raison[53], mais les restrictions

[52] Mais on peut également penser aux produits dérivés ou encore au spectacle vivant pour l'industrie musicale. Ainsi, nombreuses sont les maisons de disques qui considèrent le marché des concerts comme un relais de croissance face à la récession qui frappe les ventes de disques.

[53] http://www.stockmarketwire.com/article/4191921/Reed-Elsevier-improves-as-margins-grow.html

budgétaires publiques augurent certainement des difficultés à venir pour les bibliothèques. L'incertitude sur la viabilité et la pérennité des modèles économiques conduit à s'interroger également sur les futurs modes de financement de la création et la rétribution des auteurs. La publicité sera-t-elle capable d'apporter ce financement si elle se substitue dans bien des cas à l'achat de l'usager, selon quelle logique de soutien, quelle prise en compte de la diversité culturelle ? Si de nouveaux encadrements des œuvres s'imposent, assouplis pour soutenir une plus grande accessibilité et une plus grande innovation sur les réseaux, comment combler les revenus qui étaient assurés par l'exploitation des droits attachés aux œuvres ? Quels mécanismes de substitution pourront être crédibles et pérennes ?

2.2. Tensions entre les acteurs économiques

- Tensions entre les firmes traditionnelles (producteurs, éditeurs) et les acteurs techno-logiques (moteurs de recherche, agrégateurs, FAI, opérateurs de téléphonie mobile…)

La question du déplacement de la valeur évoquée ci-dessus se traduit évidemment par des tensions entre les différents acteurs dont le modèle d'affaire est très différent. Les producteurs/éditeurs se rémunèrent sur la vente des supports, les acteurs technologiques sur la publicité, la vente d'accès ou de matériels. Or, dans la musique par exemple, les ventes de supports (CD et fichiers numériques) ont chuté de 50% en une dizaine d'années alors que dans le même temps le chiffre d'affaires des abonnements à Internet à haut débit ou des baladeurs musicaux explosait. Parmi les risques associés à cette fragilité, il faut noter la possible disparition d'acteurs traditionnels de taille modeste et une hyper-concentration, une uniformisation de l'offre commerciale pilotée par les grands groupes et les géants du Web.

- Montée en puissance des amateurs face aux créateurs historiques

A l'inverse, le développement d'Internet et des techniques numériques permettent à des amateurs de réaliser, avec du matériel peu onéreux, des œuvres d'une qualité convenable. Ils peuvent également offrir leurs productions directement aux consommateurs.

L'autoproduction, l'autopromotion et l'auto distribution, non seulement réduisent les barrières à l'entrée mais aussi, peut-être, correspondent à une évolution des pratiques artistiques : entre l'artiste de métier et le consommateur grand-public, des amateurs éclairés, qui consomment et qui produisent, pourraient former une nouvelle couche de public, correspondant à une création de valeur spécifique (Bacache & al., 2009) (12). Ces amateurs peuvent-ils pour autant devenir réellement des concurrents des artistes/auteurs établis ? Aucune réponse réellement claire pour l'instant.

- Lobbies industriels vs. neutralité du net

Aujourd'hui, sur Internet, si les consommateurs paient un abonnement pour avoir accès au réseau plus des frais spécifiques, ou supportent de la publicité, pour pouvoir accéder à certains contenus, les éditeurs de contenus paient uniquement leur accès au réseau. Aucun frais de « terminaison » ne leur est imposé pour accéder aux consommateurs. Le réseau Internet est considéré comme neutre (Wu, 2003) (13). Cette neutralité s'exprime par le fait que les échanges entre utilisateurs dans la « couche » usages ne doivent être ni empêchés ni restreints par les pratiques des opérateurs dans la « couche » réseau et que des requêtes soumises au réseau dans des conditions équivalentes doivent être traitées par celui-ci de manière équivalente. Cette règle implicite du non paiement de l'accès aux utilisateurs par les éditeurs de contenus présente plusieurs avantages qui ont contribué au succès d'Internet (Lee et Wu, 2009) (14). D'une part, elle a facilité l'entrée de nombreux fournisseurs de contenus[54] et permis l'émergence de services (les blogs, les réseaux sociaux, etc.) qui n'auraient peut être pas vu le jour s'ils avaient eu, au préalable, à faire la preuve de leur viabilité pour justifier le paiement de droit d'accès aux abonnés des opérateurs. D'autre part, cette règle évite la fragmentation du réseau qui via des accords d'exclusivité conduirait certains contenus à n'être disponibles que sur un nombre limité de réseaux. Certains consommateurs seraient privés de contenus et certains producteurs de contenus verraient leur marché potentiel limité. Aujourd'hui cette règle implicite, et donc la neutralité du net, est remise en cause par les opérateurs de réseau. Ces derniers estiment que les nouveaux usages fortement consommateurs de bande passante, comme le téléchargement de vidéos, saturent les infrastructures de cœur de réseau et les équipements d'interconnexion. Ils demandent à ce que les fournisseurs de contenus concernés contribuent au financement des réseaux (fibre optique, réseau mobile 4G, etc.). AT&T, un FAI américain, est ainsi l'un des premiers à avoir demandé à ce que Google paye un droit pour avoir accès à ses abonnés (10). Les opérateurs de réseaux souhaitent donc pouvoir discriminer en fonction des usages du réseau, c'est-à-dire offrir un traitement privilégié, ou au contraire dégradé, à certains flux de données, en se fondant sur leur origine, leur destination ou leur teneur. Cela pourrait alors conduire à un Internet fragmenté, en raison de contrats d'exclusivité qui seraient probablement passés entre certains fournisseurs d'accès et certains éditeurs, et même à des abus anticoncurrentiels conduisant, par exemple, un acteur verticalement intégré, à la fois présent sur le marché des contenus et sur celui de l'accès, à réserver sur son réseau un meilleur traitement à ses propres contenus qu'à ceux des éditeurs concurrents. De plus, une telle exigence pénaliserait aussi, et sans doute surtout, les petits éditeurs qui ne

[54] De ce point de vue, le maintien de cette règle implicite est sûrement une condition nécessaire à l'émergence d'une longue traîne dans les industries de contenu.

seraient pas en mesure de payer les opérateurs de réseau pour obtenir une qualité convenable du transport de leurs données[55].

2.3. Tensions sur les usages

- Homogénéisation de la consommation ou développement de cultures hyper-personnalisées

Le développement de la numérisation pour les produits de contenus induisant plus de variété tant au niveau de l'offre que de la demande devrait se traduire par une réduction de la concentration de la consommation sur quelques produits. L'offre en ligne de produits de contenu dématérialisés est plus complète que celle des sites de vente en ligne comme Fnac.com ou Amazon, elle-même plus complète que celles des magasins traditionnels. De plus, la promotion décentralisée devrait faciliter la transmission d'informations sur l'existence et les caractéristiques de produits de niches. Les ventes des produits stars faisant l'objet d'importantes dépenses de promotion devraient ainsi s'éroder au profit des œuvres de moindre notoriété – produits de niches – désormais plus facilement accessibles. Ce phénomène de déconcentration des ventes est qualifié d'effet de « longue traîne » par Anderson (15). La baisse des coûts de production et de stockage augmente le nombre de références offertes aux consommateurs et allonge ainsi la traîne (la queue de la distribution). Ensuite, les plus grandes possibilités offertes par la promotion décentralisée et la distribution en ligne pour connaître l'existence et accéder aux produits de niche accroissent l'épaisseur de la traîne au détriment des produits stars.

Mais les résultats empiriques sont pour l'instant mitigés. On trouve également des signes d'une concentration encore accrue de la consommation sur un nombre plus réduit de références. Le bouche-à-oreille électronique peut en effet être particulièrement efficace en accroissant l'audience des œuvres ayant déjà un minimum de succès. A l'inverse, les œuvres confidentielles peuvent ne jamais être promu en ligne car n'entrant jamais dans les meilleures ventes ni n'étant commenté dans un nombre suffisamment grand de forums ou de blogs.

- Effet générationnel ou effet âge

Dans l'analyse des comportements, les sociologues distinguent notamment deux effets distincts pouvant influer sur les valeurs d'une variable sociodémographique : l'effet génération et l'effet âge. L'effet de génération établit un lien de causalité entre la génération de la sous-population étudiée et la variable considérée, ainsi le fait d'appartenir à

[55] En revanche, les opérateurs de réseaux pourraient proposer aux utilisateurs ou aux éditeurs des « menus de qualités » où une plus haute qualité du transport (débit, vitesse) serait associée à un prix plus élevé. Une telle différenciation de l'offre ne serait pas contraire à la neutralité puisque tous les agents se verraient offrir le même menu et qu'une fois choisi le menu, le réseau ne ferait plus de différence entre le type de paquets transportés (courriel, vidéo, voix, etc.).

une certaine génération détermine en partie la valeur de la variable pour cette génération. De la même façon, l'effet âge établit un lien de causalité entre l'âge et la valeur de cette variable, autrement dit le fait d'avoir un certain âge détermine en partie la valeur de la variable telle que constatée pour cet âge. La question pour le numérique est de savoir si le développement des pratiques de consommation numérique vont changer ou non en vieillissant ? Est-ce que les jeunes qui ont 20 ans aujourd'hui achèteront encore un journal papier quand ils auront l'âge de leurs parents et accepteront-ils de regarder une programmation télévisuelle durant leur loisir alors qu'ils préfèrent jongler aujourd'hui sur les séries en ligne ou sur les vidéos amateurs disponibles sur YouTube? Il est encore trop tôt pour savoir car les études nécessitent de suivre les mêmes cohortes d'usagers sur des périodes de vie assez longues. Par contre, les spécialistes des comportements médias pensent que l'effet génération l'emportera, que les habitudes acquises dans le numérique seront pérennes et que le système d'influence est pour la première fois renversé : les plus jeunes apprenant aux plus anciens… La dernière livraison de l'enquête du Ministère de la Culture (Donnat, 2008) montre des effets générationnels importants : baisse de la lecture d'imprimés, augmentation des consommations audiovisuelles, augmentation des pratiques culturelles en amateur, montée en puissance de la culture de l'écran…

- Complémentarité ou concurrence entre les pratiques traditionnelles et les pratiques numériques

La réponse est complexe, elle apparait différente selon les contextes d'usage. Si l'on compare aujourd'hui l'appropriation des revues scientifiques numériques et celle des ouvrages numériques de littérature générale, la différence est grande entre une acculturation générale via des grandes plateformes de revues (Chartron, Epron, Mahé, 2011) (16) et des usages encore marginaux via des appareils de lecture qui ne rivalisent encore pas avec le papier pour le confort, la commodité, l'affinité avec le support dans un contexte de loisir (IPSOS, 2011) (17). De la même façon, comme vont évoluer les rapports entre la culture d'écran et la culture de sortie ? Quel devenir de la sociabilité ? Le lien observé est multiple : si la culture de l'écran croît, elle ne fait pas pour autant disparaître la culture de sortie, on continue à fréquenter les cinémas notamment. La culture d'écran peut devenir une nouvelle culture de sortie : depuis fin 90 on sort de chez soi pour aller voir des écrans (matchs de foot dans les cafés), concerts sur grand écrans… Le rapport au temps reste aussi central, et il est à parier que les activités de lecture, d'écoute, de visionnement qui nécessitent un temps long et linéaire sont en baisse au profit d'activités plus fractionnées sauf peut-être dans un contexte de loisir, en rupture avec une activité quotidienne.

3. Conclusion : quel futur ?

De ces tensions évoquées précédemment, peuvent naître des ruptures majeures…Le futur pourrait être moins numérique que prévu si la fracture perdure, si les rejets liés à la santé, à la protection de la vie privée, à l'addiction du réseau sont massifs. Quelles seraient les conséquences pour le secteur culture-médias ? Un simple retour à la case départ ou la fragilisation de la rentabilisation des lourds investissements consentis ces dernières années ? La chaîne de valeur des industries culture-médias peut se réorganiser totalement et remettre en cause les équilibres passés et les jeux d'acteurs dominants, avec la montée en puissance des nouveaux acteurs technologiques, l'éventuelle disparition du droit d'auteur, l'émergence de nouveaux modes de tarification, l'hyper-personnalisation de la consommation et donc l'émiettement des audiences, le caractère de plus en plus actif des consommateurs, la marchandisation de la sociabilité qui ferait des réseaux sociaux des lieux de collectes de valeur… Des ruptures moins spécifiques pourraient également avoir des conséquences majeures sur le secteur culture-médias : l'émergence d'une culture monde, la dégradation des habilités intellectuelles constatées par une culture de surface au détriment d'une aptitude à l'analyse… Quelles sont donc les principales inconnues déterminantes pour l'évolution des tensions évoquées et donc des ruptures probables. Risquons-nous à en citer quelques unes pour conclure : Comment le numérique sera-t-il perçu dans vingt ans (bienfait ou danger) ? Quel sera l'impact constaté sur la diversité de la consommation et des pratiques culturelles ? Cloud et streaming seront-ils les modes de consommation dominants ? La neutralité d'Internet sera-t-elle toujours de mise, ou le Cloud sera-t-il privatisé au bénéfice des grands acteurs technologiques ? Un tel réseau dématérialisé garantira-t-il le respect de la vie privée ? Aura-t-on trouvé des modes de tarification susceptibles de financer la création ? Quelles seront les attentes, les besoins des consommateurs en matière de diversité, d'utilisation des outils numériques, de communications virtuelles ou de demande réaffirmée de lien social physique ?

Bibliographie

(1) Marc LEIBA (IDATE), Les marchés du livre numérique, bilan 2010, présenté au Salon du livre de Paris, 2011), http://www.dgmic.culture.gouv.fr/IMG/pdf/MarchesLivreNum_Idate_Salon2011_.pdf

(2) Marc BOURREAU, Michel GENSOLLEN, François MOREAU, Musique enregistrée et numérique : quels scénarios d'évolution de la filière ? Culture

Prospective , n° 1, 2007, http://ses.telecom-paristech.fr/bourreau/Recherche/scenarios.pdf
(3) Ghislaine CHARTRON, Scénarios prospectifs pour l'édition scientifique, Hermès, vol.57, 2010, CNRS Editions, p.123-129, http://archivesic.ccsd.cnrs.fr/sic_00558746/fr/
(4) Jacques BAJON (dir), Les nouveaux formats de l'audiovisuel, IDATE, 2008, http://www.ddm.gouv.fr/IMG/pdf/70134_Les_nouveaux_formats_Final-mai_2008.pdf
(5) UDECAM, Etude prospective UDECAM : Quel sera le paysage Media en 2020 ?
http://www.docnews.fr/fr/archives/etudes/etude-prospective-udecam-quel-sera-paysage-media-2020,6316.html, 2010.
(6) Roger T. PEDAUQUE, Le document à la lumière du numérique, C&F Éditions, 2006.
(7) Evelyne BROUDOUX, « Le documentaire élargi au web » in « Le(s) Multi-média(s) ». Les Enjeux de l'information et de la communication (à paraître septembre 2011).
(8) Jeremy RIFKIN, L'âge de l'accès : la vérité sur la nouvelle économie, La Découverte, 2000.
(9) Nicolas CURIEN, Innovation et régulation au service de la révolution numérique, Journal of regulation, 2011, à paraître.
(10) Olivier DONNAT, Les pratiques culturelles des français à l'ère numérique, Enquête 2008, La Découverte-Ministère de la culture et de la communication, 2009. http://www.pratiquesculturelles.culture.gouv.fr/08resultat.php
(11) François MOREAU, Numérisation, économie numérique et mise en réseau des produits de contenu, in Greffe X. et N. Sonnac (eds), Web Culture, Dalloz, 2008.
(12) M. BADACHE, M.BOURREAU, M. GENSOLLEN, F. MOREAU, Les musiciens dans la révolution numérique, Irma éditions, Paris, 2009.
(13) Tim WU, "Network Neutrality, Broadband Discrimination", Journal of Telecommunications and High Technology Law, Vol. 2, p. 141, 2003
(14) LEE, Robin S., and Tim WU. "Subsidizing Creativity through Network Design: Zero-Pricing and Net Neutrality." Journal of Economic Perspectives, 23(3): 61–76, 2009.
(15) Chris ANDERSON, "The long trail", Wired Magazine, Octobre 2004, http://wired-vig.wired.com//wired/archive/12.10/tail.html
(16) G. CHARTRON, B. EPRON, A. MAHE, Pratiques documentaires numériques dans l'enseignement supérieur (sous la direction de), Presses de l'ENSSIB, ISBN 978-2-910227-88-3, 2011, à paraître.
(17) IPSOS MEDIA CT, Notoriété et usage du livre numérique, Enquête réalisée pour le magazine Livre Hebdo entre le 21 janvier et le 7 février 2011 auprès de 3 032 personnes de la population française âgées de 15 ans et plus. http://www.ipsos.fr/sites/default/files/attachments/ipsos_livre_hebdo_salon_du_livre.pdf

Partie 6 - Edition hypertextuelle

Ré-édition de Chrestien de Lihus dans l'hypertexte

Thierry DAUNOIS
NIT - Institut polytechnique de Lorraine, France

Résumé : Cet article traite de la ré-édition d'un ouvrage ancien traitant d'agriculture dans le réseau de wikis Wicri. Après une étude de différents projets de ré-édition numérique menés en France, il ouvre une réflexion plus large sur les possibilités offertes par la technologie wiki en matière d'édition numérique. On peut en effet imaginer viser la simple mise à disposition de textes non accessibles pour permettre leur réutilisation. Mais on peut également concevoir le développement d'outils spécifiques, l'exploitation de fonctionnalités sémantiques, dans une optique de recherche. Une alternative intermédiaire consiste à accompagner des projets de recherche, sur le volet de mise à disposition et de visibilité.

Mots-clés : Edition numérique, hypertexte, wiki, agriculture, annotation.

« Le livre, comme livre, appartient à l'auteur, mais comme pensée, il appartient, le mot n'est pas trop vaste, au genre humain. »
Victor Hugo
Discours d'introduction du Congrès littéraire international de 1878

1. Introduction

Certains documents (livres et manuscrits plus ou moins anciens, numérisés mais accessibles uniquement sous la forme de pdf, parfois avec un OCR[56] de qualité variable...) ne sont pas disponibles, en version exploitable (texte brut utilisable), sur internet. Dans le même temps, des chercheurs disposent, sur leur poste de travail, d'extraits, de chapitre, et même de livres entiers qu'ils ont entièrement retranscrits dans leur propre traitement de texte. Ce constat est à l'origine de l'idée initiale, qui

[56] OCR (optical character recognition) : acronyme qui désigne la reconnaissance optique de caractère. Avec l'utilisation de plus en plus fréquente de logiciels effectuant de la reconnaissance de caractères pour exploiter des pdf, "un OCR" désigne maintenant le fichier obtenu avec ces logiciels.

consistait à tester la possibilité de mettre ces textes à disposition de tous, pour que tout le travail de re-saisie ne soit pas perdu.
Même s'il n'est pas possible de quantifier cette "ressource", il n'est pas difficile d'imaginer qu'elle est importante. La production totale de l'humanité est estimée à quelques 130 millions d'ouvrages (évaluation effectuée par Google, dans le cadre de son vaste projet de numérisation)[57]. Les livres effectivement disponibles sur internet (sans même s'intéresser à leur "exploitabilité"), bien que l'on ait assisté à une véritable explosion en la matière depuis quelques années, se comptent plutôt en centaines de milliers. Le Projet Gutenberg[58] annonce 100.000 ouvrages traités, Gallica[59] 300.000. Même le méga-projet de Google Books portait sur 15 millions de livres. La marge reste donc colossale !
L'idée initiale a donc été de tester, sur le réseau de wikis Wicri, la mise en ligne de ressources éditoriales ayant déjà fait l'objet du travail ingrat de re-saisie en format de type word. Nous étions alors dans une simple optique de ré-édition directe.
De cet exercice - commencé comme une démarche d'information scientifique et technique (IST) et d'édition de "service public" -, est née rapidement l'idée qu'il était possible d'aller plus loin. Au-delà de ce cadre initial, pourquoi ne pas proposer un enrichissement des textes, avec des annotations collectives, mais, également, dans une optique de recherche ? Comment intégrer des outils permettant l'exploitation et la capitalisation de textes - enrichissement hypertexte [CLE 2007], analyse des données textuelles, paléographie[60], codicologie[61], philologie[62]... - ?
Cet article s'attache donc à présenter, dans un premier temps, le cadre technique et les choix initiaux qui constituent le contexte de cette expérimentation. Puis nous proposons une analyse des projets d'édition

---

[57] Article disponible sur le site américain mashable.com. (http://mashable.com/2010/08/06/number-of-books-in-the-world/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+Mashable+%28Mashable%29&utm_content=Google+Reader Mashable).

[58] Données chiffrées sur le site du Project Gutenberg (http://www.gutenberg.org/wiki/Main_Page).

[59] Données chiffrées sur Gallica (http://blog.bnf.fr/gallica/?p=2991).

[60] Paléographie : "science qui traite des écritures anciennes, de leurs origines et de leurs modifications au cours des temps et plus particulièrement de leur déchiffrement - définition du Trésor de la langue française informatisé (http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?8;s=2840995800;).

[61] Codicologie : "science annexe, mais distincte, de la paléographie et ayant pour objet l'étude matérielle des manuscrits en tant qu'objets archéologiques (par l'étude des matériaux servant à la confection du livre manuscrit et leur mise en œuvre) - définition du Trésor de la langue française informatisé (http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?8;s=1988699865;).

[62] Philologie : "discipline qui vise à rechercher, à conserver et à interpréter les documents, généralement écrits et le plus souvent littéraires, rédigés dans une langue donnée, et dont la tâche essentielle est d'établir une édition critique du texte - définition issue du *Trésor de la langue française informatisé (http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?8;s=1988699865;*).

hypertexte existants, avant d'effectuer un retour d'expérience sur notre expérimentation, depuis sa phase initiale jusque dans ses développements les plus récents. Enfin nous tentons de tracer les perspectives qu'ouvre le travail effectué dans le cadre du réseau Wicri - autant dans une optique d'IST que de culture scientifique et technique, et à destination de différents publics : grand public, enseignement, recherche -, et en quoi il pourrait constituer un apport pour les projets en cours.

2. Cadre technique - Choix initiaux

L'expérimentation se déroule dans le cadre de Wicri – WIkis pour les Communautés de la Recherche et de l'Innovation -, réseau de wikis sémantiques développé au sein de l'Institut national polytechnique de Lorraine (INPL). Initié en septembre 2008, le réseau Wicri compte aujourd'hui 101 wikis, tous développés à partir de la souche logicielle libre MediaWiki créée pour "l'encyclopédie libre" Wikipédia. Cette souche de wiki, si elle bénéficie (autant qu'elle en souffre) de l'aura du wiki le plus connu au monde, permet d'effectuer des choix différents de ceux qui animent l'encyclopédie en ligne. Ainsi, afin de prendre en compte les besoins spécifiques des communautés de la recherche, le réseau Wicri ne compte-t-il aucun wiki "libre" (sur lesquels la lecture et la contribution sont possibles sans être enregistré), mais uniquement des wikis publics (lecture libre, contribution uniquement pour les acteurs enregistrés et identifiés) et privés (lecture et contribution accessibles uniquement aux utilisateurs enregistrés et identifiés). Toute intervention est ainsi précisément rattachée à son auteur : chaque donnée est "traçable".

L'une des caractéristiques innovantes du projet Wicri est de constituer un réseau. Il est classique de trouver plusieurs wikis hébergés sur un même serveur, mais nous n'avons pas identifié d'initiatives proposant un véritable fonctionnement en réseau. Cela suscite des besoins particuliers, dont, pour assurer la cohérence des données d'un wiki à un autre, une réflexion approfondie sur la gestion des métadonnées. L'expérience menée par l'équipe Wicri en la matière a fait l'objet d'une publication au colloque DCMI 2010 à Pittsburgh [DUC 2010]. Autre différence de taille : Wikipédia exige de chaque contributeur qu'il appuie ses propos de références extérieures. À l'inverse, le réseau Wicri prévoit de s'appuyer sur l'expertise de comités scientifiques, fonctionnement adapté aux communautés de la recherche.

Au sein du réseau, on peut distinguer deux grands types de wikis. Les wikis "communs" (régionaux, Wicri/Lorraine et Wicri/Alsace ou thématiques, Wicri/Eau, Wicri/Bois...), d'une part, ont vocation à être animés par la communauté à laquelle ils se rattachent, tout en s'inscrivant dans les règles communes au réseau Wicri. Les wikis "institutionnels",

d'autre part, sont rattachés à une institution identifiée, qui en assure la direction éditoriale. Les choix éditoriaux, dans ce cas, peuvent être dérogatoires par rapports aux wikis communs du réseau : ouverture plus large, ou, au contraire, plus restrictive, du wiki aux contributeurs, par exemple.

Cette structuration offre deux intérêts. Elle favorise la construction collaborative de connaissances (en public ou en privé, au travers d'une application spécifique sur un wiki institutionnel, ou lors de la rédaction d'articles collectifs...). Puis elle assure la dissémination des informations ainsi générées, en leur assurant une bonne visibilité.

L'association de wikis "communs" et "institutionnels" permet de se positionner sur les différents niveaux de la connaissance. Les données brutes ont leur place sur des wikis institutionnel et de travail. Les wikis communs du réseau sont principalement destinés à la valorisation des résultats de la recherche et à s'intégrer dans les démarches de culture scientifique et technologique. Enfin, certains wikis institutionnels peuvent proposer de la vulgarisation scientifique grand public.

La question posée par Pierre Morlon, ingénieur au département Sciences pour l'action et le développement (SAD) de l'Institut national de la recherche agronomique (Inra), sur l'éventualité de mettre en ligne dans le réseau Wicri des ressources textuelles non accessibles par ailleurs - sous une forme facilement utilisable - sur internet a soulevé de premières interrogations. Ainsi, il fallait avant tout choisir où placer cette expérimentation dans le réseau, et quelle structure lui donner. L'ouvrage choisi pour cette expérimentation, les "Principes d'agriculture et d'économie", de Chrestien de Lihus (voir la documentation éditoriale et technique de l'expérimentation[63]), publié en 1804, a toute sa place sur le wiki thématique consacré à l'agronomie, Wicri/Agronomie[64].

Deuxième choix à effectuer : comment traduire, en pages wiki, un ouvrage de 336 pages ? Nous avions déjà eu l'occasion de travailler sur des articles, mais jamais sur des livres, ce qui demandait, de fait, une répartition de diverses sections sur différentes pages, avec un outil de navigation. Le travail préparatoire a donc consisté à étudier la table des matières de l'ouvrage. Cela a été l'occasion de la première constatation : la table des matières figurant dans l'ouvrage ne correspondait pas au découpage réel du texte, certaines sous-sections semblant être au même niveau dans le texte pouvant apparaître ou non dans la table. Seule la lecture du texte nous a permis de parvenir à une table des matières réelle qui semble satisfaisante, faisant apparaître quatre niveaux de titre. La

---

[63] La documentation éditoriale et technique est accessible sur Wicri/Agronomie (*http://ticri.inpl-nancy.fr/wicri-agronomie.fr/index.php/C_de_Lihus_1804_Principes_d%27agriculture_et_d%27%C3%A9conomie_-_Doc_%C3%A9ditoriale_et_technique*).

[64] Wicri/Agronomie *(http://ticri.inpl-nancy.fr/wicri-agronomie.fr/index.php/Accueil)*.

table des matières "corrigée" comporte ainsi, outre une préface et une conclusion, trois parties, constituées, pour la première, de deux chapitres, pour la deuxième, d'un chapitre unique, et, pour la troisième, de onze chapitres. Autrement dit, chaque chapitre fait en moyenne une vingtaine de pages (dans l'édition originale), les extrêmes étant de 2 pages (pour la conclusion), et de 43 pages (pour le chapitre "Août").
L'unité de travail retenue a donc été le chapitre : l'ouvrage de Chrestien de Lihus, dans le réseau Wicri, est donc publié sur 16 pages distinctes, préface, quatorze chapitres, et conclusion. Nous avons également créé plusieurs "modèles" (équivalent, dans MediaWiki, de macros, permettant de générer, sur plusieurs pages, un même texte), dont l'un destiné à faciliter la navigation d'un chapitre à un autre.
Il était intéressant également de conserver l'indication de la pagination initiale : ainsi, si l'on recherche un extrait dont on sait qu'il figure en page 228 de l'édition originale, on peut le retrouver rapidement. A cet effet, la pagination originale est indiquée (entre crochets et en caractères de couleur). En poussant cette démarche, nous avons ajouté une page consacrée à une "table des matières inverse", dans laquelle on peut retrouver directement, en fonction de la page que l'on recherche dans l'édition originale, à quel chapitre elle appartient.

3. Les projets d'édition hypertexte

Depuis l'un des premiers projets d'édition numérique français dont on retrouve la trace sur internet - l'expérience menée à l'Institut de recherche et d'histoire des textes (IRHT) en 2002-2003, et qui avait mobilisé un groupe de travail autour du manuscrit de "La lettre volée" [BUQ 2004] - bien du chemin a été parcouru. Un nombre croissant de projets d'édition numérique se sont organisés, qu'ils visent des textes isolés ou des corpus plus vastes.
Une analyse rapide des documents accessibles concernant ces projets montre qu'il ne semble pas y avoir eu de travaux menés sur l'idée de "ré-édition de service public", telle que nous la décrivons au démarrage de ce projet. Ainsi, on ne trouve pas trace de tentatives d'évaluation de la "ressource" disponible, ni d'expérimentation de mise en ligne de textes dans l'optique qu'ils deviennent simplement exploitables par d'autres.
Mais il semble également qu'il n'y ait eu que très peu de projets visant à mettre à disposition du plus grand nombre des textes d'intérêt scientifiques. En effet, on observe plutôt, aussi bien dans la littérature consacrée à l'édition numérique [LER 2008] que dans les projets qui semblent se rapprocher de l'expérimentation décrite ici, deux grandes orientations qui diffèrent sensiblement de notre démarche.
La première orientation consiste à travailler un corpus dans le sens des travaux de recherche d'un groupe de chercheurs identifiés, et, souvent,

de disciplines proches. Ainsi, la plate-forme Dinah revendique le fait de proposer un cadre de travail pour les philologues (voir encadré). C'est également le cas avec un projet comme Sourcencyme [DRA 2009], qui vise à créer une base de travail aux spécialistes de médiévistique sur les encyclopédies de l'époque.

La plateforme Dinah

La plateforme philologique Dinah [POR 2010] est destinée à annoter, transcrire et classer des documents manuscrits. Elle vise à "permettre l'expression conjointe de points de vues différents sous la forme de reclassements et d'annotations, [et en] la mise en œuvre des procédures nécessaires à la construction collaborative de vocabulaires d'annotations". Initiée dans le cadre du Cluster 13 (allocation de recherche 2007), la plateforme est accessible depuis avril 2010[65].

Cet outil, clairement destiné à une phase de travail, peut être utilisé quel que soit le contexte d'édition envisagé. Il peut donc être employé pour préparer une édition dans le cadre du réseau Wicri.

Fonctionnalités disponibles dans le cadre de la plateforme Dinah : annotation, travail collaboratif préparatoire.

Dans de tels contextes, la priorité est donnée au travail de recherche, à l'exploration des corpus dans l'optique des projets du laboratoire concerné, sans que, dans la plupart des cas, ces corpus soient du tout accessibles ne serait-ce qu'à d'autres équipes de recherche. Cette réalité est d'ailleurs à la base de la demande croissante de l'Agence nationale de la recherche (ANR) d'un véritable volet de "mise à disposition" des corpus ainsi traités.

La seconde orientation est celle que l'on pourrait désigner comme "technology-driven". En effet, qu'il s'agisse ou non d'un choix conscient, il apparaît clairement que de nombreux projets - et par exemple ceux menés dans le cadre du "cluster 13"[66] - s'appuient fortement sur une expertise en terme de traitement des images et de numérisation. Les techniques de numérisation bénéficient ainsi d'une abondante littérature [KAL 2000].

Le Cluster Culture, Patrimoine et Création (ou Cluster 13) [RAI 2008] porté par l'Université Lumière Lyon 2 vise à "coordonner les recherches pluridisciplinaires portant sur les productions, les objets et les usages sociaux qui engagent [...] une dimension et des enjeux d'ordre culturel et patrimonial, qu'il s'agisse du passé ou du contemporain le plus actuel.

[65] Présentation de la plateforme Dinah sur le site de l'Institut Jean-Toussaint Desanti (*http://institutdesanti.ens-lyon.fr/spip.php?rubrique27*).

[66] Le Cluster 13 (*http://cluster13.ens-lyon.fr/*).

L'ensemble du dispositif concerne principalement les sciences humaines et sociales, tout en étant ouvert à des collaborations avec les sciences exactes, les sciences de la nature et, en particulier, les sciences et techniques de l'information et de la communication (STIC)."
Ainsi, le projet "Hyperdonat" [BUR 2009] ou celui consacré aux dossiers de *Bouvard et Pécuchet*, de Flaubert [DOR 2009] – sont fortement marqués par l'impact de la nécessaire numérisation des documents. L'enjeu devient alors l'exploitation la plus aboutie possible de la technologie de traitement des images, et non la mise à disposition du plus grand nombre des textes ainsi traités. Ce sont au total 15 projets d'éditions critiques qui sont menés dans le cadre du Cluster 13, dont les *Essais* de Montaigne, les *Pensées* de Pascal, les œuvres complètes de Spinoza et de Montesquieu, les *Éloges académiques* de D'Alembert, l'*Essai sur les mœurs et l'esprit des nations*, de Voltaire.
L'expérimentation menée à l'École des Chartes, dans le cadre de Theleme[67] (acronyme de "Techniques pour l'historien en ligne : études, manuels, exercices") [POU 2006], mérite également d'être citée, même si elle se distingue également sensiblement du travail mené sur le réseau Wicri. En effet, ce travail, qui a nécessité le développement d'un outil spécifique de diffusion, ne semble pas permettre de travail collaboratif, et porte (du moins en l'état actuel de ce qui est consultable librement) uniquement sur des extraits brefs de documents. 116 dossiers sont accessibles, portant chacun sur une page d'un texte plus vaste, donnant accès à différents niveaux d'annotation (paléographiques, linguistiques, diplomatiques[68] ou historiques). Theleme est conçu essentiellement comme un support d'enseignement et d'initiation aux sciences et méthodes de l'histoire.
Or il apparaît que ces orientations ne sont pas antinomiques, mais pourraient au contraire se retrouver, dans une démarche commune visant à faire bénéficier l'ensemble des communautés de la recherche d'outils complémentaires, à la fois en terme de traitement des images, d'outils spécifiques, et de mise à disposition.
Enfin, le projet le plus proche de notre expérimentation est consacré à l'*Essai sur le récit, ou Entretiens sur la manière de raconter*, édition électronique de l'ouvrage de François-Joseph Bérardier de Bataut (1776).

---

[67] Système mis en place par l'École nationale des chartes (*http://theleme.enc.sorbonne.fr/dossiers/*). Chaque document est présenté sur une page, associée à d'autres pages, sur lesquelles sont donnés des commentaires (paléographiques, linguistiques, diplomatiques ou historiques).
[68] La diplomatique est une discipline qui vise à mettre en œuvre une compréhension critique des actes écrits. La Commission internationale de diplomatique précise qu'elle englobe tout écrit utilisé ou utilisable comme titre, fondamentalement pour prouver un droit. Cette définition est extraite de la présentation figurant sur le site Theleme (*http://theleme.enc.sorbonne.fr/cours/diplomatique*).

L'*Essai sur le récit*, édition électronique de l'ouvrage de François-Joseph Bérardier de Bataut
L'édition électronique de l'*Essai sur le récit, ou Entretiens sur la manière de raconter* [SCH 2010] est un projet mené par Christof Schöch, de l'Institut de Romanistique de l'Université de Kassel (Allemagne). Idée née à l'occasion d'un travail de thèse, l'édition électronique[69] dont il est question ici reprend l'unique édition connue de l'ouvrage de Bérardier de Bataut, publiée en 1776 à Paris.
La représentation du texte proposée donne la priorité au récit au détriment de la matérialité du livre, avec un découpage en chapitres et non par pages. Deux vues alternatives du texte sont proposées : une transcription linéaire du texte de l'édition originale, et un texte de lecture modernisé. Des notes textuelles et explicatives sont ajoutées : elles sont présentées sur un seul niveau, mais des évolutions ont été annoncées, notamment sur le système de notes, par C. Schöch, pour les mois à venir.

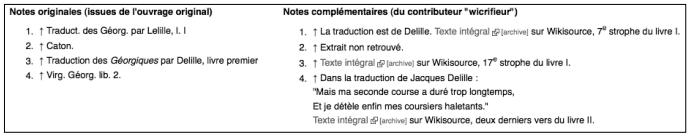
L'équipe du projet, composée d'une dizaine de personnes (outre le responsable du projet, l'équipe est composée du responsable du département d'informatique pratique, de l'un de ses collaborateurs et d'un groupe d'étudiants), a travaillé de 2008 à 2010, et prépare actuellement une nouvelle version.
Fonctionnalités disponibles dans le cadre de cette édition électronique : deux versions alternatives, annotations.
Mise en ligne effectuée sur drupal (après une version initiale sur DokuWiki).

4. L'expérimentation sur les *Principes d'agriculture et d'économie* de Chrestien de Lihus

La préface débute avec un épisode historique, rapporté par Cicéron (note originale : "Cic. de Oratore."), mettant en scène Annibal, dont il est dit qu'il fut très mécontent à l'écoute d'un philosophe, Phormion, qui discourait des devoirs d'un bon général, sans avoir jamais été militaire de sa vie. Cette anecdote sert à Chrestien de Lihus pour indiquer qu'il ne prend la plume qu'en temps qu'agriculteur lui-même, et pour apporter son expérience.
Avec l'objectif d'établir un lien vers une ressource en ligne, une rapide recherche a permis d'effectuer plusieurs observations. La première constatation est qu'aucune traduction de Cicéron disponible en ligne ne comporte le texte exact cité par Chrestien de Lihus. Soit il s'agit d'une traduction qu'il a effectué lui-même (ce dont, sans disposer d'éléments

[69] Le site dédié à l'édition électronique de l'Essai sur le récit (*http://berardier.org/*).

probants, on peut néanmoins douter), soit qu'il a repris quelque part, et qui diffère des versions que nous pouvons aujourd'hui trouver sur internet.
Dès lors, il a paru intéressant de donner des éléments plus complets sur cette citation, en apportant une note complémentaire à la note initiale : *De Oratore*, livre II, XVIII, Cicéron. Traduction consultée reprise des *Œuvres complètes de Cicéron*, publiées sous la direction de M. Nisard (1869). Texte intégral (lien) sur remacle.org". Cette première annotation du texte initial donne non seulement le lien vers une traduction en ligne mais vient également compléter la note originale, facilitant la recherche au lecteur.
Cette première note ouvrait la voie, renforcée dès la seconde : en effet, Chrestien de Lihus évoquait ensuite "L'auteur du *Préservatif contre l'Agromanie*", un ouvrage publié à Paris en 1762. Sans le citer nommément. Et pour cause, puisque ce livre était alors considéré comme anonyme, avant d'être attribué à Laurent-Benoît Desplaces. Figure ainsi la note complémentaire suivante : "Considéré un temps comme anonyme, le *Préservatif contre l'Agromanie* est attribué à Laurent-Benoît Desplaces. *Préservatif contre l'Agromanie, ou l'Agriculture réduite à ses vrais principes*, Paris : chez Jean-Thomas Hérissant, 1762, in-12, 197 p."
Il apparaissait dès lors qu'il y a un véritable intérêt à compléter, enrichir, et parfois apporter des éléments de correction aux notes originales (on parle de correction, par exemple, lorsqu'il est possible de constater qu'une citation, indiquée comme devant se trouver dans le tome II du *Voyages en France en 1787, 1788 et 1789* d'Arthur Young, se trouve en réalité dans le Tome I, page 452 (première traduction complète et critique par Henri Sée, édition Armand Colin, 1931).
Face à la constatation que la plupart des notes originales pouvaient ouvrir sur un ajout, il devenait utile d'opter pour une mise en page reprenant un double système de notes en bas de page, mettant en vis-à-vis la note originale et son commentaire (voir la figure I).


Notes originales (issues de l'ouvrage original)

1. ↑ Traduct. des Géorg. par Lelille, l. I
2. ↑ Caton.
3. ↑ Traduction des *Géorgiques* par Delille, livre premier
4. ↑ Virg. Géorg. lib. 2.

Notes complémentaires (du contributeur "wicrifieur")

1. ↑ La traduction est de Delille. Texte intégral [archive] sur Wikisource, 7e strophe du livre I.
2. ↑ Extrait non retrouvé.
3. ↑ Texte intégral [archive] sur Wikisource, 17e strophe du livre I.
4. ↑ Dans la traduction de Jacques Delille :
   "Mais ma seconde course a duré trop longtemps,
   Et je dételе enfin mes coursiers haletants."
   Texte intégral [archive] sur Wikisource, deux derniers vers du livre II.

Figure I : visualisation en vis-à-vis des notes originales et de leur commentaire.

On trouve ainsi, sur l'ensemble de l'ouvrage, matière à divers enrichissements. L'ajout de lien vers des ressources en ligne est le plus élémentaire. Parfois, il s'est avéré intéressant de comparer des sources diverses (traductions différentes, par exemple). Il a aussi été possible, parfois, d'identifier des erreurs dans des citations (sans pouvoir l'affirmer avec certitude, certaines de ces erreurs sont probablement directement

reprises des sources employées). Des sources imprécises ont également pu être éclairées : ainsi, une citation en latin, ""Delectant domi, non impediunt foris, pernoctant nobiscum, peregrinantur, rusticantur"", bénéficiait uniquement de la note suivante : "Cic. pro Archia, n°16." (figure II).

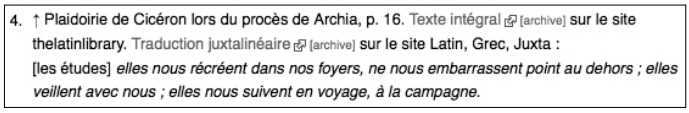
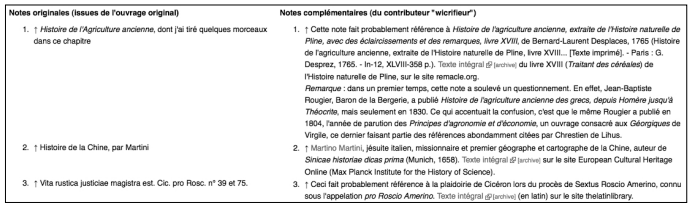

4. ↑ Plaidoirie de Cicéron lors du procès de Archia, p. 16. Texte intégral [archive] sur le site thelatinlibrary. Traduction juxtalinéaire [archive] sur le site Latin, Grec, Juxta :
[les études] *elles nous récréent dans nos foyers, ne nous embarrassent point au dehors ; elles veillent avec nous ; elles nous suivent en voyage, à la campagne.*

Figure II : les notes originelles peuvent être complétées, enrichies, rectifiées.

Au fur et à mesure de ce travail, il s'est également avéré utile d'apporter des notes sur le texte original, sur des éléments que Chrestien de Lihus n'avait pas annoté. Il reprend, par exemple, des citations latines sans les traduire. Il parle de "Rozier", sans préciser qu'il parle (probablement) de l'abbé Rozier, auteur d'un "Cours complet d'agriculture". De la même façon, lorsqu'il évoque le "chantre de Mantoue", il n'est pas forcément évident d'établir le lien (y compris en menant une recherche rapide sur internet) qu'il parle de Virgile. D'où une note : "Cette expression désigne Virgile. Voir à ce sujet les Études sur Virgile, tome III, page 132, de Pierre-François Tissot (1828, Paris). Texte intégral(lien) sur Gallica".
Enfin, pour des spécialistes de l'histoire des idées et de l'histoire de l'édition, on imagine facilement l'intérêt de ce type de démarche. Ainsi, dans le chapitre 1, partie I, la note originale [1] peut prêter à confusion, mais pourrait être intéressante dans cette optique. En effet, la note fait référence à un ouvrage employé comme source par Chrestien de Lihus, Histoire de l'Agriculture ancienne. Une première recherche fait apparaître qu'un ouvrage ainsi nommé est en effet paru, mais en 1830. Mais il n'est pas forcément totalement neutre d'observer également que ce même auteur a également publié, en 1804, justement, un autre livre, consacré aux Géorgiques, de Virgile, ce dernier étant abondamment cité par Chrestien de Lihus. Finalement, il apparaît (figure III), en se penchant plus en détail sur la question, que la note de Chrestien de Lihus fait plus probablement référence à Histoire de l'agriculture ancienne, extraite de l'Histoire naturelle de Pline, avec des éclaircissements et des remarques, livre XVIII, de Bernard-Laurent Desplaces (1765).



Notes originales (issues de l'ouvrage original)

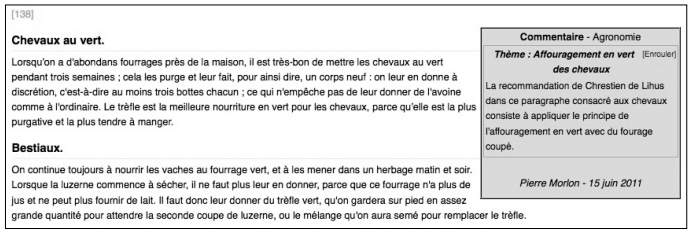
1. ↑ *Histoire de l'Agriculture ancienne*, dont j'ai tiré quelques morceaux dans ce chapitre
2. ↑ Histoire de la Chine, par Martini
3. ↑ Vita rustica justiciae magistra est. Cic. pro Rosc. n° 39 et 75.

Notes complémentaires (du contributeur "wicrifieur")

1. ↑ Cette note fait probablement référence à *Histoire de l'agriculture ancienne, extraite de l'Histoire naturelle de Pline, avec des éclaircissements et des remarques, livre XVIII*, de Bernard-Laurent Desplaces, 1765 (Histoire de l'agriculture ancienne, extraite de l'Histoire naturelle de Pline, livre XVIII... [Texte imprimé]. - Paris : G. Desprez, 1765. - In-12, XLVIII-358 p.). Texte intégral [archive] du livre XVIII (*Traitant des céréales*) de l'Histoire naturelle de Pline, sur le site remacle.org.
   Remarque : dans un premier temps, cette note a soulevé un questionnement. En effet, Jean-Baptiste Rougier, Baron de la Bergerie, a publié *Histoire de l'agriculture ancienne des grecs, depuis Homère jusqu'à Théocrite*, mais seulement en 1830. Ce qui accentuait la confusion, c'est que le même Rougier a publié en 1804, l'année de parution des *Principes d'agronomie et d'économie*, un ouvrage consacré aux *Géorgiques* de Virgile, ce dernier faisant partie des références abondamment citées par Chrestien de Lihus.
2. ↑ Martino Martini, jésuite italien, missionnaire et premier géographe et cartographe de la Chine, auteur de *Sinicae historiae dicas prima* (Munich, 1658). Texte intégral [archive] sur le site European Cultural Heritage Online (Max Planck Institute for the History of Science).
3. ↑ Ceci fait probablement référence à la plaidoirie de Cicéron lors du procès de Sextus Roscio Amerino, connu sous l'appelation *pro Roscio Amerino*. Texte intégral [archive] (en latin) sur le site thelatinlibrary.

Figure III : Une note complémentaire signale le questionnement soulevé initialement par la difficulté d'attribution de l'ouvrage cité dans la note originelle.

Ce travail sur le texte, ne nécessitant pas de compétences "disciplinaires" (bien que l'ouvrage traite d'agronomie, il n'est pas nécessaire d'être agronome pour apporter les éléments qui viennent d'être décrits), a mis en lumière l'intérêt de demander également à des spécialistes du(des) domaine(s) concerné(s) (ici, on peut imaginer faire appel à des agronomes, à des historiens, à des géographes...) de venir apporter leurs propres commentaires, afin d'enrichir encore la lecture du document. À titre d'exemple (figure IV), Pierre Morlon a accepté de se livrer à cet exercice, sur le thème de la jachère, d'une part (partie I, chapitre 2), et sur l'affouragement en vert des chevaux (partie III, chapitre Juin).



[138]

Chevaux au vert.

Lorsqu'on a d'abondans fourrages près de la maison, il est très-bon de mettre les chevaux au vert pendant trois semaines ; cela les purge et leur fait, pour ainsi dire, un corps neuf : on leur en donne à discrétion, c'est-à-dire au moins trois bottes chacun ; ce qui n'empêche pas de leur donner de l'avoine comme à l'ordinaire. Le trèfle est la meilleure nourriture en vert pour les chevaux, parce qu'elle est la plus purgative et la plus tendre à manger.

Bestiaux.

On continue toujours à nourrir les vaches au fourrage vert, et à les mener dans un herbage matin et soir. Lorsque la luzerne commence à sécher, il ne faut plus leur en donner, parce que ce fourrage n'a plus de jus et ne peut plus fournir de lait. Il faut donc leur donner du trèfle vert, qu'on gardera sur pied en assez grande quantité pour attendre la seconde coupe de luzerne, ou le mélange qu'on aura semé pour remplacer le trèfle.

Commentaire - Agronomie

Thème : Affouragement en vert des chevaux [Enrouler]

La recommandation de Chrestien de Lihus dans ce paragraphe consacré aux chevaux consiste à appliquer le principe de l'affouragement en vert avec du fourage coupé.

Pierre Morlon - 15 juin 2011

Figure IV : Une note "disciplinaire", commentaire transmis par un agronome (la note est ici déroulée, lorsqu'elle est enroulée (position initiale), seuls la discipline, le thème, l'auteur et la date sont visibles).

Ainsi s'est effectué, progressivement, le glissement d'un test de simple réédition d'un ouvrage ancien destiné à donner à chacun la possibilité de travailler sur ce texte non récupérable par d'autres moyens, vers une expérimentation plus complète de réédition commentée et enrichie, qui se rapproche davantage d'une édition critique.

5. Perspectives : le réseau Wicri et l'édition hypertexte de ressources textuelles

La plupart des projets existants d'édition hypertexte semblent, comme on vient de le voir, intégrer une étape de numérisation des données, ce

qui induit assez logiquement de se concentrer sur la question du traitement des images.
Cependant, dans le cas de la démarche qui est l'objet de cet article, la question de l'acquisition des données ne se posait pas, et pouvait être considérée comme annexe, sinon négligeable. Il était sensiblement plus important de se concentrer sur la question de la mise à disposition du résultat "final" (dans le cas d'un travail collaboratif sur un wiki, la notion de résultat final ne recoupe pas celle d'un résultat qui serait "définitif").
Ainsi, on peut parfaitement imaginer appliquer cette démarche, que les données soient déjà disponibles sous la forme d'un texte exploitable (comme c'était le cas pour les Principes d'agriculture et d'économie), qu'il s'agisse d'une source déjà numérisée et pour laquelle il n'existe pas de version exploitable autrement que par du traitement des images, ou qu'il s'agisse d'une source pour laquelle il n'existe ni texte exploitable, ni numérisation.
De fait, la question de l'acquisition des données n'a pas d'influence sur le traitement ultérieur : elle modifie uniquement – même si c'est déjà important – les questions de timing et de moyens nécessaires.
Ainsi, on peut imaginer plusieurs modes de fonctionnement, en fonction du contexte. Pour des chercheurs ou amateurs éclairés soucieux de donner accès à la communauté à des ressources "rares" dont ils disposent – et qui sont donc dans la position qui était la nôtre au début de cette expérimentation, dans une perspective d'édition de "service public" -, il est possible de proposer un espace de mise en ligne, des outils d'enrichissement et un accès à une communauté d'experts. Pour des bibliothèques, des institutions, des sociétés savantes... elles peuvent bénéficier des mêmes éléments, associés à un soutien technique renforcé sur la phase d'édition, afin d'accompagner leurs projets d'édition hypertexte. Enfin, l'équipe Wicri peut s'associer à des projets de recherche qui nécessiteraient le développement de nouvelles fonctionnalités.
La majorité des projets présente également un volet de traduction (à partir du latin, du grec). Cet aspect mérite que l'on s'y arrête un instant. En effet, cette phase de travail est en général totalement invisible pour le lecteur, qui veut accéder à la consultation parallèle du texte initial et de sa traduction, mais peut être sensiblement enrichie si le travail de traduction s'effectue de façon collaborative. De plus, le décodage même des opérations de traduction (sous la forme d'une "trace") serait potentiellement riche d'enseignement pour des lecteurs experts.
Dans cette optique, nous suggérons un travail en deux étapes : préparation et traduction du texte sur un wiki privé, accessible aux experts identifiés et apparentés au projet, puis mise à disposition des textes sur un wiki public, sur lequel se ferait alors le travail d'annotation décrit dans notre expérimentation.

On constate enfin que la plupart de ces projets ont des débouchés en terme de recherche (philologie, analyse critique, mise en perspective d'une oeuvre...) et en terme d'enseignement, offrant aux pédagogues de diverses disciplines des moyens nouveaux – et inaccessibles jusqu'ici – d'exploiter des sources anciennes, que ce soit pour les mettre en avant, ou pour en critiquer les manques. La démarche décrite ici offre, pour un coût et dans des délais particulièrement raisonnables, la possibilité à ces deux communautés d'étendre encore le champ des possibles. Notamment dans le cas de textes pour lesquels l'acquisition des données est déjà effective, la solution existante est non seulement simple mais complète. On peut ainsi estimer, pour un ouvrage de taille moyenne (500 pages) le temps total de mise en ligne à 2 mois.
La mise en place progressive (un wiki est un espace de "chantier" autant que de "versions finales") permet en outre de stimuler l'action des divers spécialistes qui peuvent être sollicités pour travailler sur les commentaires critiques disciplinaires. Notre expérimentation n'a pas étudié la possibilité d'exploiter toutes les fonctionnalités - notamment sémantiques - déjà existantes des wikis. Il est néanmoins clair que celles-ci pourraient être mobilisées afin de développer des outils d'analyse pour la recherche (indexation des termes, traitement des auteurs...). Et cela sans parler, naturellement, d'éventuels développements susceptibles de générer des fonctionnalités spécifiques.

6. Conclusion

D'une expérimentation simple et qui s'inscrivait dans un contexte d'édition de service public, destinée à donner un accès à tous à des ressources textuelles non disponibles sur internet, nous sommes, comme on peut le voir, passés à un outil permettant de fonctionner collaborativement sur différents niveaux d'annotation, pouvant intégrer une phase de travail collectif (par exemple sur la traduction des œuvres), avec une traçabilité fine des actions des divers acteurs.
Sans prétendre a priori pouvoir répondre à tous les besoins, il nous apparaît, à tout le moins, que la technologie des wikis telle que nous l'exploitons dans le cadre du réseau Wicri offre une possibilité intéressante de mise à disposition de textes, autant du fait de sa souplesse que de sa simplicité de mise en œuvre.
Il deviendrait alors possible, pour reprendre Victor Hugo, de faire en sorte que le livre, comme livre, n'appartiennent plus uniquement à l'équipe de recherche qui travaille dessus, mais bien à la communauté élargie des chercheurs de toutes disciplines.

Bibliographie

[BUQ 2004] Thierry BUQUET, « Quelques réflexions autour de la chaîne éditoriale d'un document numérique : l'exemple de La Lettre volée », *Le Médiéviste et l'ordinateur*, 43, 2004 [http://lemo.irht.cnrs.fr/43/43-04.htm].

[BUR 2009] Hyperdonat, une édition électronique des commentaires de Donat aux comédies de Térence. Bruno BUREAU, Maud INGARAO, Christian NICOLAS, Emmanuelle RAYMOND (dir.), CEROR, Université Lyon III, ENS de Lyon, 2007-2011. Accédé en ligne le 24 juin 2011, [http://hyperdonat.ens-lyon.fr].

[CLE 2007] Jean CLEMENT, L'hypertexte, une technologie intellectuelle à l'ère de la complexité, in Brossaud Claire, Reber Bernard, *Humanités numériques 1., Nouvelles technologies cognitives et épistémologie*, Hermès Lavoisier, 2007.

[DOR 2009] Stéphanie DORD-CROUSLE et Emmanuelle MORLOCK-GERSTENKORN, *L'édition électronique des dossiers de Bouvard et Pécuchet de Flaubert : des fragments textuels en quête de mobilité*, publié dans « Le patrimoine à l'ère du numérique : structuration et balisage » organisé à Caen les 10 et 11 décembre 2009.

[DRA 2009] Communication à la journée d'études : Digital Edition of Sources in Europe: Achievements, (juridical and technical) Problems and Prospects, à l'occasion des 175 ans de la Commission Royale d'Histoire. Meeting Porta Historica. [http://www.crhistoire.be/fr/partenariat/intern/portaPres_fr.html].

[DUC 2010] Jacques DUCLOY, Thierry DAUNOIS, Muriel FOULONNEAU, Alice HERMANN, Jean-Charles LAMIREl, Stéphane SIRE, Jean-Pierre THOMESSE et Christine VANOIRBEEK, *Métadonnées pour WICRI, un réseau de wikis sémantiques pour les communautés de la recherche et de l'innovation*, rapport de projet présenté au colloque DC 2010 (Pittsburgh, Etats-Unis). Version française consultable sur le wiki Wicri/Ticri [http://ticri.inpl-nancy.fr/ticri.fr/index.php/M%C3%A9tadonn%C3%A9es_pour_WICRI,_un_r%C3%A9seau_de_wikis_s%C3%A9mantiques_pour_les_communaut%C3%A9s_de_la_recherche_et_de_l%27innovation].

[KAL 2000] Enriketa KALLDRËMXHIU, *Les logiciels de numérisation des livres anciens*, Technical report, Université Claude Bernard Lyon1, 2000. [www.letterpress.ch/APINET/IMMPDF/LIVRE/gedkall.pdf] (pdf).

[LER 2008] Françoise LERICHE et Cécile MEYNARD , « Introduction. De l'hypertexte au manuscrit : le manuscrit réapproprié », *Recherches & Travaux* , 72 | 2008, mis en ligne le 15 décembre 2009. [http://recherchestravaux.revues.org/index82.html]. Consulté le 29 juin 2011.

[POR 2010] Pierre-Édouard PORTIER et Sylvie CALABRETTO. *DINAH, a philological platform for the construction of multi-structured documents*, in The European Conference on Research and Advanced Technology for Digital Libraries (ECDL), Mounia Lalmas, Joemon Jose, Andreas Rauber, Fabrizio Sebastiani, Ingo Frommholz ed. ECDL 2010 September 6 - 10, 2010, Glasgow. pp. 364-375. Research and advanced technology for digital libraries LNCS. Springer. ISBN 978-3-642-15463-8. ISSN 0302-9743. 2010. [http://liris.cnrs.fr/membres/?idn=peportie&onglet=publis].

[POU 2006] Gautier POUPEAU, *Les apports des technologies Web à l'édition critique : l'expérience de l'Ecole des chartes*, présenté à Digital philology and medieval texts, 01/2006 (Arezzo, Italie). [http://halshs.archives-ouvertes.fr/view_by_stamp.php?&halsid=p27rt2a5en6gcskfqnrpughlu0&label=S

HS&langue=fr&action_todo=view&id=sic_00137229&version=1&view=extended_view].
[RAI 2008] Ludivine RAIMONDO, *Enjeux et représentations de la science, de la technologie et de leurs usages* - rapport ENS Lyon
[SCH 2010] François-Joseph BERARDIER DE BATAUT, *Essai sur le récit, ou Entretiens sur la manière de raconter* (Paris : Charles-Pierre BERTON, 1776). Édition électronique sous la direction de Christof Schöch, 2010. [http://www.berardier.org] (Version 0.6, 12/2010).

Formalisation des processus d'éditique : Proposition d'un guide d'assistance à la formalisation de processus d'éditique à travers la transposition contextuelle de la notion de veille vue comme un système cybernétique

Sébastien BRUYERE
Responsable du pôle R&D
Custom Solutions

Vincent OECHSEL
DSI
Custom Solutions

Résumé : Aujourd'hui, les entreprises œuvrant dans les domaines du Marketing Opérationnel ont de véritables besoins en matière de production documentaire. En effet, la gestion des offres promotionnelles implique l'élaboration de nombreux documents. Dans ce cadre, les entreprises doivent conceptualiser des processus éditiques efficients afin d'optimiser la production et la distribution de documents afin de faciliter la transformation commerciale … La notion de veille, souvent utilisée pour s'informer de façon systématique sur des thématiques identifiées, de par sa structure, peut apporter un support utile à la structuration des processus éditiques. L'article a pour objectif de présenter une méthodologie dérivée de la notion de veille vue comme un système cybernétique pour formaliser les différents processus d'éditique nécessaire à la production des documents à contenu variable pour le Marketing Opérationnel.

Mots-clés : GED, éditique, veille, gouvernance documentaire, processus, cybernétique.

Introduction

Devant l'engouement des consommateurs pour les offres promotionnelles, les centres gestion et entreprises œuvrant dans les domaines du Marketing Opérationnel doivent faire face à de multiples défis. Le premier est essentiellement lié à la gestion de grandes productions de document avec des aspects liés à la performance de traitement et d'édition. Le second réside dans la mise en valeur des productions pour maximiser la transformation des affaires ou gagner en qualité de perception sur la communication au sein d'un projet client. le

troisième porte sur le choix du support de la production finale et du canal de communication adaptée pour sa diffusion.
La « révolution éditique » qui s'est traduite notamment à travers le passage de l'ère de « l'éditique de gestion » à l'ère de « l'éditique interactive » a apporté des dispositifs capables de relever ces défis. Cependant, elle nécessite d'avoir préalablement conceptualisé l'ensemble des processus « métiers » afin de les renseigner au sein de solutions d'éditique nouvelles générations. Mais cette conceptualisation des processus d'éditique peut s'avérer complexe et difficile à appréhender par les entreprises.

1. Matériel et méthode

Dans cette partie, nous reviendrons sur la notion d'éditique et les différents bouleversements qu'elle a connus. Puis nous démontrerons combien la notion de veille est structurellement semblable à la notion d'éditique. De ce constat, nous nous appuierons sur les recherches dans le domaine pour élaborer une aide à la structuration de processus formalisé pour l'éditique. Celle-ci sera modélisée à travers une approche systémique fondée sur la Cybernétique.

1.1. L'éditique, une notion devenue stratégique pour les entreprises

L'éditique est une notion qui est apparue dans les années 1990 essentiellement pour pallier à une carence forte des progiciels de gestion intégrée incapable de produire des documents en masse de manière performante, et d'aménager la structure des documents pour des mises en page de qualité. C'est alors qu'on a défini la notion comme *« les équipements matériels et logiciels mis en œuvre pour composer, imprimer, mettre sous pli et router industriellement ces documents. »* (Dupoirier, 2008). C'est ainsi que de nouveaux outils de composition industrielle capable de reprendre tout ou partie l'existant ont émergé. Cette ère est baptisé aujourd'hui l'ère de « l'éditique de gestion » (Czajka, 2010) se caractérise par une production en volume largement industrialisée et une décomposition des étapes nécessaires à l'édition sous forme de chaine éditique. Cette chaine est structurée à travers l'extraction/la réception des données, la composition et post-composition et la diffusion du document (De Montaigne, 2009). Les gains observés sont essentiellement liés à la productivité basée sur une optimisation des coûts d'affranchissement de poste *« avec des amortissements de projet parfois en moins d'une année »* (Blumereau, 2006). Au niveau des organisations, on voit émerger de nouveaux métiers comme la fonction de « Responsable Editique » ou encore des pôles métiers dédiés dans les grandes entreprises. Il devient possible de travailler sur la conception des documents à partir d'un ensemble de règles de gestion applicables aux différentes productions. La génération de documents de masse, construits à partir de différentes sources (textes, images, logos,

tableaux), est possible tout en conservant les acquis maitrisés comme la notion de publipostage.
Vers le milieu des années 2000 et devant l'apparition de « la révolution connectique » (Quoniam, Boutet, 2008), *« les documents ont un rôle essentiel dans la stratégie des entreprises et sont au cœur de la relation avec les clients et partenaires »* (Czajka, 2010). Les solutions d'éditique de première génération ne suffisent plus, elles nécessitent bien souvent l'intervention des services informatiques pour élaborer des modèles de traitement qui seront ensuite exploités par des opérateurs dédiés. Mais les entreprises en compétition doivent désormais réagir plus rapidement avec *« des temps de mise en place courts et des retours sur investissement rapides »,* la conquête de nouveau marché, l'appétence et le marketing des productions sortantes sont des facteurs clés de succès pour transformer des affaires et développer l'activité des entreprises. Les notions de temps réels et de personnalisation héritées notamment des révolutions connectiques appartiennent à cette nouvelle ère de l'éditique dite « interactive » (Alazard, 2010). Désormais les coûts directs sont extrêmement contrôlés via *« un choix rigoureux des canaux sortants les plus adaptés »* (priorité au numérique)*, « avec un contenu à jour et personnalisé pour le destinataire »* (Rémy, 2010). Les couts indirects sont aussi pris en considération avec notamment l'efficacité publicitaire et le marketing des documents. L'éditique devient un levier important pour faciliter les ventes, fidéliser les consommateurs …
Cependant, il apparait que *« les documents sont très divers et il en est de même des processus permettant de les créer, de les gérer et de les produire »* (Dupoirier, 2008), il convient donc, pour les intégrer le plus efficacement possible au sein des solutions d'éditique, de les définir le plus précisément possible en prenant en compte *« les dimensions organisationnelle et humaine dans le projet »* (Khristy, 2010). Cet aspect s'affiche par ailleurs dans un concept émergent plus large qui rejoint directement la stratégie de l'entreprise, la Gouvernance Documentaire (Boillet, 2011).

1.2. La veille, une notion structurante pour gérer l'information d'entreprise

La notion de veille est une notion plus ancienne que la notion d'éditique, elle est définie comme étant « un processus régulier de recherche, d'analyse et de sélection pertinente d'informations pouvant apporter des avantages compétitifs à une entreprise » (Pascoo, Le Ster-Beaumevielle, 2007). L'AFNOR et les experts du domaine s'accordent à définir que le processus de veille comporte cinq étapes avec l'expression des besoins informationnels, la collecte des informations, l'analyse des informations collectées, la diffusion et la mise à disposition d'une information sous la bonne forme, au bon interlocuteur et dans le format qui convient.
La « révolution connectique » qui se matérialise aujourd'hui notamment par la démocratisation d'une notion 2.0 (Quoniam, 2009) a permis de

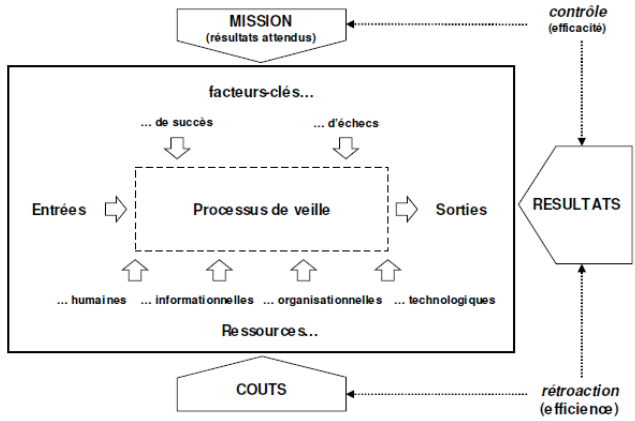
modéliser un concept de Veille 2.0 (Meingan, 2009). Celui-ci est caractérisé par un traitement de l'information 2.0 issu du travail collaboratif, du renforcement des réseaux d'entreprises à travers la constitution de communautés virtuelles, et de l'utilisation des services et des outils du Web 2.0 pour articuler les phases du processus de veille qui reste inchangé. L'information 2.0 se caractérise quant à elle comme une information « désolidarisée des applications et une accessibilité accrue par le biais de services web ». La notion de métadonnée est intégrée à l'ensemble du socle informationnel qui se caractérise désormais par un ensemble de services sécurisés, « les informations sont personnalisées, toujours disponibles et délivrées à la fois en temps réel et à la demande » (Lewis, 2009). Ces traits ont été déjà soulignés lors de la présentation de la notion d'éditique qui comme le processus de veille utilise le même fluide pour fonctionner, l'information 2.0.

De par la structure en étape itérative, le rapprochement de la veille avec l'ingénierie de méthodes informatique n'est plus à démontrer ; par ailleurs elle est souvent définie par les professionnels comme « la mise en place formalisée et organisée dans l'entreprise, d'un système d'information visant la collecte, le traitement et la diffusion de l'information concernant l'environnement de l'entreprise, ceci de façon continue et dynamique » (Bourcier-Desjardins, Mayère, Muet, & Salaün, 1990). La méthode MEDESSIE (Salles, 2000) qui est une approche de transfert des méthodes de conception des Systèmes d'Information à l'Intelligence Compétitive et les travaux de l'équipe SITE du laboratoire de recherche de LORIA qui œuvrent dans les domaines de la Modélisation et le développement de Systèmes d'Informations Stratégiques dans le cadre de l'Intelligence Economique sont de beaux exemples de transferts de méthodes. Nous considérons dans nos approches que la veille est une composante fondamentale de l'Intelligence Compétitive.

Dans ce prolongement, certains auteurs sont allés plus loin dans la modélisation en entrevoyant la veille comme un système cybernétique (Lesca & Carin-Fasan, 2008). L'objectif étant de développer les composantes de la veille en dressant les différentes missions à mener et les résultats attendus. Les entrées-sorties et les ressources techniques, organisationnelles, économiques et humaines sont abordées à travers des repères normalisés sous forme de facteurs clés de succès et d'échecs. Cette décomposition est particulièrement intéressante, car elle apporte de véritables repères en fonction des étapes et des composantes dans un processus informationnel.

Avant d'exposer les facteurs clés de succès et d'échecs, il convient de revenir sur ce qu'est un système cybernétique. Un système cybernétique est un ensemble d'éléments en interaction qui sont exprimés par des échanges d'informations et qui œuvrent à un but commun. Pour ce faire, le système en présence accepte des entrées et produit des sorties grâce à

un processus de transformation structurée qui constitue le noyau central de traitement. Les spécificités du système résident dans sa capacité à s'autoréguler et à être contrôlé (Wiener, 1948) & (Melki, 2008) & (O'Brien et al, 2001). Nous admettons ici qu'un processus est « une suite d'actions régulières et continues se déroulant d'une façon relativement bien spécifiée et aboutissant à un résultat quelconque » (Roussel & Lassalle, 2009). Dans le système cybernétique, le processus embarque donc des actions depuis l'entrée jusqu'à la sortie tout en étant soumis à des facteurs clés qui peuvent faire varier les actions et les états de ce même processus.



Représentation du système cybernétique de veille (Lesca, Caron Fasan, 2005)

Dans un système de veille cybernétique, l'entrée peut s'apparenter à une banque d'information non exhaustive et non homogène. Les étapes du processus de veille sont toujours les mêmes qu'exposés ci-dessus mais il est important d'intégrer que leurs enchainements n'est pas forcément séquentiel. En effet, en fonction des facteurs clés le système doit être capable de s'adapter, on parlera d'autorégulation. Ces facteurs clés sont liés à des influences environnementales, technologiques, organisationnelles et conditionnent les chances de succès ou d'échecs du résultat.

Lorsqu'on observe un dispositif de veille cybernétique au sein d'une organisation, on s'aperçoit qu'en entrée, on dispose d'une banque d'information importante ; la difficulté réside alors dans le choix des informations les plus pertinentes pour atteindre l'objectif.

Le traitement est souvent assisté par des systèmes informatiques capables d'orchestrer les étapes de veille pour produire le résultat. Les facteurs clés sont alors fortement liés aux paramétrages de celui-ci et aux

itérations qu'il est capable de proposer au responsable de veille au sein de l'organisation. Pour renseigner efficacement le système, le responsable de veille doit préalablement définir des processus informationnels comportant des étapes d'action, de contrôle, mais surtout des étapes alternatives en cas d'évènement soudain en provenance d'une influence externe. Ces influences conditionnent les facteurs clés de succès ou d'échecs.
En sortie, on retrouvera différents types de livrable possible, il peut s'agir de documents écrits ou électroniques comme les newsletters (Toupin, Lemaire, 2009) ou d'alertes précoces, d'hypothèses, de pistes d'action (Lesca, 2003). En somme, le résultat dépend directement du besoin informationnel, la forme et le destinataire sont définis en amont ou dans un scénario alternatif injecté dans le système.
En définitive, nous avons démontré que la veille pouvait s'apparenter à un système cybernétique (Lesca & Caron Fasan, 2005), que d'un point de vu séquentiel, les étapes du processus de veille sont des étapes certes plus abstraites, mais semblables aux processus d'éditique. Cela peut d'ailleurs s'expliquer par le fait que l'éditique et la veille utilisent le même fluide pour fonctionner, l'information. Par conséquent, les travaux de Lesca & Caron Fasan sont utilisables pour modéliser un système d'éditique cybernétique. Les facteurs clés du système de veille cybernétique seront alors utiles pour élaborer une méthodologie d'assistance à la définition de processus d'éditique. Toutefois, si la veille reste un processus informationnel plus abstrait que l'éditique, il sera nécessaire de conceptualiser davantage les facteurs clés pour élaborer un guide d'assistance adapté.

2. Résultats

Comme exposé précédemment, nous allons utiliser les facteurs clés du système de veille cybernétique pour définir un guide qui assistera un groupe projet dans la définition des processus d'éditique.

Transposition contextuelle des facteurs clés depuis un système cybernétique vue comme un système cybernétique à un système éditique

Facteurs Clés en provenance de la veille vue comme un système cybernétique (Lesca & Caron Faisan, 2005)	**Facteurs Clés transposés pour l'éditique**
Organisationnel	
Identifiez clairement les besoins en information.	Identifiez qui est chargé de créer les documents, quelles sont les règles de gestion associées, qui paramètre l'architecture du document (Oechsel, 2011).

Valoriser toutes les ressources existantes avant d'en solliciter de nouvelles.	Qui se charge implicitement des traitements éditiques au sein de l'organisation ? Quelles sont les limites de production éditique des applicatifs interconnectables à un hub dédié ?
Décentraliser et coordonner le processus de veille.	Articuler l'éditique dans la chaine de valeur de l'entreprise. Imaginer et définir les traitements éditiques au sein d'une solution dédiée et performante. Réfléchir aux interfaces techniques et humaines dans le cadre d'une centralisation des traitements éditiques.
Formaliser clairement le processus de veille.	Définir clairement les différents processus d'éditique en fonction des besoins. Utilisez des outils pour cartographier les processus et les différentes alternatives possibles et envisageables pour chacune des étapes. Qui décide, qui opère, quel est le pôle concerné.
Pérenniser le système par des dispositifs de feedback permettant d'écouter, de comprendre, de conseiller, de convaincre et faire adhérer les collaborateurs.	Paramétrer un reporting afin de pouvoir définir le potentiel éditique, la production éditique, les réalisations, la qualité (Oechsel, 2011).
Concevez un système sur mesure pour tenir compte des spécificités de l'entreprise.	Définir les connecteurs nécessaires pour une liaison avec les applicatifs de l'entreprise. Définir les processus détaillés en fonction des différentes communications à opérer. Définir les modèles de document et les zones personnalisables. Définir les publipostages et les canaux par défaut pour chacun des documents …
Matériel	
Formez le personnel aux activités de recherche et de diffusion de l'information.	Formez les équipes pour le paramétrage des documents. Formez les équipes aux paramétrages de diffusion des documents.
Valoriser le personnel œuvrant aux tâches de veille.	Valoriser le personnel affecté à l'éditique.
La technologie ne doit pas substituer la réflexion humaine.	Le responsable de l'éditique doit maitriser la stratégie de gouvernance documentaire inculpée par l'entreprise.
Humain	
Le projet doit être soutenu par la Direction et que celle-ci lui confère une légitimité.	L'éditique doit être un service supporté par les décideurs de l'entreprise.

Motiver les acteurs impliqués dans la collecte et la transmission des informations.	Valoriser le personnel et le motiver en fonction des productions. Une affaire peut être gagnée ou conservée grâce à la qualité d'un document.
Former à la collecte et/ou à l'analyse.	Le personnel doit être formé à l'acquisition des flux d'entrée et à l'analyse du traitement centralisée au sein de la solution d'éditique.
Former les acteurs à la coopération et pas seulement à l'opérationnel.	Expliquer et présenter les acteurs du pôle éditique.
Prévoir un animateur du processus.	Définir un responsable de l'éditique.

Comme exposé ci-avant, l'objectif principal de ce papier étant de définir des processus unitaires pour la conceptualisation de processus éditique efficace. Dans ce cadre, nous avons testé le guide ci-dessus lors d'une séance de travail au sein d'une société œuvrant dans les domaines du Marketing Opérationnel. Voici le résultat simplifié dépourvu d'informations confidentielles :

Modélisation de processus d'éditique simplifiée

Canal entrée	Application	Décisionnel	Opérationnel	Traitement	Structure entrée	Structure sortie	Type	Template	Doc générés
-		CP/auto	LOG (Logistique)	Flux via WS	Multi niveau	Flat		Flux via WS	Texte (CSV...)
espace depot									Texte (CSV...)
								Flux via WS	Colissimo suivi (format Expeditor)
espace depot		auto	auto	Flux via WS	Flat	Flat		Modèle spécifique	Texte brut
espace depot post script		CP	auto	Flux via WS	Flat	Flat			Texte (CSV...)
espace depot		CP	LOG	Via impression	Multi niveau	Flat			Texte (CSV...)
-		CP	LOG	Via impression	Multi niveau	Flat		EDI Transporteur)	Etiquettes Lettre Suivie (4 / page)
-		CP	LOG	Via impression	Multi niveau	Flat		EDI Transporteur	Etiquettes T 3 (70mm x 35mm, 3 x 8)
?		CP	LOG	Flux Connecteur	Multi niveau	Flat			Texte (CSV...)
-									Texte (CSV...)
-									Texte (CSV...)
-									Texte (CSV...)
-									Texte (CSV...)
espace dépôt		CP	LOG	Via impression	Flat	Flat			A4 Portrait
-									Texte (CSV...)
?		CP/auto	LOG	Flux via WS	Multi niveau	Flat		?	Fréquence C (format Logos V2)
espace dépôt		CP	PROD (Production)	Via impression	Flat	Flat			A4 Portrait
espace dépôt		auto	LOG	Via impression	Flat	Flat			A4 Portrait
espace dépôt		CP	PROD	Via impression	Flat	Flat			A4 Portrait
?								?	Texte (CSV...)
-		CP/auto	LOG	Via impression	Multi niveau	Flat		EDI Transporteur	Etiquettes T 3 (70mm x 35mm, 3 x 8)
?								?	Texte (CSV...)
espace dépôt									A4 Portrait
-									Texte (CSV...)
-									Texte (CSV...)
espace dépôt		CP	PROD	Via impression	Flat	Flat			A4 Portrait
espace dépôt		auto	PROD	Via impression	Flat	Multi niveau			A4 Portrait
espace dépôt		auto	auto	Flux via WS	Flat	Multi niveau			Texte (CSV...)
-								Flux via WS	A4 Paysage
-								Flux via WS	Texte (CSV...)
espace dépôt		CP	PROD	Fichier via EDI	Flat	Multi niveau			Standard CFONB
espace dépôt		CP	PROD	Fichier via EDI	Flat	Multi niveau			Standard CFONB
espace dépôt		CP	PROD	Fichier via EDI	Flat	Multi niveau			Standard CFONB

3. Discussion

À partir des facteurs clés définis dans le cadre des travaux portant sur la veille vue comme un système cybernétique, nous avons proposé un guide visant à assister la définition des processus documentaires et informationnels. Ceci étant, la transposition au domaine de l'éditique et aux métiers de l'entreprise demande une grande phase d'étude avec de nombreux retours d'expérience. Dans ce cadre, et au-delà de la transposition conceptuelle opérée dans cet article, il pourrait être intéressant de compléter l'étude par des questionnaires auprès de

l'ensemble des salariés ayant un rôle autour de l'éditique. De même, certaines activités qui posent le plus de problèmes en matière de gouvernance documentaire ne sont pas prises en compte dans le guide conceptualisé, l'archivage électronique, le traitement des emails, la gestion des archives papier en sont des exemples (Boillet, 2011). On pourrait aussi compléter notre guide par différents facteurs clés issues de la littérature comme l'étude du serdaLAB qui a interrogé plus de 250 entreprises sur leurs gouvernances documentaires, les préconisations de la norme ISO 30300 ou les travaux de l'APROGED (Association des Professionnels de la Gestion Electronique des Documents) qui propose une série de questions que le Directeur de Projet doit se poser à chaque étape du cycle de vie du document. Cette dernière discussion démontre aussi l'importance de deux aspects dans un projet d'éditique vue comme un système cybernétique, l'importance des processus pour gagner en efficacité, mais aussi l'organisation nécessaire pour piloter ce métier. Notre proposition traite finalement davantage ce dernier aspect et la littérature apporte principalement des facteurs clés sur la définition des processus documentaires. Leurs complémentarités peuvent s'avérer indispensables dans le cadre de la mise en place d'un projet organisationnel et technique d'éditique.

4. Conclusion

Dans cet article nous nous sommes attachés à démontrer que la structure d'un processus de veille et similaire aux processus d'Editique. Nous nous sommes ensuite appuyés sur les travaux visant à identifier les facteurs clés de succès et d'échecs dans le cadre d'un dispositif de veille vue comme un système cybernétique pour définir des facteurs clés dans le cadre d'un projet d'éditique. Après analyse, nous avons déterminé que nos facteurs clés peuvent s'avérer très utiles pour des aspects organisationnels dans un projet, mais peut-être insuffisants sur des aspects technico-fonctionnels. Pour pallier à ces carences, la complémentarité avec la littérature dans le domaine comme les référentiels de norme, le livre vert de l'APROGED ou les sondages représentatifs sur la gouvernance documentaire peuvent s'avérer utile.

Bibliographie

ALAZARD A. (2010). Donner de la valeur à vos documents. Consulté de http://www.slideshare.net/alain1965/editique-interactive-4316845

BEIGNON J.-M., & BOURMAUD, F.-X. (2005). Intelligence économique et entreprise: comprendre son environnement pour agir. Editions L'Harmattan.

BLUMEREAU B. (2003). L'heure des progiciels. Banque Magazine, (650 (Supplément)), 39-40.
BOILLET V. (2011). La gouvernance documentaire dans les entreprises françaises. SerdaLAB. Consulté de http://serda.com/fileadmin/serda/images/serdaLAB/Etudes_completes/livres_blancs/Livre_blanc_serdaLAB_gouvernance_documentaire.pdf
COTTIN M., FAURE C., FUZEAU P., JULES A., & TAILLEFER M. (2011). Livre Blanc - Introduction à la série des normes ISO 30300, Système de management des documents d'activité. AFNOR. Consulté de http://www.bivi.fonctions-documentaires.afnor.org/content/download/23222/154684/version/4/file/CG46CN11+Livre+Blanc+RecordsManagement.pdf
CZAJKA C. (2010). L'éditique interactive au coeur des processus métiers. Business Document. Consulté de http://www.industrie.com/impression/mediatheque/0/6/4/000001460.pdf
DUPOIRIER G. (2008). Enjeux et risques de la dématérialisation des documuents. Techniques de l'Ingénieur, Documents numériques Gestion de contenu. Consulté de http://www.techniques-ingenieur.fr.ezproxy.scd.univmed.fr:2048/base-documentaire/technologies-de-l-information-th9/documents-numeriques-gestion-de-contenu-ti403/enjeux-et-risques-de-la-dematerialisation-des-documents-h7602/
DUPOIRIER G. (2009). Gestion de documents numériques et de leur contenu. Techniques de l'Ingénieur, Documents numériques Gestion de contenu.
FRANÇOIS P. (2006). L'éditique: quand l'entreprise se donne les moyens de produire ses documents. ZDNet Business & Technologies. WebZine Professionnel, . Consulté juin 16, 2011, de http://www.zdnet.fr/actualites/l-editique-quand-l-entreprise-se-donne-les-moyens-de-produire-ses-documents-39363679.htm
KHRISTY J.-P. (2010). Garantir le succès d'un projet éditique. Techniques de l'Ingénieur. Consulté de http://www.techniques-ingenieur.fr.ezproxy.scd.univmed.fr:2048/actualite/informatique-electronique-telecoms-thematique_193/garantir-le-succes-d-un-projet-editique-article_7653/
LEMAIRE S. (2009). Outils et méthodes de diffusion des résultats de la veille: le cas du Centre International d'Etudes Pédagogiques (CIEP) (Mémoire de Chef de Projet) (p. 100). http://memsic.ccsd.cnrs.fr/mem_00524364/: CNAM.
LESCA N., & CARON-FASAN, M.-L. (2005). La veille vue comme un système cybernétique. Revue Finance Contrôle Stratégie, 8(4), 93–120.
MELKI A. (2008). Système d'aide à la régulation et évaluation des transports multimodaux integrant les Cybercars. Ecole Centrale de Lille, Lille.
DE MONTAIGNE J. (2009). L'éditique, ou la production en masse de documents. le CXP. Consulté de http://www.cxp.fr/gespointsed/imgbreves/Sommaire_Editique.pdf
MOURAIN J. (2011). Etat et enjeux de la gouvernance documentaire. Diaporama, Paris, La Défense. Consulté de http://lb7.reedexpo.fr/Data/kmreed_informatique/block/F_a86c5d332f75b90ad2a610ecfa9ade484d9ddebc3e2f4.pdf
REMY C. (2009). Editique, automatiser l'envoi de documents. Solutions & Logiciels, (9). Consulté de http://www.solutions-logiciels.com/magazine_articles.php?titre=Editique-automatiser-lenvoi-de-documents&id_article=156

ROUSSEL P., & LASSALE, B. (2009). Comment analyser un incident de la chaīne transfusionnelle. Transfusion clinique et biologique, 16(1), 53–60.
SEGUI M., (2011). Initiation à Documentation : création, dématérialisation, stockage, archivage ... Diaporama présenté à DOCUMENTATION, Paris, La Défense. Consulté de http://lb7.reedexpo.fr/Data/kmreed_informatique/block/F_17f4a19440349ebc02e2cedaf6bc98784d9dde0d80904.pdf
ZANON O. (2011). Comment anticiper les freins face à la dématérialisation. Diaporama, Paris, La Défense. Consulté de http://lb7.reedexpo.fr/Data/kmreed_informatique/block/F_78aa12a094de311e3a2a2c67081993e84d9dde9761501.pdf

Accès aux collections de presse ancienne : une étude exploratoire

Céline PAGANELLI
Université Montpellier 3, France ; Laboratoire Gresec, Université Grenoble 3

Evelyne MOUNIER
Université Grenoble 2, France ; Laboratoire Gresec, Université Grenoble 3

Stéphanie POUCHOT
Université de Lyon, France ; Université Claude Bernard Lyon 1, ELICO, EA 41

Résumé : Les collections de presse ancienne constituent des objets d'étude et des ressources indispensables aux travaux des spécialistes, universitaires ou professionnels de l'information. D'un accès difficile, elles font l'objet de campagnes de numérisation visant à les protéger et à les rendre accessibles au plus grand nombre. De fait, on constate un élargissement et une diversification des publics. Il reste malaisé d'apprécier l'étendue, la nature et l'impact réel sur les usages et les pratiques. L'étude présentée ici est de nature exploratoire. Elle a pour objet les collections de presse ancienne régionale de la Bibliothèque municipale de Lyon (France) dont elle propose une première approche des modes d'accès et usages.

Mots-clés : Numérisation, presse ancienne illustrée, usages des bibliothèques numériques.

Introduction

La presse du XIX^e^ siècle représente, notamment pour les chercheurs, les étudiants et les journalistes, une source d'information considérable sur la vie politique, économique, sociale et culturelle d'une époque ; l'ancrage local de certains titres présentant, lui, une source d'information sur l'histoire d'une région. Mais les collections de presse sont fragiles et leur manipulation délicate. La valeur historique de la presse et les contraintes liées à sa consultation ont tôt fait d'inciter les institutions à se préoccuper de la préservation et de la diffusion de ce type de collections. Ainsi, dès la fin des années 1950 en France est créée *l'Association pour la conservation et la reproduction photographique de la presse* qui débute le microfilmage de titres de presse en partenariat avec la Bibliothèque Nationale (Delaunay 1996). Plus récemment, la Bibliothèque nationale de France (BnF) a lancé, en

2005, un plan de numérisation de la presse quotidienne française des XIXe et XXe siècles pour la rendre accessible via la bibliothèque numérique *Gallica*. L'ensemble du corpus numérisé représente, en 2011, une trentaine de titres, soit plus de 3,5 millions de pages[70]. Parallèlement aux bases proposées par les bibliothèques et services d'archives, des outils spécifiquement dédiés à la création et la consultation de corpus de journaux anciens font également leur apparition (Schor 2008).

La numérisation de telles collections a plusieurs objectifs : une accessibilité accrue, une valorisation des richesses patrimoniales et une meilleure conservation. De plus, pour des raisons de préservation, les collections sur papier sont le plus souvent retirées de la consultation dès que la copie numérique est mise à disposition du public, la plupart du temps sur internet. L'accès aux collections papier de journaux anciens et le travail sur ces documents sont communs dans toutes les salles de lecture et les usagers habituels sont relativement bien connus. Les modes d'accès en ligne posent en revanche davantage de questions : Quels objectifs peuvent amener un usager à utiliser les collections de presse ancienne ? Quels types de contenus le lecteur vient-il chercher et sous quelle(s) forme(s) (texte, gravure…) ? Comment cherche-t-il ? Est-il indifférent de tourner les pages d'un journal ou de consulter à l'écran ? Des usagers nouveaux se manifestent-ils ? Quelles sont les pratiques émergentes liées au passage au numérique ?

Comme d'autres bibliothèques partenaires de la BnF, la Bibliothèque municipale de Lyon (BM de Lyon) s'inscrit dans cette dynamique, en numérisant depuis 2007 les collections de la presse illustrée rhônalpine du XIXe siècle. Notre étude prend place au sein du projet *CaNu XIX*, qui vise à valoriser et mettre en ligne ces collections patrimoniales. L'un des volets de ce projet concerne spécifiquement les usagers et a pour objectif d'appréhender de manière globale les attentes et usages en matière de consultation et d'utilisation des collections de presse ancienne numérisée ou sur papier.

Dans cet article, nous commençons par exposer le positionnement et le contexte de notre travail avant de présenter l'étude exploratoire que nous avons menée. Nous terminons par une discussion concernant la contribution de cette recherche ainsi que par quelques pistes prospectives.

1. Positionnement et contexte

Le travail que nous proposons, ancré en sciences de l'information et de la communication, s'inscrit dans un ensemble d'études qui appréhendent la recherche d'information comme un processus situé et contextualisé

[70] Source : site de la BnF, http://www.bnf.fr/, visité le 11/10/11.

(Fondin 2001, Boubée 2010). En effet, deux visions de la recherche d'information coexistent dans la discipline : celle qui l'envisage essentiellement comme processus technique et met l'accent sur les questions de stockage ou de traitement, approche qui a nourri la discipline de nombreux travaux comme le soulignent Rosalba Palermiti ou Claude Poissenot (Palermiti 2002, Poissenot 2002) ; et celle qui s'intéresse aux facteurs humains et considèrent la recherche d'information comme un processus de communication (Fondin 2001) ; dans cette lignée, de nombreuses études ont porté sur les usages et pratiques informationnelles (Chaudiron 2010). C'est dans ce dernier cadre que le travail présenté se situe. L'approche que nous adoptons considère d'une part que l'activité principale de l'individu et le contexte dans lequel elle prend place influencent ses attentes en matière d'information, les stratégies de recherche d'information mises en œuvre et, plus largement, l'activité informationnelle. Différents travaux ont montré que le contexte avait effectivement une influence déterminante sur l'activité d'information, par exemple (Cheuk 1999, Guyot 2002, Bartlett 2005, Staii 2006, Miranda 2007, Fabre 2010). Cette approche considère d'autre part que la prise en compte des usages et des pratiques informationnelles s'avère nécessaire non seulement pour comprendre et décrire l'activité des individus mais également pour établir des préconisations concernant le fonctionnement et le développement des interfaces. Nous considérons ainsi que la conception d'un système nécessite que soient pris en compte à la fois les caractéristiques des utilisateurs visés et les spécificités du contexte de la tâche de recherche d'information (Paganelli 2003).

Dans le cas de l'accès à l'information dans des collections de presse ancienne, qu'elles soient sur papier ou numérisées, la prise en compte des usages et pratiques pourrait donc améliorer la conception d'un dispositif d'accès à ces collections. S'il existe des travaux sur l'étude de la convivialité des outils, ceux-ci portent plus volontiers sur la comparaison d'interfaces ou sur les techniques de recherche mises en œuvre que sur la compréhension des raisons qui poussent l'usager à utiliser ces ressources ou la nature des tâches visées par l'usager (Bryan-Kinns 2000). Notre travail vise, au travers d'une étude exploratoire, à appréhender de manière globale les attentes et usages en matière de consultation et d'utilisation des collections de presse ancienne numérisée ou sur papier. Ici, l'analyse des attentes et des usages doit être entendue comme un préalable à une réflexion et à des préconisations sur les traitements documentaires et sur les modalités d'accès aux documents numérisés. Dans cet article, nous nous intéressons précisément aux usages des collections de presse ancienne ; les usages étant entendus comme « l'expression d'un processus constitué d'interactions complexes mettant en relation un individu et un dispositif » (Chaudiron 2010). Ainsi, ce sont les interactions d'un dispositif ou d'une collection qui sont étudiés ici, à

la fois dans leur dimension individuelle et cognitive, mais également dans une dimension sociale permettant de prendre en compte le contexte dans lequel l'usage se situe. Dans cette acception, l'usage apparaît alors comme restrictif par rapport aux pratiques informationnelles, terme qui désigne la manière dont l'ensemble des dispositifs, sources, outils et compétences cognitives sont effectivement mobilisés dans les différentes situations de production, de recherche, traitement de l'information (Ihadjadene 2009, Gardiès 2010).

2. Accéder à la presse ancienne : un mouvement général de numérisation et de diffusion en ligne

Les projets et réalisations de numérisation de collections, que ce soient des collections de presse, de manuscrits, d'ouvrages, sont nombreux, en France comme à l'étranger, principalement en Europe ou en Amérique du Nord (Smolczewska-Tona 2008). En France, le plan de numérisation lancé par le Ministère de la Culture en 1996 (Bequet 2000) a permis un fort développement de ces projets. Le catalogue des collections numérisées du ministère de la Culture fait état, en octobre 2011 2010, de 1868 collections numérisées et de 642 institutions concernées, dont une très forte majorité de bibliothèques. Pourtant, la situation n'en est pas moins contrastée. Ainsi, Westeel (2009, 29) fait remarquer qu'« *une observation précise de la situation des bibliothèques montre un bilan plutôt mitigé. Les projets en ligne et les véritables bibliothèques numériques sont finalement assez peu nombreux. On peut compter une trentaine de projets pour les bibliothèques municipales* ». Ce constat nuancé atteste de la difficulté des structures documentaires à maintenir ce type de projets dans la durée. Il n'est donc pas étonnant que les études concernant les usages de ces collections numérisées soient encore rares.

Les pouvoirs publics français affichent, certes, une volonté de s'intéresser à la manière dont ces fonds numérisés sont utilisés (Lesquins 2006), ainsi l'appel à projets 2010 du ministère de la Culture concernant la numérisation du patrimoine culturel portait-il « *une attention particulière [...] à l'émergence d'outils et de services favorisant des usages culturels innovants par les internautes* ». De même, les professionnels impliqués dans ces projets expriment leur intérêt pour connaître les motivations et les comportements de leurs usagers. En 2009, nous avons mené une enquête auprès de 18 structures françaises ayant numérisé des titres de presse ancienne. Les résultats montrent que, si les professionnels ont une connaissance intuitive des usagers qui utilisent leurs fonds, ils n'ont en revanche pour l'instant pas réalisé d'études qualitatives ou quantitatives précises sur ces lecteurs.

Plus largement, s'il existe un grand nombre d'études sur les usages des bibliothèques numériques (Bryan-Kinns 2000, Papy 2007), peu de

travaux s'intéressent précisément aux usages des fonds patrimoniaux numérisés. Dans ce contexte, les études sur *Gallica* (Lupovici 2003) et *Europeana* (Lesquins 2007) sont particulièrement précieuses. Elles permettent de caractériser les usages et de mettre évidence des portraits types d'usagers en ce qui concerne l'étude *Gallica*, ou d'évaluer l'utilisation d'une interface en ce qui concerne l'étude d'*Europeana*. L'étude de Lupovici (2003) montre notamment que les bibliothèques électroniques attirent un public qui n'est pas nécessairement habitué des bibliothèques mais qui y vient par le biais de recherches spécifiques. Ce public est assez différent de celui des bibliothèques classiques et le chercheur professionnel, notamment, y est peu représenté. On découvre enfin une population d'internautes seniors, gros consommateurs d'internet, avec un fort taux d'équipement et dont les centres d'intérêts gravitent autour des contenus culturels. Peu d'auteurs ont étudié précisément les pratiques de consultation des fonds anciens. Une étude menée à la BM de Lyon (Belot 2004) montre que les usagers sont majoritairement des hommes résidant en région Rhône-Alpes et d'un niveau d'études supérieur (91% ont au minimum une licence), et que les étudiants et les cadres constituent 77% de cette population. Leurs objectifs de recherche sont quant à eux répartis en trois catégories : un travail dans le cadre de leurs études (44 %), des recherches personnelles (30 %) et des recherches professionnelles (26 %).

Les bibliothèques engagées dans les chantiers de numérisation y voient, en général, une occasion de valoriser les fonds de presse ancienne qu'elles détiennent, en les mettant à disposition d'un public plus étendu. La mise en ligne de ces richesses s'accompagne donc le plus souvent d'annonces sur le site web de la bibliothèque, voire dans la presse régionale. Aussi, une fois numérisés, ces fonds sont-ils bien visibles dès la page d'accueil du site et sont facilement accessibles. De fait, la mise en ligne des collections de presse ancienne représenterait une opportunité pour tous les usagers. Ainsi, Tétu (2010) remarque un regain d'intérêt récent pour la presse ancienne illustrée de la part des historiens, mouvement probablement encouragé par les campagnes de numérisation de ces collections.

Cependant, ce type de ressources reste relativement peu utilisé parce que peu accessible, entre autres pour deux raisons :

Si la plupart des interfaces proposent de chercher par le titre de presse, selon la date de publication ainsi que par mots-clés, on peut supposer que les possibilités offertes en matière de recherche ou de consultation ne sont pas en adéquation avec les attentes des usagers.

La recherche en texte intégral n'est souvent pas pertinente en raison des limites de la reconnaissance optique de caractères. Différentes études, dont celle de Bermès (2007), ont en effet montré que la reconnaissance optique de caractères (OCR) en matière de numérisation des documents

anciens restreint sérieusement l'intérêt de la recherche en texte intégral, notamment dans le cas des collections de presse ancienne.

3. Une étude exploratoire à la Bibliothèque municipale de Lyon

3.1. Contexte géographique et institutionnel

Le site principal de la Bibliothèque municipale de Lyon est situé dans le quartier de la Part Dieu, à deux pas de la plus importante gare SNCF lyonnaise et du centre commercial le plus grand intra muros. Il s'agit d'une zone commerçante certes excentrée mais très achalandée et à laquelle il est aisé d'accéder en transports doux, puisqu'une station de métro, un arrêt de tram et une station de Velov[71] sont tout proches. Notons par ailleurs qu'aucun établissement public de l'enseignement supérieur ne se trouve à proximité géographique directe. En 2007, la BM de Lyon a démarré un important programme de numérisation et de mise en ligne de ses fonds de presse locale patrimoniale (fin du XIX^{e} – première moitié du XX^{e}). L'hebdomadaire *Le Progrès Illustré de Lyon* a été le premier des titres de périodiques choisi pour cette valorisation. Présenté comme le « supplément littéraire » du *Progrès de Lyon*, ce titre est paru entre décembre 1890 et septembre 1905.

3.2. Hypothèses et méthode de travail

Dans ce contexte, nous envisageons ici les hypothèses de recherche suivantes :

Tout d'abord, les besoins et attentes de l'usager habituel des fonds de presse ancienne ne diffèrent pas selon les supports.

Toutefois, la mise en ligne des collections induit une diversification et un élargissement des publics susceptibles de les consulter. Ainsi, d'autres besoins et attentes ont pu également naître qui ne sont pas nécessairement pris en compte par les interfaces de recherche et de consultation.

Enfin, la structure bien particulière des documents de presse induit des habitudes d'exploration et des formes d'utilisation spécifiques. Il se peut que la lecture à l'écran et/ou l'interface de recherche entraînent des pratiques ou stratégies différentes de celles mises en œuvre lors de la consultation de collections papier.

Nous avons ainsi mené une enquête par questionnaires auprès des usagers de la BM Lyon entre juin 2009 et janvier 2010, afin de mieux les connaître et de comprendre leurs pratiques et leurs motivations à consulter les fonds patrimoniaux. Nous avons choisi de recueillir les données selon deux modalités :

[71] Système de location de vélos courte durée en libre service.

Via un questionnaire en ligne visant à colliger des données auprès des lecteurs en ligne de ce titre, il était accessible depuis la page d'accueil de la base.
Via un questionnaire papier, destiné à collecter des informations auprès des usagers se déplaçant à la bibliothèque de la Part Dieu, sur rendez-vous, pour consulter des titres de presse ancienne. Ce questionnaire a été distribué à ces usagers à leur arrivée à la banque d'accueil de la bibliothèque.
Les questionnaires en ligne et papier comportent respectivement 27 et 26 questions permettant, entre autres, de cerner les motivations des répondants, leurs stratégies de recherche et de consultation ainsi que leur niveau de satisfaction par rapport aux informations trouvées. Du point de vue technique, pour le questionnaire en ligne comme pour le traitement des réponses au questionnaire papier, nous avons utilisé le logiciel libre de sondage LimeSurvey (http://www.limesurvey.org/). Cet outil libre, développé en PHP propose une interface web d'administration et permet de stocker les données dans une base MySQL.

4. Résultats

Le nombre de questionnaires colligés reste peu élevé, que ce soit en ligne ou sur place. En effet, nous avons obtenu 40 réponses au total, soit 16 questionnaires en ligne et 24 questionnaires sur papier. Ce faible nombre de répondants ne donne aucune indication sur la consultation réelle du *Progrès Illustré* en ligne et ne peut constituer un échantillon représentatif de la population concernée. Sur la période de collecte de données, le site internet a en effet été visité par plus de 30 000 visiteurs uniques[72]. Répondre au questionnaire pour accéder à la base n'étant pas une obligation, la plupart des usagers ne se sont pas sentis tenus de le remplir.
En ce qui concerne les questionnaires distribués sur place, les réponses ont pu porter sur d'autres titres que le *Progrès Illustré* puisqu'ils ont été complétés par un public spécialisé venant consulter différents titres de presse ancienne.
Bien que fondés sur un faible nombre de questionnaires et recueillis selon des moyens et canaux différents, nos résultats sont comparables à ceux des enquêtes sur *Europeana* ou *Gallica*, notamment sur l'origine géographique, les caractéristiques socioprofessionnelles, les classes d'âge.

[72] Statistiques fournies par la BM Lyon. Environ un tiers des visites sont effectuées par des robots de type crawler.

Origine géographique
Compte tenu de la spécificité du fonds (presse régionale illustrée), il n'est pas surprenant que plus de la moitié des répondants soient domiciliés en région Rhône-Alpes (27 sur 40). Toutefois, l'intérêt pour la presse ancienne rhônalpine dépasse le cadre régional : ainsi près de la moitié des répondants en ligne ne sont pas « locaux » ; on peut noter la présence de quelques répondants internautes (Europe hors France et autres pays), qui accèdent déjà à la collection (3 sur 40) en ligne.

Caractéristiques socioprofessionnelles
Les usagers des collections de périodiques anciens peuvent être étudiants, enseignants-chercheurs, bibliothécaires, journalistes, conférenciers, mais aussi retraités ou des personnes sans emploi. Des professions intermédiaires telles que les employés, les ouvriers, artisans, agriculteurs sont peu ou pas représentées. Dans la mesure où les enquêtes Gallica et Europeana aboutissent au même constat, on peut penser que ce résultat n'est probablement pas lié au nombre de répondants. Globalement, deux groupes d'usagers sont les plus nombreux : les étudiants (12 sur 40) et les retraités (11 sur 40) Mais, il semble que le public étudiant se déplace davantage à la Bibliothèque pour consulter les documents sur place (10 sur 24), alors que les consultants de la base *Progrès illustré*, seraient majoritairement des retraités (7 sur 16). Sans doute moins mobiles que les autres, ils profiteraient davantage de la mise en ligne.

Lieu de connexion habituel, équipement et pratiques habituelles du Web
En termes d'équipements, de lieu de connexion habituel, ou de pratiques sur la Toile, nos observations rejoignent également celles de *Gallica* et d'*Europeana*. Pour nos deux enquêtes, les usagers se disent équipés à la maison avec un ordinateur et une connexion à internet, leur lieu de connexion principal étant leur domicile. De tels résultats ne sont pas surprenants. En effet, déjà en 2002, les utilisateurs de *Gallica* déclaraient majoritairement se connecter de chez eux. De même, l'enquête *Europeana* de février 2007 permettait de constater que plus de 50% de la population interrogée disposait déjà d'un accès à internet, sachant que l'accès principal des répondants de l'enquête *Europeana* se faisait au domicile personnel.
De la même manière, dans les deux enquêtes, les répondants apparaissent comme des habitués du web : 13 usagers sur 16 de la collection en ligne et 17 sur 24 des collections papier déclarent utiliser internet depuis plus de 5 ans. Globalement, les usagers des deux groupes passent beaucoup de temps sur la Toile (environ deux tiers des répondants de chaque groupe déclarent y consacrer au moins 6h par semaine). Mais le groupe des consultants de la revue en ligne passerait

plus de temps que l'autre groupe sur internet, soit plus de 20h par semaine. Paradoxalement, les deux tiers des individus de ce même groupe, majoritairement constitué de personnes retraitées, s'estiment peu à l'aise dans cette activité. Dans le même temps, le groupe usagers des collections papier, globalement plus jeune, estime pour moitié être très l'aise ou assez à l'aise avec l'utilisation d'internet. De même, l'immense majorité de chaque groupe dit utiliser internet pour les loisirs (près de 9 répondants sur 10 concernant la collection en ligne et 8 sur 10 pour les utilisateurs des collections papier).

Les pratiques en matière de recherche d'information

Contexte et objectifs de la consultation de la presse ancienne régionale

S'agissant de la consultation proprement dite des collections de presse ancienne, les lecteurs se déplacent pour consulter les collections papier principalement pour des travaux universitaires (15 sur 24) avec pour objectifs la publication d'ouvrages spécialisés, la production de mémoires universitaires ; certains ont également des objectifs pédagogiques comme la préparation d'un cours ou d'autres objectifs professionnels tels que production de films documentaires. La plus part du temps, chacun a donc un objectif précis et peut formuler le thème ou le sujet qui le conduit à travailler sur un périodique ancien. Ainsi, l'un déclare faire un mémoire sur la presse lyonnaise sous La Commune, un chercheur travaille sur les fabricants de soieries, un autre encore sur les troubles politiques en Italie vus par la presse locale. Parfois le sujet de recherche est moins précis, comme pour cet usager souhaitant comparer deux visions de la guerre (celle du Préfet et celle de la presse).

Ces usagers sont avant tout des habitués des bibliothèques et des collections de périodiques anciens : ils savent où chercher, comment chercher. Ils ont coutume de travailler sur un ou plusieurs titres de journaux, peu importe le support : sur papier (12 sur 24), microfilms (14 sur 24) ou numérisés (5 sur 24).

La consultation des collections en ligne ne semble pas répondre aux mêmes caractéristiques. Le profil des consultants est différent : peu d'universitaires (2 sur 16), d'étudiants (1 sur 16), quelques professionnels autres (5 sur 16), des retraités (7 sur 16). Parfois, ce sont des motifs professionnels (3 sur 16) ou universitaires (2 sur 16) qui génèrent la démarche, mais le plus souvent on exprime des motifs personnels ou culturels (8 sur 16). Par exemple, telle personne lit *Madame Bovary* et cherche à mieux comprendre l'époque et le contexte du roman ; telle autre réalise, à titre bénévole, un site web sur les forts ; une troisième prépare une visite culturelle et cherche des anecdotes. De plus, dans ce groupe, la majorité des répondants consulte pour la première fois (13 répondants sur 16).

Enfin, nous rencontrons une forte proportion d'usagers qui n'ont pas de recherche particulière à effectuer et qui arrivent là par hasard. Il ne s'agit donc pas d'un public habituel usager de la presse ancienne, mais c'est la mise en ligne qui commence à attirer un public différent.
Ces observations recoupent donc les résultats de l'étude des usages du catalogue *Gallica* : 44% des répondants à l'étude sur *Gallica* déclarent consulter pour un usage exclusivement personnel et s'apparentent à des « chercheurs amateurs » (Assadi 2003).
Ces observations recoupent également les indications données par les professionnels des archives et des bibliothèques que nous avons sondés il y a deux ans quant à la perception de leurs usagers. Ils distinguent effectivement les usages personnels des usages professionnels. Dans le premier cas, ils envisagent essentiellement les généalogistes. Dans le second cas, ils citent, outre les chercheurs et étudiants, des journalistes de la presse locale venus consulter pour écrire leurs articles, des urbanistes utilisant la presse locale pour s'informer sur la gestion des risques naturels dans la région et, enfin, des professionnels des musées consultant ces collections dans le cadre de l'organisation d'expositions ou manifestations centrées sur le local.
Il apparaît donc que les contextes dans lesquels les usagers consultent les collections de presse ancienne sont variés, tout comme les objectifs de ces consultations.

Types d'informations recherchées

Certes, on constate que dans le cas de la recherche en ligne, certains usagers (4 sur 16) ne cherchent rien de précis, étant arrivés par hasard sur la page du *Progrès Illustré*. Hormis cette catégorie, il n'est pas sûr que les différences entre les deux groupes soient significatives. Tout d'abord, les recherches dans les collections papiers portent prioritairement sur des événements, avec mention de dates et de personnes. Mais cela n'est pas surprenant compte tenu de la forte proportion de spécialistes ayant recours à ce type de source dans l'optique de produire des travaux universitaires.
De plus, dans les deux groupes de répondants, des tendances identiques coexistent : à côté de la recherche par thématiques (12 sur 40), l'usager recherche directement des événements (23 sur 40), dates (16 sur 40), lieux (23 sur 40), personnes (19 sur 40).
D'ailleurs, la recherche d'événements recouvre des sujets aussi disparates que des élections, des procès, le décès d'une personnalité, une fête, une exposition... Cette variété est également visible dans la recherche sur les personnes qui peut concerner un peintre précis, des personnages politiques locaux ou nationaux, comme par exemple, les acteurs de la droite lyonnaise, des noms de brasseurs et assimilés : *Chanal – Janson, Trimolet*.

La recherche par date fait apparaître des années, ou des périodes plus ou moins définies : par exemple : « *De fin décembre 1856 à mars 1857* », ou encore « *depuis la Révolution à l'Empire*».
S'agissant des noms de lieux, les usagers mentionnent une grande variété de noms de localités et de villes de la région lyonnaise. Sont également recherchés des bâtiments, comme *le Grand Opéra* ou des lieux de brassage.
Ainsi, les objectifs de recherche peuvent-ils être précis ou non et exprimés précisément ou non. Globalement, ces objectifs se situent sur le terrain du factuel plus que sur celui de la thématique générale.

Modes d'exploration et de récupération de l'information

Sur les exemplaires papier des journaux comme sur leur version numérisée, la lecture en diagonale est la pratique la plus répandue (25 sur 40). Elle correspond plus précisément au comportement des spécialistes (historiens par exemple) qui travaillent sur une question en explorant rapidement les documents, le plus souvent en raison de l'importance du volume du fonds à compulser. Par ailleurs, ce type d'usager n'est pas certain de trouver d'éléments d'information pertinents et n'a pas d'idée sur la façon dont le contenu pourrait être formulé.
Dans le cas du support papier, la plupart des usagers (16 sur 24) recherchent directement une rubrique, d'autres (5 sur 24) cherchent une page précise et ne lisent que celle-ci. Pour autant, le dépouillement habituel des journaux sur papier reste parfois malcommode : l'un estime qu'il *« il manque un classement chronologique des parutions. Quand on cherche un événement on ne connaît pas, ou probablement, les titres de la presse locale »,* un autre qu'il faudrait *« Lier les articles qui parlent du même sujet mais qui sont éparpillés dans le journal »,* d'autres soulignent la difficulté à « *de sélectionner l'année où il y aura le plus de commentaires qui l'intéressent* » ou à « *ne rien rater* », d'autres enfin font remarquer des difficultés de manipulations à cause de « *la taille du livre mais le papier reste plus agréable à lire qu'un écran* ».
La consultation en ligne offre de nombreuses possibilités ; elle permet de balayer la liste des titres ou de rechercher directement l'un d'eux, d'accéder ensuite aux années de publication, puis à chaque numéro ; l'usager peut ensuite faire défiler chaque page ou encore, via un sommaire organisé en pages, choisir directement une page. La recherche par mots clé porte soit sur l'ensemble des collections soit sur un titre déjà sélectionné et permet d'accéder directement au passage voulu dans les pages pertinentes, sous réserve d'avoir bien formulé sa requête.
Globalement, l'interface est vue comme ne présentant pas de difficultés particulières. Pourtant, quelques remarques montrent qu'une partie des usagers est « déstabilisée » soit parce qu'il est nécessaire de rechercher d'abord par date puis par numéro *(« La recherche de n° par dates de publication : impossibilité de se retrouver dans la collection entière »),* soit par la lecture en diagonale devient difficile sur un écran *(« Pouvoir lire en diagonal*

tout le journal, c'est fastidieux d'être obligée d'agrandir pour lire et ensuite très compliqué le copié collé. »), soit encore parce qu'ils ne voient pas comment formuler une recherche par mots clé *(« trouver la bonne entrée ...»),* soit parce qu'il faut jongler avec les différents grossissements pour pouvoir lire. Enfin, on remarque qu'une partie non négligeable des répondants (un quart) préfère dépouiller page par page tout un numéro et lire plus attentivement, de crainte de passer à côté d'une information pertinente que l'OCR n'aurait pas détectée ou quand ils ne voient pas comment formuler leur requête précisément.

Nature de l'information recherchée

Concernant la nature de l'information recherchée, nous observons une nette différence entre les deux groupes de lecteurs (collections papier et version en ligne). Les personnes consultant les collections papier recherchent principalement de l'information textuelle (19 sur 24) alors que les consultants de la collection en ligne disent rechercher des illustrations (5 sur 16) ou indifféremment du texte ou des illustrations (12 sur 16). Cette différence est cohérente avec le fait qu'une bonne partie des internautes est arrivée sur *Le Progrès illustré* en ligne par hasard, et ne cherche rien en particulier.

Cependant, tous les éléments textuels d'un journal sont susceptibles d'intéresser les usagers. Sur les deux types de supports, sont surtout recherchés des articles (31 sur 40) mais aussi des rubriques précises (rubriques de mode par exemple, ou de cuisine) et parfois des feuilletons. Tous ces objets, caractéristiques de la structure d'un journal, constituent aussi bien des sources d'informations que des balises pour la lecture rapide. Ils conviennent particulièrement aux lecteurs qui ont à mener des travaux à caractère universitaire ou qui, d'une manière plus générale, sont des habitués des bibliothèques, familiers de la presse ancienne, c'est-à-dire des individus connaissant bien la structure des titres consultés.

Comme son nom l'indique, *Le Progrès illustré* propose un bon nombre d'illustrations, notamment des gravures, des plans ou encore des dessins humoristiques. Les deux groupes de répondants ne manifestent pas le même intérêt vis à vis des ressources iconographiques. S'agissant de la collection en ligne, tous les sujets sauf un répondent et formulent des choix ; dans le cas de la collection papier, la moitié (11 sur 24) ne donne aucune réponse.

Concernant la nature même des illustrations recherchées, les gravures sont les plus prisées par les deux groupes, surtout si elles sont situées en première page. Viennent ensuite les cartes et plans puis les dessins d'humour et enfin les graphiques. Notons qu'environ un répondant sur cinq signale que sa recherche d'image ne porte pas sur un type d'illustration en particulier.

Les objectifs des répondants peuvent être ici un facteur explicatif : dans le cadre de recherches à visée professionnelle ou universitaire, les

personnes ayant répondu au questionnaire papier recherchent des données textuelles portant sur des sujets précis (événements, dates, personnes…). Les internautes, aux objectifs majoritairement personnels, ont quant à eux moins d'attentes en la matière mais ceux ayant une préférence privilégient les illustrations.

5. Contribution de la recherche : discussion

Le présent article rend compte de deux enquêtes avec certes un nombre de restreint répondants mais dont les résultats sont en correspondance avec des enquêtes de grande envergure comme celles menées sur *Gallica* et *Europeana*. Les questionnaires avaient pour objectif de préciser les usages effectifs et les besoins des usagers des collections de périodiques anciens. Leurs résultats mettent en évidence certaines pistes sur lesquelles il convient de nous interroger.

La mise en ligne élargit les publics

*L*es résultats de notre enquête confirment des tendances déjà mises en évidence lors des études menées auprès des publics de *Gallica* et *Europeana*. La mise en ligne de collections anciennes permet un élargissement des publics. En effet, lorsque ce type de collections est numérisé, ce ne sont plus seulement des chercheurs ou spécialistes qui les consultent ; le grand public est également intéressé, que ce soit par curiosité ou pour des recherches plus ciblées en lien avec son histoire familiale ou des évènements survenus dans sa région (quand il s'agit de consulter la presse locale). Par ailleurs, ce sont la plupart du temps des personnes qui ne fréquentent pas la bibliothèque en tant que lieu physique. Cet élargissement des publics est également géographique ; là encore, comme pour l'étude *Gallica*, les publics consultent les collections à partir de différentes régions françaises ou de l'étranger. Ces résultats correspondent bien à l'idée selon laquelle la numérisation des collections est réalisée dans une perspective de valorisation, avec comme principal objectif de toucher de nouveaux publics. Pourtant, la question des modes d'accès à l'information, et notamment de l'adéquation des modes d'accès au public de ces collections, n'est pas posée de manière évidente dans les travaux en lien avec les bibliothèques numériques.

Diversité des publics, des objectifs poursuivis et des stratégies d'accès

Comme l'ont déjà souligné d'autres travaux (Guyot 2002, Jarvelin 2004, Bartlett 2005,), il apparaît que l'activité principale, le contexte professionnel voire les contraintes spécifiques de la tâche ont un impact sur l'activité d'information. Ici, nos observations suggèrent que l'activité principale ou habituelle mais aussi les objectifs immédiats des usagers

amènent ceux-ci à mettre en œuvre des stratégies qui peuvent être différentes selon ces objectifs et le type d'information recherché au moment de la consultation. Ainsi, les consultants habituels de la presse ancienne, agissent souvent dans un cadre professionnel ou universitaire et sont à la recherche d'informations factuelles mais dont le degré de précision peut varier. Ils peuvent ainsi parfois expliciter exactement ce qu'ils veulent trouver : un lieu, un événement précis, un personnage, des gravures, des cartes, etc. Dans d'autres cas, ils ne disposent que d'indices pour trouver des informations utiles et ils peinent parfois à exprimer clairement ce sur quoi porte leur besoin. Ils sont donc tenus de dépouiller les journaux et ne peuvent pas systématiquement se limiter à la recherche par mots clés dans le cas des journaux numérisés.
Toutefois, ces attentes et cette stratégie ne sont pas uniformes. Avec la mise en ligne, la proportion des visiteurs issus du « grand public » est accrue. Ainsi remarque-t-on l'importance de visiteurs du *Progrès Illustré* qui, par ailleurs, ne sont pas des habitués des bibliothèques, arrivant par hasard sur la page du *Progrès Illustré*. C'est dans ce cadre que se placent les consultations menées dans un cadre personnel, avec des objectifs différents. Le mode d'accès prédominant passe davantage par le butinage, la « promenade » dans les collections, même si, là aussi, il peut y avoir besoin d'accéder à une information précise.

Influence du support et du media

On remarque tout d'abord que les consultants de la presse ancienne en ligne ou sur papier, et plus particulièrement habitués des bibliothèques, disposent d'une stratégie habituelle appuyée sur une connaissance de la structuration des journaux : orientée principalement vers la recherche d'informations textuelles, cette stratégie consiste d'une part à lire en diagonale en prenant appui sur les rubriques, les pages, les titres et d'autre part en un dépouillement très fouillé et une lecture approfondie des articles sélectionnés. Or, dans le cadre des consultations en ligne, l'usager peut être déstabilisé parce qu'il ne peut reporter facilement ses habitudes : soit parce que les outils ne le permettent pas, soit parce que, sur une longue durée, la lecture à l'écran est jugée inconfortable. En outre, l'intérêt de la recherche par mots clés est bien sûr directement lié à la qualité de la reconnaissance optique de caractères.
Ensuite, et s'agissant de la consultation en ligne, on remarque la place occupée par le document iconographique, que l'usager occasionnel recherche le plus souvent au hasard, en parcourant les pages, en trouvant ou non. Pourtant, une typologie du matériel graphique comme une localisation dans la structure du journal constitueraient des indices de recherche pour l'usager. Il est donc sans doute possible d'imaginer des fonctionnalités et des interfaces de recherche spécifiques.

6. Conclusion

L'un des biais des enquêtes présentées dans cet article est le faible nombre de répondants. Les résultats ne sont, de fait, pas généralisables. Néanmoins, il nous semble intéressant de souligner que les objectifs des usagers consultant la presse ancienne sont variés et que les stratégies mises en œuvre le sont tout autant. Il est donc tout à fait essentiel que les interfaces de consultation prennent en compte cette hétérogénéité des besoins et des pratiques.

Cela peut se traduire par la constitution de parcours thématiques comme l'a mis en place la BM de Lyon sur son portail[73]. Cela permet en effet de guider les usagers « grand public » et de satisfaire leur curiosité (Clavier 2010), notamment via le repérage des événements, objectif de recherche d'un panel large d'usagers.

La mise en contexte du corpus (Cazenave 2004) devrait également contribuer à améliorer les interfaces, de même que le développement d'outils favorisant une lecture confortable pour les spécialistes mettant en œuvre un dépouillement précis et systématique. La modélisation du rubriquage des journaux, par exemple, permettrait aux usagers habituels de retrouver un cadre connu et faciliterait leur recherche ou la lecture en diagonale.

Enfin, la mise en place d'outils participatifs issus du web dit 2.0 (commentaires, indexation collaborative…) devrait également favoriser, d'une part, une plus grande visibilité de ce type de contenus et, d'autre part, une appropriation accrue de la part des usagers. Certaines bibliothèques l'ont déjà compris et mis en œuvre, comme le Réseau des bibliothèques publiques de Montréal qui propose sur son site internet une page dédiée aux réseaux sociaux sur lesquels il est actif[74].

La Bibliothèque du Congrès utilise également une large variété de ce type d'outils (blog, réseaux sociaux, etc.) et notamment Flickr pour ses collections patrimoniales de photographies et journaux[75]. En France, c'est par exemple le cas de la BnF pour le projet *Gallica*, dont la page Facebook rassemble plus de 7500 fans[76] et le compte Twitter[77] est suivi par près de 2300 abonnés. Même si ce type d'actions doit impérativement s'insérer dans une politique globale de communication des établissements (Leclercq 2011), c'est indéniablement par ce biais que les bibliothèques de demain renforceront leur rôle de médiation.

---

[73] http://collections.bm-lyon.fr/presseXIX/, visité le 19/09/11.

[74] http://bibliomontreal.com/reseaux-sociaux/, visité le 19/09/11.

[75] http://www.flickr.com/photos/library_of_congress/collections/, visité le 19/09/11.

[76] https://www.facebook.com/GallicaBnF, visité le 19/09/11.

[77] http://twitter.com/#!/GallicaBnF, visité le 19/09/11.

7. Remerciements

Cette recherche a bénéficié du soutien de la Région Rhône-Alpes dans le cadre du Cluster 13 « Culture, patrimoine et création ». Elle n'aurait pu se dérouler dans de bonnes conditions sans le concours de la Bibliothèque municipale de Lyon, notamment de Pierre-Yves Landron. Nous remercions les répondants, pour le temps qu'ils nous ont consacré. Merci également aux étudiants de Master Technologie de l'information ayant pris en charge les aspects techniques du questionnaire en ligne (Khaled Belalia, Youcef Benaissa, Abderahim Boukmiche, Hassaan Cherrat et Eddine Lakehal Nour).

8. Bibliographie

ASSADI Houssem (sous la dir.de). 2003. Usages des bibliothèques électroniques en ligne : projet Bibusages – rapport final, France télécom R&D, version 1.1, 25 juillet 2003 ; [en ligne] <http://www.bnf.fr/documents/bibusages_rapport.pdf> Consulté le 17 juin 2011

BARTLETT Joan, TOMS Elaine. 2005. How is Information Used? Applying task analysis to understanding information use. Actes de la conference ACSI/CAIS, London (Ontario).

BELOT Florence. 2004. Silences et représentations autour du public du patrimoine. Bulletin des bibliothèques de France, n° 5, p. 51-56 [en ligne] <http://bbf.enssib.fr/consulter/bbf-2004-05-0051-009> Consulté le 17 juin 2011

BEQUET Gaëlle, CEDELLE Laure. 2000. Numérisation et patrimoine documentaire. Bulletin des bibliothèques de France, n° 4, p. 67-72 [en ligne] <http://bbf.enssib.fr/consulter/bbf-2000-04-0067-007> Consulté le 17 juin 2011

BERMES Emmanuelle. 2007. Les moteurs de recherche. Bulletin des bibliothèques de France, n° 6, p. 5-10 [en ligne] <http://bbf.enssib.fr/consulter/bbf-2007-06-0005-001> Consulté le 17 juin 2011

BOUBEE Nicole, 2010 Qu'est-ce que rechercher de l'information ? Presses de l'Enssib (collection Papiers).

BRYAN-KINNS Nick, BLANDFORD Ann. 2000. A survey of user studies for digital libraries, RIDL Working Paper, July 2000.

CAZENAVE Jean, DAGORRET Pantxika, MARQUESUZAA Christophe, MAURO Ga. 2004. La revitalisation numérique du patrimoine littéraire territorialisé. Colloque "Le numérique : impact sur le cycle de vie du document", organisé par l'EBSI et l'ENSSIB, Montréal, 13-15 octobre 2004. [en ligne] <http://www.enssib.fr/bibliotheque-numerique/document-1213> Consulté le 17 juin 2011

CHAUDIRON Stéphane, IHADJADENE Madjid. 2010. De la recherche de l'information aux pratiques informationnelles. Études de communication, n°35, 2010. [En ligne] <http://edc.revues.org/index2257.html> Consulté le 16 juillet 2011.

CHEUK Wai-Yi B. 1999. The derivation of a "situational" information seeking and use process model in the workplace: employing sense-making. International Communication Association annual meeting, San Francisco, California, [En ligne], <http://communication.sbs.ohio-state.edu/sense-making/meet/1999/meet99cheuk.html> Consulté le 10 juillet 2011

CLAVIER Viviane. 2010. Indexer des parcours thématiques pour valoriser les collections de presse numérisée, CIDE, 13ème Congrès international sur le document électronique, Paris, 16-17 décembre 2010.

DELAUNAY Else. 1996. La sauvegarde des fonds de journaux : le partenariat des bibliothèques dans la reproduction des collections. Enrichissement et maintenance des fonds, Bulletin d'informations de l'ABF, n° 171, p. 22-25.

DOURY-BONNET Juliette. 2009. Numérisation patrimoniale : initiatives locales ou nationales, privées ou publiques, Bulletin des bibliothèques de France, n° 3, p. 78-78 [en ligne] <http://bbf.enssib.fr/consulter/bbf-2009-03-0078-004> Consulté le 17 juin 2011

FABRE Isabelle, LIQUETE Vincent, GARDIES Cécile. 2010. Pratiques informationnelles et construction des savoirs dans une communauté professionnelle. Revue Les enjeux de l'information et de la communication, supplément 2010B. [En ligne] <http://www.u-grenoble3.fr/les_enjeux> Consulté le 16 juillet 2011

FONDIN Hubert. 2001. La science de l'information : posture épistémologique et spécificité disciplinaire, Documentaliste, science de l'information, vol.38, n°2, 2001, p.112-122.

GARDIES Cécile, Fabre Isabelle, Couzinet Viviane, 2010. « Re-questionner les pratiques informationnelles », Études de communication [En ligne], 35 | 2010, mis en ligne le 01 décembre 2010. [en ligne] <http://edc.revues.org/index2241.html> Consulté le 25 septembre 2011

GUYOT Brigitte. 2002. Mettre en ordre les activités d'information, nouvelle forme de rationalisation organisationnelle, Revue les enjeux de l'information et de la communication, laboratoire Gresec, Université Stendhal, Grenoble. [En ligne] <http://w3.u-grenoble3.fr/les_enjeux/2002/Guyot/index.php> Consulté le 10 juillet 2011

IHADJADENE Madjid. 2009. La dimension humaine de la recherche d'information : pour une épistémologie des pratiques informationnelles. Habilitation à diriger des recherches en Sciences de l'information et de la communication. Université Paris Ouest Nanterre La Défense, 284 p.

JARVELIN Kalervo, INGWERSEN Peter. 2004. Information seeking research needs extension towards tasks and technology. Information Research, 101 paper 212. [en ligne] <http://InformationR.net/ir/10-1/paper212.html> Consulté le 17 juin 2011

LECLERCQ Natacha. 2011. Valorisation du patrimoine numérisé des bibliothèques françaises sur les réseaux sociaux. Mémoire d'étude DCB, enssib, 86 p. [en ligne] <http://www.enssib.fr/bibliotheque-numerique/document-49077> Consulté le 03 octobre 2011

LESQUINS Noémie, TESNIERE Valérie. 2006. La bibliothèque numérique européenne, Bulletin des bibliothèques de France, n° 3, p. 68-80 [en ligne] <http://bbf.enssib.fr/consulter/bbf-2006-03-0068-012 > Consulté le 17 juin 2011

LESQUINS Noémie. 2007. Europeana : rapport de bilan sur les usages et attentes des utilisateurs, Bibliothèque nationale de France, direction des Services

et des réseaux, département de la Bibliothèque numérique [en ligne] <http://www.bnf.fr/documents/europeana_2007.pdf > Consulté le 17 juin 2011

LUPOVICI Catherine, CLOAREC Thierry, CHARENTENAY France de. 2003. Les usages de Gallica , Bulletin des bibliothèques de France, n° 4, p. 40-44 [en ligne] <http://bbf.enssib.fr/consulter/bbf-2003-04-0040-007> Consulté le 17 juin 2011

MIRANDA Silvania.V. Arapanoff Kira.M.A. 2007. Information needs and information competencies: a case study of the off-site supervision of financial institutions in Brazil . Information Research, 132 paper 344 [En ligne] <http://InformationR.net/ir/13-2/paper344.html> Consulté le 16 juillet 2011

PAGANELLI Céline, MOUNIER Evelyne. 2003. Information retrieval in Technical documents: from the User's Query to the Information-Unit Tagging. Proceedings of ACM Sigdoc, San Francisco, Octobre 2003.

PALERMITI Rosalba, Polity Yolla. 2002. « Dynamiques de l'institutionnalisation sociale et cognitive des sciences de l'information en France », In Les origines des sciences de l'information et de la communication en France, regards croisés, sous la direction de R. Boure éd., LILLE, Presses universitaires du Septentrion, 2002, 182p.

PAPY Fabrice (dir). 2007. Usages et pratiques dans les bibliothèques numériques. Hermès Science, 364 pages.

POISSENOT Claude. 2002. De l'objet au point de vue : les bibliothèques entre sciences de l'information et sociologie. Recherches récentes en sciences de l'information : convergences et dynamiques, Colloque MICS/LERASS, Toulouse, 2002.

STAII Adrian, BALICCO Laurence, BERTIER Marc, CLAVIER Viviane, MOUNIER Evelyne, PAGANELLI Céline. 2006. Les pratiques informationnelles des médecins dans les centres hospitaliers universitaires : au croisement de la logique scientifique et de la culture professionnelle, Revue canadienne des sciences de l'information et de bibliothéconomie, vol. 30, n°1/2, p. 69-90, mars-juin 2006.

SCHOR Ralph, PEREZ Matthieu. 2008. Lire la presse ancienne à travers le logiciel d'analyse morphologique PhPress, Semen [En ligne] <http://semen.revues.org/8246> Consulté le 17 juin 2011

SMOLCZEWSKA-TONA Agnieszka, LALLICH-BOIDIN Geneviève. 2008. De l'édition traditionnelle à l'édition numérique : le cas de la presse du XIXe In Traitements et pratiques documentaires : vers un changement de paradigme ? Actes de la deuxième conférence Document numérique et société, Paris : ADBS Éditions, 2008, p. 302-303.

TETU Jean-François. 2010. L'illustration de la presse au XIXème siècle, Semen, 25 | 2008, mis en ligne le 09 juin 2010 [en ligne] <http://semen.revues.org/8227> Consulté le 11 octobre 2001

WESTEEL Isabelle. 2009. Le patrimoine passe au numérique, Bulletin des bibliothèques de France, n°1, p. 28-35 [en ligne] <http://bbf.enssib.fr/consulter/bbf-2009-01-0028-003> Consulté le 17 juin 2011