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

ISBN : 978-2-909285-49-9

Interactions & Usages autour du Document Electronique

Actes du onzième colloque international sur le document électronique

Rouen - France du 28 au 30 octobre 2008

TABLE DES MATIERES

Session 4
Usages et Référentiels

Conférences invitees

Conférences CIDE-CIFED

Atelier FractureNumérique

PREFACE

La onzième édition du Colloque International sur le Document Electronique (Cide) s'inscrit dans la Semaine Rouennaise sur le Document Numérique (Srdn) qui accueille également le Colloque International Francophone sur L'Ecrit et le Document (Cifed) ainsi que divers ateliers et tutoriels.

Le thème de Cide'11 s'attache à la notion de « consommation » de documents qui introduit un point de vue différent de celui de production qui n'est symétrique qu'à première vue. Cide'11 veut insister sur l'interdisciplinarité propre à Cide. Car « *à quoi serviraient tous les savoirs parcellaires sinon à être confrontés pour former une configuration répondant à nos attentes, à nos besoins et à nos interrogations cognitives ?* (E.Morin) » Cide'11 a souhaité inviter ses contributeurs à une réflexion sur les usages en considérant que le document numérique est dynamique (sinon actif) pour les échanges au sein de pratiques culturelles et sociales qui contraignent autant sa production que l'interprétation de ses contenus. La question de l'usage conduit notamment à :

- Interroger le cadre culturel dans lequel se déploie le document numérique et dans lequel il fait sens pour les besoins des pratiques en cours ;
- Approfondir les réflexions sur les cyber-structures qui hébergent le document électronique ;
- Problématiser l'élaboration de référentiels.

Ces thématiques visent à poursuivre les réflexions de Cide'10 sur les défis à relever pour un réel partage de l'information numérique, ce qui est la condition de l'appropriation. Cide'11 s'adresse donc à une communauté large et entend aussi renforcer les liens entre les Sciences et Technologies de l'Information et de la Communication avec les Sciences Humaines et Sociales : informatique, linguistique, psychologie, philosophie, sociologie, sciences de l'information, sciences cognitives, droit, sciences politiques...

C'est ainsi que Cide'11 propose 4 sessions de communications orales :

- Analyse Textuelle et Interaction avec les Documents ;
- Construction des Savoirs et Interprétation de Documents ;
- Usages et Référentiel**s ;**
- Recherche d'information et Annotations.

Cide'11 invite également sept conférenciers dont trois en session pleinière avec Cifed.

Cet ouvrage rassemble tous les articles de ces communications orales et des conférences invitées.

Nous tenons à remercier tous les membres du comité de programme de leur collaboration qui contribue à la qualité de ce colloque ainsi que les membres du comité d'organisation et de coordination locale et tout spécialement Patricia Gautier de l'Inist pour la préparation de ces actes. Nous remercions l'Université de Rouen de nous accueillir dans ses locaux de la Maison de l'Université. Nous remercions le laboratoire LiDiFra, le laboratoire Litis, l'Inist, l'Insa de Rouen, la Mairie de Rouen, le Conseil Régional de Haute-Normandie et l'Université de Rouen pour leur soutien.

Maryvonne HOLZEM & Eric TRUPIN
Co-présidents de Cide'11

Présidence
Holzem M. - LiDiFra Rouen France
Trupin E. – Litis Rouen France

Comité de Programme
Belaid A. – Loria Nancy France
Beust P. – Greyc Caen France
Boismenu G. – Montréal Canada
Bourdon JL. – Cergy Pontoise France
Boyer A. – Loria Nancy France
Ducloy J. – DRRT Lorraine Nancy France
Ferrari S. – Greyc Caen France
Gapenne O. - CosTech Compiègne France
Jacquet D. – Modesco Caen France
Kanellos I. – Enst Brest France
Labiche J. – Litis Rouen France
Lainé-Cruzel S. – Lyon 3 France
Lamiroy B. – Loria Nancy France
Larouk O. – Enssib Lyon France
Le Crosnier H. – Greyc Caen France
Legallois D. – Crisco Caen France
Kotowicz JP. – Litis Rouen France
Madelaine J. – Greyc Caen France
Mojahid M. – Irit Toulouse France
Mourad G. – LaLICC Paris 4 France
Prié Y. – Liris Lyon France
Raysz JP. – Jouve Mayenne France
Rousseaux F. – Ircam France
Royauté J. – LIF Marseille France
Saidali Y. – Litis Rouen France
Tazi S. – Laas Toulouse France
Toussaint Y. – Loria Nancy France
Zacklad M. – Tech-CICO Troyes France
Zreik K. – Paris 8 France

Secrétariat
CIDE 2008 - Laboratoire LITIS
Faculté des Sciences - Université de Rouen
76800 Saint-Etienne du Rouvray - FRANCE

Organisation
INIST Nancy
LITIS & LiDiFra Rouen

Session 1

Recherche d'information et Annotations

Fusion de résultats en recherche d'information : Mesure de l'impact de l'union et de l'intersection de résultats

Data fusion in information retrieval: measuring the impact of fusing by union and intersection

Désiré Kompaoré(1), Josiane Mothe(1)

(1)Institut de Recherche en Informatique de Toulouse
Université de Toulouse
kompaore @irit.fr
mothe] @irit.fr

Résumé. Cet article présente une étude que nous avons menée sur la fusion de données en recherche d'information. Nous proposons deux stratégies de fusion "systématique" des systèmes, en nous basant sur des techniques simples de fusion par union et intersection. L'objectif de ce travail est de quantifier les améliorations en termes de rappel et de précision attendues par ces techniques. Les résultats sur la collection de TREC novelty 2002 (resp. 2003) montrent que la fusion par union améliore le rappel du meilleur système de 48% (resp. 18%) si l'on considère la meilleure combinaison, et en moyenne de 23% (resp. 6,5%) si l'on considère la combinaison avec les 9 autres meilleurs systèmes. De la même façon, la fusion par intersection améliore la précision du meilleur système de 15% (resp. 9%) si l'on considère la meilleure combinaison, et en moyenne de 10% (resp. 5,5%) si l'on considère la combinaison avec les 9 autres meilleurs systèmes. La tendance de ces résultats est identique lorsque plus de systèmes sont fusionnés deux à deux.
Mots-clés recherche d'information, fusion de données, performance, évaluation, TREC.

Abstract. The focus of this paper is data fusion in information retrieval. We investigate the effect of two simple fusion algorithms: union and intersection. Union improves recall, and intersection improves precision. The goal of this paper is to quantify the improvement through a set of experiments. We show that, considering TREC 2002 (resp. 2003) novelty benchmark collection, fusing using union improve the recall of the best system by 48% (resp. 18%) considering the best fusion and in average by 23% (resp. 6,5%) when considering the 9 following best systems. In the same way, when considering fusing by intersection, precision of the best system is improve by 15% (resp. 9%) when considering the best fusion and in average by 10% (resp. 5,5%) when considering the 9 next best systems. The trend is similar when considering more systems and fusing the results 2 by 2.
Keywords. information retrieval, data fusion, performance, evaluation, TREC.

1 Introduction

De nombreux paramètres peuvent influencer les résultats qu'obtiennent les systèmes de recherche d'information : la méthode d'indexation des documents utilisée, le traitement de la requête, le modèle de recherche sous-jacent et la fonction d'ordonnancement des résultats. Les travaux dans le domaine de la recherche d'information visent donc à proposer des améliorations dans l'un ou l'autre des maillons de la chaine utilisée. D'autres travaux, partant du postulat que différents systèmes retrouvent les documents dans un autre ordre ou retrouve différents documents, visent à combiner ces systèmes. C'est dans ce dernier cadre que s'inscrivent nos travaux.

Les méthodes de fusion de résultats de systèmes [data fusion] proposées dans la littérature s'appuient sur la prise en compte de la similarité entre la requête et les documents, c'est-à-dire le degré de pertinence des documents restitués par les systèmes. Cependant, cette information est rarement disponible. Les moteurs du web par exemple ne la fournissent pas en même temps que les documents. Nous nous sommes donc intéressés aux méthodes qui ne nécessitent pas ce type d'information, de sorte que la méthode soit applicable quels que soient les systèmes à fusionner utilisés. Certaines autres techniques de fusion se basent sur le rang des documents retrouvés ; il s'agit là d'un moyen de contourner le manque d'information sur la similarité entre les documents retrouvés et la requête. Dans ces méthodes, l'objectif principal est de mieux satisfaire l'utilisateur en lui restituant d'abord (c'est-à-dire en haut de la liste) les documents pertinents. Pourtant, dans certaines activités, l'objectif de l'utilisateur n'est pas d'obtenir quelques documents pertinents, mais de bien collecter un ensemble de documents répondant à un besoin et qui seront par la suite analysés (Dousset et Mothe, 2004). Les activités de veille scientifique et technologique répondent par exemple à ce cadre. Dans ce contexte, il est donc important de pouvoir s'assurer que l'ensemble des documents retrouvés contient un maximum de documents pertinents (rappel élevé, faible silence documentaire) et peu de documents non pertinents (précision élevée et faible bruit documentaire). Une méthode simple permettant d'augmenter le rappel consiste à fusionner les résultats obtenus en effectuant une union. Par cette méthode, le rappel ne peut être qu'augmenté (ou maintenu au même niveau si les deux systèmes retrouvent les mêmes documents pertinents). De la même façon, une méthode simple pouvant permettre d'augmenter la précision consiste à considérer l'intersection des ensembles retrouvés : un document retrouvé par les deux systèmes a plus de chances d'être pertinent. Il est cependant bien connu qu'appliquer une méthode qui augmente le rappel dégrade généralement la précision et inversement.

Dans cet article, nous nous intéressons à mesurer ces effets.

2 Travaux reliés

(Fox et Shaw, 1994) ont proposé des fonctions de fusion de résultats de plusieurs systèmes basées sur une combinaison linéaire des scores des documents. Parmi les formules proposées, la formule CombSUM calcule la somme des scores de tous les documents retournés par les SRI. La formule CombMNZ prend en compte le nombre de systèmes qui ont retrouvé le même document et multiplie la valeur de CombSUM par ce nombre. Les auteurs ont montré que la combinaison de plusieurs techniques de recherche augmente l'efficacité globale de la recherche. D'après leurs conclusions, la formule Comb-SUM appliquée à la collection TREC-2 apporte des améliorations de la R-Précision (précision lorsque R documents sont

retrouvés, R étant le nombre de documents effectivement pertinents) de l'ordre de 13 %. Ces techniques de fusion ont également été utilisées avec succès par (Lee, 1997) qui a montré que CombMNZ permet d'obtenir de meilleures performances que CombSUM et produit de bons résultats dans le cas où le taux de chevauchement des documents pertinents est élevé (entre 0,75 et 0,82) et le taux de chevauchement des documents non pertinents bas (entre 0,30 et 0,40). (Beitzel et al., 2004) ont contredit ce résultat en montrant que l'amélioration n'est pas tant liée au taux de chevauchement qu'au nombre de documents pertinents qui n'apparaissent que dans un résultat de recherche. Dans (Vogt et Cottrell, 1998), les auteurs proposent d'utiliser un modèle linéaire pour combiner les différents scores obtenus. Les résultats de leurs travaux montrent que cette méthode est seulement efficace lorsque le taux de chevauchement des documents pertinents est très élevé et le taux de chevauchement des documents non pertinents est faible. Dans le cas où le score de similarité entre le document et la requête n'est pas disponible, il existe d'autres techniques qui se basent par exemple sur le rang des documents dans la fusion pour améliorer la recherche. (Voorhees et al., 1994) propose une technique simple de fusion, basée sur les rangs des documents, qui consiste à choisir les documents classés en première position dans les différentes listes retournées par les SRI, puis les documents classés en deuxième position, et ainsi de suite, après avoir supprimé les doublons. (Lee, 1997) utilise le rang des documents comme alternative aux scores de similarité, et il obtient de bons résultats en termes de précision moyenne. (Soboroff et al. 2001) se sont intéressés à la comparaison entre l'utilisation des scores de similarité et les rangs des documents, lors de la RI. Ils ont combiné les sous-listes des différents SRI en remplaçant les scores de similarité par une mesure qui prend en compte les rangs des documents Les résultats obtenus montrent que l'utilisation des rangs est plus performante que l'utilisation des scores de similarité. (Farah et Vanderpooten, 2007) propose une technique d'agrégation des rangs prenant en compte les documents ayant le même rang dans les listes de documents. Les expérimentations qu'ils ont effectuées montrent que leur méthode permet d'obtenir de meilleurs résultats que CombSUM et CombMNZ. Une des conclusions est que la fusion doit être effectuée sur les listes restituées par les meilleurs systèmes. (Lillis et al., 2006) utilise une approche de fusion probabiliste basée sur les performances passées des systèmes, pour un ensemble de requêtes de test. Les résultats qu'ils obtiennent sont supérieurs à ceux obtenus avec CombSUM. (Wu and McClean, 2006) analysent le comportement des méthodes de fusion les plus répandues (CombSUM et CombMNZ) et montrent qu'il est possible de prédire les performances des méthodes de fusion ; leurs expérimentations se basent sur les collections de TREC. (Spoerri, 2007) analysent en détail deux effets liés à la fusion de données: l'autorité (plus il y a de systèmes qui retrouvent un même document, plus le document est potentiellement pertinent) et l'effet du rang (plus un document est retrouvé haut dans les listes, plus il est potentiellement pertinent). En utilisant les données de TREC ils montrent que ces phénomènes se retrouvent quelque soit le nombre de documents retrouvés considérés, mais que si les systèmes retrouvent un grand nombre de documents, alors l'effet de rang ne débute que lorsque suffisamment de système ont retrouvé un document et/ou si le nombre de documents considéré augmente.

3 Données étudiées

3.1 Tâche TREC considérée, requêtes et caractéristiques des collections

Les méthodes de fusion sont évaluées sur les collections de TREC 2002 et 2003 utilisées dans la sous-tâche détection de passages de la tâche "nouveauté". Le choix de cette collection est motivé par le fait que cette tâche vise à sélectionner les documents ou parties de documents intéressantes, sans se soucier de l'ordre dans lequel ils sont restitués. Cela correspond à notre cadre d'étude.

La tâche détection de la nouveauté a été introduite en 2002 lors de la campagne TREC-11. Etant donné une requête et une liste ordonnée de documents pertinents, l'objectif de cette tâche est de retrouver les passages (phrases) pertinents et nouveaux répondant à la requête. Nous avons utilisé dans nos travaux les documents issus de la première sous-tâche qui consiste à détecter les passages pertinents dans les documents. La collection de départ est constituée de 50 requêtes provenant des campagnes adhoc TREC6, TREC7, et TREC8. Ces 50 requêtes ont été sélectionnées parmi celles pour lesquelles les jugements de pertinence comprenaient entre 10 et 70 documents pertinents. En 2002, TREC a choisi de sélectionner 50 requêtes parmi les besoins d'information identifiés par les numéros 300 à 450. Le NIST a sélectionné les documents effectivement pertinents pour chacun de ces besoins d'information (jugements de pertinence), avec un maximum de 25 documents par besoin, et les a fournis aux participants. Un exemple de requête est décrit dans la figure 1.

Une requête est composée d'un numéro qui identifie la requête, un titre (T) qui donne une description du sujet de la requête en quelques mots, une description (D) de la requête exprimée à travers une phrase et une partie narrative (N) qui explique plus en détail le type de documents pertinents et non pertinents qui est recherché.

Dans une seconde étape, des juges indiquent quelles phrases de ces documents sont effectivement pertinentes. Le même type de principe a été utilisé en 2003, avec 50 besoins. Les caractéristiques de ces collections sont fournies dans le tableau 1.

Topic : 35

Title : NATO, Poland, Czech Republic, Hungary

Descriptive : Accession of new NATO members : Poland, Czech Republic, Hungary, in 1999.

Narrative : Identity of current and newly-invited members, statements of support for and opposition to NATO enlargement and steps in the accession process and related special events are relevant. Impact on the new members, i.e., requirements they must satisfy, and their expectations regarding the implications for them are relevant. Progress in the ratification process is relevant. Future plans for NATO expansion, identification of nations admitted on previous occasions, and comments on future NATO structure or strategy are not relevant.

Figure 1 : *Exemple de requête TREC*

	TREC2002	TREC2003
Nombre de besoins d'information	49	50
Nombre moyen de documents pertinents par besoin	22,3	25
Nbre moyen de phrases par besoin	1321	796,4
Nombre moyen de phrases pertinentes par besoin	27,9	311,14
%moyen de phrases pertinentes	2,1	39,1

Tableau 1 *: Caractéristiques des collections TREC 2002 et 2003*

3.2 Exécutions et évaluation

Chaque participant propose la liste des éléments que le système considère comme pertinent. Un même système, paramétré différemment peut être utilisé. Dans ce cas, deux exécutions seront soumises et évaluées.

Trec_eval est généralement utilisé pour évaluer une exécution. Ce logiciel permet de calculer plus d'une centaine de mesures. Dans cet article, nous nous sommes focalisés sur les mesures en lien avec notre étude : le rappel, la précision et la mesure F. Le rappel exact mesure la proportion de documents pertinents effectivement retrouvés, en moyenne sur l'ensemble des requêtes considérées. La précision exacte (non interpolée) mesure la précision moyenne sur l'ensemble des documents retrouvés. Enfin, la mesure F combine les deux mesures précédentes (qui varient en sens inverse). Elle est calculée par la formule (2*RP/(R+P)) où R (resp. P.) est le rappel (resp. précision) exact ; faisant ainsi jouer la même importance au rappel et à la précision.

4 Etude préliminaire

L'objectif de cette analyse préalable est de détailler les caractéristiques des collections et des systèmes que nous utilisons dans les expérimentations. La figure 2 montre la distribution des performances des systèmes en termes de rappel pour les collections 2002 et 2003. Nous utilisons une représentation des données sous forme de boîte à moustaches (Tukey, 1977). Ce type de représentation permet de représenter une distribution de données. Elle utilise 5 valeurs qui résument des données : le minimum, les 3 quartiles Q1, Q2 (médiane), Q3, et le maximum. Les quartiles jouent un rôle important dans l'interprétation. La médiane Q2 divise la série en deux groupes d'effectif égaux. Le premier quartile divise le groupe de données situé en dessous de la médiane en deux groupes d'effectif égaux. Le troisième quartile divise quant à lui le groupe situé au delà de la médiane en deux groupes de taille égale. Les quartiles représentent 25 (Q1), 50 (Q2) et 75% (Q3) des données analysées.

Dans la figure 2, la première boite à moustaches décrit la répartition des mesures de rappel pour la campagne de 2002. Les valeurs de rappel qui sont présentées correspondent à la moyenne que chaque système obtient sur l'ensemble des requêtes pour lesquelles le système est évalué. En 2002 la valeur maximale de rappel est 0,597 (et de 0,999en 2003) et 75% des valeurs de rappel sont inférieures à 0,4. En 2003, plus de la moitié des systèmes obtiennent une valeur de rappel supérieure à 0,42%.

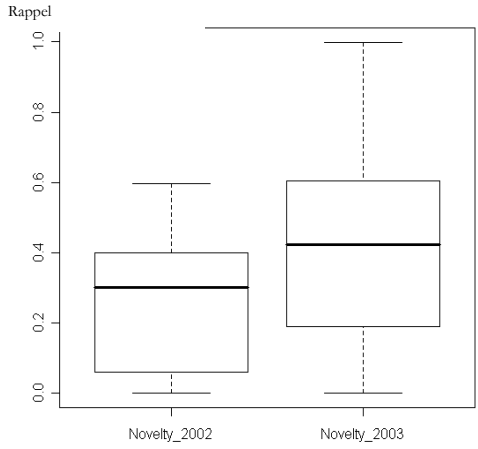


Figure 2 : *Répartition du rappel pour les collections TREC-Nouveauté 2002 et 2003.*

La figure 3 complète l'interprétation des boites à moustaches. Elle représente les variations entre le système qui obtient la plus grande valeur de rappel et les autres systèmes. Dans cette figure, les différences de rappel entre deux systèmes de rangs consécutifs ne sont pas grandes. Un certain nombre de paliers sont toutefois à remarquer aussi bien pour 2002 que pour 2003. Par exemple, en 2003, une grande différence de rappel existe entre le premier système (ISIALL03 0,999) et le deuxième système (MeijiHilF13 0,84). De la même façon, en 2003, une variation importante (0,066) du rappel est observée entre le cinquième et le sixième système en 2003 ; idem pour 2002 avec une variation entre le cinquième et sixième système de 0,049.

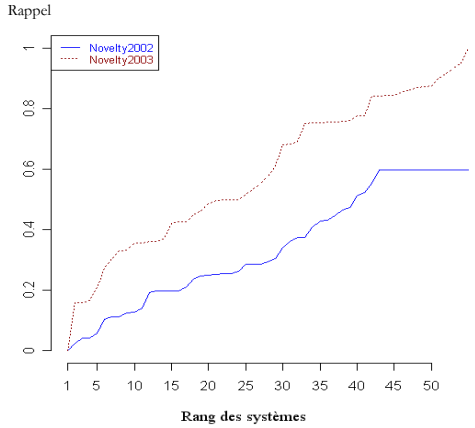


Figure 3 *: Variation entre le rappel du meilleur système et les autres systèmes*

Concernant la précision (Figure 4), quelques systèmes obtiennent des valeurs de précision en dessous de 0,5 en 2003. Par cela, ils se distinguent des autres systèmes. Il s'agit des systèmes suivants : lexiclone03 (0,484), ISIALL03 (0,411), ISIRAND03 (0,41), umbcrun2 (0,405), umbcrun3 (0,4), umbcrun1 (0,396).

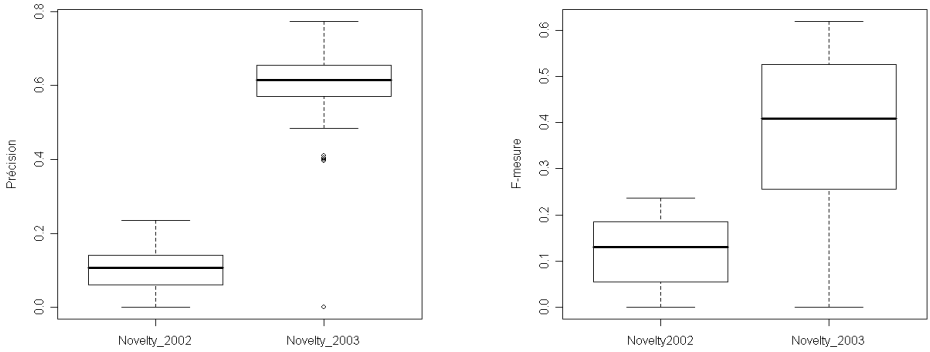


Figure 4. - *Répartition des valeurs initiales de précision et de F-mesure*

Le système ISINONE03 obtient quant à lui une valeur de précision nulle en 2003. Les performances des systèmes en 2003 sont très élevées par rapport à 2002 comme on peut le constater dans la figure 4 pour la précision et la mesure F.

La figure 5 montre la différence entre la précision du meilleur système et celle des autres systèmes ; cette différence est inférieure à 0,2 pour les 40 premiers systèmes de 2002 et 2003. Les plus grandes variations sont dues aux systèmes évoqués dans le paragraphe précédent et qui ont des valeurs de précision bien plus bases que les autres systèmes. De plus, les différences entre les performances des meilleurs systèmes et ceux qui obtiennent de mauvais résultats sont grandes. Par exemple la différence entre le meilleur et le dernier système en termes de rappel en 2003 est de 0,999.

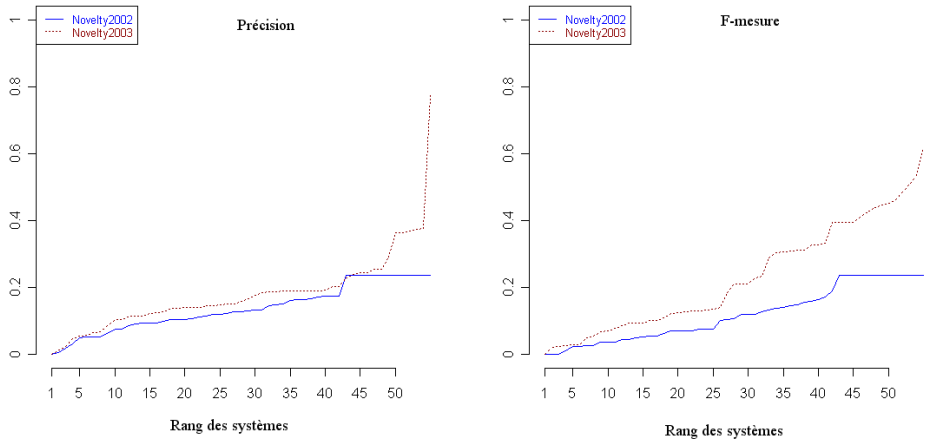


Figure 5 : *Variation entre la précision et la mesure F du meilleur système et les autres systèmes*

Cette analyse préalable indique que les performances moyenne entre les meilleurs systèmes sont proches, indépendamment des stratégies de recherche utilisées par les systèmes. Elle indique également que certains systèmes ont des performances très faibles. Ainsi, dans la suite des expérimentations, nous nous sommes focalisés sur les meilleurs systèmes (les 10 premiers en termes de la mesure combinée mesure F).

5 Fusion par union et intersection

L'objectif de cette étude est de mesurer l'amélioration en termes de précision que peut apporter la fusion par intersection, ainsi que celle en termes de rappel que peut amener la fusion par union.

5.1 Base de la comparaison

Les résultats obtenus par les techniques de fusion sont comparés à ceux obtenus par les systèmes simples. Le tableau 2 indique les performances obtenues par les 10 meilleurs systèmes.

Ainsi, en 2002, le système ayant obtenu les meilleures performances est le système thunv3 avec 0,237 (mesure F) avec un rappel de 0,404 et une précision de 0,204. En 2003, le meilleur système obtient une mesure F de 0,619 (rappel 0,792 et précision 0,597).

	Rappel	**Précision**	**Mesure F**
Thunv3	0,404	0,204	0,237
Thunv1	0,341	0,229	0,236
Thunv2	0,341	0,236	0,236
Thunv4	0,335	0,216	0,226
CIIR02tfkl	0,556	0,141	0,213
CIIR02tfnew	0,556	0,141	0,213
pircs2N01	0,486	0,161	0,211
pircs2N02	0,486	0,161	0,211
pircs2N03	0,4	0,184	0,199
pircs2N04	0,4	0,184	0,199

2002

	Rappel	**Précision**	**Mesure F**
THUIRnv0315	0,792	0,597	0,619
ISIDSCm203	0,832	0,534	0,597
Ulowa03Nov01	0,696	0,636	0,594
THUIRnv0.11	0,726	0,606	0,593
MeijiHilF13	0,84	0,52	0,589
MeijiHilF14	0,84	0,52	0,589
Ulowa03Nov02	0,637	0,649	0,568
THUIRnv0312	0,665	0,624	0,564
THUIRnv0313	0,637	0,633	0,552
THUIRnv0314	0,628	0,632	0,548

2003

Tableau 2 : *Performances initiales des 10 meilleurs systèmes*

5.2 Principe de fusion

Les méthodes de fusion que nous étudions se basent sur les principes simples d'intersection et d'union des ensembles de documents retrouvés (Kompaore et al., 2006).

L'union regroupe les documents sélectionnés par les systèmes fusionnés, après suppression des éventuels doublons. Le principe de l'union est celui de la théorie des ensembles appliqué aux ensembles de documents retrouvés. L'intersection

quant à elle permet de regrouper les documents restitués en commun par les systèmes fusionnés.

5.3 Fusion par union

Dans cette section, nous étudions l'impact de la fusion par union des résultats des systèmes. Comme nous l'avons précisé précédemment, ce type de fusion a pour objectif d'améliorer le rappel. Nous indiquons toutefois les performances en termes de mesure F afin de mieux pouvoir comparer globalement les performances.

Le tableau 3 indique les variations mesurées. Les chiffres en gras marquent une amélioration par rapport aux résultats du meilleur système. Les chiffres en italique indiquent au contraire une dégradation. Les fusions sont ordonnées par rappel décroissant.

Le tableau 3 montre que pour les données de 2002, la fusion par union de thunv1 et thunv3 n'apporte aucune amélioration par rapport au rappel initial de thunv3. Ce tableau nous indique que la fusion par union de thunv3 avec les autres versions du même système (thunv1, thunv2, et thunv4) ne modifie pas les performances obtenues par thunv3 pour les mesures utilisées. En analysant les données, on constate que les réponses des 4 versions du système thunv constituent en fait des sous ensembles des réponses du système thunv3. Concernant la fusion de thunv3 avec les autres meilleurs systèmes, le rappel est amélioré. Cette amélioration du rappel peut aller jusqu'à presque 50%, montrant ainsi que les documents restitués par l'un et l'autre des systèmes fusionnés sont différents. La fusion de ces deux systèmes permet d'obtenir un rappel supérieur à ce qu'il aurait été si les systèmes avaient été considérés individuellement (Chacun de ces deux systèmes pris individuellement : Ciir02tfkl obtenait seul un rappel de 0,556 contre 0,597 pour la fusion (soit une augmentation de 7%) et Thunv3 obtenait seul un rappel de 0,404 (soit une augmentation de 48% pour la fusion).

Systèmes fusionnés	Rappel	Mesure F
thunv3	-0,404	-0,237
Ciir02tfkl-thunv3	**0,597 (+48%)**	*0,212 (-11%)*
Ciir02tfnew-thunv3	**0,597 (+48%)**	*0,212*
pircs2n01-thunv3	**0,543 (+34%)**	*0,211 (-11%)*
pircs2n02-thunv3	**0,543 (+34%)**	*0,211*
pircs2n03-thunv3	**0,496 (+23%)**	*0,216 (-9%)*
pircs2n04-thunv3	**0,496 (+23%)**	*0,216*
Thunv1-thunv3	0,404	0,237
Thunv2-thunv3	0,404	0,237

Systèmes fusionnés	Rappel	Mesure F
thuirnv0315	-0,792	-0,619
meijihilf13-thuirnv0315	**0,938 (+18%)**	*0,613 (-1%)*
meijihilf14-thuirnv0315	**0,938 (+18%)**	*0,613*
isidscm203-thuirnv0315	**0,921 (+16%)**	*0,618*
thuirnv0311-thuirnv0315	**0,82 (+4%)**	*0,623 (-1%)*
thuirnv0312-thuirnv0315	**0,804 (+2%)**	*0,623*
thuirnv0315-uiowa03nov01	**0,797 (+1%)**	**0,62**
thuirnv0315-uiowa03nov02	**0,793 (+0,13%)**	0,619
thuirnv0313-thuirnv0315	0,792	0,619

Thunv3-thunv4	0,404	0,237		thuirnv0314-thuirnv0315	0,792	0,619
	2002				2003	

Tableau 3 : *Performance des meilleurs systèmes fusionnés par union avec le meilleur système*

Les mêmes conclusions peuvent être faites en 2003 où par exemple, le système thuirnv0315 seul obtenait un rappel de 0,792 contre 0,938 pour la fusion avec meijihilf13 (soit une augmentation de 18%) ; le système meijihilf13 seul obtenait 0,84 (soit une augmentation de 12% pour la fusion). Ainsi, dans la mesure où l'on connaît le meilleur système, il est intéressant de le fusionner avec n'importe quel autre système performant. Il est cependant important de noter que cette information n'est connue qu'*a posteriori*. La section 5 présente donc une étude plus globale des résultats de fusion par union.

5.4 Fusion par intersection

Nous présentons dans cette section les résultats obtenus concernant la fusion par intersection des résultats des meilleurs systèmes. Ce type de fusion a pour objectif d'améliorer la précision. Comme précédemment, nous indiquons toutefois les performances en termes de mesure F ; les résultats sont ordonnés par ordre de précision décroissante ; les chiffres en gras indiquent une amélioration et ceux en italique une dégradation.

Systèmes fusionnés	Précision	Mesure F	Systèmes fusionnés	Précision	Mesure F
thunv3	-0,204	-0,237	thuirnv0315	-0,597	-0,619
thunv2-thunv3	**0,234 (+15%)**	*0,235*	thuirnv0315-uiowa03nov02	**0,65 (+9%)**	*0,568 (-8%)*
pircs2n03-thunv3	**0,232 (+14%)**	*0,221 (-7%)*	thuirnv0315-uiowa03nov01	**0,639 (+7%)**	*0,593 (-4%)*
pircs2n04-thunv3	**0,232 (+14%)**	*0,221 (-7%)*	thuirnv0313-thuirnv0315	**0,636 (+7%)**	*0,552 (-11%)*
thunv1-thunv3	**0,228 (+12%)**	*0,235 (-1%)*	thuirnv0314-thuirnv0315	**0,635 (+6%)**	*0,549 (-11%)*
pircs2n01-thunv3	**0,226 (+11%)**	**0,24 (+1%)**	thuirnv0312-thuirnv0315	**0,632 (+6%)**	*0,559 (-10%)*
pircs2n02-thunv3	**0,226 (+11%)**	**0,24 (+1%)**	thuirnv0311-thuirnv0315	**0,622 (+4%)**	*0,588 (-5%)*
thunv3-thunv4	**0,216 (+6%)**	*0,226 (-5%)*	meijihilf13-thuirnv0315	**0,618 (+4%)**	*0,592 (-4%)*
Ciir02tfkl-thunv3	**0,213 (+4%)**	**0,24 (+1%)**	meijihilf14-thuirnv0315	**0,618 (+4%)**	*0,592 (-4%)*
Ciir02tfnew-thunv3	**0,213 (+4%)**	**0,24 (+1%)**	isidscm203-thuirnv0315	**0,615 (+3%)**	*0,595 (-4%)*
	2002			2003	

Tableau 4 *: Performance des meilleurs systèmes fusionnés par intersection avec le meilleur système*

Le tableau 4 montre en particulier que, en 2002, la fusion pircs2n01-thunv3 améliore la précision de 11% par rapport au système thunv3 (de 0,204 à 0,226). En outre, l'augmentation correspond à 40% par rapport à pircs2n01. De la même façon, en 2003, la fusion isidscm203-thuirnv0315 améliore de 3% la précision par rapport à thuirnv0315 seul (0,615 contre 0,597). L'amélioration de la précision est de 15% par rapport à isidscm203 seul (0,615 contre 0,534).

De façon générale, la fusion par intersection améliore la précision des deux systèmes de façon importante ; même si les pourcentages d'augmentation sont moins importants que dans le cas du rappel.

6 Expérimentations complémentaires

Dans les premières expérimentations que nous avons reportées à la section 4, nous avons choisi les meilleurs systèmes uniquement sur la base de leur performance en termes de mesure F. D'autre part, la fusion a été réalisée en considérant le meilleur système, que nous avons fusionné aux autres. Les expérimentations que nous reportons ici ne sont pas centrées sur les performances du meilleur système. Les 10 systèmes qui sont sélectionnés sont combinés deux à deux, et le résultat de la fusion par union et par intersection est analysé, indépendamment du rang initial des systèmes.

6.1 Choix des « meilleurs » systèmes à fusionner

Si l'on observe les performances des meilleurs systèmes en considérant les différentes mesures de performance pour 2002 et 2003, on remarque que plusieurs versions de différents systèmes obtiennent les meilleures performances en 2002 et en 2003 (tableau 5). Par exemple, trois systèmes et leurs versions obtiennent les 10 plus grandes valeurs de précision en 2003 (il s'agit des versions des systèmes NLPPR03n et clr03n1, ainsi que le système ccsummeoqr). Ainsi, la sélection des meilleurs systèmes est plus proche d'une sélection dans un cadre réel : si les systèmes étaient disponibles, il serait possible de les utiliser sur différentes collections d'apprentissage pour décider des meilleurs systèmes en moyenne. Le fait d'en considérer plusieurs nous affranchi d'une certaine dépendance aux collections. Pour chaque année et chaque mesure, les 10 systèmes sélectionnés sont utilisés pour réaliser la fusion. Nous avons adopté deux stratégies différentes pour la sélection des 10 meilleurs systèmes. Dans la première stratégie (stratégie1), nous utilisons la mesure F comme base de sélection. Dans la stratégie2, les 10 meilleurs systèmes sont sélectionnés en fonction de la mesure visée. Ainsi, les 10 systèmes qui obtiennent le meilleur rappel sont sélectionnés lors de la fusion par union ; ceux qui obtiennent la meilleure précision sont sélectionnés pour la fusion par intersection. La fusion est réalisée en combinant 2 à 2 les résultats obtenus par les systèmes sélectionnés. Par exemple le premier système est fusionné avec les 9 autres systèmes, le deuxième avec 8 systèmes, et ainsi de suite pour l'ensemble des 10 systèmes.

2002

Systèmes	Rappel	Systèmes	Précision	Systèmes	mesureF
nttcslabnvr2	0,597	thunv2	0,236	thunv3	0,237
UIowa02Nov4	0,574	thunv1	0,229	thunv1	0,236
CIIR02tfkl	0,556	thunv4	0,216	thunv2	0,236
CIIR02tfnew	0,556	thunv3	0,204	thunv4	0,226

2003

ISIALL03	0,999	NLPR03n1w3	0,774	THUIRnv0315	0,619
MeijiHilF13	0,84	NLPR03n1w2	0,761	ISIDSCm203	0,597
MeijiHilF14	0,84	NLPR03n1f2	0,751	UIowa03Nov01	0,594
ISIDSCm203	0,832	NLPR03n1f1	0,726	THUIRnv0311	0,593
THUIRnv0315	0,792	clr03n1d	0,718	MeijiHilF13	0,589

Tableau 5 : *Performance des systèmes en 2002 et 2003, ordonnés en fonction des performances (extrait)*

6.2 Comparaison des stratégies pour l'union (rappel)

La figure 6 compare pour chacune des collections utilisées les résultats obtenus en termes de rappel en utilisant les 2 stratégies que nous proposons. Nous retenons pour chaque stratégie les 45 meilleures valeurs de rappel après fusion (2 à 2 en utilisant les 10 meilleurs systèmes) que nous comparons avec le classement des 45 meilleurs systèmes uniques pour le rappel. Dans cette figure, Rappel_i_2002 pour 2002 et Rappel_i_2003 pour 2003 correspond au rappel de ces 45 systèmes avant fusion. Dans la figure 6, les courbes sont décroissantes car les valeurs de rappel sont ordonnées de manière décroissante. Les valeurs en abscisses correspondent au classement de la valeur de rappel considérée par rapport aux autres valeurs de rappel.

Il faut noter que toutes les courbes ne doivent pas être toutes comparées de la même façon. Les courbes notées (1) et (2), en référence à la stratégie utilisée peuvent être directement comparées. Concernant la courbe notée (i), en référence aux systèmes initiaux utilisés de façon non fusionnée, le lecteur portera d'abord son attention sur les 10 premières valeurs, correspondant aux 10 meilleurs systèmes.

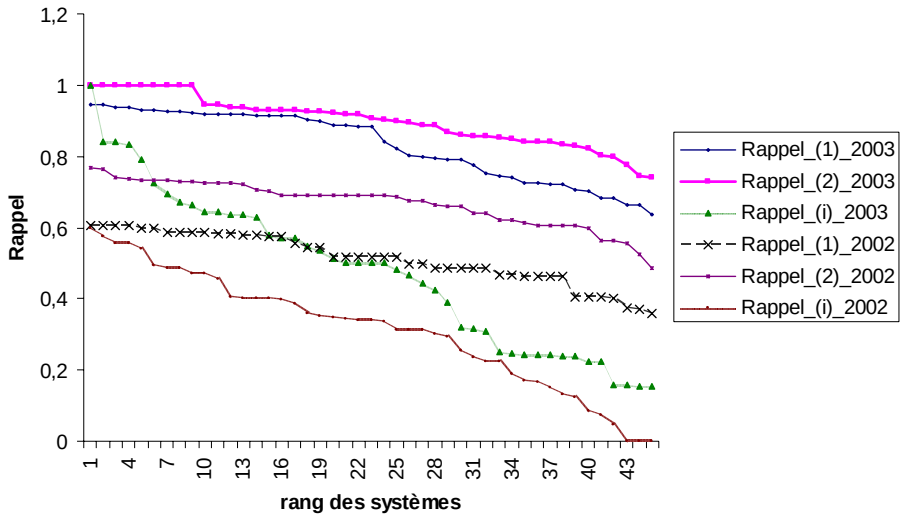


Figure 6 : *Comparaison des mesures de rappel obtenues par chaque stratégie de fusion pour la fusion par union*

La figure 6 montre que les deux stratégies proposées améliorent les performances initiales des systèmes. Les 10 premiers résultats obtenus avec la stratégie2 en 2003 obtiennent un rappel égal à 1.

Le tableau 6 indique la valeur de l'amélioration moyenne obtenue lors des fusions. Dans ce tableau, pour la stratégie 1 (sélection des meilleurs systèmes basée sur la mesure F), nous comparons la moyenne sur l'ensemble des combinaisons 2 à 2 des systèmes (45 couples de systèmes formés par la fusion des 10 systèmes sélectionnés par la stratégie 1) avec les performances du meilleur système (valeur du rappel du système étant le meilleur par rapport à la mesure F) et avec la moyenne du rappel obtenu par les systèmes utilisés isolément. La moyenne des systèmes utilisés isolément revient à calculer le rappel moyen obtenu par les 10 systèmes sélectionnés par la stratégie1. Dans ce même tableau, les performances de la stratégie2, qui sélectionne les meilleurs systèmes en termes de rappel sont comparées d'une part avec celles du meilleur système (sélectionné de la même façon) et avec celles obtenues en moyenne par les systèmes sélectionnés pris séparément. Les pourcentages indiquent l'amélioration obtenue par la moyenne des fusion 2 à 2, comparativement à meilleur système simple et par rapport à la moyenne des systèmes pris séparément.

		2002	2003
Stratégie 1	Meilleur système simple	0,404	0,792
	Moyenne systèmes simples	0,430	0,729
	Moyenne fusion 2 à 2	0,513 (27%-19,2%)	0,830 (4,9%-13,9%)
Stratégie 2	Meilleur système simple	0,597	0,999
	Moyenne systèmes simples	0,523	0,770
	Moyenne fusion 2 à 2	0,668 (11,9%-27,8%)	0,900 (0,1%-16,8%)

Tableau 6: *Valeurs de rappel moyen pour la fusion par union des systèmes 2 à 2*

Dans le tableau 6, pour la stratégie1, on remarque que le rappel moyen obtenu par les systèmes simples (0,4305) est supérieur au rappel du meilleur système sélectionné par la stratégie1 (0,404). Cela s'explique par le fait que le système détecté comme étant le meilleur avec la stratégie1, par rapport à la mesure F (Thunv3) est classé en 12ème position par rapport au rappel des autres systèmes. Dans ce cas, en appliquant la fusion par union sur les systèmes sélectionnés avec la stratégie1, on obtient une amélioration du rappel moyen d'environ 19% par rapport au rappel moyen des systèmes simple. En 2003, le rappel moyen des systèmes simples est inférieur au rappel du meilleur système avec la stratégie1 (ce système est classé en 5ème position par rapport au rappel des autres systèmes).

D'autre part, le tableau 6 montre bien que quelque soit l'année et la stratégie de fusion, en moyenne, les fusions améliorent les résultats par rapport à la moyenne des systèmes utilisés séparément. Par exemple, en utilisant la stratégie 1, pour l'année 2002, alors qu'en moyenne les meilleurs systèmes obtiennent 0,430 ; la fusion permet d'obtenir 0,531, soit une augmentation d'environ 19%. Bien évidemment, la stratégie 2 permet d'obtenir globalement de meilleurs résultats que la stratégie 1 (puisque les meilleurs systèmes sont choisis pour leur maximum de performance par rapport au rappel, mesure étudiée). L'augmentation relative est cependant plus importante. Par exemple, toujours pour l'année 2002, en utilisant la

stratégie 2, alors qu'en moyenne les meilleurs systèmes obtiennent 0,523 ; la fusion permet d'obtenir 0,668, soit une augmentation d'environ 28% environ au lieu de 19%. De la même façon, pour l'année 2003, l'augmentation par rapport à la moyenne est de 14% pour la stratégie 1 alors qu'elle est de 17% environ pour la stratégie 2.

Pour la stratégie2, le meilleur système obtient un rappel supérieur au rappel moyen des systèmes simples. On constate alors en 2003 que le rappel moyen à l'issue de la fusion des 10 meilleurs systèmes est inférieur au rappel du meilleur système. La conclusion que l'on peut tirer est que les 9 autres meilleurs systèmes retrouvent tous des sous ensembles de l'ensemble des documents pertinents que le meilleur système restitue (le deuxième meilleur système obtient un rappel de 0,84 contre 0,999 pour le meilleur système).

6.3 Comparaison des stratégies pour l'intersection (précision)

Dans la figure 7, les résultats obtenus montrent une faible différence entre les performances de la stratégie1 et de la stratégie2 en 2002 pour les 10 meilleurs résultats. Cette faible différence s'explique par le fait que 80% des systèmes sélectionnés à travers leur valeur de précision et ceux sélectionnés grâce à leur valeur de F-mesure sont identiques (cf. tableau 5). De plus, l'intersection montre un accord entre les systèmes sur les documents retrouvés.

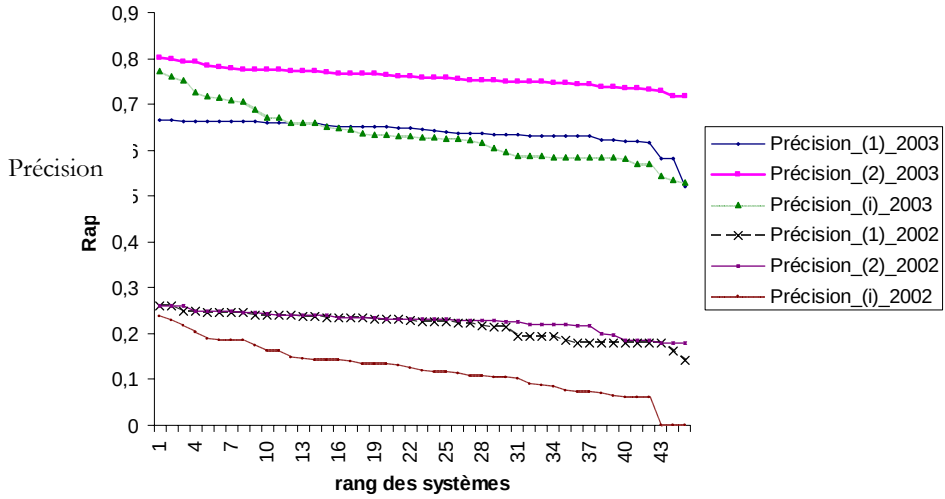


Figure 7 : *Comparaison des mesures de précision obtenues par chaque stratégie de fusion pour la fusion par intersection*

Le tableau 7 contient le même type de comparaisons que le tableau 6, mais en termes de précision.

		2002	2003
Stratégie 1	Meilleur système simple	0,204	0,597
	Moyenne systèmes simples	0,1857	0,5951
	Moyenne fusion 2 à 2	0,2169	0,6396
		(+16,8%)	(+7,5%)

Stratégie 2	Meilleur système simple	0,236	0,774
	Moyenne systèmes simples	0,1959	0,7218
	Moyenne fusion 2 à 2	0,2268	0,7598
		(+15,77%)	(+5,27%)

Tableau 7: *Valeurs de précision moyenne pour la fusion par intersection des systèmes 2 à 2*

Le tableau 7 montre qu'en moyenne la fusion deux à deux améliore les résultats. L'augmentation est plus conséquente en 2002 quelle que soit la stratégie : cela s'explique en partie par les faibles résultats de départ comparativement à 2003. La précision est augmentée de presque 17% avec la stratégie 1 en 2002 entre la moyenne des systèmes simples et la moyenne des fusions deux à deux. Cette augmentation est d'un peu plus de 7% en 2003. Concernant la stratégie 2, l'augmentation est de presque 16% en 2002 et de 5% en 2003. Ce résultat est intéressant dans la mesure où il n'implique pas d'avoir détecté « Le » meilleur système au préalable à la fusion ni au préalable à une recherche.

7 Discussions et conclusion

Les résultats des expérimentations que nous avons présentées dans cet article montrent quels sont les impacts des fusions par union et par intersection. Nous avons montré qu'il était possible, par cette simple stratégie d'améliorer fortement le rappel et donc de constituer un ensemble important de documents liés à un sujet donné. Ainsi, en 2002, en fusionnant le meilleur système en termes de mesure F avec un des 10 autres meilleurs systèmes, le rappel est augmenté de 48%. Lorsqu'un ensemble de systèmes est utilisé pour la fusion par union, pour l'année 2002, le rappel est augmenté d'environ 19% (moyenne des fusions par rapport à la moyenne des systèmes simples) pour la stratégie 1. En utilisant la stratégie 2, la fusion permet d'obtenir une augmentation d'environ 28%. De la même façon, pour l'année 2003, l'augmentation par rapport à la moyenne est de 14% pour la stratégie 1 alors qu'elle est de 17% environ pour la stratégie 2. Par la fusion par intersection, la précision est augmentée de presque 17% avec la stratégie 1 en 2002. Cette augmentation est d'un peu plus de 7% en 2003. Concernant la stratégie 2, l'augmentation est de presque 16% en 2002 et de 5% en 2003.

La mesure des améliorations ainsi obtenue permet de retenir ce type de stratégie dans le cadre de constitution de corpus. La constitution de corpus à de nombreux domaines d'application, que ce soit dans la veille scientifique (Dousset et Mothe 2004) ou dans la construction automatisée d'ontologie de domaine (Hernandez et al., 2006).

Parallèlement, nous avons montré qu'il était aussi possible d'améliorer fortement la précision. En 2002, la fusion du meilleur système avec un des meilleurs systèmes en termes de mesure F permet d'augmenter la précision de 15 % (9% en 2003).

Cette étude pourrait être complétée selon différents axes :

- en premier lieu, nous nous sommes intéressés à la fusion 2 à 2. L'étude pourrait être approfondie en fusionnant plus de deux systèmes,
- cette étude s'appuie sur une connaissance *a priori* des performances des systèmes (fusion des meilleurs systèmes). Ce choix est justifié par le fait que nous souhaitions mesurer les gains relatif à la fusion de systèmes performants. Il était donc nécessaire de les identifier. La généralisation de l'étude serait intéressante ; elle nécessiterait alors de disposer d'un

certain nombre de systèmes que l'on puisse faire fonctionner pour un échantillon plus important de requêtes. Ce n'est malheureusement pas le cas ; aucun programme d'évaluation ne met à disposition les outils.

- Cette étude fait abstraction des caractéristiques des systèmes de recherche d'information pour se concentrer sur leurs résultats globaux. Il serait intéressant d'étudier les différents paramètres de chacun des systèmes afin d'analyser s'il existe une corrélation entre leur complémentarité en termes de documents retrouvés et leurs différences en termes de modèles ou techniques utilisées.
- l'objectif de cette étude était de s'intéresser d'une part au rappel, d'autre part à la précision ; certaines tâches devant privilégier l'une ou l'autre de ces mesures. Nous souhaitons poursuivre cette étude en nous focalisant plutôt sur une mesure globale (mesure F ou précision moyenne [mean average precision]). Dans ce cas, nous souhaitons étudier une combinaison moins systématique que l'intersection et l'union telle que nous les avons appliquées. Il s'agirait de définir des critères permettant de décider *a priori* quelle stratégie (fusion par intersection ou par union) devrait être utilisée en fonction de la requête en cours de traitement. Cette stratégie pourrait s'appuyer sur les taux de chevauchement des ensembles de documents retrouvés.

Références :

Beitzel S.M., Jensen E.C., Chowdhury A., Grossman D., Frieder O., and Goharian N. (2004). *Fusion of effective retrieval strategies in the same information retrieval system*. In Journal of the American Society Information Science Technologies, 55(10), 859–868.

Dousset B., Mothe J., (2004). *Mining document contents in order to analyse a scientific domain*. In RC33 Sixth International Conference on Social Science Methodology, Amsterdam, Barbara Budrich Publishers , (support électronique).

Farah M., Vanderpooten D., (2007). *An outranking approach for rank aggregation*, In International ACM SIGIR Conference on Research and Development in Information Retrieval, 591–598.

Fox E.A, Shaw J.A., (1994). *Combination of multiple searches*, In Text Retrieval Conference (TREC-2), NIST special publication, 243–252.

Hernandez N., Chrisment C., Hubert C., Mothe J., (2006). *Mise à jour d'une ontologie de domaine à partir de l'analyse de nouveaux documents du domaine pour l'indexation de documents*, In Information - Interaction - Intelligence, Cépaduès Editions, Numéro spécial Textes et ressources terminologiques et/ou ontologiques : évolution et maintenance, Vol. Hors-série, 53–83.

Kantor P. B., Ng Kwong B., (2000). *Predicting the effectiveness of Naïve data fusion on the basis of system characteristics*, In Journal of the American Society for Information Science archive, 51(13), 1177 – 1189.

Kompaore D., Mothe J., Lemoing, E., (2006). *Fusion de systèmes pour la recherche de passages dans les textes*, Conférence francophone en Recherche d'Information et Applications (CORIA 2006), 295–300.

Lee, J., (1997). *Analysis of multiple evidence combination*, In International ACM SIGIR Conference on Research and Development in Information Retrieval, 267–276.

Lillis D., Toolan F, Peng L., Collier R., Dunnion J., (2006). *Probability-based fusion of information retrieval result sets*, In International ACM SIGIR Conference on Research and Development on Information Retrieval, 139–146.

Soboroff I., Nicholas C., et Cahan P., (2001). *Ranking retrieval systems without relevance judgements*, In Proceedings of 24th Annual International ACM SIGIR Conference, 66–73.

Spoerri A., (2007). *Examining the Authority and Ranking Effects as the result list depth used in data fusion is varied*, In Information Processing and Management: an International Journal, 43(4), 1044-1058.

Tukey J.W., (1977). *Exploratory data analysis*. EDA, Reading, MA, (Addison Wesley).

Vogt C.C., Cottrell G.W., (1998). *Predicting the performance of linearly combined IR systems*, In International ACM SIGIR Conference on Research and Development in Information Retrieval, 190–196.

Voorhees E.M., Gupta N.K., et Johnson-Laird B., (1994). *The collection fusion problem.* In 3rd Annual Text Retrieval Conférence (TREC-3), NIST.

Wu S., McClean S., (2006). *Performance prediction of data fusion for information retrieval*, In Information Processing and Management: an International Journal, 42(4), 899-915.

Instances, modèles et transformations : tout en un

Instances, Schemas and Transformations: All in One.

Catherine PUGIN(1), Rolf INGOLD(1)

(1)Département d'Informatique, Université de Fribourg - Fribourg, Suisse
Catherine.pugin@unifr.ch
rolf.ingold@unifr.ch

Résumé. La modélisation de documents et de données structurées est d'une importance fondamentale dans le monde XML. Il est essentiel de pouvoir assurer la validité des instances XML afin de garantir que les documents échangés puissent être utilisés. Dans cette contribution, nous présentons un nouveau langage de modélisation DML qui a été développé dans un contexte particulier. En effet, le projet dans lequel DML s'inscrit vise à intégrer les technologies de base XML (balisage, modélisation et transformation). De ce fait, plus que simplement un nouveau langage de modélisation, c'est toute une réflexion sur la modélisation en général et sur la notion d'héritage entre modèles en particulier qui est menée. DML considère les modèles dans une hiérarchie dont le sommet est un modèle universel et la base les modèles les plus restreints. Nous introduisons également YML, une application XML disposant d'un infoset légèrement modifié qui permet d'optimiser l'intégration et donnons un bref aperçu d'un langage de transformation simple qui bénéficie des concepts de DML.
Mots-clés. XML, modélisation, héritage.

Abstract. The modeling of documents and structured data is of fundamental importance in the XML world. It is essential to be able to ensure the validity of XML instances so that the exchanged documents can be easily used. In this paper, we present a new schema language, called DML, which is developed in a particular context. DML is part of a more global project that aims at integrating XML core technologies (markup, modeling and transformation). As a result, it is not simply a new schema language that is proposed but a whole reflection that is made on modeling in general and inheritance between schemas in particular. DML considers schemas as belonging to a hierarchy with a universal schema at its top and the most restricted schemas at the bottom. We also introduce YML, an XML application with a slightly modified infoset that aims at optimizing integration and we give a brief overview of a transformation language that benefits from the simple concepts of DML.
Keywords. XML, schemas, inheritance.

1 Introduction

Dès la publication de sa première recommandation officielle par le W3C (*World Wide Web Consortium*) en 1996, XML (Bray *et al.*, 2004 – dernière édition) est rapidement devenu un standard pour la représentation tant des documents que des données structurées. Sa simplicité et sa grande flexibilité ont permis de conquérir de nombreuses communautés scientifiques, bien au-delà de la tâche pour laquelle il fut initialement développé comme successeur du langage de balisage SGML.

De part la très large diffusion du langage, la définition de modèles pour les instances XML est apparue très tôt comme une tâche indispensable. En effet, les modèles définissent un certain nombre de règles sur le contenu des éléments ou la forme des attributs et permettent ainsi de tester la validité d'un document en plus de sa conformité. L'échange de documents XML est ainsi facilité puisque chacun connaît les règles de conception des documents.

Un premier langage de modélisation (DTD) est introduit parallèlement au langage XML dans la première recommandation de celui-ci. DTD (*Document Type Definition*) connaît un succès rapide mais doit également faire face à de nombreuses critiques dû à la relative pauvreté du langage. Dès lors, le W3C propose de nouvelles solutions beaucoup plus riches au problème de la modélisation dont XML Schema qui permet aujourd'hui de définir des modèles beaucoup plus précis et plus complets. Ce langage populaire souffre cette fois-ci d'une trop grande complexité.

En marge du W3C, d'autres propositions enrichissent le monde de la modélisation XML. Parmi celles-ci, RelaxNG est un compromis entre la complexité de XML Schema et les lacunes des DTD. DSD (Document Structure Definition) propose un nouveau paradigme de règles pour la définition de modèles.

Dans cette contribution, nous introduisons un nouveau langage de modélisation appelé DML (*Document Modeling Language*) qui, à l'image de RelaxNG, propose un compromis entre DTD et XML Schema. DML est basé sur des grammaires d'expressions régulières. Du fait que DML est développé dans un projet plus global qui tend à intégrer les technologies XML de base (balisage, modélisation et transformation), une réflexion sur la modélisation en général est menée. En particulier, l'héritage entre les modèles est étudié et mène à une hiérarchie de modèles DML. Un modèle DML dit universel décrit tous les documents potentiels et tous les autres modèles héritent de celui-ci. Au bas de cette hiérarchie se trouvent les modèles les plus restreints : les instances.

Dans ce cadre, nous travaillons avec des instances YML. YML est une application XML avec un infoset modifié qui permet de considérer les éléments et les nœuds textes de manière équivalente. YML et DML sont très bien intégrés et permettent le développement d'un langage de transformation simple et puissant appelé DGL.

La section 2 présente un bref état de l'art des principaux langages de modélisation et permet de mieux déterminer par la suite les spécificités du langage DML. DML est plus précisément introduit dans la section 3. La présentation de ses principales caractéristiques et de son modèle de hiérarchie est soutenue par différents exemples. La section 4 revient sur YML et ses spécificités puis donne quelques points d'entrées sur l'intégration de YML et DML avec un langage de transformation appelé DGL. Finalement la section 5 conclut la contribution.

2 Bref état de l'art des langages de modélisation

DTD (Bray *et al.*, 2004), XML Schema (Biron *et al.*, 2004, Fallside et Walmsley, 2004, Thompson *et al.*, 2004) et RelaxNG (Clark et Murata, 2001) sont les langages

de modélisation les plus connus et certainement les plus utilisés aujourd'hui. D'autres propositions existent (Van der Vlist, 2001), dont le langage DSD (Møller, 2002), mais elles ne bénéficient pas d'un grand support et sont dès lors moins populaires.

DTD est un héritage direct de SGML et fait partie intégrante de la première spécification de XML. Malheureusement, le langage doit rapidement faire face à de nombreuses critiques. Tout d'abord, conséquence de son héritage, il ne possède pas de syntaxe XML, ce qui dessert grandement une bonne intégration. De plus, il n'offre aucun support pour le traitement des espaces de nommage (Bray *et al.*, 2006) qui n'étaient pas encore définis lors de la première spécification du langage. Finalement, le mécanisme de typage est très pauvre. Le contenu textuel des éléments et des attributs ne peut être décrit que par la valeur unique *PCDATA* tandis que le contenu des éléments ne peut pas être défini en fonction du contexte. Tous les éléments répondant au même nom sont décrits par une seule expression indépendamment de leur position dans l'arbre.

Pour remédier à ces lacunes, le W3C propose rapidement une alternative sous la forme de XML Schema. Le langage possède une syntaxe XML. Il est très fortement typé et déterministe. Puisqu'extrêmement riche, il en devient rapidement très compliqué à utiliser. Sa spécification en est la preuve puisqu'elle est très volumineuse et difficilement accessible à des non-initiés. De plus, le langage XML Schema n'est pas décrit par un modèle XML Schema normatif, ce qui laisse encore entendre que le langage demeure trop complexe.

Parmi les propositions externes au W3C, le consortium OASIS propose le langage RelaxNG, un compromis entre DTD et XML Schema. RelaxNG est plus riche que DTD mais limite sa complexité en restreignant, par exemple, le nombre de constructions du langage. Facile à comprendre et facile à utiliser, le langage est basé sur des expressions régulières. Sa syntaxe XML permet de définir un modèle RelaxNG pour RelaxNG et rend le langage totalement auto-descriptif. Une syntaxe non-XML moins verbeuse est également disponible.

En comparaison avec XML Schema, RelaxNG ne fournit pas de système de typage de données mais intègre des librairies de types (éventuellement la librairie de XML Schema). De plus, RelaxNG ne considère pas l'héritage entre les différents contenus définis : un contenu ne peut être restreint ou étendu.

Parmi les autres langages DSD (Moller, 2002) est basé sur un système de règles plutôt que de grammaires. Son approche est la même que celle de RelaxNG, c'est-à-dire de fournir un langage plus facile d'accès aux utilisateurs.

Dans ce contexte, nous avons décidé de développer notre propre langage de modélisation qui s'inscrit dans un projet plus large d'intégration des technologies XML de base (balisage, modélisation et transformation). Ce langage est décrit précisément dans la section suivante.

3 Le langage de modélisation DML

Nous introduisons ici le langage DML et ses caractéristiques propres. DML répond aux mêmes critères de base que RelaxNG ou DSD : il cherche à offrir un langage de modélisation plus simple à appréhender que XML Schema. Mais DML va plus loin encore, car il s'inscrit dans un projet plus global où il n'est pas simplement question de modélisation mais bel et bien d'intégration. L'intégration des trois langages de base (balisage, modélisation et transformation) permet de définir une vision uniforme et cohérente des technologies XML. DML est ainsi intégré dans un langage de transformation appelé DGL où il inclut un traitement

dépendant du contenu au langage initial. Une réflexion plus large est menée et aboutit à une proposition de hiérarchisation des modèles.

La syntaxe du langage est très proche de RelaxNG mais l'approche diffère sensiblement et c'est dans cet esprit que nous introduisons un nouveau langage et non pas simplement une adaptation ou extension d'un langage existant.

3.1 Caractéristiques du langage

Expressions régulières

Le langage DML est entièrement basé sur des grammaires d'expressions régulières (Murata *et al.*, 2005) à l'image de RelaxNG ou de DSD et à la particularité de traiter les éléments et les nœuds textes de manière équivalente.

Non seulement le contenu des éléments mais également celui des attributs et des nœuds textes est décrit par des expressions régulières.

Le contenu d'un élément est une structure et les contenus des nœuds textes et des attributs sont décrits par des types.

Les structures sont composées d'éléments, de séquences d'éléments et de choix d'éléments. Les nœuds textes apparaissent aussi dans les expressions régulières. Finalement, une construction *any* permet d'inclure n'importe quel élément dans une structure.

La figure 1 présente une structure composée d'une séquence regroupant trois éléments (*name*, *phone* et *mail*) ainsi qu'un type *number* décrit par une expression régulière.

Figure 1. *Une structure et un type DML.*

```
<dml:struct name="contact">
  <dml:seq>
    <name>…</name>
    <phone> <text type="number"/> </phone>
    <mail>…</mail>
  </dml:seq>
</dml:struct>
<dml:type name="number" pattern="0 | [1-9][0-9]*"/>
```

Occurrences limitées

Pour éviter la complexité de XML Schema en matière d'occurrences, DML limite le nombre de contraintes d'occurrence à quatre. Un élément, une séquence ou un choix peuvent apparaître une fois (*once*), zéro ou une fois (*optional*), une ou plusieurs fois (*many*) ou zéro ou plusieurs fois (*free*).

La figure 2 reprend l'exemple précédent mais définit une présence optionnelle pour l'élément phone.

Traitement des nœuds textes

XML Schema traite le contenu textuel comme le type simple des éléments. Ceci génère une relation de dépendance entre éléments et nœuds textes.

Les nœuds textes et les éléments sont considérés de manière équivalente en DML. Ceci implique une construction spécifique pour modéliser la présence du texte dans une structure.

Modularité

Le langage DML intègre les espaces de nommage et bénéficie ainsi d'une grande modularité. Les structures et les types définis dans les différents modèles

peuvent être réutilisés par un simple mécanisme de référencement. Ce mécanisme se révèlera très utile par la suite lorsque nous introduirons la dérivation par restriction qui permet de hiérarchiser les modèles.

```
<dml:struct name="contact">
  <dml:seq>
    <name/>
    <dml:optional> <phone/> <dml:optional>
    <mail/>
  </dml:seq>
</dml:struct>
```

Figure 2. *L'élément phone est optionnel.*

Le contenu mixte des éléments peut dès lors être décrit de manière extrêmement précise alors que XML Schema note simplement que le contenu est mixte à la hauteur de l'élément en question.

La figure 3 illustre cette définition de contenu mixte. La structure représentée est associée à un élément *caption* dans lequel un texte en gras doit automatiquement apparaître avant le texte principal ou ne pas apparaître du tout.

```
<dml:struct name="caption">
  <dml:seq>
    <dml:optional>
      <strong> <dml:text/> </strong>
    </dml:optional>
    <dml:text/>
  </dml:seq>
</dml:struct>
```

Figure 3. *La structure associée à un élément caption.*

Méta-modèle

La syntaxe du langage DML est entièrement auto-descriptive. Un modèle particulier appelé méta-modèle décrit totalement le langage. Chaque DML peut être validé par rapport à ce méta-modèle.

3.2 Les modèles dans une hiérarchie

Dès le début du développement du langage DML, l'idée d'une hiérarchie entre les modèles est apparue de manière très claire. Dans notre vision, la hiérarchie implique qu'une instance valide par rapport à un modèle donné l'est également par rapport à toute la lignée des modèles dont celui-ci hérite.

Ce thème n'est pas traité par RelaxNG qui y voit une trop grande complexité tandis que XML Schema fournit deux systèmes de dérivation, par extension et par restriction, mais sans une vision globale de la hiérarchie.

Il semble tout à fait judicieux pour un langage de modélisation d'être capable de décrire l'ensemble des instances, par conséquent l'ensemble des modèles du langage. Ceci donne une nouvelle dimension aux instances puisque, si la conformité qui s'appuie sur les règles syntaxiques peut en tout temps être contrôlée, la validité des instances peut l'être également.

La validité d'une instance n'est donc plus un simple test qui est effectué mais est profondément ancrée dans la technologie. Une instance conforme doit également être valide par rapport à un modèle au minimum.

Modèle universel

Le modèle duquel tous les modèles DML découlent et qui décrit toutes les instances conformes est appelé *modèle universel*. La figure 4 illustre ce modèle particulier du langage DML. Pour des raisons de lisibilité, toutes les définitions liées aux attributs ont été omises dans cette figure.

```
<?xml version="1.0"?>
<yml>
 <yml:dml prefix="dml" uri="dml.dml"/>
 <dml:root>
   <dml:ref name="universal"/>
 </dml:root>
 <dml:struct name="universal">
   <dml:free>
     <dml:choice>
       <dml:any>
         <dml:ref name="universal"/>
       </dml:any>
       <dml:text type="universal"/>
     </dml:seq>
   </dml:free>
 </dml:struct>
 <dml:type name="universal" pattern=".*"/>
</yml>
```

Figure 4. *Le modèle DML universel.*

Le modèle universel se trouve donc au sommet de la hiérarchie du langage DML. Il est composé de la construction *any* et de la construction *text* dont le type est décrit pas l'expression régulière la plus générale.

Dérivation

Nous ne considérons pour l'instant que la dérivation par restriction et non pas la dérivation par extension. En effet, il est toujours possible d'utiliser le modèle universel comme modèle étendu. De plus, la dérivation ne s'applique actuellement qu'aux structures, c'est-à-dire au contenu des éléments. Aucune restriction n'est définie actuellement pour les types des nœuds textes et des valeurs d'attributs.

Comme nous l'avons introduit plus haut, les structures sont décrites par des expressions régulières. Restreindre de telles structures revient à restreindre des expressions régulières. Nous considérons qu'une expression régulière A' est une expression restreinte de A si et seulement si A' est incluse dans A. En effet, si l'expression A' décrit un spécimen S, alors ce spécimen S sera également décrit par A.

Tester l'inclusion de deux expressions régulières consiste à vérifier de l'assertion suivante (Hopcroft et Ullman, 1979) :

$$L_{A'} \subseteq L_A \Leftrightarrow L_{A'} \cap \overline{L_A} = \varnothing$$

Dans le cas de la modélisation, une instance valide par rapport à A' l'est aussi automatiquement par rapport à A. Si une instance est valide par rapport au modèle restreint, elle le sera également par rapport au modèle initial, et dans le cas absolu par rapport au modèle universel.

Grâce au système simple de contraintes d'occurrence de DML ainsi qu'à la relative simplicité de ses constructions, il est possible de définir quatre règles différentes pour restreindre les structures. Ces règles sont appliquées sur les éléments, les séquences et les choix. Elles prennent la forme de constructions DML dans la syntaxe du langage. Une cinquième construction permet de restreindre le contenu des choix de manière précise.

Restriction des éléments, des séquences et des choix

La première règle est la règle *force*. Elle permet d'imposer la présence d'un élément, d'une séquence ou d'un choix dont l'occurrence est *optional* (zéro ou un) ou *free* (zéro ou plusieurs). La figure 5 illustre cette règle : la structure de la figure 2 est restreinte.

```
<dml:struct name="contact_phone" base="contact">
  <dml:force name="phone"/>
</dml:struct>
```

Figure 5. *Forcer un élément*

La deuxième règle (*single*) permet d'individualiser un élément, une séquence ou un choix. Si l'occurrence est *many* (un ou plusieurs) ou *free*, alors les occurrences deviennent respectivement *once* (un) et *optional*. Un cas pratique est donné dans la figure 6 où une structure qui associe plusieurs adresses à un nom est restreinte pour garder une seule adresse.

```
<dml:struct name="addresses">
  <dml:seq>
    <name/>
    <dml:many> <address/> <dml:many>
  </dml:seq>
</dml:struct>

<dml:struct name="address" base="addresses">
  <dml:single name="address"/>
</dml:struct>
```

Figure 6. *Individualiser un élément.*

La troisième règle (*remove*) permet de supprimer un élément, une séquence ou un choix dont l'occurrence est *optional* ou *free*. Si l'on reprend l'exemple de la figure 2, il s'agirait de ne conserver que les contacts qui ne possèdent pas de téléphone, comme le montre la figure 7.

```
<dml:struct name="contact_nophone" base="contact">
  <dml:remove name="phone"/>
</dml:struct>
```

Figure 7. *Supprimer un élément*

La quatrième règle (*replace*) concerne la construction *any* qui permet d'inclure n'importe quel élément dans une structure. La restriction de cette construction consiste à la remplacer par un ou plusieurs – selon l'occurrence associée – éléments nommés. La figure 8 illustre ce cas.

```
<dml:struct name="any_seq">
  <dml:seq>
    <dml:any/>
  </dml:seq>
</dml:struct>

<dml:struct name="a_seq" base="any">
  <dml:replace> <a/> </dml:replace>
</dml:struct>
```

Figure 8. *Remplacement de la construction any.*

Restriction du contenu des choix

Pour restreindre le contenu d'un choix, il est possible soit d'appliquer les règles précédentes sur les éléments et les séquences qui composent le choix, soit supprimer certaines options de ce choix. Pour maîtriser précisément ces aspects, nous introduisons une construction supplémentaire *keep* qui permet de définir précisément la gestion de la restriction, comme le montre la figure 9. Dans cet exemple, nous passons de la structure *(a?|b*)* à la structure *(a?)*.

```
<dml:struct name="choice">
  <dml:choice>
    <dml:optional> <a/> </dml:optional>
    <dml:free> <b/> </dml:free>
  </dml:choice>
</dml:struct>
<dml:struct name="restricted" base="choice">
  <dml:keep/>
  <dml:remove/>
</dml:struct>
```

Figure 9. *Restriction du contenu d'un choix*

3.3 Réflexion sur la base de la hiérarchie

Comme nous l'avons vu, les modèles DML sont organisés dans une hiérarchie. Au sommet se trouve le modèle universel présenté dans la figure 4. Tous les autres modèles DML héritent conceptuellement de ce modèle. Cinq règles ont été définies sous la forme de constructions DML pour rendre la dérivation par restriction pratique. Grâce à la modularité du langage, les modèles peuvent hériter de structures préalablement définies et grâce aux règles ci-dessus, ces structures peuvent être restreintes.

Il est judicieux de s'intéresser maintenant à la base de cette hiérarchie et aux modèles qui la composent. La question à laquelle il faut répondre est la suivante : quel est le modèle le plus restreint ?

Le concept sous-jacent de DML permet de dire que le modèle le plus restreint est celui qui ne décrit qu'une seule instance. Ce modèle ne comporte aucun choix mais seulement des séquences d'éléments. Les occurrences des séquences et des

éléments sont définies comme *once* (une fois). Les types des nœuds textes suivent la même logique.

Quelle différence subsiste alors entre une instance et un modèle complètement restreint ? Ne serait-il pas possible d'identifier les deux et de considérer un tel modèle comme une instance ?

Il est évident qu'en termes d'intégration une telle approche est pertinente. Instances et modèles ne sont plus dissociés mais forment un véritable tout et l'ensemble des technologies peut être compris et appréhendé dans une véritable uniformité.

Pour parfaitement intégrer modèles et instances, il suffit donc simplement d'adapter la syntaxe du langage DML afin que le modèle le plus restreint puisse être assimilé à une instance.

4 Intégration

4.1 YML : une application XML

Afin de favoriser l'intégration des différentes technologies, nous travaillons avec le langage de balisage YML qui est une application XML avec un infoset légèrement modifié.

Dans un premier temps, ce langage est totalement dissocié de XML (Pugin et Ingold, 2006). La version actuelle de YML est à nouveau plus proche de XML. Nous considérons YML comme application XML. Ceci implique que tous les documents YML sont des documents XML conformes.

Une des spécificités de YML réside dans la présence de nœuds textes explicites. Ces nœuds textes sont délimités par des balises et permettent une meilleure gestion des caractères blancs. Les caractères blancs à l'intérieur des balises sont pertinents pour le texte, les autres sont uniquement destinés à l'indentation du document. Les retours à la ligne sont symbolisés par le caractère # qui se place après la balise fermante du nœud texte.

La définition des espaces de nommage a également été repensée. Ceux-ci sont définis au sommet du document YML dans des éléments spécifiques.

Pour être conforme à XML, un élément racine *yml* est présent au début de chaque document.

La figure 10 présente un exemple de document YML.

```
<?xml version="1.0"?>
<yml>
 <yml:dml uri="contacts.dml"/>
 <contact>
  <name>[Jean][Dupont]#</name>
  <phone>[004132569]</phone>
  <mail>[jean@company.com]</mail>
 </contact>
<yml>
```

Figure 10. *Un document YML.*

4.2 Le langage de transformation DGL

Nous introduisons brièvement le langage de transformation DGL (Pugin et Ingold, 2007) afin de compléter la vue d'ensemble sur les différentes technologies intégrées dans le cadre de ce projet.

DGL (*Document Generation Language*) est un langage de transformation simple dont le modèle de traitement s'inspire de XSLT (Clark, 1999 ; Kay, 2007). Le langage est XML. Un document DGL est composé d'un *template* principal qui initialise le traitement et de *patterns* qui peuvent être instanciés, de manière éventuellement récursive. Un pattern est un ensemble de règles qui sont appliquées aux différents nœuds.

Le traitement se base sur deux opérations simples : un nœud est copié à partir d'une entrée donnée ou est créé. Etant donné que les éléments et les nœuds textes sont considérés de manière équivalente, les deux opérations peuvent être appliquées aux uns ou aux autres indépendamment.

La troisième opération permet d'appliquer un *pattern* sur un ensemble de nœuds sélectionné. Pour la sélection des nœuds, un système propre basé sur XML a été développé. Il est moins riche que XPath (Berglund *et al.*, 2007 ; Clark *et al.*, 1999) mais peut être facilement formalisé et suffit pour l'instant amplement à l'utilisation qui en est faite.

La richesse du langage DGL est une spécification complète. Le langage a été complètement formalisé grâce à la sémantique dénotationnelle (Tennent, 1976) et les outils associés ont pu être implémentés facilement et sans risque d'erreurs.

De plus, un mécanisme de typage statique des transformations a également été défini et implémenté.

Intégration de DML

Dans sa version initiale, DGL ne propose que des transformations structurelles. Tous les aspects liés au contenu des éléments sont ignorés. Pour enrichir le langage avec ces aspects tout en évitant de le complexifier en ajoutant de nouvelles constructions, nous intégrons certaines constructions du langage DML dans DGL.

Grâce au mécanisme des espaces de nommage, cette intégration peut être réalisée de manière naturelle et le document DGL peut toujours être validé par rapport au modèle DML qui décrit le langage DGL et au méta-modèle DML.

Les structures et les types DML peuvent être intégrés dans le mécanisme de sélection des nœuds ou dans les règles des patterns. Ainsi les nœuds ne sont plus sélectionnés simplement en fonction de leur nom (pour les éléments) mais également en fonction de leur contenu. Le mécanisme est ainsi considérablement enrichi.

Les *templates* intègrent également ces constructions et permettent un typage dynamique des transformations.

La figure 11 présente un exemple simple d'un document DGL où les éléments *contact* dont le contenu contient l'élément *phone* sont sélectionnés puis copiés à l'intérieur d'un élément créé *phonecontacts*.

Intégration de YML

Une question similaire à celle de la base de la hiérarchie DML s'est posé avec DGL. Est-ce qu'un document qui n'est composé que de nouveaux éléments (des éléments créés) n'est finalement pas une instance YML simplement ? Nous répondons à cette question par l'affirmative.

Le document de la figure 10 peut donc également être vu comme un document DGL dont chaque élément est créé.

```
<?xml version="1.0"?>
<yml>
 <yml:dml prefix="dgl" uri="dgl.dml"/>
 <yml:dml prefix="dml" uri="dml.dml"/>
 <yml:dml uri="contacts"/>
 <phonecontacts>
  <dgl:apply pattern="p1">
   <dgl:select>
    <dgl:step match="contact">
     <dml:struct>
      <dml:seq><name/><phone/><mail/></dml:seq>
     </dml:struct>
    </dgl:step>
   </dgl:select>
  </dgl:apply>
 </phonecontacts>
 <dgl:pattern name="p1">
  <dgl:rule match="#"/>
   <dgl:copy>
    <dgl:apply pattern="p1">
      <dgl:select><dgl:step match="#"/></dgl:select>
    </dgl:apply>
   </dgl:copy>
  </dgl:rule>
 </dgl:pattern>
<yml>
```

Figure 11. *Un document DGL avec intégration du DML.*

5 Conclusion

Nous avons présenté un nouveau langage de modélisation – DML – qui s'inscrit dans un projet plus global d'intégration des technologies XML de base (balisage, modélisation et transformation). Ce langage va au-delà de la simple modélisation de documents ou de données structurées. Il permet d'amorcer une discussion sur l'héritage des modèles entre eux et les conséquences sur les langages de balisage et de transformation.

Le langage DML est complètement auto-descriptif et bénéficie d'un modèle universel duquel tous les modèles découlent. Ce modèle est au sommet d'une hiérarchie de modèles. Différentes règles de dérivation par restriction permettent d'organiser les modèles entre eux. Une réflexion est menée pour savoir comment doivent être considérés les modèles qui se trouvent au bas de cette hiérarchie. Est-ce que ces modèles peuvent être assimilés à des instances de documents ? Une réponse affirmative permet de favoriser l'intégration des langages mais implique une modification de la syntaxe du langage DML.

Le langage de balisage YML et le langage de transformation DGL bénéficient de l'intégration de DML. Ainsi, DGL est capable de traiter des cas complexes de sélection de nœuds dépendamment de leur contenu sans qu'aucune nouvelle construction ne soit ajoutée au langage.

L'intégration des trois langages YML/DML/DGL permet de développer une vision cohérente du traitement des documents XML. En effet, les trois langages ne sont pas développés de manière indépendante mais bien en parallèle.

La vision de la modélisation s'en est ainsi trouvée modifiée et DML propose, plus qu'un nouveau langage de modélisation, une véritable réflexion autour des modèles de documents en général.

Références :

Berglund, A., *et al.* (2007). *XML Path Language (XPath) 2.0.* Disponible à www.w3.org/TR/xpath20

Biron, P.V. *et al.* (2004). *XML Schema Part 2 : Datatypes.* Disponible à www.w3.org/TR/xmlschema-2

Bray, T. *et al.* (2004). *Extensible Markup Language (XML) 1.0 (Third Edition).* Disponible à www.w3.org/TR/REC-xml

Bray, T. et al. (2006). *Namespaces in XML 1.0 (Second Edition).* Disponible à www.w3.org/TR/REC-xml-names

Clark, J., *et al.* (1999). *XML Path Language (XPath) Version 1.0.* Disponible à www.w3.org/TR/xpath

Clark, J. (1999). *XSL Transformation Version 1.0.* Disponible à www.w3.org/TR/xslt

Clark, J., Murata, M. (2001). *RelaxNG Specification.* Disponible à www.relaxng.org/spec-20011203.html

Fallside, D.C., Walmsley, P. (2004). *XML Schema Part 0:Primer.* Disponible à www.w3.org/TR/xmlschema-0

Hopcroft, J.E., Ullman, J.D. (1979). *Introduction To Automata Theory, Languages, And Computation.* Addison-Wesley Longman Publishing Co.

Kay, M. (2007). *XSL Transformation Version 2.0.* Disponible à www.w3.org/TR/xslt20

Møller, A. (2002). *Document Structure Description 2.0.* Disponible à www.brics.dk/DSD/dsd2.html

Murata, M. et al. (2005). Taxonomy of XML Schema languages using formal language theory. *ACM Transactions Internet Technologies,* vol. 5, num. 4, 1-45.

Pugin, C., Ingold, R. (2006). YML: une version épurée de XML pour faciliter une specification rigoureuse des modèles de documents et des transformations. In *Actes du Colloque International du Document Electronique 9,* CIDE9, Fribourg, Suisse, Septembre.

Pugin, C., Ingold, R. (2007). Combination of Schema and Transformation Language Described by a Complete Formal Semantics. In *Proceedings of ACM Symposium on Document Engineering,* DocEng'07, Winnipeg, Canada, Août.

Tennent, R.D. (1976). The Denotational Semantics of Programming Languages. *Communication of the ACM,* vol. 19, num. 8, 437-453.

Thompson, H.S., *et al.* (2004). *XML Schema Part 1: Structures.* Disponible à www.w3.org/TR/xmlschema-1

Van der Vlist, E. (2001). *Comparing XML Schema Languages.* Disponible à www.xml.com/pub/a/2001/12/12/schemacompare.html

ARMARIUS- A Living Online Archive for Ancient Manuscripts

Reim DOUMAT(1), Elöd EGYED-ZSIGMOND(1), Emese CSISZÁR(2), Jean-Marie PINON(1)

(1)Laboratoire LIRIS, INSA de Lyon, Villeurbanne, France
reim.doumat@liris.cnrs.fr
elod.egyed-zsigmond, @liris.cnrs.fr
jean-marie.pinon@liris.cnrs.fr

(2)Sapientia EMTE, Tîrgu-Mres, Romania.
csiszarmeso@yahoo.com

Abstract. Many museums and libraries digitize their collections of historical manuscripts, to preserve the historic documents and to make them public. The collections are available in image format and they need annotation to be accessible and exploitable. The annotations can be created manually, automatically or semi-automatically. The problem with the manual annotation is that they are expensive and tedious. Hence the reuse of users' experiences, by tracing their actions during the annotation process, helps other users to accomplish repetitive tasks in a semi-automatic manner, and assists difficult tasks. In this article we present a digital archive model and prototype of a collaborative system for the management of online ancient manuscript. The application offers an online annotation service for this type of documents, an assistant for a semi-automatic annotation, and a tracing system that saves traces of important actions in order to reuse them later in a recommender system.
Keywords. Living digital archive, manuscript annotation, assistant, tracing system.

Résumé. Plusieurs musés et bibliothèques numérisent leur collections de manuscrits historiques pour les conserver et les rendre publiques. Les collections sont disponibles en format image et ont besoin d'annotations pour être accessibles et exploitables. La création des annotations peut être manuelle, automatique ou assistée. Le problème avec l'annotation manuelle qu'il est chère et fastidieuse, donc la réutilisation de l'expérience de l'utilisateur, en se basant sur des expériences tracées, permet d'en aider d'autres à réaliser des tâches répétitives de manière semi-automatique, ou d'effectuer des tâches non triviales de manière assistée. Dans cet article nous présentons une archive numérique de manuscrits anciens en ligne. Cette application offre un service d'annotation, un système de traçage gardant les traces de certaines actions et un système d'assistance qui exploite ces traces.
Mots-clés. Archive numérique vivante, annotation des manuscrits, système d'assistance, système de traçage.

1 Introduction

Many museums and libraries digitize their collections of historic manuscripts to protect these precious documents and to make them accessible to a large public. These collections are available online in image format and they need annotations to be accessible and exploitable.

Actually, the consultation of collections on the internet is increasing progressively, because it meets the various needs of all user types, and because it offers users with services to search rapidly the information, to mark their favourite pages and to personalize their environment (Vivarium the online digital collections of Saint John's university and the College of Saint Benedict, ContentDM collection). However, these operations might be considered as non creative operations since users do not work directly on documents. The interfaces do not allow the easy communication and publication of ideas, comments, and interpretations. The importance of the annotations according to (Bottoni and al. 2004) is that they form a support to the intellectual activities, like: a highlight of interesting parties of a text, an indication to the user reflection and an enhancement of the document with new information.

Consequently, users need to annotate documents online independently from their media type (images, audios, videos, web pages, etc.). Annotations represent primordial actions that offer to users the possibility to react directly on their documents in order to enrich them. Additionally, every annotation made by the user can generate a trace in the system in order to be reused lately. This could be beneficial for all persons who do not know the domain or for those who miss the experience. According to (Egyed-Zsigmond and al. 2003), the reuse of user's experience during the annotation process permits other users to realize repetitive tasks in a semi automatic manner, or to realize difficult task in an assisted way.

In this paper we present an online archive application to manage and annotate ancient manuscripts. We incorporate within this archive some image treatment tools and web services to annotate remotely these manuscripts. Our application is enriched with an experience capitalization layer that traces the important actions, and then it integrates traces in an assistant system to help users during the annotation procedure.

The article is organized as follows: in the next section, we expose the state of the art about some of the popular annotating systems. In section 3 we present our project, called ARMARIUS and emphasise on annotating the manuscripts online by different types of users, and then we illustrate a prototype of our web application. At the end we conclude and give some perspectives.

2 Related works

Digital annotations that are attached to digital collections represent two elements: metadata and content. The first is a group of attributes like (author, title, creation date, modification date,…) that could be defined by a standard (Dublin Core, Marc, MODS, TEI…) or by the environment of the annotation. The second element is the content that is created by users and is composed of textual information, images, hyperlinks, etc. Annotations vary depending on the system and the context where they are used. Many projects are interested in the annotations, in this section we refer to some of them and compare their characteristics.

2.1 Document annotation projects

Many annotation projects have developed diverse tools to annotate web pages, multimedia objects, or documents, the objectives of these projects varied between: creating repositories with web services that are adaptable to comprise different types of collections to form digital libraries like Fedora, offering image mark-up tool like the project UVic, integrating plug-ins in the web browsers to provide annotation tools such as Annotea (Kahan and Koivunen, 2001) that permits to exchange web annotations and bookmarks between users, TafAnnote (Cabanac and al., 2007), and MADCOW (Multimedia Annotation of Digital Content Over the Web) (Bottoni and al., 2004) for multimedia annotation over the web.

Some of the previous systems have a collaborative environment that permits different users to work on a group of documents and to share their knowledge, as mentioned in Table 1. The table summarizes the differences between the characteristics of these projects.

System Feature	**Fedora**	**UVic**	**MADCOW**	**Annotea**	**TafAnnote**
Document type	Digital collections	Images	Web pages, multimedia objects	Web content	Web content
Annotation type	Précised by the digital library	Keywords, comments	Many types of comments	Notes, explanations, bookmarks	Comments (discussion)
Collaboration work	Between systems	No	Yes	Yes	Yes
Recommender system	No	No	No	No	No
Type of the application	Web application	Standalone application and web based viewer	Plug-in client in standard web browser	Plug-in client, proxy	Plug-in client in Mozilla Firefox web browser

Table 1. ***Comparison between annotating projects***

The main disadvantage of these systems is that they handle XML documents while it is not able with images of ancients' manuscripts. In Annotea the web pages and their contents of objects (images, texts, hyper links, etc.) are identified by URLs while scanned images of the manuscripts are identified by IDs. Annotea, TafAnnote and MADCOW permit the information exchange between user groups; this service enhances the collaboration work in order to facilitate the realization of difficult user tasks. Fedora does not enhance the information exchange between users.

Other projects interested in the annotation of the ancient manuscripts like: Bambi (Calabretto and al., 1998) which is an ancient project to annotate manuscripts on a local machine; users can work in collaboration but on the same computer. Other systems are web applications that could be used to visualize and to annotate documents remotely as IPSA (Agosti and al., 2003), Scraps (documents

from the World War I) offers the access to rare books online. Annotations in Debora are extracted by image treatment tools and are classified in three levels (description, structure and contents), while IPSA works on manual image annotation of Herbal manuscripts. Table 2 summarizes the differences between these projects.

System / Feature	**Bambi**	**Debora**	**Scraps**	**IPSA**
Document type	Images	Images	Images	Images
Annotation type	Manual	Automatic extraction, predetermined	Predetermined	Manual (Textual and linking annotations)
Collaboration work	Yes	No	No	Yes
Recommender system	No	No	No	No
Type of the application	Standalone application	Standalone application	Partially web application	Web application

Table 2. ***Comparison between manuscripts annotating projects***

The listed systems do not contain assisting tools to facilitate the manuscripts annotation and the use of other services, or collaborative recommender tools to assist users in realizing difficult tasks.

2.2 Tracing and recommender systems projects

Tracing system registers important events and actions made by users while using the application, traces are used to build users experiences like (Hilbert and Redmiles, 2000), Trèfle (Egyed-Zsigmond and all., 2003) that generate an assistant system from experienced user actions, and (eMédiathèque) which is a collaborative platform for virtual classroom developed by eLycée, it includes a tracing infrastructure with a collaborative tools to help users in remote learning. Traces are also used to build intelligent applications such as recommender systems, which assist and give advice to users during his interaction with the application (Champin, 2003). Some recommender systems base on the user profile and his history to determine the interesting documents or web pages of each user, such as Personal Web Watcher (Mladenic, 1999), ITR recommender system (Semeraro and al., 2007).

2.3 State of the art conclusion

We can notice that not all of these systems are capable to organize and annotate remotely images of ancient manuscripts; stand alone applications are not useful if user groups need to work in collaboration to annotate the images. Other online projects do not offer precise annotation and collaborative space to facilitate the communication between different users (i.e. confrontations of points of views, correction to annotation done by other users). Furthermore, some projects (Annotea, MADCOW) while they have collaborative functionalities, they enable to annotate only text based web pages and not images or image fragments. Other projects concern the visualization of the rare scanned documents; they do not allow users to add annotations. All these applications do not contain recommender

systems; we think that they are important to users who perform difficult tasks. Recommender systems can be developed basing on the traces of user actions. We search to annotate images of manuscripts or fragments of them, by using a web application accessible by web browsers and providing services to annotate manually and semi-automatically the manuscripts images.

In the next section, we describe a model of online archive to manage digitized documents of ancient manuscripts. This model can handle also annotations, users, and their access rights, interactions, and preferences. Our model contains a tracing layer, an assisting system and a collaborative system that permits professional and expert users to work in groups in order to complete difficult tasks.

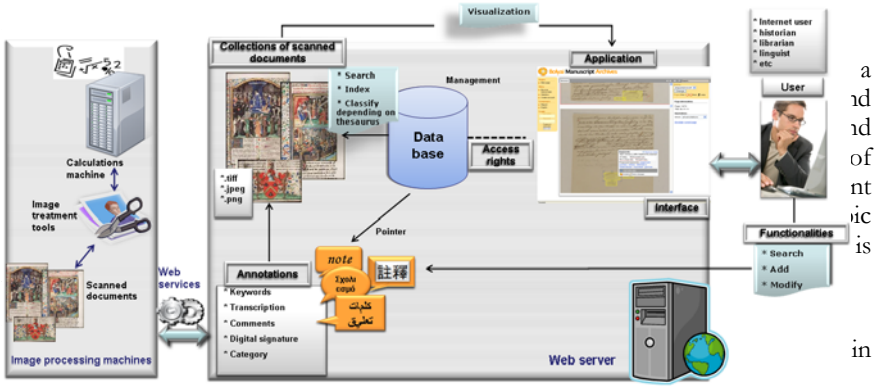


Figure 1. The system is composed of the following modules:

- *Collections of scanned documents*: digitized images of the manuscripts structured in collections/ sub-collections depending on different factors (date, theme, etc.), the images of manuscript pages are of different forms (JPEG, PNG, TIFF…) and stored in a rational database. Each image has three versions: *thumbnail* for a low resolution, *access* for an intermediate resolution, and *real* for high resolution. The advantage of this system is to provide users with *thumbnails* when they ask for an image review, and with an intermediate resolution of the image *access* when the user does not precise the image version. Each page image may contain many *document units*. Document units represent image fragments, whole images or collections. An image fragment *document unit* can be defined by the user and has coordinates linking it to its original image. However, these coordinates change in correspondence to the image size and keep the document unit in the same place in all image versions.

- *Annotations*: many types of annotations are defined (keywords, comments, transcriptions, digital signatures, administrative or descriptive metadata) with a possibility to add other types dynamically; annotations are created by users and associated to document units. We plan to add OAI-PMH and other metadata standard (Dublin Core, TEI P5, METS…) compliant annotation import/export.

- *Application*: a Web application that is accessible through web browsers

- *Web services*: we implemented a web service based image processing tool architecture. An identified user with sufficient rights can initiate an image processing treatment on a collection. An image processing treatment starts a session and lets the user to go on with his work. On the personal space interface of the *Application*, the user can consult his image processing sessions and validate the

results of the finished ones (e.g. word-spotting). In this way the system can carry out long lasting treatments.

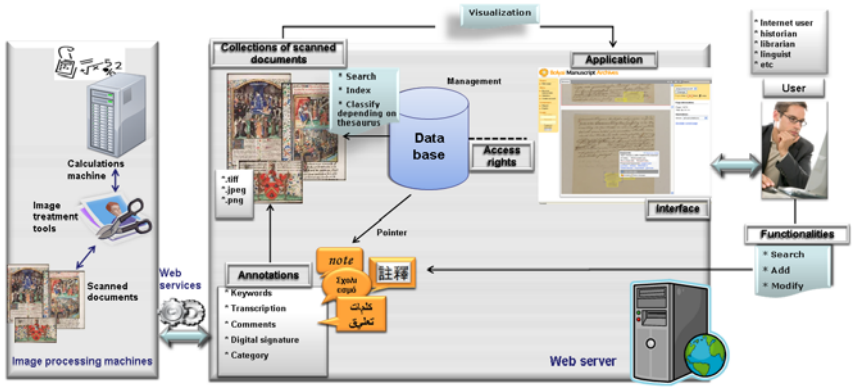


Figure 1. ***General view of ARMARIUS***

-Functionalities: Online services (research, visualization, annotation, manual transcription, adding comments…). Users have to identify themselves to access manuscript images or digitized collections, and to search images depending on their annotations, transcriptions or other metadata. Users can also annotate manually the documents with new keywords, transcriptions or comments enriching the documents with additional information.

-Users: Users in ARMARIUS are classified into three categories: non-identified users (like internet users) who can only see a demo selected by the administrator about the collections, registered users belong to groups, and the administrators who manage the system and upload images.

-Database: contains image collection information, metadata, users, users groups, and access rights.

Access rights concern collections and their content, annotations and user groups. They are defined by the system administrator. Access rights precise view and modification rights.

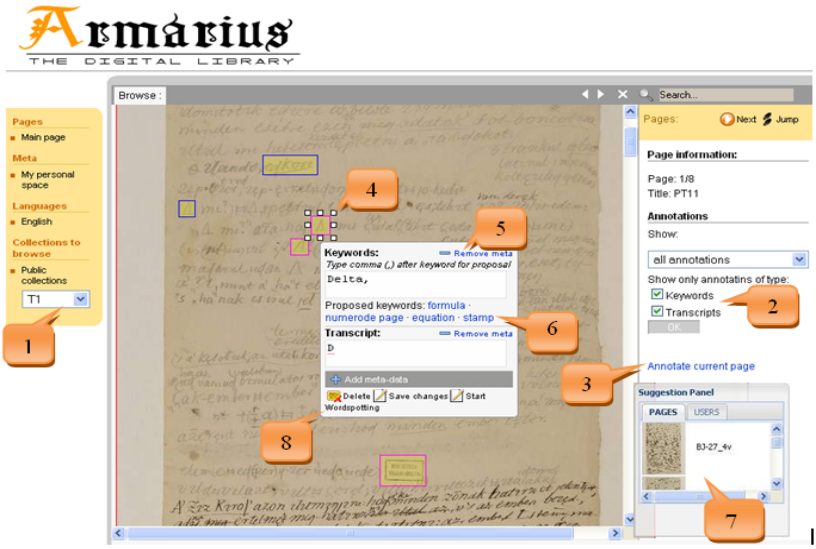
ARMARIUS registered users can create a personal space, which contains collections or pages chosen by the user. This space provides the user with all the functionalities that he needs to accomplish his work. Personal collections reflect a topic of user interests; user may organize his/her own defined collections into certain categories.

Some users' actions are registered in the system *as traces*. These traces are integrated in an assisting system in order to help other users during the search and the annotation. We created a task and object model of the application and chose from these models the tasks to be traces and the objects they modify.

3.2 ARMARIUS Functionalities

ARMARIUS application permits users to annotate remotely images and image fragments, to define objects (collections, document units, pages, keywords or other users and groups). Image annotation is done by users in a collaborative environment that permits users to work together. The collaborative system provides users with tools to see the work of other users, and to add comments on the document units of other persons or on their own work. Users can also modify the

annotations of other document units if they have the permission in their groups. ARMARIUS offers also a recommender system based on users experiences.



Select a collection; 2- filter the annotation/transcription; 3- create a new annotation on the current page; 4-draw a rectangle around a fragment; 5- add various metadata; 6- use the keywords that are suggested by the recommender system; 7- suggestion panel; 8-launch a session of Word Spotting with the selected fragment

Figure 2. *Annotation Screenshot in ARMARIUS*

Document and user management

First of all, ARMARIUS is an image management system. It enables to upload images, creates automatically different versions, and enables their classification into collections and sub collections and views. A view is composed of images from a given collection in a given order. An image can belong to several views but to only one collection. Users usually navigate through views.

Users belong to groups; a user can belong to several groups. The rights are defined between collections and groups. As an image belongs to one collection the rights are easy to be calculated. If a user is member of different groups which have different rights on a given image, the user rights are added. Annotations can be private, restricted to group members or public visible to anyone.

Each identified user has a *Personal space* on which he can select the collections to view or to annotate, set preferences, start a search, consult the image processing sessions, see favorites, manage personal views.

Annotating

Users can annotate new image fragments (document units) by creating a rectangle representing the document unit then adding annotations. The annotation is done via the web browser interface. Once the document unit has been created, a dialog box appears allowing user to add keywords and/or transcriptions or other metadata as shown in **Figure 2**.

The list of metadata types can be extended dynamically, and for each metadata type we can specify an export translation in order to be exported according to a given metadata standard syntax.

Another way to annotate documents in an assisted manner is the use of image processing services. Some of the image processing tools are implemented as asynchronous services. For example, the word spotting in ARMARIUS helps in finding the fragments that are similar to the fragment précised by the user within a collection. It is handled as sessions: a user can select a fragment of document and launches the word spotting session that could take hours to be finished. On the main page, the user has a list of current image processing sessions. A session can be in different states: launched, finished, validated. A finished session can be visualized: its results are shown and the user is asked to validate them. She can modify, delete or accept results one by one, by page, or for the whole collection. In Figure 3 we present the results of a word spotting session in ARMARIUS.

Figure 3. *The use of word spotting in ARMARIUS*

Recommender system for annotations

For better understanding of how a recommender system works: let us imagine the next scenario when a user (Anny) connects to ARMARIUS web application. If this is her first connection and she has no account in the database, she will be able to see just the demo collections proposed by the system administrator. If Anny is a regular user who has an ID, she will be able to search and browse the collections that are permitted to the groups she belongs to. After her login, the tracing system begins to register her actions (connect, search, browse, chose, create…), besides the objects that are affected by these actions.

The recommender system is based on a tracing layer that tracks the actions of identified users during their work session; traces are stored in a relational database together with the affected objects (collections, pages, metadata…). In order to create an experience based user assistance we have to go through different phases.

We consider that user manipulates objects through procedures thus the use of the system is traced according to Trèfle♣ model (Egyed-Zsigmond and al. 2003). For this tracing we need to formalize these procedures as well, so a user-task-model is built. Firstly we need to build an object model, which is composed of collections, pages, document units and metadata, a tree structure which holds the relations

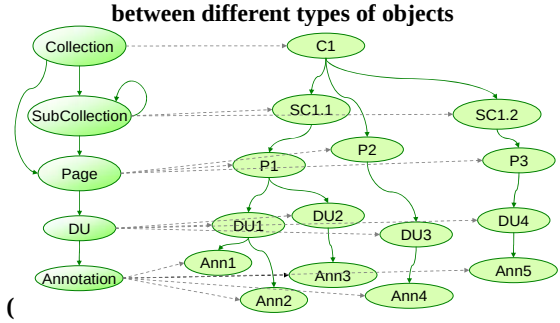


Figure 4). The instantiation of this model gives us the actual structure of a collection.

The next step in constructing our recommender system was deciding which tasks to trace in order to create the observation model. These will be the tasks, which, together with the manipulated objects will create our experience and knowledge pool. Some user tasks like registration or signing in are not relevant in future recommendations, whilst other tasks, mainly those which manipulate objects in the collection's structure, will be the basis of the user assistance methods.

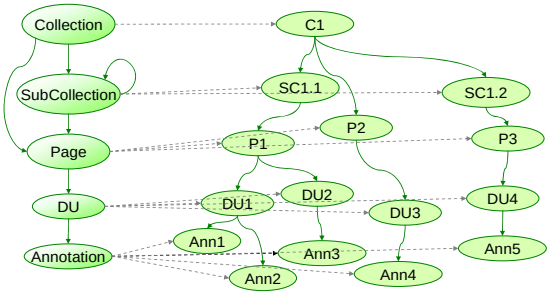


Figure 4. ***Simplified object model and an instance fragment***

In order to be able to compare these user traces among themselves, we need to formalize the distances between objects, as well as the distances between different types of tasks. This process of comparison always involves two objects of similar or different types. In case of comparing two different types of objects or two metadata we rely only on the structure specified by the collection. The similarity of the two objects will be a number, equal with the distance between the two nodes in the collection's structure tree, representing the two objects. In case of similar object types, such as document units, pages and collections, besides these physical distances we also need to take in consideration the content similarity between them. That is, for example, in case of two document units, we not only take in consideration whether they are on the same page or in the same collection, but we analyse the metadata associated with them, and their similarity. Based on these we can put together a similarity measure for calculating content distances between two different document units:

$$DC_{du}(du1, du2) = \min\left(\sum_{\substack{mdi \in A1 \\ mdj \in A2}} D_{md}(mdi, mdj) \cdot D_{mt}(mt_{mdi}, mt_{mdj}) \right)$$

where *du1, du2* are the two document units to compare, *A1* and *A2* are the metadata associated with them, D_{md} is the physical distance between two metadata, and D_{mt} tells us whether the two metadata are of the same type or not.

Analogically we c[illegible] document units positioned on the[illegible]

$$DC_p(p1, p2) = \min\left(\sum_{\substack{du_i \in B1 \\ du_j \in B2}} DC_{du}(du_i, du_j) \right)$$

where *p1* and *p2* are the pages to be compared, *B1* and *B2* are the document units associated with the pages.

The user tracing itself is a process of determining the action of the user and inserting this together with the user identifier, the objects manipulated and other parameters such as sate, and session identifier into the system's database. Each type of task has different number and type of parameters.

Upon identification, the recommender system will exploit the stored traces, to provide step by step assistance to user actions. At different points of the navigation the system provides different types of recommendations, one of them relies only on the distances between objects and it is used to suggest similar pages and document units while browsing, by recommending the objects closest to the currently browsed page or document unit, to suggest similar metadata in case of document annotation, or to add the current collection into the user's personal space. For example if the user creates an annotation on a page, and adds the metadata "paragraph" to the document unit, and there is already a document unit annotated with the metadata "paragraph", "number" and "section" the system will recommend the words "number" and "section" to the user.

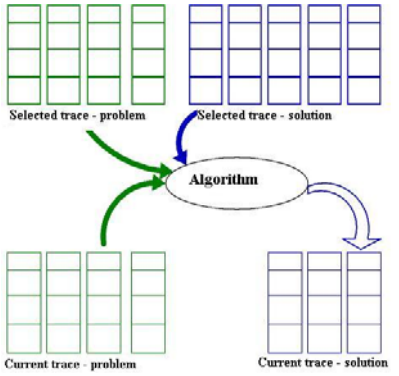


Figure 5. ***Finding and reusing similar cases***

A more complex algorithm is used for other user assisting functionalities, such as recommending search results, or recommending similar pages based on user actions. During this procedure we cut the traces in reusable and adaptable episodes according to the case based reasoning paradigm. Each case is composed of two parts: problem and solution and the current trace of the user is our current problem. The system searches for similar problems among these cases, and suggests

the solution part of the case, or those objects that were manipulated by the tasks of this solution.
This similarity is calculated based on the distances of the tasks' parameters.
For example: Anny performs a search with the keywords "stamp" and "round", she selects a page from the result set, views the suggested pages, those pages that the system finds similar to the current page. She navigates to one of these pages, and she selects a document unit on this page. From the list of similar document units she again navigates to a page holding one of the recommended document units. All of these tasks and objects are traced by the system, so when another user sings in and performs a search with the word "stamp", the recommender algorithm will extract all those pages that Anny had viewed earlier and show him.

4 Conclusion and future works

In this article, we have presented a "living" digital archive model of ancient manuscripts: ARMARIUS, with a web application prototype. Our proposed model could be also used in other domains (scientific, medical…). In this paper, we treated the following problems:

-How to represent the digitized document in a living archive? This concerns annotations creation, documents structuring, the storage in a database, and a model to access the documents. And the need for a user collaborative work space to create a discussion environment concerning the collections.

-User assistant integration, this assistant proposes different help types during the annotation, the document search, and the creation of a personal space.

-The assisting system, the collaborative system, and the discussion space are important to annotate the manuscripts. Especially that this type of document requests lot of explanations and image treatment tools are not very efficient.

In our future works, we aim at integrating technologies of type "push" and RSS to track the evolution of certain documents, themes, collections, etc. we aim also to offer and to assist the discussion space, hence users can confront straightforwardly their points of view about a document. We are also interested in developing a module that allows users to exchange messages between each others, to discuss the collections and their content. Finally, we are concerned to enrich the system with image treatment tools that are especially adapted to this type of manuscripts (other than word spotting).

References :

« CONTENTdm Digital Collection Management Software by OCLC. » http://www.contentdm.com/ (Accessed at 5 may 2008).

«Dublin Core Metadata Element Set. » http://dublincore.org/documents/dces/ (Accessed at 5 may 2008).

«Fedora Commons- Home. » http://www.fedora-commons.org/ (Accessed at 13 may 2008).

«Le protocole OAI et ses usages en bibliothèque ». http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm (Accessed 14 may 2008).

«MARC STANDARDS. » http://www.loc.gov/marc/index.html (Accessed at 13 may 2008).

«Metadata Object Description Schema: MODS (Library of Congress) ». http://www.loc.gov/standards/mods/ (Accessed at 13 may 2008).

«The UVic Image Markup Tool Project. » http://www.tapor.uvic.ca/%7Emholmes/image_markup/index.php (Accessed at 5 may 2008).

«Vivarium. » http://cdm.csbsju.edu/ (Accessed at 13 may 2008).

Agosti M., Benfante L., Orio N. (2003). "IPSA: A digital archive of herbals to support research." In Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access, Lecture Nottes in Computer Science. Springer Berlin/Heidelberg , vol. 2911/2003, 253-264

Bottoni P. And al. (2004). MADCOW: a multimedia digital annotation system." In Proceedings of the working conference on Advanced visual interfaces. Gallipomi, Italy: ACM.55-62

Cabanac G., Chevalier M., Chrisment C., Julien C. (2007). « An Original Usage-Based Metrics for Building a Unified View of Corporate Documents. » In Database and Expert Systems Applications, vol. 4653/2007, Lecture Notes in Computer Science. Springer Berlin/Heidelberg. 202-212

Calabretto S., Jean-Marie P., Bozzi A. (1998). "BAMBI: système de gestion de manuscrits anciens pour historiens." Les bibliothèques numériques, vol 2. 31-50.

Champin P.-A. (2003). « Ardeco: an assistant for experience reuse in Computer Aided Design. » In From structured cases to unstructured problem solving episodes. Trondheim (NO). 287-294.

Egyed-Zsigmond E., Mille A., Prié Y. (2003). « Club ♣(Trèfle) : A Use Trace Model. » In Case-Based Reasoning Research and Development, vol. 2689/2003, Lecture Notes in Computer Science. Springer Berlin/Heidelberg. 1056

Hilbert, D. M. and Redmiles, D. F. 2000. Extracting usability information from user interface events. ACM Comput. Surv. 32, 4 (Dec. 2000), 384-421.

Kahan J., Koivunen M. (2001). « Annotea : an open RDF infrastructure for shared Web annotations." In International World Wide Web Conference. Hong Kong: ACM. 623-632

Le Bourgeois F., Emptoz H. (2007). « DEBORA : Digital AccEss to Books of the RenAissance. » International Journal on Document Analysis and Recognition. vol. 9 193-221.

Mladenic D. (1999). « Text-Learning and related intelligent agents: a survey.» IEEE Intelligent Systems. 44-54.

Semeraro G., Basile P., Degemmis M., Lops P. (2007). «Content-based recommendation services for personalized digital libraries. » In Digital Libraries: Research and Development. Vol. 4877/2007. 77-86.

Session 2

Analyse Textuelle et Interaction avec les Documents

Méthodes d'extraction de termes basées sur une combinaison d'indicateurs

A Combination of Indicators to a Better Term Extraction

Férihane KBOUBI (1), Anja HABACHA (2), Mohamed BEN AHMED (3)

(1) RIADI, ENSI, Manouba, Tunisie
Ferihane.kboubi@riadi.rnu.tn

(2) RIADI, ENSI, Manouba, Tunisie
Anja.habacha@ensi.rnu.tn

(3) RIADI, ENSI, Manouba, Tunisie
Mohamed.Benahmed@riadi.rnu.tn

Résumé. Dans cet article, nous proposons d'évaluer un ensemble d'indicateurs d'extraction de termes. Cet ensemble regroupe les indicateurs les plus connus : *tf*, *idf*, *tf.idf*, *En* et *Pd*. Ce travail a pour objectif d'étudier l'influence de ces indicateurs pour les tâches d'extraction et de sélection de termes. Notre objectif est de prouver que la combinaison de ces indicateurs peut augmenter la performance de l'extraction des termes et conduire à sélectionner les termes les plus pertinents. Nous commençons par présenter et analyser les résultats de l'évaluation. Ensuite, nous discutons des méthodes de combinaison possibles et nous en proposons quelques unes. Toutes les méthodes proposées ont été testé sur un système de classification basé sur l'extraction de terme, et ont contribué à l'augmentation de ces performances. En effet, nous avons réalisé une amélioration qui a atteint 10%.
Mots-clés. Combinaison d'indicateurs, extraction de termes, mesure de pertinence des termes clés, tf, pertinence au domaine.

Abstract. In this paper, we propose to evaluate a set of the most used term selection measures *tf*, *idf*, *tf.idf*, *En* and *PD*. This work aims to study the influence of these indicators on the term extraction and selection tasks. Our goal is to prove that the combination of these indicators could improve the relevance of the selected terms. We start by presenting and analyzing the evaluation results. Then, we discuss the possible ways of combination and propose some combination methods. All of them enhanced the performance of a classifier system based on term extraction. Indeed we achieve an enhancement that reaches 10%.
Keywords. Combination of indicators, term extraction, key terms selection, tf, domain pertinence.

1 Introduction

L'extraction des termes clés est une étape très importante et utile pour de nombreuses applications, citons à titre d'exemples l'extraction de connaissances, la classification thématique de documents, la recherche d'information, etc.

Dans la littérature, les méthodes d'extraction de termes se classent en deux grandes catégories à savoir: les méthodes statistiques (Calvo et Ceccatto,2000) (Vercoustre et al,2006) (Zou et al,2003)(Bellot et El-Bèzel,2001) et les méthodes linguistiques (Forest et Meunier,2004) (Hernandez,2005) (Lame,2002) (Radhouani et al,2006). Les méthodes statistiques se basent sur le calcul de la fréquence des termes et sont indépendantes des langues. Pour cette raison, elles sont souvent qualifiées par multilingues. De l'autre côté, les méthodes syntaxiques sont basées sur l'analyse du rôle grammatical des mots. De ce fait, elles ont besoin de processus syntaxiques qui sont spécifiques à la langue traitée.

Dans cet article, nous nous intéressons aux méthodes statistiques d'extraction de termes, en vue d'une classification thématique de documents. Ceci en partant du principe qu'une bonne méthode d'extraction de termes pourrait améliorer considérablement le résultat de classification. "Le modèle de classification contient trios parties: le traducteur, l'extracteur de caractéristiques et le classifieur. Le traducteur capture les données et les converties dans un format approprié pour le traitement automatique. L'extracteur de caractéristiques (appelé aussi récepteur, filtre de propriétés, détecteur d'attribut ou préprocesseur) extrait les informations présumées importantes à partir des données d'entrées. Le classifieur utilise ces informations pour assigner les données en entrées à une des catégories. Généralement, la tâche d'extraction des caractéristiques est beaucoup plus problématique que la classification." (Duda,1976).

Dans la plus part des cas, avant la sélection des termes clés, les méthodes statistiques commencent par une étape préliminaire de prétraitement qui peut être aussi utilisée dans les méthodes syntaxiques. Cette étape procède à la filtration des termes. Par exemple, dans (Lame,2002), les auteurs ont éliminé tous les termes contenant des caractères non alphabétiques ou des caractères en majuscules. Il ont aussi éliminé les termes qui n'appartiennent pas à l'une des catégories grammaticales suivantes : nom, verbe, adjectif et adverbe. La sélection des termes clés se fait en utilisant des indicateurs statistiques. Les indicateurs les plus utilisés dans la littérature sont *tf* (fréquence du terme), *tf.idf* (fréquence du terme * inverse document frequency), *En* (entropie) et *Pd* (pertinence au domaine).

Selon (Hernandez,2005), il y plusieurs conclusions contradictoires sur la pertinence comparative de ces indicateurs. Quelques travaux (Koo et al,2003) ont montré que *tf* génère des résultats meilleurs que *tf.idf*. Alors que d'autres travaux (Maedche et Staab,2004) recommandent l'utilisation de *tf.idf*. Il n'y a aucun travail de combinaison proposé. En se basant sur ces observations, nous proposons dans cet article d'évaluer ces indicateurs dans l'intention d'étudier leurs comportements pour la sélection de termes et leurs pertinences pour la classification thématique de documents. Yiming Yang et Jan O. Pedersen ont présenté une étude (Yang et Pedersen,1997) similaire à la nôtre. Les auteurs se sont intéressés à cinq indicateurs tous différents des nôtres. En effet ils ont évalué les indicateurs suivants : la fréquence à l'intérieur des documents (*DF*), le gain d'information (*IG*), le CHI statistique (χ^2), l'information mutuelle (*MI*) et le Term Strength (*TS*). Les auteurs ont utilisé uniquement deux critères pour l'évaluation qui sont la précision et le taux de réduction de l'espace des caractéristiques.

La suite de cet article est organisée comme suit : dans la section 2, nous commençons par présenter les indicateurs auquel nous nous sommes intéressés. Ensuite, dans la section 3, nous expliquons notre méthode d'évaluation. Dans la section 4, nous présentons et analysons les résultats de l'évaluation. Dans la section 5, nous discutons des résultats obtenus et nous proposons quelques directives pour choisir une bonne méthode de combinaison. En réalité, nous avons proposé deux méthodes de combinaison qui toutes les deux ont amélioré les résultat du système de classification. Finalement, dans la section 6, nous résumons notre contribution et discutons des éventuelles perspectives de notre travail.

2 Les indicateurs de sélection de termes

Nous avons expérimenté cinq indicateurs classiques que nous avons tenté de combiner :

- *tf*: la fréquence du terme, qui peut représenter, selon la formule utilisée, soit le nombre d'occurrences du terme dans le document, soit le nombre d'occurrences du terme dans tous les documents du domaine, soit le nombre d'occurrence du terme dans tous les documents du corpus.

- *idf* (inverse document frequency): classiquement, cet indicateur représente la répartition du terme dans le corpus et est calculé par la formule suivante: $idf = \log\left(\frac{N}{n_i}\right)$

Où N est le nombre de documents dans le corpus et n_i est le nombre de documents contenant le terme *i* dans tous le corpus.

Nous avons remarqué qu'avec la même équation nous pouvons définir d'autres formules ayant des significations différentes. Notons *idf1* l'indicateur représentant la distribution du terme dans le domaine. Avec *idf1*, N représente le nombre de documents dans le domaine et n_i représente le nombre de documents contenant le terme *i* dans ce domaine. Avec *idf1*, les termes ayant les plus petites valeurs sont les plus pertinents.

Notons *idf2* l'indicateur représentant la pertinence du terme par rapport à un domaine. Avec *idf2*, N représente le nombre de documents dans tous les autres domaines (sauf celui en cours d'étude) et n_i représente le nombre de documents contenant le terme *i* dans tous les autres domaines. Avec *idf2*, les termes ayant les plus grandes valeurs sont les plus pertinents. Finalement, notons *idf3* l'*idf* classique.

- *tf.idf*: remarquons qu'en littérature l'*idf* n'a pas été utilisé tout seul mais [illegible] fréquence du terme pour former l'indicateur *tf.idf*.

$$tf.idf = tf \times \log\left(\frac{N}{n_i}\right)$$

D'une façon analogue à *idf*, nous formons trois formulas à partir de *tf.idf* qui sont: *tf.idf1*, *tf.idf2* et *tf.idf3*.

- Entropie: cet indicateur représente la répartition du terme *i* dans le corpus. Il est calculé comme suit : $En = -\sum_{i,x} r(i)_x \log\left(r(i)_x\right)$;

$r(i)_x = \frac{tf_i(x)}{TF_i}$ où représente le nombre d'occurrences du terme *i* dans le document *x* divisé par le nombre d'occurrences du terme *i* dans tout le corpus.

- La pertinence au domaine *Pd*: soit $(D_1, D_2 \ldots D_n)$ un ensemble composé de *n* domaines. La pertinence du terme *t* par rapport au domaine D_i est calculée comme suit : $PD(t, D_i) = \frac{P(t/D_i)}{\sum_{i=1..n} P(t/}$ Où une estimation de $P(t/D_i)$: $P(t/D_i) = \frac{freq(t \text{ in } D_i)}{\sum_{i=1..n} freq(t \text{ in } D_i)}$

Cet indicateur permet de sélectionner les termes apparaissant uniquement dans le domaine considéré.

3 Notre méthode d'évaluation

Pour évaluer ces indicateurs, nous avons utilisé un système de classification base sur l'algorithme de simple Bayes. Nous avons mené plusieurs expérimentations, où à chaque fois nous appliquons le classifieur sur l'ensemble de termes sélectionnés par chaque indicateur. Ensuite, nous avons évalué les résultats de classification pour déduire une estimation de la pertinence de chaque indicateur utilisé. La classification est basé sur une phase d'apprentissage automatique qui permet la constitution de l'ensemble des termes clés de chaque domaine. Cette phase d'apprentissage débute par une étape de prétraitement durant laquelle les mots vides sont supprimés et les restants sont lemmatisés. Pour la lemmatisation, nous avons utilisé l'algorithme de Porter (Porter, 1980). Après cette étape, nous avons procédé à la sélection des termes clés en utilisant l'un de ces indicateurs : *tf*, *idf*, *tf.idf*, *En* ou *PD*. Les termes clés sélectionnés sont utilisés pour la classification du corpus de test. Ceci nous permet d'évaluer la pertinence de chaque indicateur en fonction de son résultat de classification. Pour l'évaluation, nous avons utilisé les critères classiques : *précision*, *recall* et *F-mesure*.

$$recall = \frac{nombreDocumentsBienClassés}{nombreDocumentsCorpus}$$

$$precision = \frac{nombreDocumentsBienClassés}{nombreDocumentsClassés}$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}$$

Nous avons ajouté en plus deux autres critères qui sont le taux de confusion et le taux de rejet. Ceci nous permet d'analyser le comportement de chaque indicateur en fonction du type d'erreurs qu'il produit.

$$ConfR = \frac{nombreDocumentsMalClassés}{nombreDocumentsCorpus}$$

$$RjR = \frac{nombreDocumentsRjetés}{nombreDocumentsCorpus}$$

4 Evaluation

Dans cette section, nous commençons par présenter les corpus d'apprentissage et de test que nous avons utilisés. Ensuite, nous discutons des résultats obtenus.

4.1 Corpus d'apprentissage et de test

Notre corpus est formé par un ensemble de documents web au format HTML. Il contient 709 documents distribués en cinq domaines à savoir: architecture des ordinateurs, bases de données, réseaux, ontologie et systèmes multiagents (MAS). Après l'étape de prétraitement, le corpus contient 28005 termes distincts. Nous avons divisé le corpus en deux parties : une base d'apprentissage contenant 80% des documents et une base de test contenant les 20% restant. Notre principal objectif n'est ni d'évaluer des méthodes de classification, ni d'évaluer les indicateurs de sélection de termes. Notre objectif est plutôt d'étudier et de comparer les caractéristiques des indicateurs de sélection de termes à travers leurs résultats de classification.

4.2 Les résultats de l'évaluation

Le tableau 1 présente les performances obtenues des différents indicateurs. Les diagrammes des figures 1, 2, 3, 4 et 5 représentent graphiquement ces résultats. Les figures 1 et 2 montrent respectivement les valeurs de recall et de précision obtenues pour les différents indicateurs de sélection de termes. Nous avons remarqué que *tf*, *tf.idf3* et *En* sont les indicateurs qui produisent les meilleurs résultats. *idf3* quant à lui a généré de mauvais résultats, mais lorsqu'il a été combiné avec *tf* nous avons obtenu de bonnes performances qui ont dépassé celles obtenues avec l'indicateur *tf*.

Table 1: *Les résultats de performances des différents indicateurs*

	tf	idf1	idf2	idf3	tf.idf1	tf.idf2	tf.idf3	PD	En	tf.PD	tf.En
Precision	74.82	44.93	96.30	44.44	58.99	96.30	78.99	96.30	81.29	77.70	74.82
Recall	74.29	44.29	55.71	2.86	58.57	55.71	77.86	55.71	80.71	77.14	74.29
ConfR	25.00	54.29	2.14	3.57	40.71	2.14	20.71	2.14	18.57	22.14	25.00
RjR	0.71	1.43	42.14	93.57	0.71	42.14	1.43	42.14	0.71	0.71	0.71
F-mesure	74.55	44.60	70.58	5.37	58.77	70.58	78.42	70.58	80.99	77.41	74.55

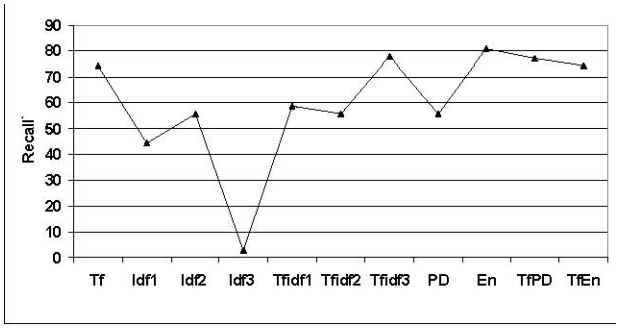


Figure 1. *Les valeurs de Recall obtenues*

Nous avons aussi constaté que *idf2*, *tf.idf2* et *Pd* ont exactement les mêmes performances. Ceci peut être expliqué par le fait que ces indicateurs ont la même signification malgré qu'ils soient calculés par des formules différentes. En effet, ils estiment tous la pertinence d'un terme par rapport à un domaine. Puisque la valeur de *idf2* tend vers l'infini quand le terme est pertinent au domaine, alors les termes sélectionnés par *idf2* ne devront par être modifiés lorsque nous utilisons *tf.idf2*. Nous n'avons pas le même problème avec *Pd* parce que son domaine de valeurs est [0..1]. $P(t, D_i) = 1$ si le terme *t* apparaît uniquement dans les documents du domaine *Di*.

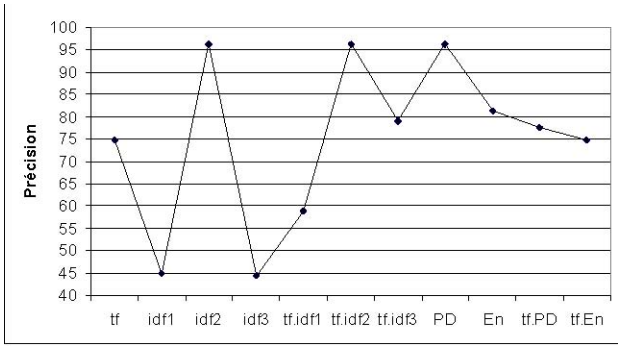


Figure 2*: Les valeurs de Précision*

La figure 3 présente les taux de confusion produits par tous les indicateurs. Nous avons constaté que les taux les plus faibles sont produits par les indicateurs estimant la pertinence par rapport à un domaine (*idf2*, *idf3*, *tf.idf2* et *Pd*). Ainsi, nous pouvons supposer que la pertinence au domaine a un effet positif pour la minimisation des erreurs de confusion.

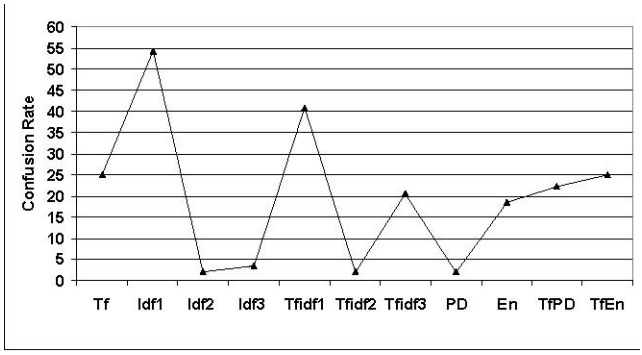


Figure 3. *Les taux de confusion obtenus*

La figure 4 présente les taux de rejet produits par les indicateurs de sélection de termes. Nous remarquons que les taux les plus faibles sont ceux obtenus par les indicateurs qui estiment la fréquence et la répartition des termes (*tf*, *idf1*, *tf.idf1*, *tf.idf3*, *En*, *tfPd* et *tfEn*). Alors, nous pouvons conclure que la fréquence et la répartition des termes a un effet positif sur la diminution des erreurs de rejets.

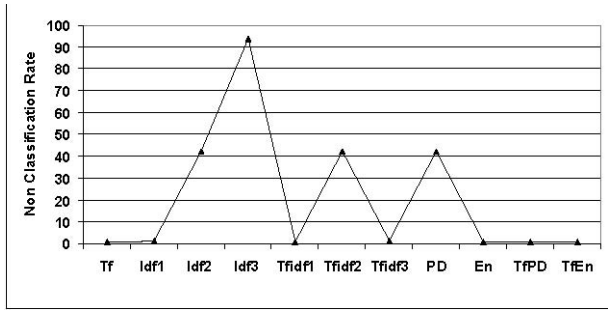


Figure 4. *Les taux de rejet obtenus*

La figure 5 montre les valeurs obtenues de la F-measure. Nous avons observé que les meilleurs résultats sont ceux produits par les indicateurs *En*, *tf.idf3* et *tf.Pd* (respectivement égaux à 80.99%, 78.42% et 77.41%).

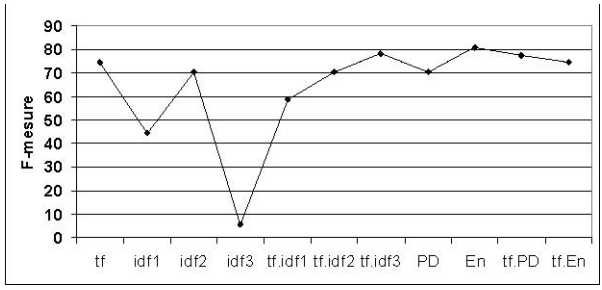


Figure 5. *Les valeurs de F-mesure*

5 Les méthodes proposées de combinaison

À partir des deux figures 3 et 4 nous pouvons constater qu'aucun de ces indicateurs n'a minimisé ni maximisé en même temps les deux types d'erreurs. Chaque indicateur donne ou bien un taux de confusion faible et un taux de rejet élevé ou bien l'inverse : un taux de confusion élevé et un taux de rejet faible. Cette observation est très importante dans la mesure où elle nous permet d'espérer améliorer les résultats de classification en combinant ces indicateurs. Une bonne combinaison peut en effet permettre de diminuer en même temps les erreurs de confusion et de rejet.

Le problème est maintenant de déterminer une bonne formule de combinaison. Pour ceci devons nous utiliser tous les indicateurs? Ou devons nous nous contenter de choisir un indicateur de chacun des deux groupes (un indicateur de pertinence au domaine et un indicateur de fréquence)

Pour répondre à ces questions nous devons étudier la similarité entre ces indicateurs de sélection de termes.

5.1 Directives pour le choix d'une méthode de combinaison

La similarité entre les indicateurs de sélection de termes peut être estimée en comptant le nombre de termes identiques retenus par ces indicateurs. Si deux ou plusieurs indicateurs permettent de sélectionner le même ensemble de termes, alors nous pouvons utiliser l'un d'entre eux et ignorer les autres. Par contre, si chaque

indicateur forme un ensemble de termes différents des autres alors nous pouvons qualifier ces indicateurs de « complémentaires » et nous devons les utiliser tous dans la méthode de combinaison.

Soient:

- M_1 : l'ensemble des indicateurs représentant la pertinence au domaine {*idf2*, *idf3*, *tf.idf2* et *PD* },

- et M_2 l'ensemble des indicateurs représentant la fréquence et la répartition des termes {*tf*, *idf1*, *tf.idf1*, *tf.idf3*, *En*, *tf*PD et *tf*En}.

Pour déterminer la similarité entre les indicateurs de M_1 (respectivement M_2), nous avons calculé le taux de termes identiques sélectionnés par chaque couple d'indicateurs. Les tableaux 2 et 3 représentent les résultats obtenus pour les deux ensembles M_1 et M_2.

À partir du tableau 2, nous avons remarqué que le taux d'intersection entre les ensemble de termes de *idf* 2 et *Pd* - $IR(idf\,2, PD)$ - est 100%. Ceci signifie que les deux ensembles générés par *idf* 2 et *Pd* sont identiques. Par conséquent, nous pouvons choisir l'un d'entre eux puisqu'ils se remplacent mutuellement. Nous proposons alors de choisir *Pd* puisqu'il génère un meilleur résultat que *idf2* quand il est combiné avec *tf*. Nous avons raisonné de la même façon avec chaque couple d'indicateurs. Les indicateurs ayant un taux d'intersection supérieur à 50% sont considérés comme complémentaires. Ceci est appliqué notamment à et $IR \geq 50\% IR(idf\,2, tfidf\,2)$ et $IR(tfidf\,2 IR(tfidf\,2, PD)$.

Ainsi, à partir de M_1 nous avons retenu uniquement *idf3* et *Pd*. Nous pouvons éliminer *idf3* puisqu'il donne des mauvais résultats.

Table 2. *Similarités entre les indicateurs de M1*

	Idf3	*Tf.idf2*	*PD*
Idf2	8	100	100
Idf3	-	8	8
Tf.idf2	-	-	100

Table 3. *Similarités entre les indicateurs de* M_2

	Idf1	*Tf.idf1*	*Tf.idf3*	*En*	*Tf.PD*	*Tf.EN*
tf	28	65	64	29	68	87
Idf1	-	55	22	14	26	28
tf.idf1	-	-	44	23	51	62
Tf.idf3	-	-	-	12	78	53
En	-	-	-	-	20	33
Tf.PD	-	-	-	-	-	60

Nous avons fait de même pour M_2. Ainsi, nous avons retenu uniquement *idf1*, *tf.idf3* et *En*. Nous pouvons éliminer *idf1* puisqu'il a des faibles performances. Par conséquence, les ensembles d'indicateurs susceptibles d'améliorer les résultats de classification sont M_1={*Pd*} et M_2={*tf.idf3, En*}.

Pour prouver expérimentalement ces conclusions, nous avons entrepris quelques expérimentations pour combiner ces trois indicateurs.

5.2 Les méthodes de combinaison proposées

Les tâches de sélection et d'attribution des poids aux termes clés, dans les documents, sont similaires aux tâches de sélection et d'attribution des poids dans le domaine de datamining. Le problème de sélection du meilleur sous-ensemble de caractéristiques ou d'attributs représente une part importante de la conception d'un bon algorithme d'apprentissage. Des caractéristiques non pertinentes peuvent dégrader considérablement les performances de ces algorithmes. Dans la littérature, il y a plusieurs conclusions contradictoires concernant l'efficacité des indicateurs de sélection de termes, aucune combinaison n'a été proposée (Hernandez, 2005).

C'est pour cette raison que nous proposons de concevoir une stratégie qui combine ces indicateurs afin d'améliorer leurs performances en évitant leurs faiblesses respectives. Dans cette section, nous proposons deux méthodes de combinaison basées sur les trois indicateurs : *Pd*, *tf.idf* et *En*.

La première méthode consiste à combiner les trois indicateurs en utilisant les deux formules suivantes

- nous utilisons cette formule pour la sélection du sous-ensemble des termes pertinents. $tf.idf3 \cup En \cup PD$
- $tf.idf3 \times En \times PD$ nous utilisons cette formule pour calculer le poids de chaque terme.

Avec la deuxième méthode, nous avons d[illegible] ensemble des termes clés d'un domaine *i* en utilisant la formule suivante : $(U_i \cap F_i)_{iiii}$. Où I_i est l'ensemble de termes occurant dans tous les documents du domaine *i*. U_i représente l'ensemble de termes résultant de l'union des ensembles des termes les plus fréquents dans chaque document du domaine *i*. F_i représente l'ensemble de termes les plus fréquents dans le domaine *i*.

Nous avons évalué ces deux méthodes de combinaison avec le même procédé que nous avons utilisé avec les indicateurs classiques. Nous testons les ensembles de termes sélectionnés par chaque méthode de combinaison avec le système de classification. Puis nous étudions les variations de ces performances pour déduire une appréciation sur la réelle pertinence des termes sélectionnés par chaque méthode de combinaison.

La première méthode de combinaison nous a donné un *recall* = 83.51%, un taux de confusion *Conf*R=15% et un taux de rejet *NCR*=1.42%. Ainsi, avec cette méthode nous obtenons une amélioration de l'ordre de 2.8% par rapport au meilleur résultat obtenu par les indicateurs classiques (le résultat de *En*).

La deuxième méthode de combinaison nous a donné un *recall* = 90.71%, un taux de confusion *Conf*R=8.57% et un taux de rejet *NCR*=0.71%. Ce qui nous donne une amélioration de 10% par rapport au meilleur résultat obtenu par les indicateurs classiques.

Nous avons obtenu des résultats encourageant avec les deux méthodes de combinaison, spécialement avec la deuxième qui nous a permis d'atteindre une amélioration de 10%.

6 Conclusion

Dans cet article, nous avons évalué les indicateurs de sélection de termes les plus utilisés, à savoir : *tf*, *idf*, *tf.idf*, *En* et *PD*. Nous avons montré que les mesures représentant la pertinence au domaine permettent de diminuer les erreurs de confusion. Analogiquement, nous avons montré que les mesures représentant la fréquence et la répartition des dermes permettent de minimiser les erreurs de rejets.

Nous avons aussi prouvé que la combinaison de ces indicateurs peut augmenter leurs capacités de discrimination entre les termes. Nous avons proposé quelques méthodes de combinaison et nous avons obtenu à chaque fois des résultats de classification meilleurs que ceux obtenus par le meilleur indicateur individuel. En effet, nous avons obtenu une amélioration de 2.8% avec la première méthode et de 10% avec la deuxième méthode.

Références :

Calvo, R.A. and Ceccatto, H.A.. (2000). Intelligent document classification. *Intelligent Data Analysis*, vol. 4 N°5.

Duda, R.O., Hart, P.E. (1976). *Pattern Classification and scene analysis*, by John Wiley & Sons, Inc, ISBN 0-471-22361-1, USA.

Vercoustre A.M., Fegas, M., Lechevallier, Y., Despeyroux, T. (2006). Classification de documents XML à partir d'une représentation linéaire des arbres de ces documents. EGC : pp 433-444.

Forest, D., Meunier, J.G., (2004). Classification et catégorisation automatiques: application à l'analyse thématique des données textuelles, *JADT: 7ème Journées internationals d'Analyse statistique de données Textuelles.*

Hernandez, N. (2005). Ontologies de domaine pour la modélisation du contexte en recherche d'information, PhD Thesis, Institut de recherche en informatique de Toulouse, université Paul Sabatier, 06 décembre

Lame, G. (2002). Construction d'ontologie à partir de textes une ontologie du droit dédié à la recherche sur le web, PhD Thesis, Ecole des mines de Paris.

Maedche, A., Staab, S. (2004). Ontology Learning, Handbook on Ontologies, S Staab, R. Stubers (Eds.), pp 173-190

Koo, S., Lim, S.Y., Lee, S.J. (2003). Building an Ontology based on Hub Words for Informational Retrieval, *In Proceedings of the IEEE/WIC International Conference on Web Intelligence.*

Porter,.M.F. (1980). An algorithm for suffix stripping, *Program*, vol. 4, N°3, pp 130−137

Radhouani, S., Maisonnasse, L., Lim, J.H., Le Thi-Hoang-Diem, Chevallet, J.P. (2006). Une Indexation Conceptuelle pour un Filtrage par Dimensions, Expérimentation sur la base médicale ImageCLEFmed avec le méta thésaurus UMLS, *Actes de CORIA*, Lyon, France

Yang, Y., Pedersen, J.O, (1997). A comparative study on feature selection in text categorization. *In Douglas H. Fisher, editor, Proceedings of ICML-97, 14th International Conference on Machine Learning, Nashville*, US, pp 412–420

Zou, Q., Chu, W.W., Morioka, C., Leazer, G.H., Kangarloo, H. (2003). IndexFinder : A Method of Extracting Key Concepts from Clinical Texts for Indexing, *Annual Symposium on biomedical and health informatics*, pp. 763–767

Bellot, P., El-Bèze, M.. (2001). Classification et segmentation de texte par arbre de décision, *Technique et science informatiques*. Vol. 20 – n° 1, pp 107-134

Réflexions sur les liaisons entre passages d'ouvrages en philosophie

Reflections on links between texts of philosophy

Benoit HUFSCHMITT(1)

(1)LASELDI, antenne de Montbéliard, Université de Franche-Comté, France.
Hufschmitt.benoit@wanadoo.fr

Résumé. A partir d'un exemple : les références appelées dans les commentaires du Discours de la Méthode de Descartes, cet article tente de spécifier les intertextes générés autour d'un texte rationnel et argumentatif rédigé en langue naturelle. Les réponses fournies par les théoriciens de l'intertextualité étant inadaptées, cet article propose une spécifisation adaptée à ce genre de texte.
Mots-clés. Hypertexte, intertexte, philosophie, argumentation, ontologie.

Abstract. From an example : the references called in notes added to *le Discours de la Méthode* of Descartes, this paper tries to specify the *intertextes* generated around a rational and argumentive text written in natural language. The answers provided by theorits of intertextuality supply being unsuited, this paper proposes a specification adapted to this kind of text.
Keywords. Hypertext, intertext, philosophy, argumentation, ontology.

1 Introduction

Depuis quelques années, nous avançons, très empiriquement, dans une réalisation éditoriale qui se complique petit à petit (présentation d'un prototype fonctionnel lors de CIDE 10 à Nancy en juillet 2007) (HUFSCHMITT, B. et alii 2007). Il s'agissait à l'origine de proposer, sous la forme d'un hypertexte, une version numérique du *Discours de la Méthode* de Descartes mettant à disposition, par des liens, les références (puis les ouvrages eux-mêmes) qui auraient pu inspirer Descartes dans son écriture du *Discours* : d'abord celles qui sont explicites, évidemment (le *Discours* n'en contient qu'une !), mais aussi et surtout celles qui sont vagues, implicites ou cachées. Pour les dégager, il avait été décidé non de les découvrir par nous mêmes, mais de les puiser dans les notes explicatives et critiques des éditeurs du *Discours*. Très vite, un document éclipsa les autres : la volumineuse édition commentée de Gilson (en 1926) avec ses 1213 notes sur 400 pages suite aux 78 pages du *Discours* lui-même, citant 4367 passages à travers plusieurs centaines d'ouvrages (331 notices hors le corpus cartésien, lui-même cité en toutes ses parties) (DESCARTES, 1967).

Cette récollection fut complétée par l'intégration des références venant d'autres jeux de notes dans d'autres édition du *Discours* Edition de F. Alquié (DESCARTES, 1963), de G. Rodis-Lewis (DESCARTES, 1966), de J. M. Fataud (DESCARTES, 1973), de A. Bridoux (DESCARTES, 1953), de J. Simon (DESCARTES, 1842) …,

mais le résultat fut décevant, les références non pointées par Gilson étant très rares. La seule exception fut le jeu de notes d'un nommé Minos quelques 20 ans avant Gilson (DESCARTES, 1907) dans son édition du *Discours* : strictement limitées à des citations, sans aucun commentaire, qui concernent indifféremment l'antériorité et la postérité du *Discours*.

La limitation de notre recollection aux seules sources ou influences du Discours devint très vite insoutenable. Les notes de Gilson nous montrèrent que les cas litigieux concernant les sources étaient nombreux : un grand nombre des sources ne sont que des hypothèses, sans que l'on puisse savoir si Descartes a lu le texte lui-même, en a reçu un commentaire ou une compilation, l'a étudié lors d'un cours, voire en a simplement entendu parler. Par ailleurs, si les références internes au corpus cartésien et antérieures au *Discours* peuvent, à la rigueur, être assimilées à des sources, ce n'est pas possible pour celles qui sont postérieures, alors que leurs fonctionnalités explicatives sont très proches. De plus, Gilson a souvent besoin de consolider une source par un commentaire tardif qui explique le texte en même temps qu'il justifie la source postulée. Enfin, et sans même prendre leçon de l'intertextualité, nous ne pouvions ignorer que la considération des seule sources nous liait à des choix philosophiques, certes honorables, mais discutables. Cela nous poussa alors à étendre notre récollection de documents à l'ensemble de ce que les notes citaient.

Nous n'avions jusque là retenu que les références issues des notes, mais au nom de quoi aurions-nous exclu les références présentes dans un ouvrage, ou plutôt un extrait d'ouvrage, qui explique un passage ou la totalité du *Discours* (une biographie raisonnée par exemple qui consacre un chapitre à l'élaboration du *Discours*).

Enfin, mais en comprenant clairement que nous glissions vers un autre registre, nous nous trouvions quelquefois indécis pour savoir si l'ouvrage expliquait le *Discours*, ou si c'était le *Discours* qui illustrait ou cautionnait une thèse de l'ouvrage. En acceptant ces références, nous comprenions bien que nous nous engagions dans une quête sans fin (alors que les notes et études attachées au *Discours* forment un ensemble important mais a priori maîtrisable) ; et pourtant, en un sens large, n'est-ce pas encore le commenter, voire l'expliquer, que de dégager sa postérité quelle qu'en soit la genèse (ce qui d'ailleurs nous renvoyait aux notes de Minos !).

Pour conclure le tout, nous ne pouvions non plus ignorer que d'autres modes de liaisons inter-textuelles méritaient d'être retenus pour comprendre un passage de l'ouvrage. Avant tout les propositions issues de traitements lexicologiques, des plus directes (recherche de chaînes de caractères) aux mieux élaborées (en termes de lemmatisation ou de classification automatique). Ces travaux nous conduisirent enfin, à l'origine pour des objectifs de validation, à considérer les *index rerum* concernant le corpus cartésien comme autant de systèmes, intellectuellement établis, de classification des passages, ce qui, en nous limitant aux descripteurs pointant, entre autres, un passage du *Discours*, nous fournissait autant de liens avec les autres passages du corpus pointés par le même descripteur.

Tout ceci nous amena finalement à présenter une application très éclectique de tous les moyens découverts au hasard pour lier, à des fins explicatives, les passages d'un ouvrage de philosophie (voire un texte à visée rationnelle en général) à des passages d'autres textes. Un peu de rigueur réflexive s'imposait alors afin de classer cet amas de références. Nous pensâmes d'abord à utiliser les analyses de la transtextualité, un peu passées de mode il est vrai, mais qui, dans un contexte certes beaucoup plus large que le nôtre, ont fourni des distinctions devenues communes ;

mais ces analyses se révélant peu pertinentes pour notre projet, nous nous orientâmes alors vers la récollection des différents cas d'appels à références qui servent à un lecteur (de philosophie) soucieux de travail réflexif, rationnel et critique, que l'on peut abstraire de l'étude des notes au *Discours de la Méthode* ou plus simplement lire dans les méthodologies d'explications de texte.

Par ailleurs, enfin, il allait de soi que l'application de cette classification sur les cas concrets devait pouvoir se prêter, au mieux, à un traitement procédural, et s'intégrer alors à des formalisations informatiques (peut-être en termes de langages de description et d'ontologies au sens de l'informatique), permettant d'espérer une automatisation, au moins partielle de la mise en catégories (en classes approximativement partitives plutôt) des ouvrages appelés.

Pour ce petit travail, notre objectif est donc de proposer une partition des ouvrages jugés pertinents à être appelés par liens hypertextuels afin d'expliquer, commenter, critiquer… en un mot comprendre, un ouvrage philosophique et pouvant conduire, partiellement au moins, à une normalisation documentaire et informatique des liens à des fins de vérification, puis de production.

Si nous prenons un ouvrage de philosophie comme illustration, ces remarques concernent tous les ouvrages, que nous nommons, par extension au cas de la philosophie, doctrinaux, qui se caractérisent de la sorte : énoncés dans la langue naturelle, ils possèdent cependant un cœur conceptuel fixé selon des règles de type terminologique qui posent elles-mêmes un ordre ontologique (ce que sont les choses, les concepts et leurs relations) ; ils énoncent des thèses organisées entre elles selon des schémas démonstratifs et des règles de validation ; ils possèdent des principes et des concepts premiers qui forment le soubassement de l'ensemble. En conséquence, ils sont soumis à l'obligation de clarification et de justification, ainsi qu'à la consistance formelle. On peut les réfuter, soit de manière intra-doctrinale (selon la règle du jeu mise en place), soit de manière extra-doctrinale, en refusant de valider les principes ou concepts premiers, ou en refusant la règle du jeu elle-même.

Après avoir constaté que l'intertextualité pour le type d'ouvrages qui nous intéresse (argumentatif rationnel) diffère de ce qui est nommé telle dans l'analyse littéraire, puis cerné les fonctionnalités attendues de l'appel à un autre texte, lors de l'explication d'un texte de philosophie, nous chercherons à tirer de ces résultats une typologie utilisable.

Pour éviter les quiproquos nous proposons de fixer ainsi les deux mots :

Intertexte : l'ensemble des textes liés au texte étudié (ouvrage, texte ou passage). C'est le sens générique habituel (vs. Celui de Genette). Dans notre illustration, au début du *Discours de la Méthode*, on y trouvera par exemple : Montaigne : *Essais*, Livre II, ch. XVII, De la praesumption ; Sénèque ; *De Vita beata*, XII, 1 ; etc.

Texte noyau (*ouvrage*, ou *passage noyau)*, le texte étudié lui-même, ce qui est l'occasion de l'intertexte (ici le *Discours* ou un de ses passages). Nous évitons donc l'orientation des analyses selon un réseau acentré.

A ces deux termes, nous ajouterons par la suite : *hypertexte* au sens informatique du terme (vs le sens posé par Genette) ; *cotexte* : la partie de l'intertexte qui peut avoir influencé l'écriture de l'ouvrage noyau ainsi que tout ce qui peut donner au lecteur idée de ces influences, *doctritexte* ou *intertexte du corpus doctrinal* : la partie de l'intertexte qui peut être considérée comme faisant partie intégrante du corpus doctrinal dans lequel est inclus l'ouvrage noyau ; *métatexte* (suivant, ici, la terminologie de Genette) : la partie de l'intertexte qui a comme objet d'étude ou d'analyse (que ce soit explicatif, commentatif ou critique) l'ouvrage ou un de ses

passages ; *extratexte ou réappropriation* : la partie de l'intertexte qui utilise le texte noyau comme moyen pour une fin autre que la compréhension ou l'analyse de ce noyau.

2 Transtextualité philosophique

2.1 Points communs avec l'intertextualité

Notre projet dans son cheminement empirique s'inscrit, à première vue, dans le cadre des pratiques de l'intertextualité après les années 60. Nous utiliserons, pour ce qui suit, la belle introduction de Sophie RABAU, et les extraits d'ouvrages qu'elle présente (RABAU, S. 2002).

1- Le texte (ici le *Discours*) ne peut se comprendre que par ses liens à d'autres textes qui le clarifient, le prolongent, voire le contredisent.

2- L'intertextualité est donc d'abord le résultat d'un travail de commentaire, une généralisation du travail érudit de la théorie littéraire et de la littérature comparée en ce qu'elle s'affranchit de limites de la problématique des sources et des liens historiques (RABAU, S. 2002, p. 45).

3- Cet environnement ne doit en effet pas être entendu comme ce qui aurait pu inspirer Descartes, selon la problématique des sources, mais plutôt ce qui rend clair, à nous lecteurs, un contexte devenu bien brumeux, ou ce qui évite un contresens rétrospectif (vocabulaire, évidences pour l'époque…), ainsi que ce qui analyse, approfondit, explique, critique, interprète l'ouvrage. L'originalité, le calcul de la dette envers des prédécesseurs, les éléments biographiques… n'importent que s'ils aident à la compréhension et à l'analyse du texte.

4- De même l'auteur s'efface derrière l'œuvre, et l'écriture derrière les lectures.

2.2 Divergences d'avec l'intertextualité

Mais la présomption que le texte tend à la rationalité discursive, et l'objectif de lecture en vue d'analyse critique d'un discours qui prétend à l'établissement d'une vérité partagée en raison (i.e. respectueux des règles minimales d'une inter-discursivité rationnelle : non contradiction, rigueur terminologique, limitations des intuitions, inter-connexion cohérente des propositions soutenues…) présente l'intertextualité sous un jour particulier.

1- Le texte noyau n'est pas réduit à une abstraction et transformation de son intertexte (RABAU, S. 2002, p. 21), encore moins à une mosaïque de citations, ainsi que le soutient Julia Kristeva (KRISTEVA, J. 1969, p. 145 ; RABAU, S. 2002, p. 57), mais est posé comme unité autonome (ou plutôt inscrit dans l'autonomie d'un corpus doctrinal) qui intègre, digère, son environnement intertextuel, à l'image d'un organisme envers ses aliments. C'est l'intertexte lui-même qui apparaît comme une mosaïque de fragments, puisque chaque élément y est lu dans le cadre du texte noyau [1].

2- Le primat du lecteur conduit en général à concevoir qu'un ensemble intertextuel est toujours subjectif, provisoire et relatif, dans des approches herméneutiques et interprétatives, que l'on s'en accommode ou que, comme RIFFATERRE, on cherche à s'en dégager (RIFFATERRE 1983, p. 161 ; RABAU, S. 2002, p 161). Les exigences de lecture que nous posons font du lecteur un lecteur

[1] Les conséquences en sont importantes : alors que la lecture du texte noyau (suivie et raisonnée) est adaptée à la lecture traditionnelle d'un ouvrage papier (le transfert sur écran pose avant tout le problème de la simulation optimisée de cette lecture), la lecture de l'intertexte, fragmenté, est mal adapté aux documents papiers et beaucoup mieux aux fichiers informatiques. C'est la raison de notre travail éditorial sur le *Discours*.

universel, ou qui cherche à l'être, ce qui marque tout élément de subjectivité comme une défaillance, et repousse le relatif et le provisoire au seul niveau des règles mêmes de la discursivité rationnelle.

3- L'extension du domaine intertextuel reste certes indéfini, jamais achevé (BUTOR, M. 1968, p. 111, RABAU, S. 2002, p. 213) : une nouvelle lecture pouvant toujours compléter l'ensemble de compréhension, mais on doit, par défaut, considérer que ce développement est asymptotique ; la mise à jour de textes nouveaux qui bouleverseraient la compréhension est une hypothèse envisageable, mais vaine car indéterminable dans ses effets tant qu'elle n'est pas advenue.

4- Il n'est pas exclu que coexistent cependant des ensembles intertextuels différents pour une même œuvre, mais c'est de manière régulée, selon des hypothèses explicites sur la genèse des oeuvres de philosophie, ou selon des orientations interprétatives différentes. Notre texte est ici un bon exemple : il est difficile de trancher sur le point de savoir si le *Discours* doit être lu, ainsi que le demande Descartes, comme une rupture générale envers toute philosophie précédente (les textes antérieurs étant tout au plus des repoussoirs, des hasards de rencontre, voire des cheminements parallèles), ou si, au contraire, à l'instar de Gilson, il faut révéler toutes les influences, de la scolastique tardive en particulier, que Descartes cacherait (volontairement ou non) dans son ouvrage. Mais en ces cas, c'est l'union mêmes des choix interprétatifs qui fait l'intertexte (ainsi que l'indique Ioannis Kanello (KANELLO, I. 1999, p. 53).

5- L'auteur a une importance première, non pas parce que lui seul détient les clés de bonne interprétation, mais parce qu'il est la marque la plus immédiate de l'unité doctrinale qui se développe dans le texte noyau (nonobstant toutes les nuances à apporter que l'on ne peut développer ici). Ainsi, l'ensemble du corpus qu'il a construit forme le fond le plus solide des éléments intertextuels que l'on peut développer. Plus généralement, une échelles de pertinence de participation à l'intertexte peut être construite : les documents de l'auteur lui-même et ceux qu'il cite, puis ce qui concerne les biographes et les disciples, les commentateurs ensuite, ceux qui enrôlent la doctrine dans leur propre doctrine (les utilisateurs) enfin.

6- En conséquence, la problématique des sources ne peut être éliminée dans notre approche, car la source fournit un indice de pertinence privilégié, Ce qui est éliminé ici est la problématique de l'exclusivité explicative des sources. On peut noter que les théories intextuelles ont rapidement abandonné la thèse radicale du rejet de l'étude des sources (FUSILLO 1991, p. 19, RABAU, S. 2002 p. 88).

7- L'intertextualité au service des catégories ou des genres textuels, ou encore celle au service de l'invention ne concernent guère cette approche, puisque le genre est déjà fixé a priori (texte argumentatif rationnel) et puisque la seule invention éventuellement convoquée est interne à ce genre (régulée par les règles de la rationalité discursive).

2.3 Analyse de Genette

L'analyse princeps de Genette (GENETTE G. 1982, p. 7-12) à propos de ce qu'il nomme le transtextuel : "tout ce qui le <le texte> met en relation, manifeste et secrète, avec d'autres textes", nous est, en particulier, à peu près inutile :

1- L'intertextuel, selon Genette, concerne la coprésence de texte, selon la citation, le plagiat ou l'allusion. Il n'y a pas de citations explicites dans le Discours (une référence seulement), pas de plagiat non plus, mais les allusions y sont nombreuses, parfois vagues, souvent proches de la citation. Seulement, la plus grande partie des sources intertextuelles dégagées par Gilson occupe une zone mal déterminée : réminiscences non reconnues comme telles relevant de la citation

vague ou de l'allusion savante. Genette identifie ce domaine à celui des sources, mais l'incertitude même que nous avons sur la nature de ces relations, nous permet de concevoir cet intertexte de manière plus vague, incluant tout autant les sources que les énoncés précurseurs, peut-être inconnus par l'auteur mais qui peuvent l'avoir atteint indirectement (par l'enseignement, les conversations, les présences allusives dans des textes lus), ou encore les lieux communs de l'époque, de la culture, du milieu, les idées qui flottent dans l'air du temps etc. Il faut y ajouter aussi, sans craindre le paradoxe, tous les textes (en général postérieurs) qui exposent au lecteur contemporain ces "évidences" du XVIIième oubliées depuis, par exemple, sur le vocabulaire : le *Dictionnaire universel* de Furetière (1690) ou encore le récent *Petit Glossaire des Classiques français du XVIIième* de Huguet (1907). A la fois pour éviter les quiproquos et tenir compte de ce que nous venons d'indiquer, nous délaissons l'usage restrictif du mot intertextuel selon Genette, que nous conserverons dans son sens générique habituel et nous proposons de parler ici de *cotexte*, voire de conserver le mot de "sources", à condition de l'entendre dans un sens large.

2- Le paratexte : "titre, sous-titre, intertitres ; préfaces, postfaces… notes marginales, infrapaginales… illustrations ; prière d'insérer, bandes, jaquette… qui procurent au texte un entourage (variable) et parfois un commentaire, officiel ou officieux…", est, dans notre perspective, un fourre-tout malcommode. Par exemple, les titres, sous-titres… ont une fonction d'analyse intra-textuelle importante, celle du plan ; alors que les illustrations, préfaces, notes… ont plutôt une fonction explicative métatextuelle (selon la terminologie de Genette infra).

3- La métatextualité est identifiée au commentaire qui "unit un texte à un autre texte dont il parle, sans nécessairement le citer (le convoquer), voire à la limite sans le nommer". Ainsi le début du *Discours* sur le bon sens serait en position métatextuelle vis à vis des *Essais* de Montaigne, ou bien la dernière partie de la *Logique de Port Royal* vis à vis de la seconde partie du *Discours*. Cette catégorie est particulièrement inadaptée à notre propos car elle mélange deux approches qu'il faut minutieusement distinguer : les écrits portant sur le *Discours* (pour prendre notre exemple) se répartissent d'une part en textes explicatifs et textes de commentaires, et d'autre part en textes qui utilisent le discours dans leur propre logique argumentative (comme exemple, repoussoir, argument d'autorité), qui sont donc externes à l'environnement doctrinal du texte noyau. Entre ces deux types, on trouvera le texte interprétatif partial et orienté vers des fins étrangères à ce que l'on sait (ou suppose) des fins originaires. Nous conserverons, pour notre part, le mot *métatexte* pour la partie de l'hypertexte qui est clairement explicative et commentative, c'est-à-dire qui analyse, prolonge ou critique le texte noyau dans l'environnement doctrinal de ce texte noyau, à des fins de perfectionnement éventuellement, et proposons *extratexte* ou *réappropriation* pour ceux qui suivent leurs propres fins doctrinales.

4- L'hypertextualité, qui est la raison pour laquelle Genette produit son analyse, consiste à imiter, sous forme de parodie, pastiche… un autre texte (l'hypo-texte). Le livre polémique de Daniel (DANIEL, G. 1690) qui caricature la philosophie cartésienne est un (rare) bon exemple… jamais cité par nos commentateurs. Le soin que Genette met à distinguer ce cas du précédent explique pourquoi ces catégories sont mal adaptées à notre domaine de travail. Ce qui différencie, dit-il le métatexte de l'hypertexte est que le premier relie deux textes de genres différents (le texte en position de méta est explicatif ou argumentatif), alors que les deux textes sont de même genre pour le second. Or, dans notre cas, les textes liés ont toujours l'un et

l'autre valeur argumentative et relèvent donc du même genre. Ne conservant pas cette catégorie, nous pouvons donc conserver le sens habituel d'*hypertexte*.

5- Enfin, l'architextualité, qui unifie les textes selon le genre, ne nous concerne guère, ainsi que nous l'avons déjà mentionné, alors qu'une branche importante des recherches de l'intertextualité s'est focalisée sur cette orientation.

Il faut mentionner, pour être honnête, que Genette précise que ces catégories ne sont pas disjonctives, et que, au contraire, il faudrait mieux les considérer comme les diverses facettes de toute relation d'un texte à un autre (voire d'un texte à lui-même) ; il faut aussi ajouter que le succès de cette analyse a débordé les raisons pour lesquelles elle a été posée : dégager la relation hypertextuelle entre ouvrages de type littéraire. Ces deux remarques confortent les réticences que nous avons manifestées précédemment concernant l'usage de telles catégories, ce qui nous pousse donc à proposer notre propre typologie pour les textes du genre précis qui nous intéresse : les textes argumentatifs à prétention rationnelle.

2.4 Résultats provisoires

Nous concluons donc que les différenciations dans les corpus intertextuels construites dans le cadre de l'analyse littéraire par les théoriciens de l'intertextualité ne sont pas adaptées aux corpus que nous étudions. Nous posons la nécessité d'élargir le domaine des sources (cotexte) à tout ce qui explique, pour nous lecteurs, le contexte de production de l'ouvrage ; le paratexte ne nous concerne pas : certains de ses éléments participent du commentaire (introductions, notes…), d'autres s'intègrent dans une catégorie limite de l'intertexte : les indications de plan et d'organisation du texte noyau ; l'hypertexte (selon Genette : le texte qui en pastiche un autre) se fond, selon notre approche, dans le métatexte (le texte qui en discute un autre) ; le métatexte, par contre, relève de deux genres différents : le texte qui commente dans le cadre de la doctrine du texte noyau (qui, peut-on dire, le finit au sens de la finition d'un produit) et celui qui l'utilise dans d'autres perspectives (qui en fait un matériau de travail). Enfin, la question des genres littéraires (architexte) est extérieure à notre problématique.

Par ailleurs si nous partageons totalement la thèse que le domaine intertextuel est relatif au lecteur, nous évitons toute approche subjective et relativiste, étant donné que le caractère rationnel argumentatif du texte noyau est l'enjeu des lectures que nous considérons. C'est la raison pour laquelle, nous rejetons aussi l'idée que le texte noyau éclate et se dissémine dans son intertexte. Au contraire, une partie majeure de l'intertexte a comme fonction de travailler (justifier ou critiquer) l'unité doctrinale dans laquelle le texte noyau évolue, et ce sont les éléments de l'intertexte lui-même qui se manifestent en fragments.

Reste alors à valider (ou contester) mais aussi approfondir, ces résultats, ce que nous envisageons de deux manières ; en appliquant ces analyses à des intertextes réels (ici les références indiquées dans les jeux de notes liés à diverses éditions du *Discours de la Méthode*) ; mais aussi en les confrontant aux exigences usuellement demandées dans l'explication et le commentaire rationnels d'un texte (ici un texte de philosophie). Faute de place, notre approche sera extrêmement schématique.

3 Appel à des textes pour l'analyse d'un texte argumentatif rationnel (philosophique).

3.1 Tour d'horizon des éditeurs de notes

Les notes explicatives apportées au *Discours de la Méthode* sont assez différentes d'une édition à l'autre, en nombre et en contenu, cela vaut également pour les références qu'elles fournissent.

1- Dans l'édition princeps d'Adam et Tannery, il n'y a qu'une note, la référence fournie en marge par Descartes lui-même : à Harvey, *De Motu Cordis*.

2- Les notes dans l'édition de la Pléiade, ou celles de l'édition ancienne de Jules Simon, très rares, ne concernent que des sources qui éclaircissent une allusion de Descartes. Ce sont toujours d'autres textes cartésiens (*Météores, Dioptrique, Géométrie, Traité de l'Homme* …) à la seule exception du *Grand Art* de Lulle.

3- Les références fournies par les notes de G. Rodis Lewis sont avant tout soit des appels à un dictionnaire d'époque (le *Furetière*), soit des indications d'autres passages du corpus cartésien, soit enfin des renvois à *La Vie de Monsieur Descartes* de Baillet (à entendre comme une extension du corpus cartésien). Ces derniers ont un caractère particulier car les passages cités sont présentés en annexe dans cette édition. Quelques notes citent des auteurs antérieurs pour des clarifications de vocabulaire ou levées d'obscurité du texte (Saint Thomas sans références précises, le *Manuel* d'Epictète, les *Essais* de Montaigne, *Le grand Art* de Lulle, et évidemment Harvey).

4- Celles que fournit J. M. Fataud, sans éliminer totalement les appels au corpus cartésien, sont beaucoup plus riches en appel à des ouvrages antérieurs (à des fins de levée d'obscurités) et surtout à des commentaires à valeur explicative, en raison de l'orientation pédagogique et peu spécialisée de l'édition proposée.

5- Celles de F. Alquié relèvent presque exclusivement du cotexte. L'essentiel des références concerne le corpus cartésien (95 références, dont 6 internes au *Discours*, sur un total de 114) ; Pour les 19 restantes, nous trouvons surtout des sources (Charron, Montaigne, Epictète, Sénèque, mais aussi Porphyre, Saint Anselme, Saint Thomas, et encore le père François), et des clarifications de vocabulaire (*Dictionnaire de Furetière*, *de l'Académie*, Corneille) ; il n'y a que deux citations de commentateurs contemporains (Milhaud et Lefèbvre) qui donnent des informations contextuelles ; enfin, étranger au cotexte, Kant est cité 2 fois.

6- Les notes de Gilson, imposantes, nombreuses et complexes, semblent très hétérogènes. Environ 120 pour l'antiquité, 130 pour le moyen-âge, 220 pour le XVIième relèvent d'auteurs dont Descartes aurait pu s'inspirer ; on en trouve 280 pour le XVIIième dont 30 pourraient être des sources, 70 relèvent des premiers biographes et 100 des dictionnaires et ouvrages littéraires ; s'y ajoutent 370 références pour les commentateurs postérieurs, la plupart après 1880 ; enfin 2200 références concernent le corpus cartésien lui-même dont 380 pour un autre passage du *Discours* et 300 pour sa version latine, la *Dissertatio*. Ces références concernent environ 150 notices antérieures au *Discours* et 100 postérieures, la plupart de la fin du XIXème et du début du XXème (l'actualité de la recherche au moment où Gilson écrit). A peu près tout le corpus cartésien est utilisé, dont environ 200 lettres différentes. Leibniz, Spinoza ou Malebranche sont assez peu cités. Enfin, il y a très rarement appel à des grands auteurs postérieurs : Kant deux fois seulement. Il est intéressant d'observer (nous ne pouvons l'illustrer, faute de place) que ces références suivent généralement le même schéma, avec plus ou moins de parties manquantes :

Définitions et clarifications de vocabulaire (faisant appel aux dictionnaires et à la littérature d'époque, ou à des compilations récentes), ou de concepts (faisant appel au corpus cartésien et à sa version étendue : biographes, premiers disciples).

Considérations contextuelles qui fournissent, outre des indices extraits du corpus, des témoignages d'époques ou des ouvrages récents d'étude sur l'époque.

Prolongements, éclaircissements, variations à l'intérieur du corpus cartésien, puis des textes les plus proches (biographies d'époque, ouvrages des premiers cartésiens).

Analyses d'objections et réponses appelant le corpus cartésien et d'autres textes contemporains.

Indication des sources, nombreuses et documentées. Il faut savoir que Gilson a toujours soutenu que la rupture cartésienne envers la scolastique tardive était surfaite. Ces sources sont parfois doublées, voire remplacées par des ouvrages récents qui les ont exposées (par extraits, récollections, synthèses…).

Commentaires et explications développés à propos du *Discours* ou de la doctrine en général, avant tout les analyses les plus récentes (par rapport à 1926).

Très rarement enfin, postérité du passage ou du thème. L'exemple le plus marquant consiste dans les deux appels à Kant : le *je transcendantal* à propos du *Cogito* et l'argument ontologique à propos de la troisième preuve de l'existence de Dieu.

7- Enfin, les notes de Minos, très surprenantes pour un lecteur contemporain, consistent à indiquer les références qui lui viennent en tête à la lecture du *Discours*, sans qu'il y ait aucun filtre méthodologique. Il propose 818 références, dont 220 sont de courts extraits issus du corpus cartésien (89 fois le *Discours*), les autres références sont très éclectiques, dont beaucoup surprenantes : le Baghavata, Le Cardinal de Richelieu, Confucius, Napoléon… mais il cite aussi Locke, Pascal, Marc Aurèle, Malebranche, Bossuet, Spinoza, Taine, Condillac… On notera une présence importante d'Epictète : 83 citations et de Platon : 186 citations.Cette approche, peu appréciée par les philosophes, ressemble beaucoup à ce que propose l'intertextualité littéraire. Ces références relèvent à l'évidence de l'analogie, sauf, peut-être, lorsque c'est le corpus cartésien qui est appelé.

En laissant de côté le cas de Minos, ces différents exemples marquent assez nettement une même manière d'envisager l'intertexte : fournir de quoi rendre clair ce qui est obscur. Ce n'est que secondairement, dans le cas de Gilson, quand les notes sont étendues, que sont indiquées des références qui appuient une argumentation, ou qui font des objections, puis y répondent. Et ce n'est, hors Minos, que très exceptionnellement qu'est mentionnée la portée du texte noyau hors de la doctrine qu'il supporte. Tous s'accordent donc à reconnaître que le corpus doctrinal est le noyau de l'intertexte de clarification (ce que nous nommerons *doctritexte*), et personne ne discute le besoin de convoquer parfois des dictionnaires d'époque, ni de clarifier une allusion. Par contre, l'extension de cet intertexte est très variable et les appels aux sources, aux commentateurs, à la biographie… sont plus contestés. Enfin, il semble que Gilson suive une méthode qui n'est pas sans rappeler la méthode proposée généralement pour réaliser une explication de texte.

3.2 Appel selon les règles habituelles données pour une explication.

L'approche méthodologique d'analyse d'un texte philosophique est assez claire et peu contestée, elle est même, pour quatre de ses grandes parties, institutionnalisée dans une épreuve d'explication de texte du baccalauréat. On questionne en effet le candidat sur l'organisation du texte (et son idée générale), sur le vocabulaire, sur les justifications, puis on lui demande les développements critiques qu'il peut apporter.

L'organisation

L'énonciation isolée de cette première exigence explicative pousse à distinguer l'intertexte qui lui est lié de l'ensemble métatextuel. On y trouve, hors les liaisons entre marques explicites de chapitres (ici parties, dont celles de la version latine) et les indications du préambule (ce que Genette place dans le paratexte), les textes qui explicitent cette organisation. On peut espérer qu'une classification automatique du lexique permette la distinction des parties principales [2].

Les clarifications

Les clarifications de termes ou d'expressions (voire de phrases, de paragraphes..) concernent tout ce qui est susceptible d'améliorer la compréhension, ce que nous entendons précisément soit comme enrichissements des dénotations, soit comme ajouts de liens (de choses, de concepts, de mots) validés à travers des compositions de vocabulaire. On y trouve avant tout les définitions et descriptions, mais aussi, les exemples, les levées d'obscurités. On peut distinguer :

1- Des clarifications de vocabulaire ou de concepts, qui permettent de lever l'ignorance ou l'équivoque du lecteur sur le sens des mots ou des expressions :

Clarifications de vocabulaire dans le cadre de la langue d'usage du lecteur, soit en raison d'évolution de la langue ou de difficultés de traduction, soit parce que la langue est équivoque et le contexte insuffisant pour dissiper tout risque de malentendu. On considère que le vocabulaire en cause n'est pas spécifique à la doctrine (ou plutôt que la doctrine suit sur ce point des usages adoptés communément à l'époque de sa mise en place).

Clarifications de concepts et de thèses doctrinales. Le mot, ou l'expression, est clairement défini, ou du moins décrit, dans le cadre du vocabulaire (terminologique) de la doctrine en jeu.

Clarifications inter-conceptuelles. Elles définissent ou décrivent, comme précédemment, un concept de la doctrine, mais de manière externe, soit à l'aide du vocabulaire commun, soit dans le cadre d'un autre ensemble doctrinal, soit dans le cadre d'une discipline théorique ou pratique.

Les clarifications par exemples, illustrations etc. sont, dans une optique rationnelle, une partie d'un des cas précédents(ou sa propédeutique).

2- Des clarifications d'ellipses et de sous-entendus qui complètent le texte là où il est obscur ou incomplet. Il importe peu de savoir s'il s'agit d'une imperfection, d'un effet de style, d'une volonté ésotérique ou élitiste de la part de l'auteur, ou, ce qui est le plus courant, d'un effet de distanciation entre l'auteur et le lecteur, ce qui conduit au cas suivant.

3- Des clarifications de contextes enfin, qui fournissent tous les éléments que le lecteur ne connaît pas et que l'auteur n'a pas jugé utile de faire connaître. Si on laisse de côté ce qui relève aussi des cas précédents, on trouvera ici essentiellement des éléments historiques, culturels et biographiques.

Les justifications, commentaires et critiques intra-doctrinaux

Les justifications visent à asserter (ou conforter) les propositions soutenues par la doctrine, les commentaires les discutent, les critiques les mettent en doute. Ils sont intra-doctrinaux dans la mesure où ils sont conformes aux règles d'assertion validées par la doctrine. En fait, soit ils apparaissent quelque part dans le corpus

[2] Un essai fait sur le *Discours de la Méthode* à l'aide du logiciel *Neuronav*, permet de repérer non seulement les six parties de l'ouvrage, mais aussi les divisions très nettes à la lecture (ou exposés par les commentateurs) de la première, de la cinquième et de la sixième partie.

doctrinal, soit ils sont produits dans l'esprit de la doctrine. Ils peuvent avoir été produits explicitement pour le texte visé, ou être rattachés par un commentateur. Les types de justifications sont multiples ; on distingue habituellement les arguments déductifs ou démonstratifs, les arguments d'autorité, les arguments d'induction, de transduction, d'exemplification, les arguments d'analogie… tous n'étant pas forcément assumés par la doctrine, et leur valeur de validation différant selon la doctrine. Les commentaires et critiques mettent aussi en place : des indications de présupposés, sous entendus…, des objections fondées soit sur la contestation des arguments justificatifs, soit sur de nouveaux arguments comme indiqué ci-dessus, soit sur des révisions limitées de l'espace conceptuel.

3.3 Les développements et critiques extra-doctrinaux

A la différence du cas précédent, la critique, (voire justification) est extérieure à la doctrine (autres règles justificatrices, principes, liaisons conceptuelles, autorités, exemples...)

A la limite des conseils et consignes de méthodes d'explication, on propose parfois, enfin, de développer la postérité d'un texte dans un nouveau cadre doctrinal. Les concepts ou les thèses sont réinterprétés, réorganisés en tant que précurseurs ou germes d'une pensée nouvelle. Considérées de peu d'intérêt explicatif quand les doctrines sont très différentes, ces analyses sont relevées avec attention en cas de liens de filiation entre les doctrines (Spinoza, Malebranche ou Leibniz relativement à Descartes) en tant qu'indices des limites de la doctrine, voire évolution doctrinale.

3.4 Typologie de base pour un intertexte à visée argumentative.

Les distinctions que nous avons opérées ci-dessus n'ont cependant directement guère d'efficacité pour notre objectif, étant difficiles à caractériser par des traitements généraux, humains ou informatiques.

Mais elles sont grosses d'une analyse, finalement assez évidente (on nous reprochera peut-être de ne pas l'avoir énoncée directement) qui se révèle être la matrice des spécifications que nous cherchons. C'est ainsi que, en inversant les genres et les différences, nous pouvons distinguer trois ensembles dans l'intertexte :

L'ensemble intra-doctrinal : tous les textes qui relèvent de la même unité de doctrine, partageant la terminologie, les méthodes et les principes avec le texte cible. C'est d'abord l'ensemble des écrits d'un auteur, mais il peut être réduit quand la doctrine d'un auteur est reconnue avoir varié ; il peut aussi, c'est le cas général, être plus étendu, incluant alors les disciples les plus proches et les diverses compilations et synthèses de doctrine. Il ne faut pas cacher que le cadre d'une doctrine est sujet à discussion, en raison des variations qui y sont tolérées : le cartésianisme peut aussi bien être attribué à Descartes seul, qu'à tous ceux qui ont pris nom de cartésiens (Spinoza, Malebranche, Leibniz), qu'à la pensée mécaniste, voire qu'à la tradition qui se réclame d'un rationalisme selon les règles de la méthode etc. Mais, sauf déclaration explicite, c'est l'usage commun qui doit valoir en la matière, Descartes et ses disciples proches pour notre exemple. Il y a là une garantie intersubjective suffisante, qui n'interdit cependant pas une évolution historique des contours d'une doctrine. Cet ensemble concerne principalement : les clarifications conceptuelles, les justifications, les objections et réponses, les présupposés.

L'ensemble inter-doctrinal : tous les textes qui relèvent d'une autre doctrine, à des fins de caution ou de critique, d'analogie ou d'alternative etc. On peut y distinguer : les sources, y compris les sources qui servent de repoussoirs, les autorités, les clarifications inter-doctrinales, les débordements de doctrine et les

critiques de la doctrine dans ses principes et concepts, les réinterprétations selon une autre doctrine.

L'ensemble extra-doctrinal : tous les textes qui se rapportent au texte noyau sans en référer à un arrière plan doctrinal, du moins déclaré. Ils exposent des faits, évidences, lieux communs… ou s'appuient sur des théories positives à prétention scientifique ou objective, pouvant aussi bien éclairer un passage précis que construire une approche systématique (organisation, genèse...). C'est à ce niveau qu'il faut placer les textes dont la fonction est de lier le texte cible à d'autres textes selon des procédures objectives ou automatisées. On y trouvera : les clarifications extra-doctrinales évidemment, les incitations à penser, les approches scientifiques.

4 Typologie proposée

Nous tirons donc de ces analyses l'organisation suivante pour l'intertexte d'un ouvrage de philosophie (ou rationnel critique en langue naturelle).

1- Cotexte (*sources* au sens large) : les ouvrages et surtout passages d'ouvrages qui ont pu déterminer l'écriture de l'ouvrage, directement : selon la thématique de l'inspiration ou celle des influences, ou indirectement : tous les autres textes qui permettent à un lecteur contemporain d'appréhender le contexte d'écriture de l'ouvrage noyau, sans que nécessairement l'auteur y ait accédé. Le cotexte du *Discours* peut être lui-même organisé en séparant les époques, ou bien en distinguant :

Les sources au sens strict : celles issues de l'éducation de Descartes, celles qui tiennent de la culture d'un honnête homme vers 1636, celles qui relèvent de ses préoccupations et de son milieu intellectuel spécifique ;

Les ouvrages dont l'auteur (Descartes) n'a peut-être pas eu connaissance, mais qui témoignent cependant des influences sur le texte noyau (imprégnations par les études, les discussions…) ;

Les ouvrages dont Descartes n'avait nul besoin, mais qui sont nécessaires à notre compréhension en raison de la différenciation des contextes (état commun de la langue à l'époque, propos obligés, explication de l'arrière plan idéologique, social, moral…, contexte historique…).

2- Doctritexte (*intertexte doctrinal*) : les textes extraits du corpus cartésien, antérieurs ou postérieurs, y compris d'autres textes de l'ouvrage lui-même (si le texte noyau est un passage). Ce corpus inclut aussi les ouvrages qui témoignent ou résument la doctrine. Le corpus peut être distingué selon les époques (avec toutes les difficultés et exceptions que cela présente) mais aussi, selon les domaines théoriques explorés.

3- Métatexte (*explications et commentaires*) : les textes postérieurs qui ont le *Discours*, ou un de ses passages, comme objet explicite d'étude et d'analyse. On y inclut tout autant les explications au sens strict (avec comme cas particulier les notes d'une édition particulière du *Discours*), les reprises (y compris les parodies et plagiats) voire les synthèses ou prolongements non retenus comme appartenant au corpus doctrinal, que les commentaires critiques ou les objections et réfutations. On peut y introduire la distinction traditionnelle entre organisation, explication, analyse critique, commentaire, objections et réfutations, pamphlets… On peut aussi penser à ce que leur organisation se fasse plutôt par un jugement de valeur sur leur pertinence, qui allie le sérieux à l'originalité des analyses.

4- On peut concevoir que l'intertexte qui expose l'organisation du texte noyau soit une catégorie à part (*plan, organitexte* !).

5- Extratexte (*réappropriations*) : les textes, parfois mal démarqués des précédents, qui utilisent le *Discours* ou un de ses passages, dans une optique qui n'a rien à voir avec la compréhension du *Discours*, que ce soit comme exemple, autorité ou repoussoir.

6- Analotexte (*analogies*) enfin : les différents textes qui présentent tout autre rapport à notre texte noyau, qui sont à la limite de nos préoccupations. Ils renvoient le plus souvent à l'orientation vers la création, et l'invention de nouvelles œuvres, ce que promeut le plus généralement l'hypertexte littéraire. Cette catégorie peut être considérée comme vide, ou du moins assimilée à la précédente, dans le cadre d'un intertexte à vocation rationnelle discursive.

Le résultat paraîtra peut-être superficiel, et semble aisément amendable. Seulement, des déterminations plus précises se feraient au détriment soit de la possibilité de l'exporter hors de notre exemple, soit de la facilité à déterminer le type.

5 Quelques pistes d'application pratique

En effet, cette partition, au premier niveau du moins, est aisée à construire, à partir de données objectives, par grands ensembles (nonobstant quelques erreurs) sur un intertexte existant, la mise en place de l'intertexte lui-même fournissant une première série d'indications. La disponibilité informatique (sous forme ascii) des documents étant supposée acquise, un certain nombre de procédures élémentaires semblent pouvoir être mises en place pour l'une ou l'autre opération.

5.1 Récupération de l'intertexte latent

L'extraction automatisée des références de textes cités dépend de leurs conditions de présentation : immédiate pour un document numérisé aux normes actuelles, encore simples dans le standard des notes en bas de page (en raison de la qualité de leur présentation formelle), complexe et peu sûre pour les ouvrages anciens. On ne peut guère proposer alors que du bricolage sur le vocabulaire contextuel et des appels à des listes d'auteurs ou d'ouvrages (pour repérer la référence ou la compléter). La qualité de la référence dépend évidemment de ce que fournit le document numérisé lui-même (chapitres, pages, voire chaîne de caractères) 3.

1- Les références fournies par le texte noyau lui-même doivent évidemment être toutes conservées et relèvent toutes du cotexte.

2- Les références qui sont citées dans les notes et commentaires critiques doivent aussi être toutes conservées, mais leur catégorisation (typologie) ne peut être fixée a priori. Des réponses de bonne probabilité peuvent être cependant fournies au cas par cas, en connaissant les objectifs du commentateur (dans notre exemple, Alquié privilégie le cotexte, Minos l'analotexte), voire en raison, comme chez Gilson, de la présence d'une organisation des notes conforme à nos desseins : cotexte (définitions), puis doctritexte, à nouveau cotexte (sources) et enfin métatexte.

3- Les références d'ouvrages qui citent le noyau sont aisés à repérer par parcours sur le plein texte, mais leur typologie est indéterminable. Elles excluent seulement le cotexte. On peut cependant postuler qu'un appel fréquent au noyau

[3] Sur cette question des références cf. l'article, non publié, *Gérer les références bibliographiques. Le cas des références fournies par les commentateurs du Discours de la Méthode de Descartes*.in *http ://pagesperso-orange.fr/hufschmitt.benoit/textes/refbiblio.pdf.*

caractérisera plutôt un métatexte, et un appel ponctuel un extratexte (hors texte interne à la doctrine). Remarquons en passant qu'un commentaire relève en lui-même du métatexte, ce sont les références qu'il contient qui s'éparpillent selon les différents types. De la même manière, tout ouvrage qui cite, ensembles, le texte noyau et un autre texte, est ponctuellement dans la position d'une note d'édition commentée.

4- La détermination de l'intertexte par coréférencement depuis des mots clefs ou tout autre index a la qualité et le type fournis par la constitution des mots clefs ou index. On distinguera :

Le cas des index intra-doctrinaux, manuels ou automatisés, qui doublent idéalement le doctritexte fourni par les commentateurs.

Les cas des index interdoctrinaux, dont la pertinence est variable et les types sont non fixés a priori, si ce n'est pas considération des dates d'édition des ouvrages.

Notons, à partir d'un cas particulier issu de notre illustration, les index dont l'objectif est de construire un intertexte spécifié avant la lettre, tel l'Index scolastico-cartésien de Gilson (GILSON, E. 1912), qui lie des références de passages du corpus cartésien à des extraits de textes de la scolastique tardive que Descartes aurait pu connaître, fournissant donc directement du cotexte, et indirectement du doctritexte.

5- L'appel aux dictionnaires d'époque suivant le vocabulaire du texte noyau fournit directement (article du dictionnaire) ou indirectement (citations dans l'article) un cotexte extrêmement important, ainsi que ce peut être remarqué par l'usage fréquent qu'en font les commentateurs dans notre illustration.

6- Enfin, par extension du cas précédent, l'appel aux références fournies, pour un même vocabulaire, par des index d'autres corpus doctrinaux méritent attention si ces corpus sont antérieurs ou contemporains au noyau, d'autant plus si le corpus a déjà fourni de l'intertexte. Le type est évidemment aussi cotextuel.

Ce tour d'horizon étant très empirique, justifié seulement par ce que relèvent les notes des commentateurs, peut sans doute être complété. S'il montre que la typologie de base peut être facilement extraite, il indique aussi qu'une typologie plus fine est difficile à établir (hors cas des dictionnaires : définitions).

5.2 Contrôle des types de l'intertexte

Des procédures complémentaires pour fixer les types d'intertexte manquants sont aisées à imaginer et mettre en place. Par exemple :

1- Les documents antérieurs au noyau (non co-doctrinaux) relèvent du cotexte.

2- Les sources rétrospectives (ouvrages postérieurs au texte noyau qui indiquent cependant le contexte d'écriture) peuvent être dégagées, en général, par considération des indications explicites dans leurs titres ou têtes de chapitres.

3- A ce niveau de généralité, mais pas de manière plus précise, les textes d'un même corpus partagent généralement le même type intertextuel…

4- Les titres, sous-titres et titres de chapitre citant le texte noyau (ou son allusion), voire la doctrine en général, cernent assez exactement les explications et commentaires, relevant donc du métatexte.

5- Par contre, les citations ou références isolées en plein texte du texte noyau pointent plutôt les réappropriations et les analogies, extratextes ou analotextes.

5.3 Propositions pour une mise en place élaborée de l'intertexte

Les manipulations précédentes ne font que reprendre et organiser de l'intertexte déjà dégagé, certes non spécifié. On peut envisager aller plus loin et proposer des procédures qui créent l'intertexte lui-même. La lexicologie est

l'instrument privilégié pour cela, à travers les listes de co-occurrences, à partir d'une partie du lexique du texte noyau, celle qui est censée repérer les unités conceptuelles, le corpus parcouru fournissant le type (doctritexte pour le corpus cartésien, co-texte pour celui de Montaigne ou Sénèque, métatexte pour les œuvres de Gilson concernant le cartésianisme etc.). Même si l'on néglige la difficulté à cerner les concepts par le lexique (et surtout des concepts interdoctrinaux analogues par le même lexique), les résultats sont généralement très insuffisants : le bruit submerge tout, sauf quelques cas précis :

1- Un choix très limité du lexique, et avant tout celui que l'on soupçonne d'être emprunté, tel "bon sens" dans le Discours. C'est la base de l'Index scolastico-cartésien de Gilson cité ci-dessus. Seulement, ni la sélection du lexique, ni le choix du corpus d'étude ne relèvent d'un traitement automatisé.

2- Un travail sur le seul lexique cartésien est par contre plus efficace, mais les silences liés aux synonymies (dont celles des expressions composées) et aux indicateurs anaphoriques, et le bruit issu des homonymies (surtout grammaticales) ne sont pas négligeables. Les logiciels de lemmatisation, de détermination des expressions régulières… parent en partie ces difficultés.

3- La procédure qui nous semble réellement efficace, pour le corpus d'où est issu le texte noyau du moins, consiste non tant à repérer la présence d'un même lexique, mais à calculer par analyse factorielle des données la proximité lexicale des différents passages. Nous en avons fait l'expérience pour le Discours et la partie française du corpus cartésien, à l'aide d'un logiciel de classification automatique (Neuronav de Diatopie). Le résultat, contrôlé par les propositions parallèles fournies par les commentateurs, est prometteur, en dépit de toutes les inadaptations du logiciel (prévu pour l'archivage de documents administratifs) à notre entreprise 4.

6 Conclusion

Cotexte, doctritexte, métatexte, extratexte et analotexte fournissent une première partition de l'intextexte d'un texte philosophique (plus généralement textes à visée rationnelle discursive), laquelle est aisée à mettre en place de manière massive., alors que leurs spécifications, qui atteignent une organisation fine de l'intertexte, demandent la considération particulière de chaque document.

Les références d'un intertexte sont présentes, de manière plus ou moins larvée, dans les divers documents qui citent et analysent le texte noyau (commentaires, index…) et peuvent donc en être extraites. Une production automatisée de la partie doctritexte d'un intertexte est à la portée d'une application en analyse factoriel du lexique (classification automatique), mais que faire pour le reste de l'intertexte ?

Références :

Butor, M. (1968) : "La Critique et l'Invention" in *Répertoire III*, Minuit, Paris.

Daniel, G. (1690). *Le Voyage du Monde de Descartes*, Vve de St Bénard, Paris.

Descartes, R. (1891-1912). *Œuvres,* édition Adam et Tannery (13 vol.), Cerf, Paris (le *Discours* est en début du tome VI, l'index rerum dans le tome XIII).

[4] *Neuronav* (http://www.diatopie.com). Sur ce point voir, *Génération automatique de liens dans un texte de philosophie : l'exemple du Discours de la Méthode de Descartes.* non publié, in *http ://pagesperso-orange.fr/hufschmitt.benoit/textes/generationautomatiquelienstextephilosophie.pd*f.

Descartes, R. (1963). *Œuvres philosophiques,* édition de F. Alquié, tome I, Garnier Frères, Paris.

Descartes, R. (1967). *Discours de la Méthode.* Texte et commentaire par Etienne Gilson, Vrin, Paris.

Descartes, R. (1966). *Discours de la Méthode* suivi d'extraits de la *Dioptrique*…, Garnier-Flammarion, Paris.

Descartes, R. (1973). *Discours de la Méthode.* Avec des aperçus… par J. M. Fataud, Bordas, Paris.

Descartes, R. (1953). *Œuvres et Lettres.* Textes présentés par André Bridoux, Paris.

Descartes, R. (1842). *Œuvres de Descartes*, nouvelle Edition collationnée… par M. Jules Simon, Charpentier, Paris

Descartes, R. (1907). *Discours de la Méthode*, Cerf, Paris.

Fusillo (1991) : *Naissance du Roman*, Seuil, Paris.

Genette G. (1982) *Palimpsestes. La Littérature au second Degré*, Seuil, Paris.

Gilson, E. (1912) *Index scolastico-cartésien*, Alcan, Paris.

Hufschmitt, B. et alii (2007). "Un Projet numérique autour du *Discours de la Méthode* de Descartes" in *CIDE 10, Le Document numérique dans le Monde de la Science et de la Recherche, Actes du dixième Colloque international sur le document numérique*, Europia éditions, Paris, 2007, 117-131.

Kanello, I. (1999). "De la Vie sociale du Texte…" in *Cahiers de Praxématique : Sémantique de l'Intertexte*, N° 33.

Kristeva, J. (1969). "Le mot, le Dialogue, le Roman" in *Semiotiquè. Recherche pour une Sémanalyse*, Paris.

Rabau, S. (2002). *L'Intertextualité*, Introduction, choix de textes, commentaire, vade mecum et bibliographie par Sophie Rabau, G. F., Paris.

Riffaterre (1983).*Sémiotique de la Poésie*, Seuil, Paris.

Détection automatique des types de structures textuelles énumératives

Automatic detection of types of textual enumeration structures

Khaldoun AL FARAJ(1), Mustapha MOJAHID (2)

(1) IRIT, Université Paul Sabatier Toulouse III, Toulouse, France
alfaraj@irit.fr

(2) IRIT, Université Paul Sabatier Toulouse III, Toulouse, France
mojahid@irit.fr

Résumé. Cet article porte sur l'étude d'un objet textuel particulier : l'énumération. Cet objet possède certains phénomènes textuels spécifiques, comme la différence entre la fonction et la présentation de ses constituants « items ». Cela a conduit à distinguer les énumérations parallèles et non-parallèles. Nous proposons un système qui permet de détecter le type d'énumération à l'aide d'indices discursifs. Le principe du système consiste à repérer, à l'aide d'un algorithme d'apprentissage, les relations de coordination et de subordination existantes entre les différents items d'énumération. Deux types d'indices discursifs ont été considérés : le suivi thématique et les connecteurs.
Mots-clés. Analyse de texte, structure énumérative, indices discursifs, modèle d'apprentissage.

Abstract. This article focuses on the study of particular textual object: the enumeration. It has some specific textual phenomena as the difference between function and presentation of its various components "items". These phenomena led to propose a classification to parallel and no-parallel enumerations. We propose a system that detects the type of enumerations through discursive clues. Using learning algorithm, it detects the coordination and subordination relations between different items within enumeration. Two types of discursive clues were considered: thematic sequence and connectors.
Keywords. Text analysis, enumeration structure, discursive clues, machine learning.

1 Introduction

Nous nous situons dans un cadre où l'accès au contenu d'un document textuel est réalisé via un écran d'ordinateur. Il peut prendre plusieurs formes suivant la tâche à laquelle se destinent le système et les caractéristiques du document (de la visualisation de grand document à la production d'une hiérarchie conceptuelle à partir de documents). Mais les approches rendant compte d'un résultat textuel pour décrire un document sont généralement regroupées sous l'étiquette de résumé automatique (Spark-Jones, 1999).

Ainsi, l'analyse linguistique des objets textuels (les énumérations, les définitions, les titres, etc.) permet de fournir aux systèmes, voire directement à l'utilisateur, des informations de description et d'organisation du contenu des documents nécessaires pour un traitement efficace d'un texte. Dans cet article, nous nous sommes intéressés à analyser un objet textuel particulier : l'énumération.

En comparant l'énumération avec d'autres objets textuels, on remarque que les paramètres typographiques, dispositionnels et discursifs de la structure énumérative sont plus importants, notamment les constituants de sa structure interne, ce qui justifie notre choix de cet objet d'étude.

Nous ciblons notre étude ici sur deux points principaux : le premier est de repérer les constituants de la structure énumérative au sein du texte et le second est d'identifier les relations entre ces constituants. Ainsi, nous proposons un algorithme qui permet de Détecter la Structuration des Enumérations (DES) en nous basant sur les deux indices discursifs : le suivi thématique et les connecteurs.

Pour présenter notre contribution, nous commençons d'abord par donner une description analytique des structures énumératives, y compris la classification des énumérations. Nous présentons ensuite une caractérisation des différents indices des structures énumératives à partir desquels nous avons conçu l'architecture d'un système qui permet de détecter les relations entre différents constituants des structures énumératives. Enfin, nous présenterons notre système de reconnaissance automatique des types de structures énumératives.

2 Les énumérations

2.1 Les composants d'une énumération

La Structure Enumérative (SE) comprend une amorce, une énumération et parfois une conclusion (Bouraoui, 2000), (Luc, 2000), (Péry-Woodley, 2000).

Une amorce est une phrase introductrice précédant l'énumération. Cette amorce est caractérisée par une ou (des) combinaison(s) de marques lexicales, par exemple « les suivant », typographiques « : », dispositionnelles « saut de ligne » ou syntaxiques.

Une énumération est un ensemble d'items (au moins 2).

Un item est une entité énumérée (ou plutôt co-énumérée) perceptible par variation de la MFM[1]. Il est caractérisé par diverses marques pouvant être typographiques (tiret, numérotation, etc.), dispositionnelles (espacement vertical ou horizontal), lexico-syntaxique (organisateurs textuels, schémas syntaxiques des items, etc.), ou toute combinaison de ces marques.

2.2 Typologie de structures énumératives

Luc (2000) a proposé une classification pour les énumérations en s'appuyant sur les caractéristiques et propriétés du Modèle de l'Architecture Textuelle (MAT) (Pascual, 1991), et de la Rhetorical Structure Theory (RST) (Mann et Thompson, 1987). Il a identifié trois catégories : (1) des énumérations dont les items entretiennent des relations diverses entre eux, (2) des énumérations dont les items ne sont pas structurellement équivalents et enfin (3) des énumérations dont un ou des constituants entretiennent des relations avec un ou des objets textuels extérieurs à la structure énumérative. Les types d'énumérations dans chaque catégorie sont

[1] MFM : La mise en forme matérielle est un ensemble de propriétés de réalisation de texte, ces propriétés peuvent être de nature lexico-syntaxiques, typographiques, ou dispositionnelles. Ces propriétés entretiennent entre eux des relations de dépendance variées (Pascual, 1991).

exclusifs, mais une énumération relève des trois types, chacun des types appartenant à une catégorie différente.

Première catégorie[2]

Enumération subordonnée : énumération dont tous les items successifs entretiennent des relations de dépendance (syntaxique ou rhétorique).

Enumération coordonnée : énumération dont tous les items sont fonctionnellement équivalents (des points de vue syntaxique et rhétorique).

Enumération hybride : énumération dont (au moins) deux items sont fonctionnellement équivalents et dont (au moins) un item dépend d'un autre.

Deuxième catégorie

Enumération visuellement homogène : énumération dont tous les items sont visuellement équivalents.

Enumération visuellement hétérogène : énumération dont un (au moins) des items est visuellement différent des autres items.

Par exemple, une énumération hétérogène possède deux items dans un paragraphe et un troisième item dans un paragraphe disjoint du premier, alors qu'une énumération homogène possède tous ces items dans un seul paragraphe.

Troisième catégorie

Enumération liée : énumération dont au moins un item est en relation avec un objet textuel extérieur à la structure énumérative ou bien une énumération dont un membre de la structure énumérative contient une autre énumération.

Enumération isolée : énumération dont aucun item n'est en relation avec un objet extérieur à la structure énumérative et dont aucun membre de la structure énumérative ne contient une autre énumération.

À partir de cette typologie, Luc a pu définir les énumérations parallèles et non-parallèles :

une énumération est *parallèle* si elle est coordonnée, homogène et isolée.

une énumération *non-parallèle* est une énumération qui n'est pas coordonnée, ou qui est hétérogène ou liée.

Des travaux récents (Porhiel, 2007) ont proposé une analyse des structures énumératives à deux temps. Il s'agit d'une structure composée d'une structure énumérative avec classifieur dans le premier temps et d'une énumération thématique dans le deuxième temps. Nous trouvons que ce type de structure énumérative correspond à la troisième catégorie de la classification Luc, i.e. la structure énumérative à deux temps est une structure énumérative non parallèle liée.

La détection automatique des types de structures énumératives parallèle ou non-parallèle, qu'elle soit le fait de l'homme ou d'un système automatique, consiste essentiellement à prendre en compte les régularités correspondant aux trois catégories mentionnées plus haut.

Dans cet article nous focalisons notre attention sur la première catégorie : la distinction d'ordre syntaxique et/ou rhétorique entre énumération coordonnée et énumération subordonnée.

La restriction a été faite car les énumérations de la première catégorie sont plus rencontrées, dans la plupart des textes techniques et scientifiques représentant notre corpus d'étude, que celles des autres catégories.

---

2 Les termes utilisés par C. Luc sont énumération syntagmatique et paradigmatique. Nous préférons utiliser les termes d'énumération subordonnée et coordonnée empruntés à Choi (2002), qui nous semblent mieux adaptés.

Notre but consiste ainsi à repérer les relations de subordination et coordination existantes entre les items des structures énumératives. Pour cela, il faudrait d'abord déterminer aussi précisément que possible la nature et les caractéristiques de ces relations, afin de mieux connaître la structure de notre objet d'étude, et de mieux pouvoir la repérer.

3 Les propriétés de structure énumérative

3.1 Corpus d'étude

L'importance du choix du corpus pour relever d'éventuelles variations liées aux configurations, genre et domaine est évidente. Notre corpus comportant des exemples anglais et français proviennent d'articles et d'ouvrages scientifiques. Ainsi, nous avons constitué le corpus d'une part du recueil des 75 énumérations non parallèles de Virbel (1999), et d'autre part de 22 exemples d'énumérations parallèles pris dans la thèse de Luc (2000). Ce corpus nous a amené à dresser un ensemble exhaustif des indices de MFM, et également des indices discursifs afin de mettre en place une détection automatique de la typologie des énumérations.

3.2 Coordination des items de structure énumérative

Dans le cas d'énumération coordonnée, nous constatons que la propriété de symétrie sur l'ensemble des items doit être nécessairement vérifiée et nous qualifions d'items symétriques les items marqués par des similarités lexico-syntaxiques et/ou sémantiques. Ces similarités peuvent concerner tous ou une partie des items, et certains traits similaires peuvent présenter plus de poids que d'autres, ou présenter une certaine équivalence entre eux.

3.3 Subordination des items de structure énumérative

Dans le cas d'énumération subordonnée, nous remarquons que les propriétés de cohérence et complétude doivent être nécessairement vérifiées. Ces propriétés concernent deux relations.

La première est la précédence sur l'ensemble des items. I1 ***P*** I2 a pour interprétation : dans une énumération subordonnée, l'item I2 requière la présence préalable de l'autre I1. La relation ***P*** est antisymétrique, transitive et partielle. Elle est donc une relation d'ordre strict partiel sur l'ensemble des items subordonnés d'une énumération. ***P*** est ainsi une propriété qui permet d'éviter un certain nombre d'incohérences.

La seconde est l'obligation sur l'ensemble des items. I1 ***O*** I2 signifie que : dans une structure énumérative subordonnée, la présence de l'item I1 d'énumération entraîne obligatoirement la présence de l'autre I2. La relation d'obligation est également antisymétrique, transitive et partielle. Elle est donc une relation d'ordre strict partiel sur l'ensemble des items d'énumération. ***O*** est ainsi une propriété qui permet de vérifier la complétude de structure énumérative.

Pour illustrer ces deux relations, prenons l'exemple suivant :

Le Lindy Hop est :

- une danse swing ;
- dont la naissance remonte aux années 20.

I1 ***P*** I2 : l'item I2 nécessite la présence préalable de l'item I1, et l'ordre inverse rend la structure incorrecte.

I2 ***O*** I1 : l'absence de l'item I1 rend l'item I2 incorrect syntaxiquement et incompréhensible du point de vue sémantique.

Enfin, l'existence de la relation d'obligation au sein de l'énumération implique la relation de précédence, la réciproque n'étant pas forcément correcte.

4 Marques discursives

Nous nous appuyons sur les marques discursives des structures énumératives afin de mettre en place une détection automatique des types de structures énumératives parallèles et non parallèles. Les relations entre items d'une énumération sont soit de type *subordination* (relation de dépendance entre items), soit de type *coordination* (relation d'équivalence fonctionnelle entre items). L'enjeu que nous soulevons est d'identifier les marques qui soutiennent chaque relation.

4.1 Suivi thématique

Nous considérons la structure d'un item (I) selon un découpage de celui-ci en deux parties : le thème (T), i.e. ce dont il est le sujet, et le rhème (R), i.e. ce qui est dit à propos du thème. Les deux principales séquences thématiques distinguées sont les deux suivantes :

1. Quand un thème d'un item est lié directement avec celui de l'énumération se situant souvent dans l'amorce.
2. Quand un thème d'un item est lié avec le thème ou le rhème de l'item précédent.

Les séquences thématiques peuvent être relevées selon trois configurations : d'un item vers un autre (thème (I2) = rhème (I1) ou thème (I2) = thème (I1)), d'un item vers plusieurs (éclatement), ou encore de plusieurs items vers un seul (synthèse).

Afin de repérer les séquences thématiques, nous utilisons la notion de chaîne lexicale [9], qui correspond au rapprochement d'expressions linguistiques selon deux types de relations :

1. Les relations morphologiques : deux mots appartenant à la même morphologie, indépendamment de leur catégorie lexicale ;
2. Les relations sémantiques utilisées pour l'identification d'entités discursives, telles que la synonymie, l'hyperonymie et l'hyponymie, la méronymie et l'holonymie.

4.2 Connecteurs

Nous avons opté pour une pré-classification de connecteurs en trois classes en fonction de leur comportement pour structurer l'énumération et réduire ainsi la complexité du nombre d'indices. Les classes que nous avons définies sont les suivantes :

1. Coordonne : toute expression signifiant la conjonction ou la disjonction des deux items telle qu'il n'y a pas de contraintes sur l'ordre de présentation de l'information : « et, ou, ou bien, etc. ».
2. Subordonné : toute expression exprimant ce qui est élaboré, apparaît en début d'un item subordonné : « qui, que, où, cependant, par exemple, en conséquence, etc. ».
3. Subordonnant : toute expression exprimant ce qui est élaborant, apparaît en fin d'un item subordonnant : « en tant que, tel que, ci-après, ci-dessous, etc. ».

4.3 Exemple

Nous illustrons ces marques discursives sur l'exemple suivant dans la figure 1 :

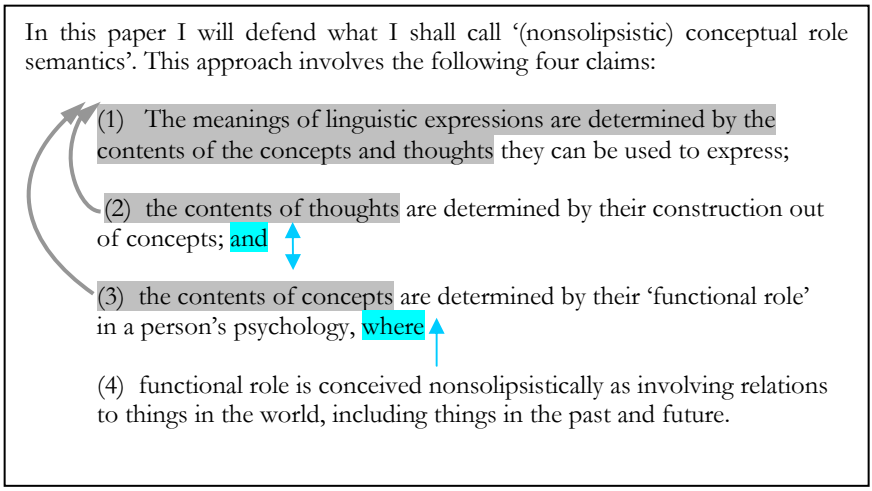


Figure 1. *Illustration des indices discursifs à partir d'un exemple tiré du corpus.*

Dans cet exemple, nous avons d'abord un suivi thématique : le thème du deuxième item est lié directement au rhème du premier item (et idem pour le troisième item). Ensuite nous avons un connecteur coordonné « and » entre le deuxième et le troisième item noté ↕. Enfin nous avons un connecteur subordonnant entre le troisième et le quatrième item « where » noté ↑.

5 Système de détection des types de structure de l'énumération

Le système que nous décrivons dans cet article offre une analyse fine de la structure énumérative en tenant compte des relations existantes entre les items de l'énumération. Pour construire notre système DSE, nous nous sommes basés sur deux études (Choi, 2002) et (Hernandez, 2005).

DSE possède deux composantes :

1. Un algorithme de structuration qui, parcourt une énumération donnée et qui, pour chaque item entrant, détermine son point d'attache dans la structure en cours de construction.
2. Un modèle de dépendance qui, pour deux items donnés et en fonction des indices discursifs de ceux-ci, prédit le type de relation qui existe entre ces items. La relation est soit de type de subordination, soit de type de coordination.

5.1 Algorithme de structuration de l'énumération

Admettant que la structure de l'énumération en arbre unique est une simplification de la réalité, nous adoptons cependant une modélisation *hiérarchique* essentiellement pour deux raisons : d'abord, parce que c'est la plus communément rencontrée, ensuite parce que pour l'instant aucune méthode automatique ne permet de la détecter.

Nous interprétons les relations inter items en termes de relations « parentale » de la manière suivante :

1. Une relation de subordination (père - fils) ;
2. Une relation de coordination explicite (relation de fratrie) ;
3. Une relation de coordination implicite (même père, mais non frère).

L'algorithme utilise deux structures de données : une pile qui stocke la branche « frontière droite » de l'arbre en cours de construction (le dernier élément empilé est

le point d'attache le plus prioritaire), et une file qui contient la liste des items tels qu'ils sont ordonnés dans l'énumération et analysés successivement. Notre choix de ces deux structures de données n'est pas arbitraire, les prospérités des relations de subordination et coordination expliquées précédemment dans la section 3 justifient ce choix.

L'objectif est d'identifier les items qui sont liés et les relations qu'ils entretiennent. Le principe repose sur la réalisation de certaines opérations de construction suivant la reconnaissance de telle ou telle configuration de dépendance entre l'arbre en construction et les items entrants. La section suivante décrit notre modèle de dépendance et explicite les différentes configurations de dépendance.

Les étapes de l'algorithme retenues sont simples. La pile joue un rôle de mémoire dont chaque empilement correspond à une granularité inférieure obtenue dans la structure de l'énumération :

1. Lorsque la pile est vide, on défile la file et empile la pile (état initial)
2. Tant que la pile et la file ne sont pas vides, nous effectuons une mesure de subordination ou de coordination explicite éventuelle entre l'élément au sommet de la pile et le premier élément de la file.

- Si une relation de subordination est détectée, alors l'élément de la file est défilé et empilé (on descend dans la granularité de l'énumération) ;
- Sinon si une relation de coordination est détectée, alors l'élément au sommet de la pile est dépilé et remplacé par l'élément de la file ;
- Sinon (cas de relation de coordination implicite) l'élément au sommet de la pile est dépilé et écarté (l'idée étant de remonter jusqu'au niveau de dépendance de l'élément en tête de file).

Cet algorithme possède de plus la capacité d'identifier en quoi une structure énumérative n'est pas parallèle.

5.2 Modèle de dépendance

Le modèle de dépendance a pour objectif d'identifier le type de relation qui unit deux items donnés en fonction des indices de ces items (nous parlerons aussi d'attribut pour désigner un indice). Dans ce travail, nous prenons la subordination et la coordination comme types de relation existant entre items.

Nous désignerons par IS, un item appartenant à la structure en cours de construction, et IN, le nouvel item apparaissant dans la linéarité de l'énumération et non encore attaché dans la structure de l'énumération. Le couple des deux items IS et IN est alors décrit suivant un ensemble d'indices {indice1,…, indicen}, supposés effectivement pertinents pour la description de relations discursives. Chaque indice possède un état d'activation relatif à l'observation d'une certaine relation entre IS et IN. En général, des indices binaires rendent de meilleurs résultats que ceux ayant plusieurs valeurs (Choi, 2002), aussi nous avons décrit nos indices suivant deux valeurs discrètes : 1 active et 0 non-active (respectivement pour repérée et non repérée). Etant donnés IS et IN, l'indicei est tel que indicei : IS * IN → {0,1}.

Voici des comportements qui pourraient être attendus :

- I_N : début d'item Λ élément de (Coordonne) $\Rightarrow$ $I_S \updownarrow I_N$

Si une marque de classe coordonne apparaît en début de l'item entrant I_N, il est fort probable que la relation qu'il indique avec l'item de la structure en cours de construction I_S soit de type coordination.

- I_S : fin d'item Λ élément de (Coordonne) $\Rightarrow$ $I_S \updownarrow I_N$

Si une marque de classe coordonne apparaît en fin de l'item de la structure en cours de construction I_S, il est fort probable que la relation qu'il indique avec l'item entrant I_N soit de type coordination.

- I_N : début d'item Λ élément de (Subordonné) $\Rightarrow$ $I_S \leftarrow I_N$

Si une marque de classe subordonnée apparaît en début de l'item entrant I_N, il est fort probable que la relation qu'il indique avec l'item de la structure en cours de construction I_S soit de type subordination.

- I_S : fin d'item Λ élément de (Subordonnant) $\Rightarrow I_S \leftarrow I_N$

Si une marque de classe subordonnant apparaît en fin de l'item de la structure en cours de construction I_S, il est fort probable que la relation qu'il indique avec l'item entrant I_N soit de type subordination.

La Figure 2 illustre l'architecture de notre système DES en appliquant sur l'exemple de structure énumérative mentionné précédemment dans la section 4. La première relation détectée est une relation de subordination, et ainsi de suite pour l'item suivant.

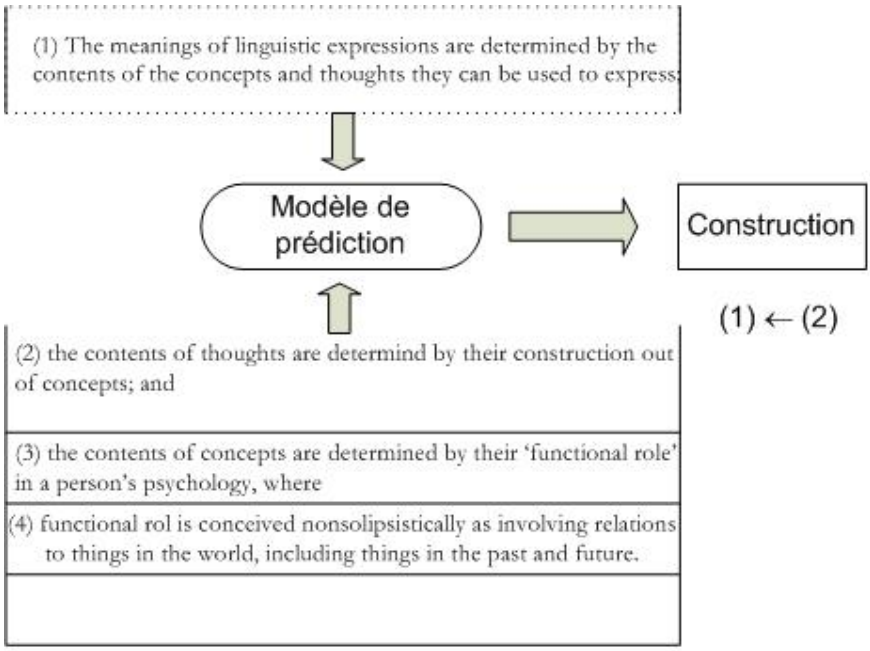


Figure 2. *Illustration de notre système DSE.*

6 Evaluations

Nous avons effectué deux évaluations sur l'ensemble des exemples constituant notre corpus d'étude. Dans la première, nous avons appliqué notre algorithme de structuration d'énumération sur toutes les structures énumératives de types parallèles et non parallèles du corpus. Dans la seconde, afin de tester la capacité de notre modèle de prédiction, nous l'avons évalué également sur l'ensemble des exemples du corpus.

6.1 Evaluation algorithmique

L'algorithme réussit à trouver pour toutes les énumérations du corpus (75 + 22) si elles sont parallèles ou non parallèles ; ce qui répond à notre premier objectif. De plus, l'algorithme arrive à caractériser en quoi les énumérations ne sont pas

parallèles sauf pour deux exemples. Ces deux cas posent en effet un problème au DSE, car le modèle hiérarchique que nous avons adopté s'avère ici insuffisant.

6.2 Evaluation du modèle de prédiction

L'ensemble des indices discursifs, INDICES = {$indice_1$, …, $indice_n$} que nous avons décrits permettent de représenter une configuration linguistique existante entre deux items I_S et I_N. L'objectif est alors de construire un modèle qui, pour une configuration linguistique donnée, détermine la relation correspondante,

D = {subordination, coordination} = {←, ↕}. A cette fin, nous avons utilisé le logiciel WEKA[3] pour faire apprendre ces deux relations à la machine à l'aide des indices discursifs.

Afin de constituer des échantillons d'apprentissage, nous avons manuellement annoté trois exemples de notre corpus d'étude, le premier correspond à une énumération coordonnée, le deuxième à une énumération subordonnée et le dernier à une énumération hybride. Cette variété sert à diminuer le taux d'erreur du classifieur, puisque sa mesure de la performance se fait généralement en termes de taux d'erreur qui correspond à la proportion d'erreurs faites sur l'ensemble des exemples d'apprentissage au niveau de la prédiction.

L'annotation a consisté à indiquer pour chaque item de l'énumération les relations de subordination et de coordination existantes avec un autre item.

Subordination et coordination	Suivi thématique	Connecteurs
Enumérations coordonnées	57.68%	69.10%
Enumérations subordonnées	64.33%	66.21%
Enumérations hybrides	52.63%	63.25%

Tableau 1. *Précision de DSE pour la prédiction des relations de coordination et de subordination*

Le tableau ci-dessus donne les précisions que nous obtenons dans la détection des différents indices discursifs pour les différentes sous-classes des énumérations de la première catégorie de la classification proposée en 2.

Nous obtenons des bons résultats pour les connecteurs grâce aux différents patrons définis. Des améliorations restent nécessaires pour les patrons concernant le suivi thématique, et plus particulièrement pour les énumérations hybrides.

Dans cette première expérience, nous avons opté pour un choix des échantillons d'apprentissage très petits, qui contrôle le taux d'erreur, car l'annotation manuelle coûte chère.

7 Conclusions et perspectives

Il en résulte que l'objet textuel – l'énumération – apparaît comme extrêmement plus complexe (que ce soit dans sa forme interne ou dans les relations qu'il peut entretenir avec des objets environnants) que l'idée intuitive que l'on peut s'en faire. Nous avons donc focalisé notre attention à détecter sa typologie en

[3] http://www.cs.waikato.ac.nz/ml/weka/

proposant le système DSE. Il s'agit d'un algorithme de recherche de point d'attache optimal pour l'item en cours d'analyse (en se basant sur les structures de pile et de file). Il utilise pour cette finalité un modèle prédictif basé sur un algorithme d'apprentissage qui pour une configuration d'indices donnés prédit le type de relation de dépendance existant entre deux items. Dans ce travail initialisé ici, nous avons obtenu des résultats satisfaisants et encourageants qui méritent d'être poursuivis.

Parmi nos perspectives nous envisageons d'enrichir notre modèle de nouveaux indices comme ceux de la mise en forme visuelle qui caractérisent la présentation de la structure énumérative, ainsi que de patrons de détection plus précis. Enfin nous proposons d'appliquer notre étude sur d'autres objets textuels comme les titres.

L'ensemble des études présentées regroupe deux champs de recherche : l'informatique et la linguistique. Ce type de démarche pluridisciplinaire mené dans cet article est une étape nécessaire pour pouvoir appréhender le plus complexe des objets textuels : le texte.

Une dernière perspective importante à ce travail est de mettre à profit ces connaissances au service de recherche en psycholinguistique. En effet ce travail peut servir dans l'élaboration de protocoles expérimentaux portant sur des énumérations en s'appuyant sur une description aussi fine qu'elle a été présentée dans cet article. On pourra évaluer ainsi le comportement de sujet humain dans des tâches cognitives (mémorisation, compréhension, prédiction, …). Le retour de telles expériences pourraient aider bien utilement des systèmes en TALN.

Références :

Barzilary, R., Elhadad, M. (1997). Using lexical chains for text summarization. In ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid.

Bouraoui, J. L. (2000). Les structures énumératives : caractérisation linguistique et reconnaissance automatique. Mémoire DEA, Université Toulouse Le Mirail.

Choi, F. Y. (2002). Content-Based Text Navigation. Ph.D. Thesis. Computer Science, University of Manchester.

Hernandez, N., Grau, B. (2005). Détection Automatique de Structures Fines de Texte, TALN'05, Dourdan.

Luc, C. (2000) Représentation et composition des structures visuelles et rhétoriques du texte. Thèse de doctorat, Université Paul Sabatier.

Mann, W. C., Thompson, S. A. (1987). Rhetorical Structure Theory: A theory of Text Organization. Rapport technique, ISI-RS-87-190, Information Sciences Institute, Marina Del Rey, Ca.

Pascual, E. (1991). Représentation de l'architecture textuelle et génération de texte. Thèse de doctorat, Université Paul Sabatier.

Péry-Woodley, M. P. (2000). Une pragmatique à fleur de texte. Mémoire réalisée pour l'obtention de l'habitation à diriger les recherches, Université de Toulouse Le Mirail.

Porhiel, S. (2007). Les structures énumératives à deux temps. Revue Roman, Vol. 42, num. 1, 103-135, Mai.

Sparck-Jones, K. (1999). Automatic summarizing: factors and directions. In I. Mani and M. Maybury, editors, Advance in Automatic text Summarization, MIT Press., Cambridge MA.

Virbel, J. (1999). Structures textuelles – planches. Fascicule 1 : Enumération. Rapport technique, IRIT.

Elaboration d'une ressource lexicographique informatisée pour les patients atteints de cancer : LexOnco, Lexique d'oncologie

Elaboration of an electronic dictionary for patients and relatives : LexOnco, dictionary of oncology.

Valérie Delavigne (1)

(1) Institut national du Cancer, Paris
v-delavigne@wanadoo.fr

Résumé. Face à l'augmentation de la demande d'information des personnes malades et à leur rôle croissant dans la prise de décision médicale, l'accès à une information validée, compréhensible et systématiquement actualisée, en correspondance avec leurs besoins, est un enjeu majeur de Santé publique. Améliorer la qualité de la prise en charge des patients passe par l'appropriation des principaux termes en lien avec la maladie. Dans le cadre du programme pluridisciplinaire SOR SAVOIR PATIENT, le projet LexOnco (LEXique d'ONCOlogie) vise à offrir aux patients un dictionnaire d'oncologie, validé sur le plan médical et qui tient compte des besoins d'information et des préférences des personnes concernées par le cancer.
Mots-clés. Dictionnaire électronique, lexicographie, cancer, information, patient

Abstract. In response to the evolution of the information-seeking behaviour of patients and concerns from health professionals regarding cancer patient information, the French National Federation of Comprehensive Cancer Centres (FNCLCC) introduced, in 1998, an information and education program dedicated to patients and relatives, the SOR SAVOIR PATIENT program. LexOnco project is a dictionary on oncology adapted for patients and relatives and validated by medical experts and cancer patients.
Keywords. Electronic dictionary, lexicography, cancer, information, patient

1 Un contexte en mouvement

1.1 La circulation des savoirs

Devenus plus accessibles par la montée en puissance de l'internet, la science et ses « experts » (Reboul, 2004), entrés de plain pied dans le champ médiatique, remodèlent la circulation des savoirs.

La médecine, sphère d'activités particulière, n'est pas exempte de cette transformation. Tout un chacun : usagers, patients, proches, personnel soignant et… linguiste ou terminologue, est concerné.

L'expertise circule, le savoir se transfère et ce, de plus en plus largement.

1.2 Une évolution de la prise en charge du patient

Sous le joug de la demande sociale, la relation médecin-patient est en mutation, passant progressivement du modèle « paternaliste » traditionnel, dans lequel le médecin décide du traitement, au paradigme idéal participatif d'une décision médicale partagée. Née des mouvements de malades, approfondie par des réflexions sur les relations entre science et démocratie (Latour, 1989 ; Stengers, 2002), cette évolution a fait émerger les besoins des patients, mis en évidence par diverses enquêtes sur leurs préférences et les obstacles à leur information et à leur participation active à leurs traitements.

La prise en compte de cette demande sociale s'est accompagnée d'une juridicisation : l'information du patient constitue aujourd'hui non seulement une demande, mais aussi un droit pour le patient, et elle est devenue une obligation pour le médecin par le biais de contraintes réglementaires. Divers lois et règlements font ainsi acte ainsi cette évolution. Ils soulignent le droit du patient à l'information et à une prise en charge conforme aux données actuelles de la science.

La prise en charge des patients en cancérologie a notamment connu une évolution importante : le patient est aujourd'hui censé détenir les éléments d'information nécessaires à son implication dans les choix thérapeutiques ; il est censé être désormais acteur des soins, qui lui sont *proposés*, et non plus imposés.

Dès lors, cette évolution suppose une information précise, complète, technique et fiable.

1.3 S'approprier des connaissances ?

Cependant, afin que le patient puisse détenir le savoir nécessaire à son implication dans ses choix thérapeutiques, l'expertise doit circuler. Cette transmission, pour être de qualité, doit être conforme aux données actuelles de la science (Coulter, 1998 ; Entwistle, 2003).

Cet élargissement du transfert de connaissances fait évoluer les usages langagiers. La transmission terminologique est partie prenante dans les relations interindividuelles, médicalisées ou non, intervenant tout à la fois dans l'interlocution avec les médecins et les équipes médicales, et dans les échanges autour de la maladie avec les proches ou d'autres patients.

Améliorer la qualité de la prise en charge des patients passe par l'appropriation des savoirs en rapport avec la maladie et partant, des terminologies en circulation. Car l'écueil essentiel réside bien évidemment dans le « jargon » médical, jargon que nous assimilons aux termes et à la phraséologie propre à cette activité.

Parler la même langue est la condition fondamentale du dialogue. La notion d'appropriation des terminologies circulantes, support des connaissances, est donc

centrale : il s'agit de faire en sorte non seulement que les patients puissent comprendre leur maladie, leurs examens, leurs traitements, mais qu'ils puissent aussi dialoguer avec les professionnels de santé, communauté inhomogène. Les patients doivent donc s'approprier les formes foisonnantes utilisées par la communauté langagière médicale pour comprendre ces mots et les manipuler au mieux en fonction de leur désir de dialogue.

La terminologie peut être aussi le lieu d'affrontement des rôles sociaux dévolus au médecin et au malade. Posséder l'usage du jargon - et il est bien question d'usage car une terminologie maladroitement utilisée révèle les failles d'un savoir mal assis – est aussi une façon de dénouer une forme de pouvoir symbolique. Dans un contexte d'évolution de la relation médicale dans lequel le patient est censé se prendre en charge, les rôles se reconfigurent, tout en étant dans le même temps aux prises avec les modèles antérieurs. Tiraillés entre les anciennes et les nouvelles valeurs, ces rôles sont bousculés et rejouent les usages, les interprétations et l'appropriation des terminologies.

Cette appropriation des terminologies médicales est à corréler au niveau socioculturel des patients, à leur culture et, bien évidemment, à leur désir de savoir. Cependant, c'est seulement à cette condition, s'il y a appropriation, que peuvent être réellement intégrées les données sur la maladie.

Dès lors, comment se fait-cette transmission ? Comment le patient et ses proches s'approprient-ils le vocabulaire qui leur est nécessaire pour discuter de leur maladie et de ses traitements ? De quels outils disposent-ils pour tenter de s'approprier les connaissances nécessaires à leur implication tout au long de leur parcours de soin ?

2 Elaboration d'une ressource terminologique pour les patients atteints de cancer

2.1 Une recherche active d'informations

Le désir d'informations conduit certains patients à une recherche active de données sur des supports divers.

Des informations sur le cancer sont diffusées de plus en plus largement, que ce soit dans la presse écrite, les guides ou les brochures édités par diverses associations ou institutions ou encore, sur les sites internet. Le web rend accessibles des renseignements de tout ordre. Les sources d'information y sont multiples et il n'est aujourd'hui pas rare de voir les patients arriver en consultation avec des documents imprimés à partir d'internet.

Cependant, il faut constater que ces documents répondent rarement aux besoins des patients et ne sont pas toujours en cohérence avec les données scientifiques (Bensing *et al.*, 2000 ; Coulter *et al*, 1999). Se précipiter sur Wikipédia à l'annonce d'un diagnostic de cancer n'est guère à conseiller. Les informations disponibles en ligne sont souvent orientées, fréquemment parcellaires (Gourdain *et al.*). A l'inverse, certaines sources scientifiques consultées par des patients à la recherche d'une information objective et fiable sont difficilement compréhensibles pour le profane et, non décryptées, sont couramment génératrices d'interprétations inadéquates et d'angoisses inutiles (Jadad et Gagliardi, 1998).

Si le Conseil de l'Europe préconise que les patients doivent « avoir facilement accès à une information pertinente au sujet de leur santé et des soins de santé les concernant sous une forme et dans une langue compréhensible par eux », ceci est particulièrement important en cancérologie. Les schémas thérapeutiques sont souvent lourds et complexes, associés à un langage technique spécialisé, difficile à

appréhender pour des non-experts (Hadlow et Pitts, 1991). En réponse à ces constats, des travaux menés avec des groupes de patients et d'usagers ont permis d'identifier leurs préférences à propos des caractéristiques des informations écrites (Coulter *et al.*, 1998).

C'est aussi ce qui pousse aujourd'hui la Haute Autorité de Santé à envisager d'attribuer des labels de qualité des sites en santé fondés sur des critères de fiabilité en matière d'information. Piste d'action qui reste à mettre en œuvre, et qui pose encore des questions de méthodes de mesure de la qualité à utiliser...

2.2 Le programme SOR SAVOIR PATIENT : élaborer des documents pour les patients atteints de cancer et leurs proches

Le programme « Standards, Options et Recommandations » (SOR) en cancérologie, initié par la Fédération nationale des Centres de Lutte contre le Cancer (FNCLCC) en 1993, a pour objectif d'améliorer la qualité des soins pour des patients atteints de cancer en fournissant aux praticiens un résumé et une analyse critique des données scientifiques actuellement disponibles. Ce travail national est mené avec la coopération d'experts des secteurs public et privé, et de sociétés savantes. Il débouche sur l'élaboration et la publication de recommandations pour la pratique clinique en cancérologie, disponibles en version papier et sur l'internet.

Depuis 1998 s'adosse à ce programme un second, dénommé SOR SAVOIR PATIENT, « Standards, Options et Recommandations pour le SAVOIR des PATIENTS », qui vise à mettre à la disposition des personnes malades sous forme de guides une information médicale validée, compréhensible et régulièrement actualisée, fondée sur ces recommandations destinées aux professionnels de santé (Carretier *et al.*, 2004).

L'élaboration des outils textuels produits par cette équipe[1] : guides ou fiches d'information, est fondée sur une méthodologie stricte. Les informations médicales sont issues des SOR ; elles sont validées par un groupe de travail composé de professionnels de santé spécifiques pour chaque thème abordé ; des patients, anciens patients et proches collaborent tout au long du processus d'élaboration, alimentant les documents en fonction de leurs besoins spécifiques, par le biais de groupes thématiques, d'entretiens individuels et de questionnaires. En bout de course sont édités des écrits destinés aux patients, à leur entourage et à toutes les personnes concernées par la maladie. Chaque guide est doté d'un glossaire.

Ces documents sont publiés sous forme de guides ou de fiches papier, ou en accès libre sur l'internet (www.fnclcc.fr).

2.3 LexOnco, un dictionnaire pour les patients

Dans le cadre de ce programme a été initié un projet lexicographique spécifique, LexOnco (LEXique d'ONCOlogie). Ce projet a pour objectif de proposer aux patients un dictionnaire sur le cancer, validé sur le plan médical et qui tient compte des besoins d'information et des préférences des personnes malades et de leurs proches.

Une ressource terminologique porte au cœur de son élaboration la problématique de son adaptation au public visé. Toute énonciation, qu'elle soit lexicographique ou autre, prend en charge la question de son co-énonciateur, quand bien même elle reste implicite. Or le « patient » est un destinataire multiple et multiforme. Confronté à la variabilité, on se retrouve démuni, sans pouvoir réellement définir un usager modèle.

[1] L'équipe SOR SAVOIR PATIENT se compose de Julien Carretier, Line Leichtnam-Dugarin, Sylvie Brusco, Marie Déchelette, méthodologistes, et Valérie Delavigne, linguiste.

On peut tenter de résoudre cette aporie en utilisant divers outils de recueil de données : « focus groups », entretiens semi-directifs et questionnaires. En esquissant une modélisation des utilisateurs potentiels, on peut ainsi dégager des indicateurs sur leurs préférences.

En fonction du public et de ses attentes sont ensuite convoquées des questions portant aussi bien sur la microstructure que sur la macrostructure du dictionnaire. Une méthodologie d'élaboration plaçant les usagers au cœur du processus a été mise en place, garante de la qualité des définitions produites (cf. Delavigne, 2008). Mais afin d'optimiser l'utilisation du produit fini, le volet consultation doit être attentivement considéré.

3 Un dictionnaire peut-il permettre une appropriation terminologique ?

Nous souhaitons proposer un outil qui puisse permettre aux non spécialistes de s'approprier les terminologies médicales dont ils ont besoin. Nous travaillons à construire une ressource qui, au-delà des aspects lexicographiques « traditionnels », puisse autoriser un réel partage de l'information.

3.1 La nomenclature

Sans décrire ici l'ensemble de la grammaire du dictionnaire, explicitons néanmoins quelques-unes des options théoriques et méthodologiques qui guident le projet.

En ce qui concerne le choix de la nomenclature, disons rapidement qu'il s'agit de repérer les termes en lien avec la maladie qui sont réellement utilisés. Ces termes sont saisis à travers différents types de discours où s'opèrent les « réglages » de sens d'unités cernables en contexte. Tout discours est le lieu de « normes terminologiques », variables selon les locuteurs et les objectifs, qui influent sur l'usage des termes et leurs paradigmes désignationnels. Notre approche se centre sur la notion d'usage et tente de prêter attention aux variations présentes : d'un médecin à l'autre, d'un établissement à l'autre, les pratiques terminologiques ne sont pas les mêmes, les sens se modifient.

Les *focus groups* sont un lieu de récupération d'usages oraux *déclarés* des patients. Ces entretiens permettent de vérifier si les terminologies convoquées dans les guides et qui, ensuite, intègrent la nomenclature de LexOnco, sont en adéquation avec celles qu'ils ont entendues et qu'ils utilisent.

Notre perspective sociolinguistique vise ainsi à repérer les mécanismes d'usage et la dynamique du fonctionnement des termes, afin de construire un objet lexicographique qui tente de prendre en compte la diversité des pratiques langagières situées.

La nomenclature recueille des termes sans préjuger de leur appartenance à un domaine spécifique (on y trouve par exemple le terme de *curatelle*, réputé appartenir plus spécifiquement au domaine du droit ou du social) ou à un quelconque réseau conceptuel. La nomenclature actuelle comprend l'ensemble des termes des glossaires des guides SOR SAVOIR PATIENT. Elle est alimentée par les termes extraits des guides à venir en fonction des besoins.

3.2 Le traitement lexicographique

L'ensemble de ces termes sont répertoriés dans une base de données Access. La base contient actuellement (2008) 940 termes. Tous ont fait l'objet de définitions brèves pour les besoins des glossaires des guides SOR SAVOIR PATIENT. Or il s'avère que ces définitions ont souvent été construites au coup par coup. L'objectif

du projet est d'en revoir l'ensemble, de les transformer en définitions utiles pour les patients, de les valider, pour ensuite les publier sur l'internet.

Là où un dictionnaire de langue n'a pas pour but direct l'apprentissage, un dictionnaire de vulgarisation peut modestement l'envisager. Dans le cas qui nous occupe, ce qui est visé est une certaine efficacité didactique. La description des entrées doit permettre à l'utilisateur du dictionnaire non seulement de comprendre le sens du terme, mais aussi de produire des énoncés « corrects » en empruntant ce terme, donc de saisir le fonctionnement discursif du signe.

Pour ce faire, l'on se doit de fournir des informations d'ordre lexical et paradigmatique d'une part, et d'autre part, des informations d'ordre syntagmatique, proposant des éléments sur la combinatoire du signe et sa phraséologie, visant ainsi la production, l'encodage.

Mais il est nécessaire d'aller au-delà.

Un exemple. Définir *chimiothérapie* sans signaler les effets secondaires de ce traitement reviendrait à négliger de prendre en compte les besoins des patients. Car encore plus que de répondre à « qu'est-ce que cela veut dire ? », c'est à des questions du type « est-ce que ça fait mal ? », « vais-je perdre mes cheveux ? », « puis-je continuer à vivre normalement ? » qu'il convient d'apporter une réponse. Au-delà d'offrir un accès à la signification, il s'agit aussi, et surtout si l'on vise une quelconque utilité pour ce public spécifique, de répondre à ses demandes et interrogations.

Le patient n'est pas à la recherche de définitions de mot. On dépasse ici les données strictement linguistiques pour en intégrer d'autres, que l'on peut repérer et relever. Sans revenir à l'opposition si malaisée à tracer entre dictionnaire de langue et dictionnaire encyclopédique, signalons juste que le dictionnaire, en tant que « genre » textuel particulier, est subordonné à des normes, et qu'un dictionnaire de langue contournerait volontairement ce type de données. Le nôtre ne saurait s'en passer, sauf à manquer son but.

Une collection de cotextes

Le corpus contient des discours repérés comme « médicaux » : brochures pour les patients, ouvrages et sites médicaux, cours, articles, recommandations pour la pratique clinique, revues ou sites de vulgarisation, ressources terminologiques, électroniques ou non etc. Il rassemble aussi des documents portant sur des aspects psychologiques, sociaux, le droit, etc., reflet du multipartenariat de la cancérologie qui, comme bien d'autres sphères d'activité, est pluridomaniale. Il est complété de ce qu'on peut désigner par des corpus « opportunistes », autrement dit, des corpus construits *ad hoc* pour les besoins de recherche autour d'un terme, correspondant à l'état des ressources disponibles.

À partir de ce corpus sont recueillis dans la base des définitions ou des cotextes définitoires, collecte qui, soulignons-le, n'est pas sans poser bien des problèmes de choix et de délimitation...

Des fiches descriptives

Le traitement lexicographique se mène à l'aide de fiches décrivant le fonctionnement sémantique et distributionnel de chaque terme à partir du corpus. Cet outil de description a montré son opérativité à plusieurs reprises (Bouveret et Gaudin, 1996 ; Tran, 1999 ; Gaudin, Holzem et Wable, 1999 ; Delavigne 2001). Cette fiche a été ici constituée à la fois en outil d'aide à l'analyse des emplois des entrées et à la rédaction des définitions.

Les fiches rassemblent un certain nombre de rubriques destinées à faire apparaître le « mode d'emploi » du terme : sa place dans le système de la langue, ses

cooccurrents possibles en discours, sa combinatoire, ses référents habituels. Trois types de fiches ont été construites, les relations convoquées n'étant bien évidemment pas les mêmes pour les noms, les adjectifs et les verbes. À chaque fois, les entrées sont décrites sous leurs dimensions morphologique, syntaxique, sémantique et combinatoire.

Chaque fiche descriptive est remplie en puisant dans le corpus, une analyse locale des contextes collectionnés permettant d'extraire les relations et de compléter les rubriques correspondantes.

Construire une définition adaptée et validée

L'analyse de la structure ainsi obtenue permet de mettre en évidence les relations saillantes que le terme entretient avec les autres mots de la langue et d'élaborer une définition adaptée.

La définition construite est ensuite engagée dans un processus de validation, tant sur le plan du contenu médical que sur celui de la lisibilité et de l'accessibilité sémantique. Le protocole mis en place se laisse décrire en deux étapes : une validation par les experts et une validation par les patients.

1- La « validation experts »

Le rôle des experts est d'être témoin de l'usage qui confirme ou infirme le choix de telle ou telle unité et prescrit son sens. Comme il ne peut être question de ne recueillir le témoignage que d'*un* spécialiste qui reflète seulement son propre usage (Depecker, 1997, Delavigne et Gaudin, 1996 ; Gaudin et Delavigne, 1997, Baudouin *et al*, 2003), la validation passe par le biais de collèges d'experts afin de bénéficier de la complémentarité des expertises.

2- La « validation patients »

Une autre validation se tourne vers un autre type d'expertise : les patients eux-mêmes en tant qu'utilisateurs finaux. Les patients sont détenteurs d'un savoir médical « non formel » (Jacobi, 2001) et de compétences linguistiques, méta- et épilinguistiques. Ils sont aptes à juger des termes proposés et de la qualité des définitions construites. Sont consultés des patients et d'anciens patients touchés par des cancers différents, ainsi que des proches et des personnes non malades.

De l'analyse du corpus à la construction des définitions, de l'expertise des cliniciens à celle des utilisateurs, chaque acteur du protocole est ainsi engagé dans un rôle qui doit garantir la qualité de l'objet fini, et assurer la validité médicale et l'adaptation des définitions à ses usagers.

Nous espérons ainsi réunir les possibilités d'une appropriation terminologique optimisée, en observant tout à la fois une certaine rigueur dans la description et en prenant en compte les exigences des utilisateurs.

4 Des glossaires sur papier à un dictionnaire sur l'internet

La méthodologie d'élaboration, qui place les usagers au cœur du processus d'élaboration, est garante de la qualité des définitions produites. Mais l'optimisation du produit fini doit prendre en compte l'usage (les usages) qu'en feront les utilisateurs.

Nulle enquête formelle n'a encore été menée. Néanmoins quelques pistes se dessinent.

4.1 Dictionnaire papier vs dictionnaires électroniques

En tant que dictionnaire électronique, LexOnco ne se veut pas une version numérisée des glossaires sous forme papier dont il est issu et pour laquelle on aurait

opéré un simple transfert de contenu. Il est conçu pour exister hors de tout référentiel imprimé.

Pour tout document, dictionnaire ou autre, le support pèse sur le contenu. L'environnement informatique procure une grande souplesse d'utilisation qu'un support papier n'offre guère et ce, sur différents plans.

La technologie numérique ouvre de belles perspectives au lexicographe, tant au niveau de l'élaboration lexicographique que de l'exploitation par les utilisateurs (cf. Kilgarriff, 2005 ; Pruvost, 2006 par exemple). Outre le fait qu'elle permette de s'affranchir de l'ordre alphabétique, de s'arranger des variantes orthographiques et des siglaisons grâce à des repérages préalables, etc., l'informatisation permet de reconsidérer les problématiques liées à la diffusion et à l'accessibilité du contenu.

Du point de vue du concepteur, au-delà de la possibilité d'une lexicographie « instrumentée » pour reprendre le terme de Benoît Habert, un dictionnaire électronique présente des avantages indéniables au nombre desquels on peut compter l'intérêt évolutif. En effet, une version électronique reste ouverte à des modifications et à toute information qui permet de compléter les descriptions existantes. Notre propos n'est pas de dresser une liste de tous les développements potentiels, mais de souligner ceux par lesquels on se retrouve avec un document « interactif ». Ainsi l'ajout d'entrées au fur et à mesure de leur traitement dissout les problèmes de taille de nomenclature par exemple, celle-ci n'étant plus à relier qu'à la capacité du lexicographe à les traiter… Ou encore, il est possible d'ajouter un champ jusqu'alors non repéré dont l'analyse a révélé le manque.

Du point de vue de l'usager, un dictionnaire informatisé permet une consultation directe et sélective des entrées. Le format actuel des définitions est généralement court : les articles les plus longs font 500 signes. Chaque terme de la définition jugé « technique » se voit doté d'un lien qui doit conduire à sa propre définition. Les liens entre les différents champs autorisent un accès direct aux unités inconnues ou oubliées. Afin de répondre à la demande de patients, certaines entrées bénéficient d'un développement plus important, permettant d'approfondir des notions centrales en cancérologie : *cancer*, *chimiothérapie*, *radiothérapie*, etc. Les articles, proposés en déroulé, permettent de répondre aux différents degrés de besoins d'information des patients et de faire du dictionnaire un outil interactif.

Au-delà de données linguistiques, l'informatisation permet l'intégration de données de différente nature : illustrations plus nombreuses, voire documents audio ou vidéo.

Tout ceci n'est guère original, si ce n'est à l'intégrer à un projet lexicographique. Nous voudrions plutôt insister sur la perspective de mettre en ligne les fiches descriptives et les cotextes qui ont guidé la description afin de les mettre à disposition des usagers.

4.2 Un dictionnaire en fiches

Un dictionnaire de vulgarisation vise une certaine efficacité didactique. Si la description des entrées doit permettre à l'utilisateur de comprendre la signification du terme décrit (décodage), elle devrait aussi lui permettre de produire des énoncés (encodage). Il lui faut pour cela percevoir le fonctionnement discursif du signe.

En outre, nous avons insisté sur la nécessité de proposer aux patients une définition élargie adaptée à leurs besoins d'information. Cependant il n'est pas sûr que la sècheresse d'une définition, jamais complète, parvienne à réaliser cet objectif…

En indiquant les relations que le terme-entrée entretient avec les autres unités de la langue, la fiche descriptive (celle-là même qui permet l'élaboration des définitions) offre des informations d'ordre lexical et paradigmatique d'une part, et

des informations syntagmatiques, proposant des éléments sur la combinatoire du signe, d'autre part. Dès lors, ces fiches peuvent être adaptées et constituées en outils mis à disposition des usagers, leur procurant des éléments pour *s'approprier* le terme et partant, pour l'employer (cf. Delavigne, 2001).

4.3 Un dictionnaire cotextuel

La valeur d'une unité varie en fonction de plusieurs éléments : sa distribution, ses constructions syntaxiques, les classes dans lesquelles il entre, etc. Intégrer ces différents éléments à la description permet l'accès à la signification et au fonctionnement d'un mot nécessite. Cette prise de position, qui insiste sur la nécessaire prise en compte de l'interaction du cotexte et du mot dans la production de sens, entraîne des conséquences méthodologiques, notamment lorsqu'il s'agit de construire un dictionnaire. Un dictionnaire est un objet qui, dans l'imaginaire culturel, est censé produire des définitions (cf. sens 1 de *dictionnaire* dans le *Petit Robert*), ce que nous nous attachons à faire.

Cependant, toute définition reflète le sentiment linguistique du descripteur, parfois moins pertinent que celui d'énonciateurs « réels ». Par ailleurs, la définition bloque la construction du sens dans une direction, celle choisie par le rédacteur, sans révéler la pluralité des « points de vue » possibles et la labilité des emplois. C'est un métadiscours soumis à des contraintes éditologiques fortes et qui résulte de choix contingents. Comme tout métadiscours, il est critiquable et nous aurions mauvaise grâce à en démontrer les imperfections. Nous n'irons même pas jusqu'à émettre l'idée d'une « impossible définition ». Malgré leurs défauts, le succès commercial des dictionnaires montre que les définitions restent de quelque utilité…

Cependant, une solution peut venir combler les manques de la définition

En mettant l'unité-entrée en action, des cotextes soigneusement choisis (cf. la typologie qu'offrait Josette Rey-Debove, 2005) la rendent souvent plus accessible et plus aisément appropriable. A côté de la définition, ce mode d'accès complémentaire au sens offre là une autre voie d'accès à l'appréhension du signe.

Les cotextes présentent en effet bien des avantages. Les vertus du cotexte tiennent à plusieurs choses. Il montre le signe avec ses cooccurrents, ses différents modes de combinaison et la façon dont le sens est saisi par différents types de locuteurs et donc, introduit à une plus grande gamme d'usages et à la variation. Le cotexte présente l'atout d'éclairer non seulement le sens du signe, mais aussi les *points de vue* qui peuvent être adoptés dès lors qu'il en est question, son fonctionnement en discours et quelques-uns de ses cooccurrents les plus fréquents. L'acte illocutoire à l'origine de leur énonciation offre fréquemment un ensemble de traits précieux pour l'appropriation du terme. Et les cotextes mettent souvent en œuvre des termes hyperonymiques ou isonymiques qui permettent de mieux appréhender les catégories.

Ainsi une sélection d'illustrations discursives - qui ne se réduisent pas à l'exemple – peuvent-elles permettre une interprétation du signe en requérant non seulement les compétences linguistiques de l'usager, mais aussi ses compétences *herméneutiques* qui participent à la production *des* sens, lui permettant d'en saisir les représentions sous-jacentes.

Il faut cependant résoudre d'épineuse question des droits d'auteurs…

Cette double stratégie dictionnairique, cotextes et fiches - qui ne sont, somme toute, que des modélisations des cotextes recueillis -, offre donc une certaine « ergonomie cognitive » (Delavigne, 2001). En faisant en sorte que le locuteur se familiarise avec le terme placé en situation et en exposant clairement ses relations avec les autres mots de la langue, elle en facilite ainsi l'appropriation.

5 Conclusion

LexOnco se veut un outil en ligne qui offre à l'usager la possibilité de personnaliser son accès au contenu en fonction de ses besoins.

Pour un véritable partage de l'information, l'on se doit cependant de veiller à ce que cet outil soit en accord avec les pratiques des usagers. Il reste encore à construire une interface ergonomique et en optimiser l'utilisation pour faire de LexOnco un outil utile pour les patients.

Les premières définitions devraient être publiées sur le site internet de l'Institut national du cancer au cours du deuxième semestre 2008.

Il nous faut encore souligner un point, loin d'être accessoire : à côté du dictionnaire en ligne, il est indispensable, à terme, d'envisager une publication papier.

Il peut sembler paradoxal après avoir montré les avantages d'un dictionnaire électronique de plaider pour un recours (retour ?) au papier. C'est simplement rappeler que la fracture numérique et les disparités dans l'usage d'internet sont une réalité (cf. Holzem, 2006, par exemple). C'est pourquoi une diffusion optimale doit laisser envisager un autre type de publication que la seule diffusion en ligne. Ce n'est pas emprunter le chemin à rebours, mais seulement prendre en compte la variété des usages et des échanges.

Références :

Baudouin, N., Holzem, M., Saidali, Y., Labiche, J. (2003). Acquisition itérative de connaissances en traitement d'images : consultation d'un collège d'experts. Actes des 14èmes journées francophones d'ingénierie des connaissances (IC 2003), Laval, 101-116.

Bensing, J.M., Verhaak, P.F., Van Dulmen, A.M., et al. (2000). Communication: the royal pathway to patient-centered medicine. Patient Educ Couns 39, 1-3.

Bouveret, M., Gaudin, F. (1996). Pistes de description sémantique : le cas de Biolex, dictionnaire des bio-industries. Actes du colloque Lexicomatique et dictionnairiques, AUPELF-UREF, Montréal, 349-357.

Carretier, J., Leichtnam-Dugarin, L., Delavigne, V., Brusco, S., Philip T., Fervers, B. (2004). Les SOR SAVOIR PATIENT, un programme d'information et d'éducation des patients atteints de cancer et de leurs proches, Bulletin du Cancer, 2004-91(4), 351-361.

Castagnoli, S. (2008). Corpus et Terminologie : raisons d'un mariage réussi. In Corpus et dictionnaires de langues de spécialité, Maniez F. et al. (dir.), PUG, Grenoble, 213-229.

Coulter, A. (1998). Evidence based patient information. is important, so there needs to be a national strategy to ensure it. BMJ 317, 225-226.

Coulter, A, Entwistle, V, Gilbert, D (1998). Informing Patients: An Assessment of the Quality of Patient Information Materials. King's Fund.

Coulter, A., Entwistle, V., Gilbert, D (1999). Sharing decisions with patients : is the information good enough ? BMJ 318, 318-322.

Delavigne V. (2001). Les mots du nucléaire. Contribution socioterminologique à une analyse des discours de vulgarisation, thèse de doctorat, Université de Rouen, 3 vol.

Delavigne V (2008). Construire un dictionnaire d'oncologie pour les patients : aspects méthodologiques. In *Corpus et dictionnaires de langues de spécialité*, Maniez F. *et al.* (dir.), PUG, Grenoble, 153-173.

Delavigne V. et Gaudin F (1996). À propos d'implantation terminologique. Questionner l'usage ou le sentiment linguistique ? *Le questionnement social. Cahiers de linguistique sociale* n°28-29, 131-140.

Depecker L. (dir.) (1997). *La mesure des mots : cinq études d'implantation terminologique*, PUR, Rouen.

Entwistle V. (2003). Patient's information environments : deserts, jungles and less hostile alternatives. *Health Expect* 6, 93-96.

Gaudin F. et Delavigne V. (1997). L'enquête en terminologie : point de la question et propositions. In *Terminologies Nouvelles* 16, 37-42.

Gaudin F., Holzem M. et Wable T. (1999). *Aménagement terminologique à partir des thèses soutenues devant l'université de Rouen*. Rapport à la DGLF, 2 tomes.

Gourdain P. *et al.* (2007). *La Révolution Wikipédia*. Mille et une nuits, Paris.

Grosjean M., Lacoste M. (1999). *Communication et intelligence collective. Le travail à l'hôpital.* PUF, Paris.

Hadlow J, Pitts M. (1991). The understanding of common health terms by doctors, nurses and patients. *Soc Sci Med* 32, 193-196.

Holzem M. (2006). Suppôt ou supporter des TIC, y a-t-il une fracture numérique ? *SDN* 2006.

Jacobi D. (2001). Savoirs non formels ou apprentissages implicites ? In *Recherches en communication* 15, 169-184.

Jadad A.R., Gagliardi A. (1998). Rating health information on the Internet: navigating to knowledge or to Babel? *JAMA* 279, 611-614.

Kilgarriff A (2005). Informatique et dictionnairique. In *Revue française de Linguistique appliquée* 2005-2.

Latour B. (1989). *La Science en action*. La Découverte, Paris.

Ligas P. (2008). Définition et exemple : quelle complémentarité ? L'illustration du concept dans le « Dictionnaire alphabétique et analogique du français des activités physiques et sportives » (à paraître, 2009). In *Lexicographie et Informatique. Bilan et perspectives. Colloque international*, Nancy, janvier 2008.

Pruvost J. (2006). *Les dictionnaires français. Outils d'une langue et d'une culture.* Ophrys, Paris.

Reboul-Touré S. (2004). Écrire la vulgarisation scientifique aujourd'hui. In *colloque Sciences, Médias et Société, 15-17 juin 2004, Lyon, ENS-LSH*, Disponible à : http://sciences-medias.ens-lsh.fr/article.php3 ?id_article=65

Rey-Debove J. (2005). Statut et fonctions de l'exemple dans l'économie du dictionnaire. In *L'exemple lexicographique dans les dictionnaires français contemporains. Actes des « Premières Journées allemandes des dictionnaires », Klingenberg am Main, juin 2004*, 15-20.

Stengers I. (2002). *Sciences et pouvoirs*. La Découverte. Paris.

Tran D. T. (1999). *La standardisation de la terminologie médicale vietnamienne. Une approche socioterminologique*, Thèse de Doctorat, Université de Rouen, 2 vol.

Session 3

Construction des Savoirs et Interprétation de Documents

Instrumenter la lecture savante de documents multimédia temporels

Instrumenting temporal multimedia documents scholarly reading

Thomas BOTTINI(1), Pierre MORIZET(1), Bruno BACHIMONT(1)

(1)Laboratoire Heudiasyc, UMR CNRS 6599, Université de Technologie de Compiègne (UTC), France
tbottini@hds.utc.fr
pierre.morizet-mahoudeaux@hds.utc.fr
bruno.bachimont@utc.fr

Résumé. L'évolution des technologies de numérisation et de diffusion documentaires confronte le lettré à des contenus et des méthodes de travail qui excèdent le cadre théorique et technique hérité de la tradition de la lettre et de l'imprimé. Prenant acte du caractère matériel et spatial de l'activité critique, où la manipulation des connaissances est conjointe à la manipulation des objets qui les incarnent, cet article propose une réflexion sur la nécessité de prolonger l'action de l'ingénierie documentaire dans le champ des interfaces homme-machine. C'est en effet par l'espace – autant celui de son environnement que celui de l'écriture – que le lecteur se rend maître de la temporalité de ses documents et de son projet interprétatif même. Notre enjeu est alors de comprendre les fondements de l'activité critique pour esquisser des directions structurantes pour son instrumentation informatique. Suite à l'exposition de ces aspects théoriques, nous présenterons quelques outils et instruments articulés dans un environnement de lecture savante multimédia, actuellement en cours de développement et de test.
Motsclés. Lecture savante, documents multimédia, temps, spatialisation, écriture.

Abstract. The evolution of document digitalization and diffusion technologies confronts the scholar to contents and reading methods which exceed the tradition inherited from the theoretical and technical frame of the letter and the print. Taking into consideration the material and spatial nature of the critical activity, where the manipulation of knowledge goes with the manipulation of the objects they are embodied in, this article proposes a reflection on the necessity to prolong the document engineering action in the field of computer human-interaction. It's indeed by the space – the one of the environment and the one of the writing – that the reader brings the temporality of his documents and interpretative plan under control. Our stake is then to understand the critical activity foundations so as to lay down directions for its computer instrumentation.
Keywords. Scholarly reading, multimedia documents, time, spatialization, writing

1 Introduction

Le travail documentaire savant se remodèle alors que les technologies de numérisation, de manipulation et de diffusion documentaires se développent. Cette évolution confronte notamment les lettrés à des contenus graphiques et temporels qui excèdent le cadre théorique et technique fourni par la tradition de la lettre et de l'imprimé (Ingraham, 2000). Pour que le sens, l'interprétation et la critique de tels contenus puissent se déployer, il faut que ceux-ci aient été « incarnés » dans des formes manipulables. Littéralement, le « lettré » est celui qui dispose d'une technique – la lettre, et tous les artefacts sur lesquels repose notre civilisation de l'écrit – lui permettant l'exercice du travail critique sur des connaissances. La numérisation de celles-ci ne les ancre plus dans une forme matérielle statique ; le but de cet article est alors de fournir des pistes réflexives sur l'importance des formes d'appropriation pour les usages savants. Plus largement, notre projet est de comprendre quelles sont les conditions techniques qui président à l'émergence d'une figure du « lettré du numérique ». Nous verrons dans un premier temps en quoi le travail intellectuel s'appuie avant tout sur des manipulations matérielles opérées par le lecteur. Nous aborderons alors l'articulation des aspects temporels et spatiaux de la lecture. La question de l'appropriation critique des documents – qui est un processus ou le rapport temps / espace est déterminant – nous amènera à insister sur la nécessité d'une réflexion sur les interfaces homme-machine dans l'instrumentation des pratiques savantes.

Les réflexions exposées dans cet article gouvernent le développement d'un environnement informatique de lecture savante multimedia au sein duquel peuvent être convoqués et manipulés des contenus textuels, graphiques et temporels. Les communautés de lecteurs concernées sont à ce jour les chercheurs en SHS, toujours plus fréquemment amenés à travailler sur des images ou des enregistrements sonores, et les musicologues. La section quatre sera ainsi consacrée à la présentation de cet environnement et à certains des enjeux technologiques soulevés par sa conception et son utilisation auprès des lecteurs.

2 La lecture savante : manipuler la connaissance

2.1 Une lecture active et spatiale

Clef de voûte des mondes lettrés, la lecture savante est l'activité par laquelle une communauté accroît son vivier documentaire par un travail critique effectué sur les connaissances existantes. Observer l'histoire des pratiques savantes fait ressortir leur caractère éminemment matériel et manipulatoire : le lecteur savant est avant tout un lecteur actif, plongé dans un corps-à-corps perpétuel avec les documents de son corpus et avec ses propres traces, qu'il saisit, déplace, empile, organise, compare, annote, fragmente, recombine... Cette forme de lecture est donc indissociable d'une écriture (Stiegler, 1994). Le jeu sur les connaissances s'accompagne d'un jeu sur les inscriptions qui les incarnent, sur la matière et la configuration des supports qui les accueillent. Cette idée charpente le récent – et imposant – ouvrage dirigé par Chirstian Jacob, Lieux de savoir (Jacob, 2007). Le

lecteur « savant » est d'ailleurs toujours savant dans la manipulation de ses outils et instruments. Des outils de lecture et d'écriture comme le codex, les livres de médecine pliants du Moyen-Âge, la roue à livres d'Agostino Ramelli, l'utopique Memex de Vannevar Bush, jusqu'aux récents logiciels de composition textuelle et hypermédia sont avant tout de nouveaux moyens de disposer et manipuler des inscriptions matérielles auxquels se rapportent autant d'élargissements des possibles cognitifs.

La condition fondamentale de cette « puissance manipulatoire » sur les connaissances est leur grammatisation, processus de discrétisation d'un flux en unités afin de le rendre contrôlable (Auroux, 1995). La grammatisation repose sur une spatialisation : en discrétisant et en spatialisant le flux temporel de la parole, l'écriture rend possible des opérations de classement, sélection, hiérarchisation, comparaison, etc. propres à toute activité critique. La synopsis spatiale offerte par des structures d'organisation matérielle telles que la liste ou le tableau permet de comparer et de revenir infiniment sur les inscriptions (Goody, 1979). Une réflexion sur l'instrumentation informatique du travail intellectuel portant sur des connaissances ne peut alors faire l'économie d'un questionnement sur la nature de leur substrat matériel et de leurs modalités de réappropriation sensible dans l'espace.

2.2 Des opérations cognitives et matérielles

En vue d'instrumenter informatiquement une lecture savante multimedia, il nous apparaît important de comprendre les opérations qui sous-tendent l'activité critique documentaire en général. Au-delà de l'important polymorphisme que présentent ces pratiques selon le rattachement disciplinaire du lecteur et des idiosynchrasies gestuelles et méthodologiques propres à celui-ci, nous pensons qu'il est possible de circonscrire un socle opératoire commun. Critiquer un corpus documentaire suppose avant tout de pouvoir en appréhender la forme, y naviguer,y localiser et identifier les éléments de contenu pertinents. Les documents doivent également pouvoir être annotés, qu'il s'agisse de rendre plus saillant un élément (par exemple, pour faciliter les opérations de navigation) ou de les enrichir (glose). L'objectivation et la représentation d'une structure documentaire – matérielle ou logique – permet de faciliter l'appropriation et l'exploitation (par exemple, Faure and Nicole, 2007). Annotations et saillances structurelles, en tant que discrétisations, peuvent alors servir de base à une fragmentation, geste qui constitue le pendant matériel de l'analyse. L'activité critique repose en définitive sur les possibilités qu'a le lecteur de faire varier les rapports spatiaux qui structurent son matériel lectorial pour en faire émerger des configurations sémantiques nouvelles.

Le fondement commun à ces opérations est la nécessité pour le lecteur de disposer d'espaces, de lieux, qu'il pourrait manipuler, articuler et reconfigurer pour marquer son cheminement interprétatif. L'ingénierie documentaire, en opérant une grammatisation des contenus (structuration et qualification) au niveau de leur forme d'enregistrement, ne fait que rendre potentielles les opérations évoquées supra. Pour s'actualiser, l'activité critique suppose que les contenus, structures et descripteurs documentaires soient incarnés dans des formes d'appropriation plongées dans le flux perceptif et gestuel du praticien. Si dans le monde physique l'étendue spatiale est la condition même de l'existence des choses, dans le « monde numérique », elle est évacuée a priori, car celui-ci sépare forme d'enregistrement et forme d'appropriation (Bachimont, 1998). La notion d'interface témoigne de cet écart entre ces deux niveaux de matérialité, auxquels correspondent deux régimes sémiotiques. L'interface homme-machine (par la suite, IHM) est alors le milieu technique où s'opère la reconstruction de la dimension matérielle et manipulatoire.

3 Le temps et l'espace des documents : une question d'IHM

Pour penser le rapport du lecteur savant aux documents sur lesquels il travaille, nous nous appuierons sur des concepts tirés des travaux de l'historien jésuite Michel de Certeau. Dans (Certeau, 1990), celui-ci étudie le rôle de l'espace dans des pratiques temporelles telles que la marche ou la lecture en faisant usage du couple de concepts stratégie / tactique. La stratégie « postule un lieu susceptible d'être circonscris comme un propre », celui-ci étant « une victoire du lieu sur le temps ». La réussite d'un projet lectorial se caractérise par la capacité qu'a le lecteur d'éviter la dispersion, la désorientation et l'oubli. Dans les termes de Michel de Certeau, il s'agit alors de doter le lecteur informatisé d'un « lieu propre » où il pourrait « capitaliser » ses traces pour « préparer des expansions futures ».

3.1 Le « corps » des documents temporels

Le premier « oubli » auquel est confronté le « lecteur multimédia » est celui qui est consubstantiel à l'écoulement temporel d'un document sonore. Privé de lieu, l'écoute impose en effet un rapport tactique au contenu qu'elle vise : selon les termes de Michel de Certeau, « du fait de son non-lieu, la tactique dépend du temps, vigilante à y « saisir au vol » des possibilités de profit. Ce qu'elle gagne, elle ne le garde pas. » Or pour critiquer un contenu temporel, il faut que l'écoute puisse « capitaliser » les traces interprétatives afin de pouvoir revenir dessus. Autrement dit, elle doit pouvoir s'écrire, et donc disposer d'un support spatial. Des documents sonores tels que des interprétations musicales peuvent par exemple s'appréhender par l'intermédiaire de leur transcription spatiale traditionnelle qu'est la partition. Si l'annotation d'un document textuel peut se faire dans son corps même, le document sonore suppose des outils d'annotation, éventuellement en cours d'écoute, articulant leur temporalité propre à leurs enrichissements spatiaux textuels ou graphiques. L'IHM devient alors le lieu où peut se construire une appropriation multimédia reposant sur la synchronisation d'éléments de contenus de natures hétérogènes. En offrant réversibilité et synopsis aux contenus temporels, les représentations spatiales permettent les opérations d'annotation, de structuration ou de fragmentation, et transforment ainsi le rapport tactique imposé par l'écoute seule en un rapport stratégique pouvant s'articuler avec un projet lectorial dépassant les simples frontières du document isolé.

3.2 Critique et spatialisation

Des entretiens avec des chercheurs en SHS et des musicologues ont mis en exergue l'impérieux besoin de spatialiser dès lors que plusieurs documents doivent être étudiés, ce que Certeau désigne comme étant une « pulsion scopique ». Comme nous l'avons vu supra, l'activité critique n'a pas le regard collé à ses objets. Au contraire, elle les considère dans un contexte au sein duquel elle peut les reconfigurer à loisir. Le rôle de ce contexte d'accueil – qui est un lieu, un espace – est traditionnellement joué par le bureau. Le recul synoptique exige alors que l'IHM permette la coprésence de plusieurs documents dans un même espace, et sache articuler différents point de vue sur le corpus. La conduite d'une stratégie interprétative doit s'accompagner d'une « maîtrise des lieux par la vue ». Ceci est bien souvent négligé par les outils logiciels de lecture et d'écriture qui tendent à enclore les contenus dans des espaces isolés, interdisant ainsi la construction de liens sémantiques, de réseaux d'annotations, de mise en listes ou en tableaux et autres techniques de mise en relation critiques.

3.3 Vers une lecture augmentée

Le lecteur doit également négocier avec la temporalité de sa lecture, qui articule différentes phases, de l'annotation du corpus à la rédaction finale en passant par les configurations intermédiaires que reçoit son matériel documentaire. Les IHM modernes permettent une grande multiplicité des modes de représentation et de manipulation : le lien entre le contenu et sa forme d'appropriation matérielle n'étant plus figé, il devient pensable de matérialiser spatialement dans l'environnement de lecture l'intégralité des objets, « cheminements », paramètres et opérations mobilisés par le lecteur alors qu'il accomplit son objectif interprétatif. Ainsi objectivées, les « trajectoires interprétatives » empruntées peuvent à leur tour faire l'objet d'une stratégie, d'un processus de navigation. C'est d'ailleurs ce qu'offre la carte (spatiale) au voyage (temporel) : maîtrise, prévision et communicabilité. L'environnement du lecteur informatisé n'est pas seulement un espace d'inscription flexible et polymorphe des entités documentaires, c'est également un espace d'inscription des gestes savants présidant à l'élaboration de celles-ci. Par essence, la technique informatique repose sur la « mise en espace » de processus temporels (c'est le concept de programme). L'objectivation, la grammatisation, des gestes savants par le truchement de l'interface homme-machine permet leur manipulation, leur communication, leur mémorisation, et donc leur mise à disposition pour la communauté.

4 Eléments d'application

Cette section vient illustrer la mise en pratique de certains des aspects théoriques exposés précédemment. Nous nous appuierons sur le prototype d'environnement logiciel conçu et développé dans le cadre du projet Poliesc (Pratiques Ordinaires, Lectures Intensives et écritures Structurées de Contenus numériques multimédias). Celui-ci ambitionne de définir les conditions cognitives et technologiques d'une lecture active, intensive ou savante, effectuées sur des ensembles documentaires hétérogènes (texte, images et sons). Notre démarche de conception technique prend ses sources dans une enquête historique et philosophique qui, de l'invention de l'écriture aux dispositifs cognitifs qui nous sont contemporains, entreprend d'élucider les rapports entre le déploiement d'une pensée critique et les possibilités – et contraintes – matérielles offertes par les différents supports documentaires. À la lumière de ce cheminement, les sections précédentes se veulent être un premier compte-rendu de nos observations et prescriptions pour la conception d'outils et instruments de lecture numériques « multi-documents » et multimédia. Il s'agit là d'un des problèmes de recherche exposés dans (Bachimont, 1998) (où la question de l'appropriation personnelle de documents audiovisuels est considérée via l'opération centrale d'annotation).

Précisons que nous héritons de la distinction entre outil et instrument de Gilbert Simondon, qui, dans (Simondon, 2001), précise que l'outil est « l'objet technique qui permet de prolonger et d'armer le corps pour accomplir un geste » et l'instrument, « l'objet technique qui permet de prolonger et d'adapter le corps pour obtenir une meilleure perception ». Comme nous allons l'exposer infra, cette distinction se révèle particulièrement structurante dans le contexte d'une instrumentation (et donc également, d'un outillage...) de l'activité critique.

Il est également nécessaire de souligner que notre approche informatique repose sur une volonté de considérer la lecture savante dans sa globalité, dans l'étendue des opérations matérielles sur lesquelles elle repose (annotation,

comparaison, (re)structuration, mise en relation de fragments prélevés sur les documents du corpus, en vue de produire une glose toujours plus riche et structurée), ainsi que dans l'articulation de celles-ci au sein d'un même espace logiciel. Il ne s'agit pas de s'attarder sur telle ou telle question relevant de l'ergonomie de lecture « mono-documentaire », cette problématique étant déjà très largement balisée, ce autant dans le champ de l'analyse des pratiques et des usages en Sciences de l'Information et de la Communication que dans celui des IHM. Dans cette optique, des entretiens ont été réalisés avec des chercheurs et enseignants dans différentes disciplines des Sciences Humaines et Sociales. Il a été demandé aux sujets d'expliciter leurs pratiques de lecture et d'écriture (préparation de cours, rédaction d'articles, veille, prise de notes, etc.) pour identifier les dispositifs qu'ils utilisent et l'articulation qu'ils en font (ce en vue de mettre à jour d'éventuels détournements ou frustrations). Ces entretiens ont pu mettre à jour certains facteurs clefs liés à la perception et à la maîtrise spatiale d'ensembles documentaires multimédia, qui nous serviront de guides dans ce qui suit. Nous achèverons cette section applicative avec des enjeux et problématiques liés à l'articulation du spatial et du temporel soulevés par l'informatisation de pratiques d'analyse musicologique.

4.1 La nécessité d'un modèle documentaire souple et générique

La première constatation que nous tirons des entretiens est que l'« instrumentarium » de lecture et d'écriture le plus couramment utilisé est fragmenté et spécialisé. Dans le cas de la rédaction d'un article scientifique, l'utilisateur doit en effet jongler entre un outil de mind-mapping pour organiser des idées dans la phase précédent celle d'écriture, un traitement de texte pour composer le document final, un simple système de gestion de fichiers pour s'orienter dans le corpus de références bibliographiques, etc. Cet éclatement fonctionnel a pour conséquence de partitionner l'espace global de travail en autant d'espaces clôts, chaque entité documentaire obéissant alors à des règles hétérogènes de codage et de manipulation. La première conséquence négative de cette organisation est la très grande difficulté – ou l'impossibilité – d'exploiter le travail accompli au sein d'un autre environnement que celui qui lui a donné naissance (par exemple, un traitement de texte est aveugle à un plan complexe bâti au sein d'un logiciel de mind-mapping, ce qui l'ampute de toute possibilité de réutilisation). Cette observation nous a amené à concevoir un modèle de données générique susceptible de représenter l'ensemble des objets et relations mobilisés dans une lecture savante.

Un premier niveau, constitué d'Entités de Contenu se déclinant en Entités Textuelles, Entités Graphiques et Entités Sonores, représente la matière documentaire selon les spécificités propres à chaque média. Ainsi, les Entités de Contenu organisent des Sélections opérées sur les ressources documentaires, qui servent de base à des opérations d'annotation, de mise en exergue, de qualification locale ou de mise en relation avec d'autres éléments.

Ce niveau matériel s'accompagne d'un niveau sémantique, constitué d'Entités Sémantiques susceptibles de recevoir un ensemble de métadonnées textuelles et de catégories, ainsi que de faire l'objet d'une mise en relation par l'intermédiaire de liens typés[1]. C'est parce qu'une Sélection est associée à une Entité Sémantique qu'elle peut être sémantiquement reliée à d'autres objets ou enrichie textuellement (une

[1] Ces objets permettent ainsi ce que D. Cotte désigne comme étant un « marquage », qui est « endogène » au document (Cotte, 2000).

annotation est, précisément, la délimitation d'une portion matérielle dans la continuité d'un document, enrichie d'informations).

Un troisième niveau propose des Entités Structurelles pouvant être assemblées selon certaines contraintes pour donner naissance à des structures d'organisation documentaire telles que les listes, arbres, tableaux, treillis, etc. Ces briques structurelles sont des Entités Sémantiques (au sens de l'héritage, en programmation orientée-objet), et peuvent être associées à une Entité de Contenu (au sens de la composition). Ces relations structurelles sont généralisées du plus infime fragment à l'ensemble documentaire dans sa totalité, chaque élément prenant place dans une ou plusieurs Entité(s) Structurelle(s) de niveau supérieur. La navigation dans l'intégralité des éléments de contenu est ainsi facilitée par l'homogénéité structurelle de leur organisation. Transversalement à ces liens structurels – qui témoignent de la composition des documents en sous-éléments – peut être déployé un réseau hyperdocumentaire, dont les liens sont librement tissés et typés entre Entités sémantiques.

Cette incursion informatique est nécessaire pour comprendre les fondations de l'interopérabilité des différents modules accessibles dans l'environnement.

4.2 Spatialisation de structures et de réseaux documentaires

Comme le font justement remarquer F. Ghitalla et C. Lenay (Ghitalla et Lenay, 2001), « Là où les documents papier imposaient leur format matériel, avec un ordinateur il incombe au lecteur de gérer, plus qu'avant, l'apparition des documents et l'organisation « phénoménale » de l'espace-écran. ». Cette observation se révèle d'autant plus criante lorsque l'on considère des tâches relevant de la critique et mobilisant donc explicitement l'activité du lecteur. S'il ne peut être immédiatement rattaché à la figure de l'auteur, en tant que producteur de documents ex-nihilo, celui-ci est, a minima, le bâtisseur d'un édifice interprétatif reposant sur une construction inter-documentaire. Repérer, délimiter, annoter, extraire, mettre en relation, rapprocher, etc. sont autant d'opérations qui requièrent un outillage de création et non seulement de lecture (au plus pauvre des sens que celle-ci peut recevoir)[2] Par exemple, bien des travaux ont montré que les premiers temps de l'activités critique ou rédactionnelle reposent sur une forte plasticité structurelle (par exemple, (Nakakoji et al., 2005)), et ne peuvent être encloses dans des structures purement arborescentes. La figure 1 expose une vue de l'environnement où, au sein d'un même espace bidimensionnel libre, cohabitent listes, tableaux et agrégats informés de fragments documentaires. Les modalités d'interaction induites par une libre maîtrise des relations structurelles et spatiales entre objets confèrent une spontanéité certaine dans leur réorganisation et comparaison. De plus, taille, position relative, transparence ou couleur des fragments documentaires sont – au même titre que l'indication d'un auteur, d'une date, d'un titre ou que la présence d'un ensemble de tags sémantiques – des propriétés qui participent directement à la

[2] Pour donner au lecteur des moyens d'action et d'organisation sur les objets et sur l'environnement de son activité, nous pensons que la conception de dispositifs techniques pour les lectures « savantes » exige de dépasser le simple champ des outils de « perception » (visualisation, navigation, etc.) documentaire. En tant que construction matérielle, ces pratiques doivent pouvoir bénéficier des concepts développés dans le champ des outils de création (musicale, graphique, cinématographique), dont un facteur d'effectivité est précisément la possibilité d'une fine maîtrise de l'espace de l'œuvre en cours d'élaboration.

constitution du sens, et à ce titre ne doivent donc pas être écartées du processus de grammatisation.

Les entretiens ont également mis en lumière deux besoins dont la nécessité se ressent davantage à mesure que l'ensemble documentaire de travail croît en complexité (multiplication des annotations et des relations interdocumentaires, enrichissement du corpus, structuration toujours plus précise, création de nouveaux documents, etc.).

Le premier d'entre eux est la possibilité de partitionner l'espace en zones (Workspaces), où peuvent être rassemblés des éléments mobilisés dans une tâche lectoriale identifiée. Par exemple, rédiger un état de l'art demande de jeter un regard simultané sur des extraits d'articles abordant le domaine étudié, ce à quoi peut-être destiné un Workspace libre tel que celui de la figure 1. La réappropriation d'un document peut quant-à-elle supposer une concentration plus affirmée sur celui-ci, évacuant toute présence de contenus extérieurs. La figure 3 témoigne d'une telle situation, où l'utilisateur a assigné à un Workspace un outil de structuration et d'annotation sonore (voir sous-section 4.3). En matérialisant une macro-structuration de son espace documentaire, les Workspaces constituent ainsi pour le lecteur un moyen d'organiser le déroulement de son objectif global de lecture ou d'écriture.

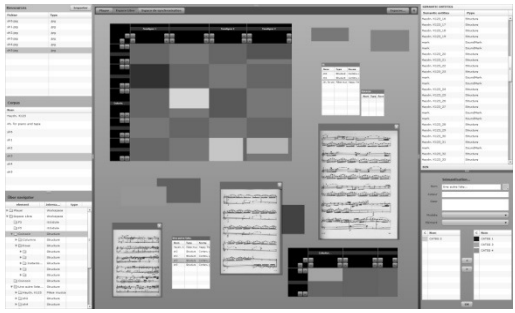


Figure 1. *Un espace libre bidimensionnel. Autour, des outils de sémantisation et de visualisation structurelle accessible à chaque endroit de l'environnement.*

Le second besoin est la nécessité de synthèse, et donc de synopsis spatiale. La « forme » d'un espace documentaire peut devenir difficilement appréhendable quand les éléments qui le composent sont structurellement très ramifiés et connectés par un réseau de liens hyperdocumentaires entremêlés. Le concept de carte, telle que l'illustre la figure 2, satisfait alors à cette exigence.

Figure 2. *Cartographie dynamique d'un espace hyperdocumentaire.*

Remarquons que ce type de cartes est à un niveau inter et extra documentaire ce que la mise en forme matérielle (propriétés graphiques et spatiales des blocs de texte) est au niveau intra documentaire : une incarnation spatiale et matérielle dont les propriétés graphiques se veulent autant le reflet du sens que des « poignées » qui en permettent une manipulation active. Ainsi, les outils de structuration et d'« hyperconnexion » libres doivent être couplés à des instruments d'orientation aptes à satisfaire toute « pulsion scopique ». Cet équilibre des dispositifs d'action et de perception est la condition d'un « jeu » efficace sur la matière documentaire et de l'édification intelligible de l'espace dans lequel elle se déploie. Un point capital de la conception de technologies pour le travail intellectuel réside alors en l'articulation entre outils d'action et instruments de perception. La critique et la production documentaire reposent sur une perpétuelle variation du point du vue et des modalités d'action, du corps à corps avec un document au surplomb complet du corpus.

4.3 Temps et espace dans la construction et l'étude de documents multimédia et hypermédia

Parallèlement à une réflexion sur les aspects inter-documentaires d'une lecture savante, notre environnement propose des outils de structuration et d'annotation de contenus textuels, graphiques et sonores. Il ressort des entretiens avec les « lecteurs savants » que ce dernier point est particulièrement problématique. Du fait des fréquentes fragmentation et spécialisation fonctionnelles évoquées supra, les documents sonores ne sont pas intégrés aux outils de composition textuelle traditionnellement utilisés dans la rédaction de contenus scientifiques ou pédagogiques. Les chercheurs interrogés voient donc dans le fichier audio un document temporaire, dont la possibilité d'exploitation réside dans une transcription textuelle. Certains d'entre eux expliquent d'ailleurs l'écartement de ce type de ressources par le manque d'outils critiques graphiques basé sur la spatialisation d'éléments sur lesquels il est possible de faire varier des paramètres de couleur ou de taille, comme ils ont coutume de le faire au sein de documents textuels ou picturaux pour s'y repérer et y souligner les passages pertinents. Nous avons généralisé la portée de tels propos dans les sections précédentes, en énonçant le rôle des manipulations spatiales dans le travail sur un flux temporel. La figure 3 expose un composant permettant d'effectuer sur de tels contenus des opérations de structuration hiérarchique et d'annotation libre. Par un geste simple de définition d'« instants clefs » déterminés à l'écoute, il est possible de matérialiser graphiquement des segments temporels pertinents. Chaque nouvel élément ainsi défini gagne alors une certaine autonomie, et peut être manipulé selon les possibilités du modèle (sémantisation, hyperliens, structuration, réutilisation, etc.). La rationalité spatiale permet ainsi d'aborder facilement un objet temporel, et de l'exploiter au sein d'un ensemble documentaire via des dispositifs présentant une certaine homogénéité logique et manipulatoire.

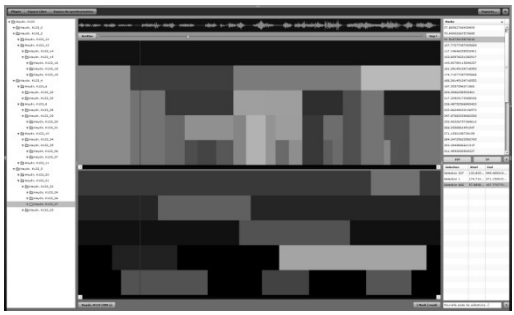


Figure 3. *Segmenteur audio : structuration et annotation d'un contenu temporel*

L'étude des rapports entre temps et espace dans les pratiques de lecture savante nous a amené à considérer le cas des partitions musicales synchronisées à leurs interprétations. Cette réflexion technologique s'inscrit dans le cadre d'un travail de recherche plus large sur les apports de l'informatisation des méthodes d'analyse musicologique basées sur des opérations de segmentation et de mise en tableau de partitions. L'informatique multimédia permet de réintroduire l'écoute dans des pratiques traditionnellement basées sur les possibilités du support papier, offrant ainsi à l'analyste un accès à des informations que les formalismes d'écriture mobilisés dans les partitions ne peuvent entièrement capter. Nous nous intéressons aux cas, largement majoritaires, où les documents disponibles aux musicologues sont de simples images de partitions et fichiers audio numérisés « manuellement », c'est-à-dire n'ayant jamais fait l'objet d'un processus d'indexation ou de synchronisation en amont par un système d'information. Intégré au processus critique, le geste de synchronisation consiste alors à définir des points d'équivalence entre des instants temporels au sein des différentes interprétations et des points spatiaux correspondant sur les partitions. La réunion de ces points selon la temporalité de l'œuvre constitue un document hypermédia, dont la perception et la manipulation par l'utilisateur nécessite une représentation spatiale donnant à voir sa composition. Si l'organisation interne des documents hypermédia est une thématique désormais classique de l'ingénierie documentaire, celle-ci ne dit rien quant aux modalités sensorielles de leur construction et de leur exploitation critique. La figure 4 présente un Workspace où peuvent être importées interprétations audio et partitions graphiques. Guidé par ses yeux et ses oreilles, l'utilisateur peut ainsi, par un jeu intuitif de la souris et du clavier, définir des points de synchronisation. La partie droite de l'écran expose alors une vue synoptique des relations inter-documentaires ainsi créées entre la partition graphique (préalablement structurée en pages, systèmes de portées et portées) et l'interprétation en cours de synchronisation. Lorsque cette opération est répétée sur différentes transcriptions et interprétations de la même œuvre, la carte synoptique de la figure 5 permet alors d'appréhender la forme et la constitution de l'hyperdocument ainsi construit.

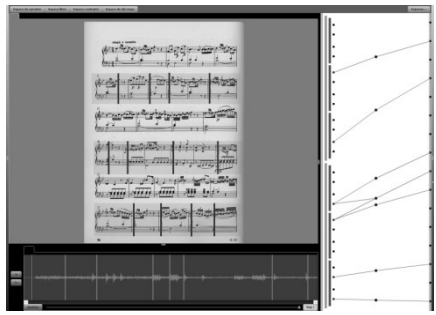


Figure 4. *Espace de synchronisation image / son.*

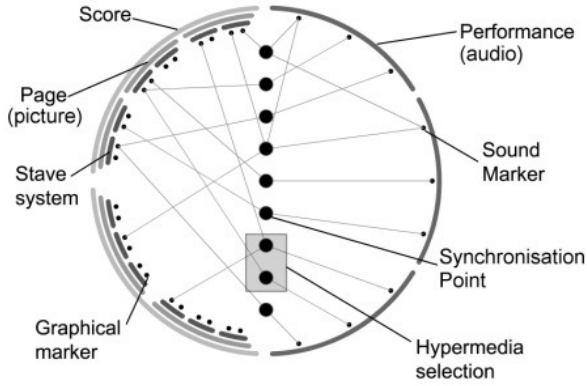


Figure 5. *Vue synoptique des points de synchronisation entre deux partitions graphiques structurées et trois interprétations audio.*

De telles représentations ambitionnent donc de matérialiser et spatialiser des entités documentaires qui n'existeraient pas sans la calculabilité informatique, afin de rendre possibles les opérations savantes fondamentales. Ces opérations sont jusqu'ici l'apanage des technologies purement textuelles, ou des outils destinés à la création. L'enjeu d'un environnement destiné à des utilisateurs savants dans leurs pratiques documentaires, mais non nécessairement savants dans le maniement de l'outil informatique, est alors de proposer des IHM au sein desquels les objets documentaires complexes s'offrent à la perception et à l'action de manière claire et intuitive. La question du « tactile » émerge souvent dans les entretiens que nous avons conduis. La réappropriation de corpus documentaires hyperliés et structurés ne peut en effet passer que par leur incarnation sensible.

5 Conclusion

Ces quelques réflexions sur l'importance des représentations spatiales, qui aboutit à la nécessité de conduire une réflexion dans le champ des IHM pour des contextes d'usage savant, nous amène à redéfinir ce que peut être le document numérique. Afin que celui-ci puisse renseigner (docere) et être effectivement manipulé, il doit être incarné et instrumenté. Cette instrumentation dépasse la

simple question de la modélisation et de l'enregistrement de son contenu, et côtoie alors la notion de logiciel. La calculabilité numérique amène au travail intellectuel de nouvelles possibilités manipulatoires et représentationnelles, conduisant notamment à l'émergence d'une Raison Computationnelle (Bachimont, 2006). Tirer partie du « supplément du numérique » dans l'instrumentation de pratiques documentaires demande une réflexion sur les modalités esthétiques des rapports que tisse le lecteur avec les connaissances.

Notre démarche de conception et de développement repose sur l'idée que pour aborder ces questions d'« incarnation » d'ensembles documentaires complexes et multimédia dans le cadre d'une lecture savante, on ne peut faire l'économie d'une réflexion sur le modèle de données qui les organise et les « inscrit ». Leurs modalités d'exploitation reposent alors sur l'articulation d'outils et d'instruments, dont l'intégration fine au sein d'un environnement est la condition d'un travail critique efficient.

Une première prolongation de notre travail consiste en une réflexion sur les conditions cognitives et techniques de la rédaction d'articles non plus seulement textuels mais multimédia. Adossé à des systèmes de chaînes éditoriales permettant une publication multi-support, nous envisageons à moyen-terme de faire de notre environnement un espace d'écriture répondant à une telle nécessité. Cette question est fréquemment posée par les revues en ligne, notamment celles-ci qui ont à faire à des documents temporels. Force est de constater que les environnements d'écriture hypermédia idéalisés depuis Vannevar Bush ne font toujours pas partie de notre quotidien de lettrés.

Les quelques outils présentés dans cette section illustrent ce que peut être un environnement logiciel de travail documentaire personnel où le paramètre spatial est pris en compte par le modèle et offert au lecteur comme support supplémentaire de son interprétation. Ainsi « écrite » et mémorisée, l'organisation et la configuration spatiale que lecteur confère à son espace de travail et aux éléments qui le peuplent deviennent transmissibles. Cette possibilité trouve un écho particulier dans le contexte de nos travaux d'applications musicologiques. Il est en effet fréquent que lors de son travail sur une œuvre le musicologue fasse appel à des analyses passées. Nous comptons ainsi prolonger notre travail sur l'inscription spatiale vers une réflexion sur la transmission du geste lectorial, de l'environnement au sein duquel les choix, pistes abandonnées, hésitations et autres errements du lecteur sont matérialisées.

Remerciements

Ce travail s'inscrit dans le cadre du projet POLIESC soutenu par la région Picardie et le Fonds Social Européen.

Références :

Auroux, S. (1995). La révolution technologique de la grammatisation. Mardaga, Paris.

Bachimont, B. (1998). Bibliothèques numériques audiovisuelles : des enjeux scientifiques et techniques. Document Numérique, vol. 2, num. 3-4, 219-242.

Bachimont, B. (2006). Support de connaissance et intelligence collective : héritage et individuation technique. In Actes des Rencontres Intelligence Collective 2006, Nîmes, Mai.

De Certeau, M. (1990). L'invention au quotidien, tome 1 : Arts de faire. Gallimard, Paris.

Cotte, D. (2000). Représentation des connaissances et convergence numérique : le défi de la complexité. Document Numérique, vol. 4, num. 1-2, 167-182.

Faure, C., Nicole, V. (2007). Document Image Analysis for Active Reading. In Proceedings of the 2007 international workshop on Semantically aware document processing and indexing . ACM, New York.

Ghitalla, F., Lenay, C. (2001). Largeur et profondeur des espaces de compréhension dans l'exploration des réseaux numériques. Colloque Interdisciplinaire en Sciences Cognitives.Disponible à : http://www.isc.cnrs.fr/ARCo2001/VOdesVF/ghitalla.doc.

Goody, J. (1979). La raison graphique. Les Editions de Minuit, Paris.

Ingraham, B. (2000).Scholarly Rhetoric in Digital Media. Journal of Interactive Media in Education. Disponible à : www-jime.open.ac.uk/00/ingraham/.

Nakakoji, K., Yamamoto, Akaishi, M., Hori, K. (2005). Interaction design for scholarly writing: Hypertext representations as a means for creative knowledge work. New Review of Hypermedia and Multimedia, vol. 11, num. 1, 39-67.

Jacob, C. (2007). Lieux de savoir , tome 1 : espaces et communautés. Albin Michel, Paris.

Simondon, G. (2001). Du mode d'existence des objets techniques. Aubier, Paris.

Stiegler, B. (1994). Machines à écrire et matières à penser. Genesis, num. 5, 25-49.

Conception et usages d'un environnement numérique de travail pour une aide à l'interprétation de documents juridiques.

Design and uses a digital environment working for an aid to interpreting legal documents.

GROUPE ν

(1) GREYC, Université de Caen, Caen, France
Pierre.Beust@info.unicaen.fr , Stephane.Ferrari@info.unicaen.fr ,
Fabrice.Maurel@info.unicaen.fr , Serge.Mauger@unicaen.fr

(2) LiDiFRa, Université de Rouen, Rouen, France
Maryvonne.Holzem@univ-rouen.fr, n.baudouin3@laposte.net

(3) LITIS, Université de Rouen et INSA de Rouen, St Etienne du Rouvray, France
Eric.Trupin@univ-rouen.fr , Youssouf.Saidali@univ-rouen.fr,
Jacques.Labiche@univ-rouen.fr , Jean-Philippe.Kotowicz@insa-rouen.fr ,
Nathalie.Chaignaud@insa-rouen.fr ,

(4) Laboratoire Psychologie des Actions Langagières et Motrices (PALM)
et PPF Modelisation en Sciences Cognitives (Modesco)
Denis.Jacquet@unicaen.fr

Résumé. Cet article présente un projet de recherche en cours visant à améliorer les interactions d'utilisateurs de différentes catégories professionnelles avec un système d'information dédié au droit du transport et de la logistique qui repose sur un corpus de textes réglementaires et de compte rendus de jurisprudence. L'objectif vise à concevoir et à mettre au point un environnement numérique de travail (ENT) destiné à un public professionnel (entreprises de la filière logistique, juristes, risk managers, assureurs, avocats, etc.) et non professionnel (usagers ou salariés des transports). Après avoir posé la question de l'appropriation des contenus dans le cadre des documents numériques, nous décrirons les spécificités de notre corpus de travail. Nous placerons alors notre projet dans un cadre théorique actuellement novateur au sein des sciences cognitives, celui de l'énaction. Ceci nous amènera à proposer une approche résolument centrée utilisateur dans la conception de l'ENT. Nous terminerons par une description des premières spécifications du futur ENT.
Motsclés. Environnement numérique, usages, corpus, jurisprudence.

Abstract. This article presents a research project in progress aiming at improving

the interaction of users of different professional groups with an information system dedicated to transport law and logistics based on a corpus of regulations and reports Court. The goal is to design and develop a digital working environment (ENT) for a professional public (companies in the logistics industry, lawyers, risk managers, insurers, lawyers, etc.) and nonprofessional (users or employees in the field of transportation). After having placed the question of acquisition of content in the context of digital documents, we will describe the specifics of our corpus. Then, we will place our project in a theoretical innovator framework in cognitive science, the one of enaction. This will lead us to propose a resolutely user focused approach in the design of the ENT. We conclude with a description of the first specifications for the future ENT.
Keywords. Digital environment, uses, corpus, jurisprudence.

1 Le passage au numérique : prouesses techniques versus appropriation des contenus

Avec l'apparition des techniques numériques et de l'internet, la distance entre la population et l'information économico-juridique tend apparemment à se réduire. De nombreux sites proposent aujourd'hui une large gamme d'informations générales ou spécialisées. Sur le plan juridique, par exemple, il est aujourd'hui possible d'accéder à un large pan de la réglementation et de la jurisprudence, qu'elles soient françaises ou étrangères. Mais ce rapprochement technique n'est pas pour autant signe d'une meilleure maîtrise et appropriation de l'information, bien au contraire. Les spécialistes en sciences de l'information qui s'intéressent à la gestion des flux ainsi qu'aux stratégies de mise en forme et de mise en circulation de l'information font le constat des écarts (des fossés) entre prouesses technologiques et appropriation des contenus par des utilisateurs de plus en plus hétérogènes au sein d'une économie mondialisée. La question de l'appropriation ne se résout pas en effet par le simple ajout de métadonnées aux documents numériques comme dans le projet du Web Sémantique où l'objectif annoncé par Tim BernersLee (1998), initiateur du projet et directeur du W3C, est d'enrichir (notamment au moyen des technologies développées autour du langage XML) les documents (à l'aide d'ontologies normalisées, soit automatiquement, soit en assistant leurs auteurs) avec « des informations sur leur propre sémantique qui soient directement interprétables par des agents logiciels sans la supervision d'une interprétation humaine ». Ce positionnement fait l'hypothèse que la valeur sémantique d'un passage de document n'est le fait que de son auteur (alors que c'est tout autant celui de son lecteur confronté à ses pratiques professionnelles). Il s'agit d'une voie aristotélicienne déjà fort ancienne, reprise par F. Bacon puis G. W. Leibniz au 17ème siècle, persuadés de la nécessité d'un système universel d'organisation des connaissances de nature métaphysique et donc indépendant des points de vue particuliers. Ces universaux cognitifs seraient alors à même de pourvoir à la circulation de l'information, celle-ci occultant la question liée à son interprétation. Comme l'a souligné R. T Pédauque[1], « la réponse du Web sémantique est d'indexer les textes avec les concepts d'une ontologie partagée par une large communauté » sans que celle-ci ne soit d'ailleurs clairement définie, ni surtout que soient prises en considération la diversité et l'évolution des pratiques langagières au sein de sphères d'activités hétérogènes.

[1] Acronyme développé en analogie avec le Réseau Thématique Pluridisciplinaire sur le Document (RTP DOC) sous lequel des chercheurs issus des départements STIC et SHS ont tenté d'approfondir collectivement la réflexion sur le document numérique.

Nous constatons que les systèmes actuels (interface de dialogue, bases de données, etc.) conduisent à une interaction Système/Utilisateur forcément appauvrie, parce qu'ancrée dans un environnement prédéfini (Peschard, 2004), celui de réponses du système exprimées sous la forme de thésaurus qui réorientent la question de l'utilisateur[2]. Aussi, nous jetterons ici les bases de la conception d'un environnement numérique de travail (E.N.T.) capable de s'enrichir d'apports successifs dus aux interactions de plus en plus denses et complexes au sein de sphères d'activités devenues numériques. Cela nous conduit à sortir de la problématique du mot-clé, ou du figement lexical (représentation de connaissances), pour celle de la thématique des textes et de l'interprétation située. Celle-ci s'ancrera dans l'alternance d'innovation et de sédimentation (le substrat culturel), laissant libre cours à l'imagination réglée (Ricoeur, 1986) de l'utilisateur. Notre démarche pose donc la question des signes avant-coureurs des connaissances partagées. Elle fait en effet l'hypothèse que la valeur sémantique d'un passage est d'abord le fait de son lecteur (entité pouvant être collective) qui grâce à cette étrange faculté de l'esprit qui est de relier (Vico, 1744) tracera ses thématiques (constituant des molécules sémiques en regroupant plusieurs unités de signification dans une même unité sémantique) en fonction de son environnement[3] en même temps qu'il constitue un corpus de textes par sa navigation intertextuelle.

C'est dans ce contexte de la révolution du passage au numérique qui modifie les usages professionnels et privés et qui interroge le statut même des media que nous abordons cette question Les réflexions pluridisciplinaires menées autour du document numérique (Réseau Thématique Pluridisciplinaire sur le Document[4] du CNRS de 2005 à 2007, Semaine du Document Numérique en 2004[5] et 2006[6], etc.) nous invitent en effet à nous interroger sur les usages du document induits par le passage au numérique, tout particulièrement sur l'interprétation textuelle à l'œuvre en navigation intertextuelle (herméneutique numérique) et à définir de nouvelles approches pour les échanges de contenus ainsi que pour les interfaces cognitives et interactives à mettre en œuvre pour l'accès à ces contenus, surtout lorsque la collection de documents est importante et augmente en masse. Ces nouveaux corpus devenus numériques sont ouverts à une lecture discontinue et à une navigation intertextuelle. Ils nous invitent à ne plus relier le sens aux textes, dépôts de connaissances à partir desquels doivent opérer des outils d'extraction, mais aux situations de production et d'interprétation. C'est dans ce contexte que nous nous inscrivons, car le passage au numérique invite à une lecture extensive[7] à partir de textes fractionnés puis recomposés à dessein. C'est dans cette discontinuité devenant à la fois réticulaire (Adam, 2006) et réflexive (les textes se réfléchissant les uns dans les autres) que nous aborderons la question des documents concernant le droit en transport et logistique pour la gestion des risques.

Pour cela nous nous intéressons plus particulièrement à la conception d'un environnement numérique de travail (E.N.T.), sorte d'extranet dédié aux usages de

[2] A des requêtes en langue naturelle le système répond en terme de requêtes acceptables par le système d'information.

[3] Les sujets se constituent en même temps qu'ils constituent leur environnement (l'Umwelt de Von Uexkhüll 1934)

[4] http://rtp-doc.enssib.fr/

[5] http://www.univ-lr.fr/sdn2004/

[6] https://diuf.unifr.ch/event/sdn06/accueil.html

[7] En opposition à une lecture scolastique dite intensive focalisée sur un corpus limité de textes

la filière transport et logistique. Certes, cet ENT ne recèlera guère de fonctionnalités inédites. En revanche, l'intégration d'un ensemble de ressources et de services interopérables en son tout, dédiés non pas à une collection de cas d'usages particuliers, mais justement à une sphère d'activité (transport et logistique) large et en évolution rapide, constitue une réelle nouveauté, voire une singularité. La mise en œuvre de ce dispositif est susceptible de contribuer à des évolutions notables de l'usage de documents réglementaires et, plus généralement, des activités des acteurs de la filière transport et logistique.

2 Un corpus stratégique mais encore difficile d'accès

Le corpus réglementaire est encore difficile d'accès malgré une forte demande sociale et économique tout particulièrement en transport et logistique. A ce jour, le corpus et la base documentaire de l'Institut du Droit International du Transport (IDIT) sont accessibles en ligne. Cette base s'adresse à des adhérents spécialistes du droit, mais elle est difficilement utilisable par un novice dans le domaine juridique comme un transporteur, qui chercherait des informations pour la mise en place de conditions de transport de marchandises conformes à la législation en vigueur, par exemple. Cette base documentaire est associée à un thésaurus hiérarchisé « maison » pour améliorer son interrogation. Elle est renseignée manuellement à partir de décisions rendues par diverses juridictions françaises et étrangères depuis 1971, de revues papier ou en ligne ou d'après l'interrogation d'autres sources de données en ligne auxquelles l'IDIT a accès. Elle impose aussi la saisie manuelle de comptes-rendus (CR) de jurisprudence et de réglementation sous la forme de fiches. Cette captation de l'information et la veille présentent des difficultés majeures pour renseigner et mettre à jour le système d'information. Nous envisageons, suite à la numérisation (en cours) des collections papiers (des milliers d'articles et de décisions de justice relatifs à des risques et litiges en matière de transports), une aide à l'interprétation de contenus textuels. Pour cela, une analyse préalable du corpus sera nécessaire. Nous détaillons ci-dessous les points de vue théoriques linguistiques que nous comptons éprouver dans cette analyse à commencer par l'étude des spécificités propres aux contraintes normatives qui pèsent sur un genre textuel à forte valeur perlocutoire [8] . C'est donc par une typologie de la structure argumentative en lien avec la présentation matérielle des données et les attentes d'adhérents hétérogènes (juristes, transporteurs, etc.) que nous abordons cette étude de corpus.

Le SI de l'IDIT est conçu pour diffuser des informations aux adhérents (services payants) afin qu'ils puissent gérer dans les meilleures conditions leurs entreprises et sécuriser leurs activités. Ces acteurs du transport et de la logistique se doivent de rechercher et d'analyser des informations de plus en plus nombreuses. Ainsi le suivi et l'anticipation des cadres juridiques communautaires et nationaux sont des éléments de gestion incontournables. Or, la fragmentation de l'information relative au droit des transports et de la logistique qui couvre des domaines aussi variés que le droit commercial, le droit des sociétés, le droit de l'environnement, le droit administratif, le droit pénal, le droit social, rend difficile l'accès à l'information (information éparse, réglementation pléthorique, accès difficile et coûteux, etc.), d'où la nécessité d'une mise en relief (signalement pour interprétation) de celleci.

[8] destiné à faire acte et à normer des activités futures en fonction de précédents (loi et interprétations contextuelles de la loi).

3 Une approche nouvelle en intelligence économique : herméneutique juridique et énaction

Consécutivement à une étude économique des services que peut rendre l'IDIT, portée par le pôle de compétitivité Logistique Seine Normandie (devenu Nov@log), deux types d'outils sont alors apparus comme nécessaires : un outil de veille globale pour renseigner le SI et un outil de diffusion. Le premier doit permettre, par exemple, de faire un état des lieux hebdomadaire selon un domaine précis (avec une requête) par interrogation automatique (à l'aide d'un moteur de recherche). Le second doit permettre une veille personnalisée sur le SI pour un adhérent, spécialiste du droit ou non. Il s'agit de créer des alertes pour les adhérents hétérogènes de la base dans le domaine du droit des transports et de la logistique et plus particulièrement de la sûreté dans le transport.

Il convient de signaler le statut exemplaire de l'herméneutique juridique du point de vue de l'interprétation en contexte. La tâche d'interprétation d'une loi consiste à concrétiser cette loi dans chaque cas particulier. Cette concrétisation (nous dirions aujourd'hui l'interprétation contextualisée ou située) est d'ailleurs le thème central de la jurisprudence. Une loi ne demande pas en effet à être comprise historiquement mais doit se concrétiser dans sa valeur juridique (tant que sa fonction subsiste) à travers ses cas particuliers d'interprétation. Le juriste cherche en effet, comme le rappelle Gadamer (1976), à être fidèle à l'intention juridique de la loi en la mettant en rapport avec le présent. Il s'agit bien ici d'une démarche d'actualisation, de représentation au sens de rendre présent, qui confère à l'herméneutique juridique une valeur alliant philologie et histoire (le sens d'un extrait de texte, d'un terme, relevant de facto de l'histoire de ses interprétations actualisées en contexte). Ce contexte ne saurait être anticipé tant la situation de chaque utilisateur de l'ENT lui est propre. Etant donné l'unicité de chaque cas de litige ou de mise en conformité avec la réglementation, nous ne pouvons disposer de connaissance a priori pour renseigner l'utilisateur de cet ENT. Son besoin d'information dans son intention globale doit être mis en rapport avec sa navigation locale en rapport avec son problème posé. Dans cet esprit, il nous semble intéressant de nous positionner scientifiquement avec une posture issue de la théorie de l'énaction[9] c'est-à-dire une nouvelle méthodologie de conception qui place l'interaction (le couplage) avec l'utilisateur au centre de la démarche et qui est basée sur un environnement numérique permettant de faire émerger des usages non prescrits a priori au cours des expérimentations à partir des différentes fonctionnalités proposées. Ceci est en rupture avec les méthodes de conception pour lesquelles les cas d'usages sont premiers. L'utilisateur disposera d'un ensemble de fonctionnalités ou d'outils qui lui permettra de construire son propre parcours interprétatif à partir d'informations rendues disponibles par cet ENT. Les outils dont nous disposons pour partie et que nous envisageons de développer (cartographie documentaire, traitement automatique de la langue, visualisation et navigation dans un grand ensemble de documents) seront intégrés et mis à disposition dans l'ENT afin que l'utilisateur puisse formuler ses requêtes et être aidé

[9] L'enaction : théorie établie à partir de l'observation du vivant par deux biologistes, Maturana et Varela, repose sur la propriété d'autopoïèse (auto constitution) propre au vivant, et sur l'affirmation que la connaissance est incarnée. donc indissociable du vivant et de l'histoire du sujet pensant. La cognition n'est alors pas affaire de représentations mais d'actions incarnées. La représentation non d'un monde pré-donné mais rendu présent par l'action.

dans l'interprétation des résultats (textes réglementaires) qui lui sont proposés pour affiner et/ou reformuler sa recherche d'information. Finalement, la réalisation de cet ENT repose bien sur une nouvelle approche dans la conception d'interfaces interactives et cognitives où l'engagement cognitif de l'utilisateur sera concentré dans son activité et non plus déplacé dans la machine, par exemple par apprentissage par celleci de cas d'utilisation selon des finalités connues a priori.

3.1 Emergence de nouveaux usages des outils

L'approche herméneutique et énactive dans la conception et l'intégration d'outils de Traitement Automatique des Langues (TAL) marque une différence de point de vue avec les méthodes classiques en favorisant une démarche scientifique expérimentale. Nous mettons en avant l'expérimentation comme une boucle de conception où la modélisation n'est pas une étape initiale, pas plus que les évaluations (non nécessairement comparatives) sont des étapes finales. L'objectif ici n'est pas de chercher à faire mieux certaines tâches déjà réalisées avec des méthodes éprouvées mais plutôt d'inventer de nouveaux usages du TAL ainsi que de nouvelles façons d'utiliser des outils informatisés dans des recherches sur le langage. Il ne s'agit donc pas, comme dans la plupart des systèmes de Traitement Automatique des Langues, de proposer une fonctionnalité complexe (extraction de termes, de relation, classification automatique, annotation automatique) orientant le parcours interprétatif du lecteur sur le texte, mais d'utiliser des fonctionnalités élémentaires qui permettent à l'utilisateur de faire émerger des fonctionnalités de plus haut niveau par combinaison. Des fonctions atomiques proposées émergeront alors de nouvelles fonctionnalités actualisées par l'interaction entre les utilisateurs et le corpus. L'idée est, par exemple, de proposer à l'utilisateur des rapprochements de contextes syntagmatiques (alignement) et de le laisser en inférer des classes sémanticolexicales (apparaissant selon des patrons, des contextes particuliers).

Le but serait ici de combiner plusieurs outils de TAL existants au sein d'un ENT, en particulier :

- des outils centrésutilisateurs (Themeditor, Lucia, Proxidocs développés au laboratoire GREYC à Caen),
- des outils de statistique textuelle tels Lexico 3 (Salem), Hyperbase (Brunet), NooJ (Silberztein), permettant de construire une topographie des segments de textes (Viprey, 2005), (Mayaffre, 2007, Brunet, 2007)[10].

Cette expérimentation permettra ainsi de tester différents modes de présentation des données textuelles qui supportent une navigation interactive dans ces textes.

3.2 L'interprétation comme « énaction de »

D'un point de vue expérimental, il s'agira de savoir comment un environnement numérique de travail, et le couplage qu'il induit, permettent l'émergence par énaction d'une perception sémantique du corpus et ainsi un meilleur accès aux documents juridiques. Si on considère que le sens provient de la démarche outillée de l'interprétant face à un texte et à son intertexte, alors, une expérimentation mettant en oeuvre un environnement numérique de travail qui permet d'effectuer des traitements sur des documents électroniques participe à la coproduction de sens pour l'expérimentateur. Dans le couplage personne-système les interprétations des utilisateurs et les traitements des machines ne sont pas en concurrence. Au contraire, nous les pensons comme complémentaires dans la

[10] voir *Lexicométrica :* Topographie et topologie textuelles, 2007 [consultable en ligne]

mesure où l'activité d'une machine a pour objectif de produire dans l'interaction des traces qui vont participer aux interprétations du, ou des utilisateurs. Nous poursuivons ici l'idée de Dionisi et Labiche (2006) qui consiste à caractériser des processus logiciels impliqués dans des processus expérientiels, eux-mêmes impliquant des processus cognitifs.

Nous nous situons dans l'opérationnalisation de protocoles expérimentaux. Il s'agit de compléter notre état de l'art des outils de navigation textuelle et intertextuelle existants, de les implanter, de permettre leur appropriation (ergonomie linguistique) et de faire une analyse précise et détaillée de leurs apports pour jeter les bases d'un environnement numérique de travail ENT. Cet ENT doit permettre de mettre en interaction des outils linguistiques et des outils de navigation, utilisables selon la démarche adoptée pour la construction dynamique d'une ressource termino-ontologique personnalisée.

Cette recherche s'insère dans le champ très large de la représentation d'information par navigation et du traitement automatique de la langue naturelle pour lequel il existe de multiples outils opérationnels. Notre positionnement, novateur et ambitieux, ne cherche pas développer de nouveaux outils mais bien à combiner, séquencer, relier, rendre plus interactifs ceux qui existent déjà, en renouvelant les hypothèses et voies de recherche grâce aux apports combinés de l'herméneutique matérielle et de l'énaction. La charge cognitive de production de sens résulte alors de l'histoire et du couplage des diverses actions qu'accomplit un être dans le monde (comprendre peut alors s'appréhender comme un agir avec). Il s'agit là, comme le remarque François Rastier (2005) d'un courant de pensée qui, comme l'herméneutique matérielle, ne se présente pas comme une théorie globale mais comme voie de recherche conduisant à un questionnement permanent, d'une part, des textes (domaine de l'interprétation comme énaction de) et, d'autre part, de la place qu'il convient de réserver aux outils informatiques dans le traitement des données.

4 Une approche centrée utilisateur

Notre approche de l'accès aux documents se situe à l'opposé de celles défendues dans le cadre du Web Sémantique. Là où le Web Sémantique cherche à rendre le plus possible partagées de vastes ontologies qui synthétisent une connaissance pensée comme objective et devant convenir à tous les utilisateurs, nous préférons manipuler des ressources termino-ontologiques (bases de données terminologiques, représentations du contenu lexical etc.) propres à un utilisateur ou un petit groupe d'utilisateurs et liées à leur tâche, leurs besoins et de leurs centres d'intérêt. Il en découle une certaine légèreté sémantique de ces ressources, au sens de Perlerin (2004), dans la mesure où elles ne représentent que ce qui est important du point de vue de l'utilisateur et restent ainsi de taille raisonnable (par exemple une centaine de termes) ce qui les rend moins complexes à construire, à maintenir et à enrichir.

Cette approche centrée utilisateur conduit à opérer un certain renversement scientifique relativement aux ressources qu'utilisent les modèles de TAL. Premièrement, d'un point de vue très pratique, force est de constater que des ressources très généralistes, valables pour tout type de traitement envisagé ainsi qu'à destination de tout utilisateur potentiel, ne sont pas facilement disponibles (sous forme électronique pour des traitements automatiques) et encore moins gratuites. Deuxièmement, nous soutenons que l'idée même d'une ressource généraliste est

illusoire car elle dépend inévitablement du contexte qui lui préexiste (le but recherché par le ou les auteurs ainsi que leurs spécificités socioculturelles). Le rapport de l'Action Spécifique 32 du CNRS/STIC en 2003 (Charlet et al., 2003) va également dans ce sens en précisant un obstacle au projet du Web Sémantique : la détermination et l'ajout, même de simples métadonnées, n'est pas une activité naturelle pour la plupart des personnes.

La tradition logico-grammaticale et plus précisément la sémantique formelle et computationnelle cherchent à représenter et à produire, automatiquement ou pas, des formes le plus possible objectivées des significations et du sens. Dans la démarche centrée utilisateur, nous partons d'une position duale où nous considèrerons que les traitements sémantiques appliqués à l'accès aux contenus des documents ont tout à gagner à être le plus possible subjectivés, tant du point de vue des ressources que du point de vue des résultats opératoires. Cette démarche nous paraît être une réponse au constat que dressent Didier Bourigault et Nathalie Aussenac-Gilles à propos de la variabilité des terminologies qu'il y a autant de ressources termino-ontologiques que d'applications dans lesquelles ces ressources sont utilisées (Bourigault et al., 2003).

Les ressources qui sont les plus importantes pour un utilisateur dans une instrumentation informatique pour l'accès aux documents sont celles qui doivent être produites de manière endogène dans une boucle d'interaction entre un outil logiciel, un utilisateur et des corpus. Dans cette boucle, chaque pôle est déterminant. Il en découle une importance significative des corpus utilisés qui du coup ne peuvent plus être considérés uniquement comme un réservoir de formes attestées sur lequel on tenterait de mettre en œuvre un calcul à base de ressources exogènes. Le corpus utilisé est à l'origine des ressources lexicales construites et constitue en même temps le matériau d'expérimentation. L'accès personnalisé au contenu s'inscrit dans un processus interprétatif en allerretour entre des outils (des logiciels d'étude), des corpus (des corpus d'étude) et des ressources personnelles, les uns étant conditionnés par les autres.

Dans notre approche herméneutique et énactive du langage et dans le but d'une intrumentalisation en TAL, nous mettons l'accent sur l'interprétation plus que sur les connaissances. Ainsi la priorité est donnée aux spécificités sociolinguistiques des utilisateurs (par exemple leurs centres d'intérêt, leurs habitudes terminologiques, leurs parcours interprétatifs). Ce qui a du sens pour les utilisateurs ne se réduit pas à une représentation et encore moins à une formalisation. Ce n'est pas le résultat d'un calcul, c'est une activité au centre d'une interaction hommemachine (activité qui de plus n'est pas forcément finalisée dans le temps). Ainsi, on remet en cause l'idée qu'un mot, une phrase, un texte ou un corpus ait du sens (ou non) pour défendre plutôt l'idée qu'ils font sens (ou pas) dans un couplage personnesystème.

4.1 L'exemple de la cartographie documentaire

L'expérience de la cartographie documentaire dans le projet Proxidocs (Roy, 2007) est un exemple de mise en place d'un système interactif centré utilisateur avec ressources endogènes. Cette expérience sera reprise et mise à contribution pour la conception de l'ENT.

Le but du logiciel d'étude ProxiDocs[11] est de plonger son utilisateur (ou un petit groupe d'utilisateurs) dans des interactions qui offrent la possibilité de notamment mieux discerner l'homogénéité thématique d'un corpus, de mettre en évidence sa densité, d'en extraire les principales tendances thématiques de chaque

[11] http://www.info.unicaen.fr/~troy/proxidocs

document et de permettre un accès rapide à tel ou tel document ou passage de document. Au sein de ces interactions l'outil permet de produire et de naviguer dans des représentations graphiques personnalisées que l'on appelle des cartes.

Au préalable, l'utilisateur doit décrire un ensemble de termes qui sont ceux qui l'intéressent et les fournit en entrée au logiciel ainsi que son (ou ses) corpus. Ces termes peuvent être représentés selon deux modèles, soit sous forme d'une liste de graphies, soit sous la forme d'un dispositif de représentation sémique différentielle des significations des termes. A ces deux formes de ressources lexicales correspondent des outils interactifs qui permettent à des utilisateurs de les constituer de manière incrémentale. Avec ces ressources, ProxiDocs construit des cartes dynamiques et interactives (en 2 ou 3 dimensions statiques ou bien animées) ainsi que des visualisations des textes agrémentées d'un coloriage des isotopies.

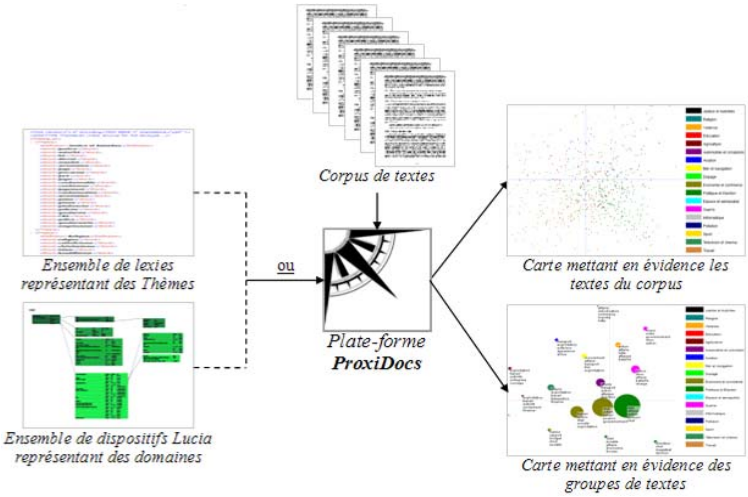


Figure 1. *Utilisation de Proxidocs*

Ces outils logiciels seront redéveloppés pour être intégrés à l'ENT et devront être complétés par de nombreux outils de TAL, de manipulation de corpus et de navigation que nous détaillons par la suite. Dans le couplage personnesystème les interprétations des utilisateurs et les calculs des machines ne sont pas en concurrence car les uns n'ont en aucun cas le but de supplanter les autres. Au contraire, nous les pensons comme complémentaires dans le sens où l'activité d'une machine a pour objectif de produire dans l'interaction des traces qui vont participer aux interprétations du ou des utilisateurs.

5 Conclusion et perspectives : conception de l'ENT

D'un point de vue expérimental, il s'agit de savoir comment un environnement numérique de travail, et le couplage qu'il induit, permettent l'émergence par énaction d'une perception sémantique du corpus et ainsi un meilleur accès aux documents. Si l'on considère que le sens provient de la démarche outillée de l'interprétant face à un texte et à son intertexte, alors, une expérimentation mettant en oeuvre un environnement numérique de travail qui permet d'effectuer des

traitements sur des documents électroniques participe à la coproduction de sens pour l'expérimentateur.

Dans notre stratégie d'amélioration de la navigation intertextuelle, nous proposons à l'utilisateur plusieurs approches pour naviguer dans l'ensemble des documents, visualiser, manipuler et organiser le résultat de ses recherches. Il pourra notamment s'appuyer sur l'historique de sa navigation, ses propres traces, mais aussi celles qui sont liées à sa sphère d'activité (collectif de travail). Il s'agira donc d'observer l'utilisateur dans son activité, et de lui permettre d'exploiter dynamiquement cette observation. Avec ses traces (volontaires ou involontaires), nous ne cherchons pas à modéliser un comportement pour faire de la prédiction, mais à disposer d'outils de description et d'analyse de la navigation intertextuelle en situation réelle.

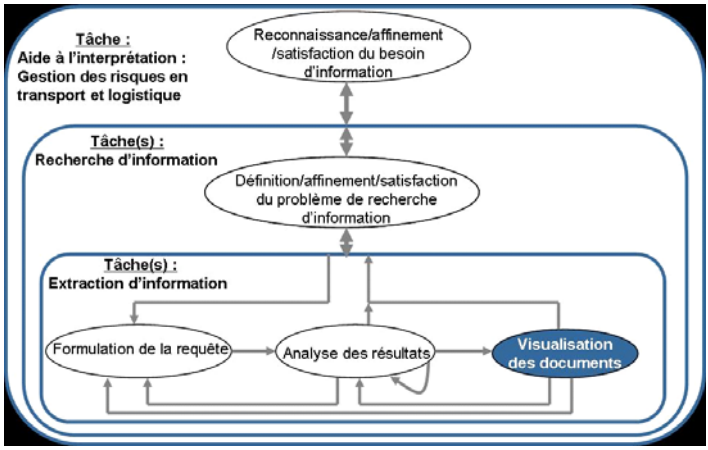


Figure 2. *Visualisation dans un processus d'aide à l'interprétation*

La visualisation et l'analyse des résultats de la recherche sont des étapes nécessaires qui s'inscrivent dans le processus global de recherche d'information. La perception de l'information est liée à la prise de décision dans le contexte d'utilisation de l'ENT proposé. Elle est donc initialisée par une tâche métier ou une motivation personnelle, elle reflète une culture et des contraintes organisationnelles, ou encore des normes sociales. Les tâches d'interprétation évoluent avec l'avancement du processus de reconnaissance/satisfaction en besoin d'information. Dans ce contexte l'utilisateur identifie les sources documentaires, formule des requêtes et examine les résultats ; il se retrouve ainsi au centre d'une boucle itérative « formulation-analyse-visualisation-reformulation » dans une représentation globale du processus comme celle de la figure 2 (inspirée de Kules et al., 2008).

Le processus est donc initialisé lorsqu'un utilisateur identifie un besoin informationnel et tente de le satisfaire en entreprenant une ou plusieurs tâches de recherche. Il prend des décisions sur la ou les stratégies à adopter, les outils à exploiter et le corpus ou partie du corpus à consulter. Chaque unité d'information découverte peut déclencher de nouvelles idées, suggérer de nouvelles directions et changer la nature même du besoin d'information[12]. On émet alors l'hypothèse que,

[12] Bates M J : *The design of browsing and berrypicking techniques for the online search information.* Online review, 13, 407,-431, (1989).

la gestion sous forme d'historiques de traces (incluant point de blocages et retours arrière) laissées par les différents utilisateurs peut aider à la découverte de nouvelles stratégies et de nouvelles informations.

Pour ce qui est de l'extraction d'information, chaque action implique un engagement cognitif et physique, et peut induire une évolution dynamique de l'interface ou des connaissances en émergence. Nous cherchons à faciliter le couplage et l'engagement de l'utilisateur en lui proposant des outils simples pour la manipulation/sélection/déplacement des documents résultats, ainsi que pour l'expression dynamique des requêtes. L'environnement numérique de recherche et de visualisation d'information que nous proposons de réaliser présente alors les modalités suivantes :

- Définition interactive (graphique ou non) des requêtes
- Classification suivant des attributs et items définis dynamiquement par l'utilisateur ou le groupe d'utilisateurs de la sphère d'activité :
    - A partir d'une liste hiérarchique que l'utilisateur a la possibilité de reconfigurer (avec l'ajout ou la suppression d'éléments).
    - A partir de différents points de vue (Thèmes, catégories, dates d'édition, taille, numéro de fiche, etc.) exprimés et réorganisables par l'utilisateur. Ces points de vue permettent à l'utilisateur d'ajouter des contraintes appropriées à son contexte de recherche (problèmes à résoudre et compétences). Pour simplifier la recherche, un passage rapide de la souris sur un point de vue peut donner le nombre de documents pouvant être sélectionnés à partir de ce point de vue.
    - A partir de tout ou partie de documents visualisés, l'utilisateur demande des documents similaires.
- Visualisation de l'ensemble des résultats de la recherche
    - Liste : Il s'agit d'une approche classique pour visualiser un ensemble de documents. L'ordre dans lequel les résultats sont présentés peut être déterminé ou non par des mesures liées à la pertinence du document au vu du contexte utilisateur.
    - 2D : On applique ici les techniques de visualisation de grands ensembles documentaires (Jacko et Sears, 2006). La classification proposée par l'utilisateur sert de base une présentation animée sous forme de clusters, treemap, arbres hyperboliques, Zoom 2D, fisheye, matrice, etc. L'utilisateur peut naviguer dans chaque catégorie, à chaque niveau de la hiérarchie et visualiser les résultats associés à la classe, en les organisant suivant ses propres paramètres.
    - 3D. Il y a ici un risque de confusion par rapport à ce que peut apporter une visualisation 2D. Toutefois on espère l'émergence de connaissances en proposant de positionner les documents dans un espace 3D comme les arbres coniques ou les Zoom Sémantiques
- Outils d'extraction et analyse de tout ou partie d'un document du corpus
- Affinement dynamique du contexte de la recherche dans une zone de filtrage

Notre démarche consiste donc à utiliser des représentations spatiales dynamiques pour concevoir et mettre en œuvre une plateforme (figure 3) générique en personnalisation de la visualisation et en intégration de modalités variées et hétérogènes.

A travers cet article et notre projet de mise au point d'un ENT dans le domaine juridique, nous avons voulu ici poser d'une manière particulière la problématique du document électronique. Les sciences humaines et les sciences cognitives nous incitent à envisager le document électronique de manière indissociable à l'activité du ou des humains qui les produisent, recueillent, indexent et recherchent. Ainsi le champ de l'herméneutique nous montre à quel point les rapports intertextuels sont déterminants dans les interprétations des documents, et le champ de l'énaction nous montre en quoi ces interprétations sont situées et subjectives. Il en découle à notre avis que les instrumentations nouvelles autour du document électronique ne pourront rester indépendantes des utilisateurs et sans prise en compte des paliers d'intertextualité comme le sont actuellement par exemple les moteurs de recherche ou encore certains projets dans le contexte du web sémantique. Il en va de même dans le cadre d'autres situations de couplage pour lesquelles l'expérimentation est également une piste intéressante de recherche pour éprouver nos hypothèses ; en particulier les situations sous contraintes (handicaps sensoriels ou situationnels) : précisément, ces discussions alimentent de nouvelles perspectives pour l'amélioration de l'accès tactilooral des nonvoyants à l'information textuelle, en permettant à cette population d'usagers une réappropriation dans de nouvelles modalités (et l'émergence de nouveaux usages ?), de « traces » spécifiquement visuelles telles que la mise en forme des textes (Maurel, 2004).

Plus que jamais la problématique du document électronique requiert des collaborations pluridisciplinaires pour mettre au point, expérimenter et évaluer les conditions d'une relation entre documents et interprétants d'où puisse émerger du sens. Notre groupe ν se donne comme objectif d'y apporter sa contribution.

Références :

Adam, J.M. (2006). Autour du concept de texte. Pour un dialogue des disciplines de l'analyse de données textuelles. In *Actes du colloque JADT 2006*, Besançon, Disponible à : www.cavi.univparis3.fr/lexicometrica/jadt/JADT2006PLENIERE/JADT2006_JMA.pdf.

Bates, M. J. (1989). *The design of browsing and berrypicking techniques for the online search information.* Online review, 13, 407431. Disponible à : www.si.umich.edu/~rfrost/courses/SI110/readings/InfoFinding/Bates_on_Berrypicking.pdf.

BernersLee, T. (1998). *What the semantic web can represent ?* W3C, Disponible à : www.w3.org/designissues/rdfnot.html.

Charlet, J., Laublet, P., Reynaud, G. (2003). *Web sémantique*. Rapport de l'Action Spécifique 32 CNRS/STIC.

Dionisi, D., Labiche, J. (2006). Enaction et informatique : les enjeux de l'opérationnalisation technologique d'une théorie de la cognition. In *Actes du colloque ARCo 2006*, Bordeaux, 68 Décembre 2006.

Engisch, K. (1953) *Die Idee der konkretiesierung.* Abh der Heidelberger Akademie.

Gadamer, H.G. (1976) *Vérité et méthode : les grandes lignes d'une herméneutique philosophiques.* Seuil, Paris.

Jacko, J. A., Sears, A (2006). *The HumanComputer Interaction Handbook : Fundamentals, Evolving Technologies and Emerging Applications*. 2nd Edition, Lawrence Erlbaum Associates.

Kules, B., Shneiderman, B. (2008). *Users can change their web search tactics : Designs guidelines for categorized overviews.* Information Processing and Management.

Maurel F, (2004). *Transmodalité et multimodalité écrit/oral : modélisation, traitement automatique et évaluation de stratégies de présentation des structures « visuoarchitecturales » des textes.* Thèse de doctorat, Université Paul Sabatier.

Perlerin, V. (2004). *Sémantique légère pour le document*. Thèse d'informatique. Université de Caen.

Peschard, I. (2004). *La réalité sans représentation, la théorie de l'enaction et sa légitimité épistémologique.* Thèse de Philosophie. Ecole Polytechnique.

Rastier, F. (2005). Sémiotique du cognitivisme et sémantique cognitive : questions d'histoire et d'épistémologie. *Texto,* mars 2005. Disponible sur : www.revuetexto.net/Inedits/Rastier/Rastier_Semantiquecognitive.html.

Ricoeur, P. (1986). *Du texte à l'action : essais d'herméneutique.* Point Seuil, Paris.

Roy, T. (2007). *Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires*. Thèse d'informatique. Université de Caen.

Varela, F. (1989). *Invitation aux sciences cognitives*. Seuil, Paris.

Vico, G. (1744). *Principes d'une science nouvelle.* (trad JL. Lemoigne) Nagel 1986.

Viprey, J.M. (2005). Philologie numérique et herméneutique intégrative. In *Sciences du texte et analyse de discours*, Adam, J.M., Heidmann, U. (eds), Slatkine, Genève, 5168.

Architecture de documents et économie des vecteurs

Document architecture and vectorialism

Hervé Le Crosnier(1)

(1)Laboratoire GREYC, UMR 6072, Université de Caen
herve@info.unicaen.fr

Résumé. Nous assistons à une nouvelle mutation du web, le passage d'une métaphore liée au « document » vers une architecture centrée sur les services et les usagers. On retrouve cette question dans la normalisation du web, avec le débat au sein du W3C entre XTML 2.0 et HTML 5, qui recoupe celui entre les normalisateurs et les web designers. On retrouve aussi cette question dans le domaine juridique, avec le nouveau statut des sites de diffusion d'information et la question de l'archivage. On retrouve enfin cette révolution copernicienne dans les conséquences économiques, avec l'émergence de vecteurs qui draînent toutes les productions d'information vers leurs univers de services. En ce domaine, l'analyse sur les multiples facettes des sciences humaines et des sciences de l'informatique peut éclairer tant le citoyen que le programmeur du web.
Mots-clés. normalisation, métaphore du document, services sur le web, vectorialisme, architecture du web.

Abstract. A new mutation of the web is at hand, with changes from a web based on the document metaphor to a web of services. This question is active in the standardisation process in W3C, with two competing standards of XHTML 2.0 and HTML 5, representing the two sides of the web, standardization and web design. Legal conceptions of editorship, and of archival of the web are also concerned by this new approach. This copernician revolution also impact the economy, with the emergeance of vectors, who attracts all document production into their universe of services. To understand this phenomenom, we must use knowledge from humanities, social and économical sciences as well as computer science, to share those concepts with the citizen and the webmaster.
Keywords. standardisation, document metaphor, web services, vectorialism, web architecture.

1 Introduction

On a souvent décrit le passage au « web 2.0 » comme la transition entre la publication et la conversation. Dans cette nouvelle architecture des relations sociales médiées par les documents, l'immédiateté, la capacité à publier sans entraves, la

forme des rebonds (trackback, liens, commentaires,...) ont induit une autre approche de l'écriture et de la réception. Un document n'est plus une « somme » qui engage l'auteur, mais une « proposition » de débat ou une « réaction ». Le « tribunal de la Raison », cher aux philosophes des Lumières et qui sert de référence dans l'organisation des publications scientifiques, laisse ainsi la place à la « sagesse des foules ». Le calculable (nombre de liens, nombre d'amis, nombre de commentaires,...) a pris le dessus sur l'argumentation et la démonstration (Ayres, 2007). Parce que ce « calculable » permet des architectures de services au dessus des documents publiés, il est devenu la norme de référence. Notamment parce qu'il est la base d'une nouvelle « économie de la recherche », qui associe publicité, lecteurs et documents (Salaün, 2008).

Mais cette évolution globale des pratiques sociales du web (mode de production et de lecture) et des pratiques économiques des web-média (Salaün, 2006), notamment la place centrale de la recherche documentaire et de la gestion de la « base de données des intentions » (Battelle, 2006), masque une autre évolution : la notion de « document » n'est plus homomorphe à la notion de page web. L'architecture ajax + web services incite les développeurs, et ce faisant les utilisateurs, à concevoir les pages web comme des regroupements d'applications. Une page est un grid sur lequel sont posés des widgets, ou des documents distants incorporés (vidéos provenant de sites de dépôt tels *YouTube* ou *DailyMotion*). La notion de « média », au sens social d'une structure diffusant de l'information auprès de lecteurs, en organisant un financement par un tiers-acteur, en général la publicité, est elle-même mise à mal. Le « diffuseur » (techniquement l'organisme qui gère un serveur de pages web sous une marque commune) est de plus en plus souvent un opérateur de ré-organisation d'informations et de services disponibles ailleurs sur le web. Le développement de la syndication, ou du phénomène des « marques blanches », ou encore la façon dont sont intégrées des vidéos provenant de ressources tierces accentue ce phénomène d'industrialisation. Avec la confusion entre le document et l'application vient le délitement des frontières entre l'éditeur et l'hébergeur. Ce qui provoque des débats juridiques et économiques importants.

Le numérique tend à transformer tout produit en service, et la vente de produits (logiciels, documents, artefacts support comme les CD ou les DVD,...) en une économie de l'accès. Qu'en est-il du document quand il devient « application » ? Quelles en sont les conséquences sur l'archivage ou sur la rémunération des créateurs ? Comment l'informatique de l'instant, caractérisée par les widgets et les « flux » (RSS, vidéo, radios web,...) s'accorde-t-elle avec la construction collective du savoir et de la culture et leur mise à disposition de tous ?

2 Documents et applications

Notre notion du document découle de l'histoire du livre, un objet qui enserre le texte entre ses deux couvertures (Melot, Taffin, 2006). Le document est une entité auto-complète, que l'on peut échanger aisément, pour en faire partager le contenu, que l'on peut lire et relire sans crainte qu'il ait été modifié, que l'on peut archiver et retrouver, et qui traite d'un sujet défini (ou qui parle d'une seule voix, comme le roman).

Avec le numérique, cette notion du « document » porteur en lui-même de sa preuve s'évanouit. Il n'est que de constater la transformation des « billets » (d'avion, de spectacle,...) qui de preuve disponible dans la main de leur possesseur, deviennent simplement la marque de l'enregistrement dans une base de données.

Ceci a des effets sur l'objectif des normes d'interopérabilité, permettant de retrouver et d'échanger des documents « *standalone* » au travers du réseau informatique. Le premier web a considéré ce modèle historique du « document » comme la source de sa normalisation, notamment pour les trois premiers éléments de l'architecture du web :

- HTPP, un protocole pour échanger des « pages web » (i.e. des « documents »). Rappelons que le mode non-connecté de HTTP répond justement à cette notion d'un « document » que l'on va chercher sur un serveur, à la différence de la notion de flux temporel des médias diffusés. Aujourd'hui, http est devenu un support/canal pour de multiples applications, en mode connecté (streaming), comme asynchrone (ajax) ;
- URI, une norme pour désigner une instance unique d'un document. La pérennité des URI a toujours été une question posée par l'interprétation libre et laxiste de la norme : dans la littérature du premier web, l'URI désignait le chemin à parcourir pour retrouver un document sur le web. Dans la norme elle-même (Berners-Lee, 1998), un URI est avant tout une stratégie de nommage d'un document, indépendamment de sa place sur le web (dans la hiérarchie du système de fichiers) ou de la méthode d'accès (extension liée au langage de programmation utilisé ou utilisation de la query string pour contrôler des programmes). La clarification est maintenant bien connue, et les stratégie de nommage l'emportent dans l'attribution des URI (CoolUri, 2008). Avec toutefois des excès : l'URI est composé pour satisfaire l'appétit des moteurs de recherche plutôt que de respecter la règle de base : le support principal d'un URI est la nappe d'une table de restaurant, on doit pouvoir désigner simplement un document pour le transmettre ;
- HTML, un langage de balisage qui procède de la forme d'organisation des articles scientifiques. Dans ses premières versions, HTML reprenait les éléments structurels de ces articles, depuis les niveaux de titre jusqu'aux énumérations. Le livre est venu ensuite, avec les éléments logiques de liens entre documents (élément link et attribut rel) . La bibliothèque a suivi avec l'intégration de métadonnées dans les en-têtes des « pages HTML ». Aujourd'hui, HTML est devenu un langage d'intégration de parties dans une « page web », et l'objet de deux conceptions alternatives, l'une portées par les tenants de la normalisation au travers de XML (évolutions vers XHTML 2.0) et l'autre par les designers pragmatiques qui veulent un langage de présentation d'applications (HTML5).

Dans sa première architecture, le web était à l'opposé des médias de flux (radio, télévision, vidéoconférence,...) : le lecteur devait aller chercher un document repéré par un URI. L'image de l'entrepôt mondial, ou de la bibliothèque universelle s'appliquait beaucoup mieux. C'est d'ailleurs cette dernière qui a été choisie par le CERN pour la première application massive, l'ouverture d'une « web library ».

Mais très vite la notion de « site » l'a emporté sur celle de document. Les bibliothèques elles-mêmes ont constitué des répertoires de « signets » ciblant des « entités éditoriales » repérées et ayant une compétence sur les sujets spécifiques de

la classification. Le modèle des « annuaires » mettait en avant la « marque » médiatique des sites, en non chaque contenu (document) particulier. Un phénomène accentué par l'usage des « frames » qui réduisent les « documents » à des adjuvants dans la promotion de la vitrine, et brisant au passage le modèle de nommage des URI. Cette évolution vers les sites a attiré les médias diffusés sur le web.

Avec les médias, nous sommes passé d'une logique d'entrepôt de documents, de « bibliothèque virtuelle » à celle de source d'information en continu. Une source d'autant plus prisée qu'elle laisse l'utilisateur décider du moment de sa lecture, et qu'elle permet de cibler les informations recherchées. Le 11 septembre 2001 a marqué de ce point de vue l'adoubement de l'internet comme média à part entière : devant la saturation des réseaux de téléphone et le désir des lecteurs d'aller plus profondément dans l'information (soit auprès de ses proches pour avoir des témoignages – communication -, soit en cherchant des informations sur les répercussions mondiales, ce qui en général n'était guère diffusé par les médias mainstream, reposant sur l'émotion et la proximité). L'internet est devenu une nouvelle source de « news ». On a constaté depuis un phénomène d'attirance aux nouvelles et à l'actualité, qui a eu comme support le développement des flux RSS et l'apparition de Google News... deux applications significatives du web dit 2.0.

La syndication des informations par les flux RSS a eu deux conséquences :

- l'apparition des « pages personnelles » regroupant les flux à usage privé (le modèle étant *Netvibes*, aujourd'hui imité par *iGoogle* ou *Bubbletop*). L'agrégateur de flux permet de suivre l'actualité et devient le « média » personnalisé annoncé durant les années 90 (le DailyMe de Nicholas Negroponte (1995))
- l'insertion dans les pages des sites web de zones dédiées à des nouvelles provenant d'un autre éditeur, ou prestataire d'information. Depuis la météo jusqu'aux nouvelles des agences de presse, un éditeur de site délègue à une source tierce une partie de ce qu'il donne à voir à son lecteur.

L'effet général est à double détente :

- l'information circule plus vite, mais devient aussi une « information circulante », diminuant la part de l'analyse et de l'interprétation, qui est le propre du « document ». On assiste à une « industrialisation » de l'information au sens où la société industrielle est celle de la ré-utilisabilité des parties (importance de la normalisation industrielle, et des contrats de sous-traitance pour des composants partagée par plusieurs marques, comme en automobile) ;
- l'éditeur d'un site maîtrise de moins en moins ce qui est transmis comme information à ses lecteurs, y compris dans le domaine de la publicité. Un flux RSS, un widget, une carte peuvent contenir des publicités émanant d'une source différente et présentée au sein de la page.

3 De XHTML 2 à HTML 5

XHTML a été une ré-écriture de HTML 4 pour respecter le formalisme de XML et séparer encore plus nettement l'architecture de document et la présentation

sur un terminal (écran, papier, téléphone,...), XHTML 2 s'affronte à la modularisation du document, l'intégration de jeu de balises provenant d'autres espaces de noms, permettant l'intégration d'applications XML (SVG, RSS, RDF,...) et de plugins.

Pour sa part, HTML 5 se situe d'emblée dans un autre cadre : la conception d'un web-média, ou de sites de service (commerce électronique, forums, grandes bases de données). Dans cette optique, les éléments de base sont des applications qu'HTML 5 permet de ré-organiser dans une « page web ». Le document de travail au sein du W3C précise ainsi : « *XHTML2 définit un nouveau vocabulaire de type HTML qui améliore les fonctionnalités des liens hypertextes, l'insertion des contenus multimédia, l'annotation et l'édition de documents, la richesse des métadonnées, la construction déclarative des formulaires interactifs, et permet de décrire la sémantique des œuvres littéraires tels que des poèmes et des articles scientifiques.*

Cependant, il lui manque des éléments pour exprimer la sémantique d'un grand nombre de contenus non-document, qui sont pourtant fréquents sur le Web. Par exemple, les forums, les sites d'enchères, les moteurs de recherche, les boutiques en ligne, et autres, ne trouvent pas leur place dans la métaphore du "document" et ce faisant ne sont pas couverts par XHTML 2. »

On peut aussi regarder ces deux visions divergentes de la normalisation des documents/pages web en partant des groupes ayant poussé ces deux logiques (Lecarpentier, Le Crosnier, Madelaine, 2008) :

- pour la normalisation suivant la métaphore du document, on trouve les informaticiens, les documentalistes et archivistes, qui définissent le « cycle de vie » du document, de sa création à son archivage. Pour ces catégories, il est clair que les difficultés propres à la normalisation d'un document (métadonnées, insertions dans le web sémantique et usage des classifications et ontologies) sont inhérentes à la complexité du mode documentaire et surtout de sa pérennité. Publier (rendre public) un document n'est qu'un moment dans son cycle de vie. Certes essentiel, parce qu'il est le moment de la réalisation économique. Mais stocker, rendre disponible à tout moment (indexation, recherche documentaire) et finalement archiver sont des activités sociales liées au document qui sont de tout temps abstraites et échappent au commun des lecteurs. Au point que se sont développés des secteurs de spécialités et des emplois spécifiques pour assurer ce service de la lecture (archives, bibliothèques, documentalistes,...)
- pour la conception d'un web « de service », agrégeant des informations, des flux ou des outils (commerce électronique, forums,...), on trouve derrière HTML 5 les « concepteurs de sites web » et autres « web designers ». Leur problématique n'est pas de produire de nouveaux documents, mais de permettre à leur client de mieux utiliser le potentiel du web et la généralisation des terminaux web (sur tous les types d'outils, du navigateur visuel des ordinateurs aux téléphones mobiles) pour fidéliser et influencer les utilisateurs.

Ces tendances ont aussi des conséquences qui méritent d'être évaluées en regard de l'archivage des sites web. Autant « archiver un document » est une continuité avec les pratiques antérieures, autant archiver des visions personnalisées d'un ensemble d'informations nécessite une nouvelle réflexion sur le type d'échantillonnage et la forme de l'observable (i.e. ce que le lecteur aura vu) que la société doit conserver et indexer pour une consultation ultérieure.

4 Responsabilité juridique de l'éditeur

Le basculement du document à l'application dans le cadre du web est bien évidemment un processus complexe, ayant des ramifications dès les premiers sites web. L'archivage des premiers sites « dans leur état initial » tel que peut le réaliser « Internet Archive » restait possible au travers de robots transformant les URL et stockant les images et autres éléments associés. Ceci devient plus complexe avec des applications construites autour de l'interaction (communication) et non du contenu (information). Les techniques d'archivage doivent changer.

Les bibliothèques sont-elles habilitées à archiver le web ? Même si la Loi du 2 août 2006 (dite Loi DADVSI) organise le cadre juridique d'un tel archivage, il reste à définir les formes et les négociations juridiques. Les approches changent selon les pays. La Bibliothèque nationale de France par exemple, veut négocier avec les éditeurs cet archivage, et en faire une excroissance du dépôt légal. *A contrario*, aux États-Unis, les bibliothèques ou les organismes privés comme Internet Archive s'appuient sur le *fair use* pour engager des travaux d'archivage à partir de ce qui est actuellement disponible à tout lecteur du web. Négocier avec l'éditeur implique de choisir la partie de contenu que l'éditeur propose au travers de sa base de données et non l'image de la page telle qu'elle est lue par un lecteur, avec les informations produites par un tiers juxtaposées... par exemple des cartes avec des marques et des traces issues de Google Map. Pour les documents d'entreprise, le format pdf a su s'imposer parce qu'il représentait au mieux la volonté typographique de l'éditeur et n'était pas sensible au terminal de lecture. Comment inventer l'équivalent pour les sites web... tout en gardant l'interactivité permise par les liens ? C'est là un nouveau chantier absolument nécessaire, qui échappe aux pratiques des web designers, plus axées sur l'immédiateté des échanges que la pérennité de « documents ».

Mais l'inscription dans les normes du basculement de la métaphore du document à celle du service, tout comme les nouvelles pratiques sociales de la lecture ont aussi des impacts sur le statut juridique du document numérique. Des exemples récents permettent de mieux préciser les conséquences d'un web inscriptible dans lequel le « document » et la relation auctoriale dont il garde la trace historique sont en train d'exploser sous nos yeux.

- les procés fait à des blogueurs concernant les « commentaires » publiés en marge de leurs posts. Le « commentaire » est-il de la responsabilité de l'éditeur/auteur du blog ? La logique de la Loi française dite LCEN (Loi sur la Confiance dans l'Économie Numérique) serait plutôt tolérante, considérant le statut « d'hébergeur » de forum. Il importerait alors seulement de savoir si tout a pu être mis en œuvre pour retirer les commentaires et inscriptions diverses quand une demande légitime a pu être émise. Mais on peut aussi arguer de la différence de « responsabilité éditoriale » entre l'hébergeur d'un forum et l'éditeur d'un blog, qui au titre de directeur d'une publication se doit de respecter les termes de la Loi sur la liberté de la presse de 1884. En particulier en ce qui concerne la diffamation. Être

diffamé dans le corps d'un post ou dans un commentaire ne fait pas de différence pour la victime.

- le statut des agrégateurs d'information, comme Vikio ou Fuzz en France ou Digg aux États-Unis. Ceux-ci permettent aux lecteurs de classer les documents et de valoriser leurs lectures dans un système de promotion social (par la force du nombre et du calcul). Cette opportunité fait-elle pour autant changer le statut de la personne qui a organisé ce service d'un « éditeur » de site à un « hébergeur » des informations et commentaires de ses lecteurs, ou de diffusion des flux RSS de tiers-acteurs ? Dans le cas de Fuzz (Parody, 2008), le juge estime que le site est un travail éditorial, notamment parce que les flux RSS y sont classés, et qu'à ce titre présenter des liens portant atteinte à la vie privée d'une vedette est une faute professionnelle... même si le lien est simplement celui d'un fil RSS vers un autre site, de surcroît un lien placé sur Fuzz par un lecteur. Cette situation ou le modèle « application » l'emporte largement sur celui du document, a incité le GESTE, Groupement des éditeurs en ligne en France, à proposer une clarification entre le statut d'éditeur de contenu et celui d'éditeur de service (GESTE 2008)... sachant que le numérique, en intégrant toutes les activités conceptuelles, ne favorise pas des distinctions si tranchées.

Tant pour l'archivage, activité hautement socialisée, que pour la mise en flux permanent de l'information, on voit qu'une interférence existe avec les logiques éditoriales (ou du moins de prestataires de services) des média-web et leur attachement à la protection des sources et des personnes et la logique orientée vers le lecteur qui est le propre d'une « application ». Les juristes, considérant l'équilibre économique des deux types d'activités (la gestion/édition de contenu et l'insertion dans la « conversation mondiale » du web participatif), n'ont aujourd'hui pas tranché.

5 Édition et production de contenu

Dans l'édition traditionnelle, comme dans le monde des médias, l'éditeur (ou diffuseur) acquiert du contenu (par achat ou en signant un « contrat d'édition » avec un auteur), puis se charge de le rentabiliser économiquement au travers de ses artefacts, ou de la marque de son flux médiatique. Le diffuseur établit un accord moral et financier avec le producteur.

Or ceci est en train de changer à grande vitesse. Le diffuseur-éditeur agrège des sources, des flux, des documents, des commentaires, des reviews,... et n'a plus lui-même une vision organisée de ce qui circule sur son site. La conception des pages web comme un ensemble de services renforce le caractère industriel du web : non seulement les informations sont réparties, mais les outils de lecture ou de suivi des usagers proviennent de sources tierces :

- bibliothèques de widgets
- analyse de fréquentation grâce à des sites centraux, tels *Google Analytics*
- apports cartographiques (*Yahoo! maps, Google maps*,... et la possiblité de cibler des lieux ou d'inscrire des parcours sur ces cartes fournies par des tiers)
- liens vers des boutiques en ligne, comme *Amazon* et son programme d'association

- intégration de vidéos hébergées sur des nuages de serveurs externes tels *YouTube* ou *DailyMotion*

L'authentification des usagers auprès des sites devient elle-même sous-traitée aux fournisseurs de certificats de type *OpenId* (Openid, 2008), qui sont aussi souvent les grands opérateurs du web (*Orange, Google, Yahoo!,...*).

L'éditeur de sites web s'éloigne chaque jour de la conception de l'édition issue de l'histoire du livre et des revues ou journaux, mais se rapproche du modèle des médias : il s'agit d'organiser et de promouvoir des contenus produits en dehors de l'organisme de diffusion, souvent exploités par plusieurs sites simultanément. Un site web devient un outil pour créer de l'audience à des productions de contenu, et en sens inverse, d'espérer que l'audience générée par les contenus va permettre de déclencher un modèle économique de rentabilisation (en général la publicité, ou plus largement l'industrie de l'influence).

Ce qui distingue les web-médias des médias en général est le mode non-linéaire de la lecture. Non seulement par l'usage des liens hypertexte, mais aussi par le circuit qui conduit un lecteur devant le document présenté par un site :

- usage de connexions « sociales », des recommandations de groupe ou du « vote » de promotion comme sur les sites Digg-like ;
- recherche documentaire sur les moteurs de recherche
- veille informationnelle et usage des flux RSS
- liens conseillés par un ami ou une lecture

Le lecteur arrive alors directement sur des éléments de contenu, sans même consulter l'organisation globale du média (i.e. la « page d'accueil »). C'est le phénomène de délinéarisation, qui touche l'écrit comme l'audiovisuel. Ce phénomène est accentué par la disponibilité sur un temps plus long du contenu sur les serveurs de stockage du web. Les écrits restent... mais les programmes télévisés aussi, par exemple pour la « catch-up TV ».

Une nouvelle contradiction risque d'émerger entre les diffuseurs de contenus et les auteurs-producteurs de ces contenus. L'éditeur-diffuseur n'est comptable que de son public, et le producteur-auteur peut de moins en moins compter sur l'éditeur (ou le média de diffusion) pour pré-financer son travail. L'exemple des documentaristes ou des photographes de presse et des nouvelles relations qui s'établissent pour eux avec les médias qui utilisent leurs travaux est significatif (McDonald, 2008). Pour l'éditeur, il s'agit maintenant de « faire son marché » de contenus qui vont lui permettre de générer une audience. Et pour le producteur, il s'agit de trouver des formes de financement nouvelles, indépendamment des ventes unitaires réalisées *a posteriori*. Les modèles vont du pré-financement (bourses, pré-achats,...) au support publicitaire, en passant par les abonnements et les forfaits (telle l'idée d'une « licence légale » en musique), ou l'insertion dans des nouveaux produits (par exemple la musique utilisée en sonnerie de téléphone ou dans les jeux vidéo).

Les web-médias ne sont plus comptables devant leurs propres producteurs, par les pré-achats, ou par les multi-diffusions, mais doivent agréger des applications (interaction contenu-publicité, réseaux sociaux autour du contenu, voire gains

secondaires dans la commission en cas de vente de produits – modèle de *Amazon associate*). La fluidification de ce marché des œuvres est un des objectifs recherché par les grands éditeurs. Et contradiction ou ironie, les licences d'usage telles Creative commons servent aujourd'hui de tremplin pour ces nouvelles formes de marché mondial des documents pour nourrir les sites de service. Ces licences vise à l'origine à élargir la diffusion d'œuvres que les auteurs veulent voir circuler en dehors de la sphère commerciale. Loin de construire des biens communs, les producteur de contenus dit « User generated content » servent en réalité à alimenter la machine des web-médias.

La généralisation de ce mode de diffusion met en péril l'équilibre économique général de la création de documents et d'information. Les documents déjà rentabilisés vont trouver une plus grande diffusion. Ce qui devient rare, c'est le lecteur et non le document (Piotrr, 2008). Ceci donne naissance à une économie de l'attention, qui cherche à valoriser la capacité des acteurs industriels à mettre des lecteurs en face des documents. D'évidence, la gratuité est un élément central dans ce dispositif. La multimodalité aussi : il faut que le lecteur puisse rencontrer le document en tout lieu, n'importe quand et sur tous les outils électroniques possibles (ATAWAD : *AnyTime, AnyWhere, AnyDevice*).

Cela va-t-il favoriser la création autonome, inventive, parfois ardue, ou bien va-t-on simplement voir se généraliser la circulation des produits d'audience ? (Le Crosnier, 2008) La capacité du web des documents à rendre disponible une masse de plus en plus grande d'informations va-t-elle se briser sur les écueils du web-média ?

6 Concentration et vectorialisme

La concentration des acteurs proposant les « services » (depuis les widgets jusqu'aux contenus en marque blanche) aux éditeurs-diffuseurs de sites web est à la fois le produit du modèle économique en œuvre, et de l'industrialisation des processus de production de l'information. Cette situation fait peser une menace sur la diversité culturelle, sur la véritable ouverture de l'offre. Il devient de plus en plus difficile de mettre en avant des stratégies éditoriales, telles la sélection des contenus, la construction de la notoriété des auteurs -et la marque de l'éditeur – et la captation d'un lectorat stable,... Il existe d'ores et déjà des sites qui concentrent les attentes des lecteurs. Par exemple, un musicien se doit aujourd'hui d'avoir sa page sur *MySpace*, quel qu'en soit pour lui le bénéficie économique ou même de notoriété.

Cette concentration est un effet secondaire de l'économie de la « longue traîne » (Anderson, 2006) : les travaux peu lus (ou vus, ou entendus) sont devenus disponibles sur les étagères du web, mais seuls de gros opérateurs peuvent financer un tel service car les revenus unitaires sont trop faibles pour des indépendants, et d'autre part, c'est la proximité de tous ces documents peu lus au sein d'un même site majeur qui favorise leur découverte par sérendipité. C'est ainsi que seul *Amazon* peut se permettre d'ouvrir une boutique pour diffuser des disques « introuvables » ou « non-réédités »

Or ces gros opérateurs vivent de leur capacité à tracer leurs usagers/lecteurs, à leur modèles de calcul de profils et à leur capacité à revendre cet ensemble de profils auprès d'acteurs intéressés par l'influence que cela leur ouvre (publicité, mais aussi États ou militaires). Ces opérateurs que je propose d'appeler « vecteurs » peuvent alors proposer d'organiser les page web selon les « goûts » du lecteur, ciblant les publicités ou plus largement les zones d'influence.

C'est cette jonction entre le web-média et l'économie de compteur qui constitue le cœur du vectorialisme, nouveau moteur économique du web. Par compteur, on entend généralement un système physique qui calcule les consommations de chaque usager. La possession du « compteur » est le sésame qui permet d'utiliser les flux (eau, électricité, téléphone, carte SIM,...). Or avec le numérique, le compteur n'est plus un cadre contraignant (comme un abonnement, un forfait,...), mais représente la nécessité pour un lecteur de revenir régulièrement rafraîchir ses données personnelles. On trouve ainsi parmi les nouveaux dépositaires de « compteurs » les réseaux sociaux, les prestataires d'outils personnels sur le web (mail, albums photos,...), et même les moteurs de recherche, avec l'usage des cookies.

Hal Varian, économiste en chef de *Google* parle ainsi « d'effet d'expérience » (Lohr, 2008) pour souligner les évolutions du modèle de compteurs et le distinguer des « effets réseaux ». Les services aux personnes au travers du web deviennent des formes de compteurs « virtuels ». Les entreprises qui peuvent faire cohabiter la sélection/présentation des documents et les informations obtenues par les passages réguliers des usagers devant des compteurs virtuels, peuvent seules prétendre à tirer bénéfice de la création et diffusion de documents sur le web. Documents que l'on dit « gratuits », mais qui sont en réalité des moteurs économiques pour les autres formes de valorisation de l'audience des vecteurs.

La construction de biens communs, et l'invention d'une nouvelle forme de rémunération de la création devient d'autant plus urgente que le laminage par l'audience pourrait redevenir une conséquence de cette concentration des vecteurs. Or nous sommes dans une phase où les vecteurs proposent de plus en plus des services ou des aides pour les développeurs de sites et les web-designers. En s'engageant à faire valoriser son lectorat auprès des vecteurs (publicité *AdSense* de *Google* par exemple), les éditeurs de site participent aussi de cette nouvelle architecture du web qui transforme le document en une représentation d'applications et l'appel à des services distants (ici le « service de la publicité »)... sur lesquels le contrôle de l'éditeur-diffuseur devient chaque jour plus faible. Quand *Google* décide d'ajouter des publicités dans les vidéo de *YouTube*, ce ne sont pas les éditeurs de sites ou des blogs qui re-diffusent ces vidéos qui vont en bénéficier. Le contrat d'usage des *Google maps* réserve la possibilité pour *Google* d'ajouter des informations (publicité géolocalisée par exemple) que le diffuseur n'a pas l'autorisation de contester ou de retirer.

7 Comprendre les enjeux de la distinction document/service

Le web est très jeune, mais diverses révolutions ont agité sa courte histoire. Le passage d'une « bibliothèque mondiale » constituée de documents à un réseau de

services reliant des pages considérées comme des « applications » est une de ces nouvelles révolution qui va provoquer des séismes dans les relations entre les producteurs de contenu et les éditeurs de site... au grand bénéfice des prestataires de services concentrés, les vecteurs qui deviennent les nouveaux maîtres du jeu.

Les logiques du document et du service sont distinctes. On l'a approché avec l'émergence de la télévision comme média majoritaire. Mais le web porte cette distinction à un degré supérieur : les « gagnants » économiques sont ceux qui peuvent réduire la part « documentaire » (contenu stable et pérenne) au profit de l'offre de zones applicatives... elles mêmes accessibles à bas coût pour le diffuseur, voire potentiellement susceptibles de générer des revenus (micro-publicité ciblée).

Comprendre la distinction entre les deux modes d'usage de l'information permet de mieux définir les activités collectives comme l'archivage, la défense de la diversité culturelle, ou la liberté d'expression, au sens fort du terme, c'est-à-dire la capacité à produire des informations susceptibles d'être entendues et conservées. Cette distinction permet aussi de mieux explorer la construction d'un nouveau mode d'organisation industriel et politique qui se met en place autour de nous (concentration des acteurs, vectorialisme). Enfin, cette distinction, qui se retrouve dans les deux tendances divergentes dans la normalisation technique du web, nous permet de mieux définir les points fondamentaux permettant de garantir la pérennité des activités collectives. Jusqu'à présent, les « documents » représentaient ce consensus. L'appât économique du modèle de service se double de la participation active des créateurs de sites, qui bénéficient des offres de widgets ou de services, sans disposer du recul nécessaire pour en maîtriser les conséquences à terme.

La première modernité est née avec le livre... la seconde modernité qui prend son essor sous nos yeux devra faire avec les stratégies du calcul, des traces, des compteurs physiques et « d'expérience », et toutes les méthodes de captation des individus et non de proposition de documents. C'est la recherche au carrefour des sciences humaines et des sciences informatiques, tenant compte des approches économiques et sociales pour les usagers comme pour les programmeurs et les concepteurs, qui doit nous aider à éclairer le chemin.

Références :

Anderson, C. (2006) The long tail, Random House Business books, juillet 2006

Ayres I.. (2007). Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart. Bantam, 2007.

Battelle, J. (2006). The Search : How Google and its rivals rewrote the rules of business and transformed our culture. Portfolio, 2006

Berners-Lee et al. (1998) Uniform Resource Identifiers (URI): Generic Syntax. RFC 2396 http://www.ietf.org/rfc/rfc2396.txt

CoolUri, 2008. Cool URIs for the Semantic Web, W3C Interest Group Note 31 March 2008 http://www.w3.org/TR/cooluris/

GESTE 2008 Édition de contenus et de services en ligne : Mode d'emploi, Victoires editions, janvier 2008 http://www.geste.fr/pdf/DP-Geste-ECSLME.pdf

Lecarpentier, JM ; Le Crosnier, H. ; Madelaine, J. Évolutions de l'architecture du web et des documents numériques In : Document et Société 2008, Chartron G, Broudoux, E (eds), ADBS ed, 2008).

Le Crosnier, H. (2008) Pour un regard politique sur la « courbe d'audience » Bibliothèque(s), num 39, juillet 2008.

Lohr, S. (2008). Google, Zen master of the market. The New York Times, 7 juillet 2008 http://www.nytimes.com/2008/07/07/technology/07google.html

McDonald (2008) Selling Photographs in the Digital Age, FUMSI, Août 2008 http://web.fumsi.com/go/article/use/3155

Melot, M. , Taffin, N. (2006). Livre, L'oeil neuf ed., 2006

Negroponte, N. (1995). L'homme numérique, Robert Laffont, 1995

Openid, 2008. http://openid.net/

Parody, E. (2008) Affaire Fuzz vs Olivier Martinez: le crétinisme au service d'un fiasco, Ecosphere le 27 mars 2008 http://www.zdnet.fr/blogs/2008/03/27/affaire-fuzz-vs-olivier-martinez-le-cretinisme-au-service-d-un-fiasco/.

Piotrr (2008) L'édition en ligne : un nouvel eldorado ? Un nouveau paradigme pour l'édition de sciences humaines, Blogo-Numericus, 22 mai 2008 http://blog.homo-numericus.net/spip.php?article154

Salaün, JM (2008). Le coeur du métier de Google. http://blogues.ebsi.umontreal.ca/jms/index.php/2008/09/15/533-le-coeur-du-metier-de-google-suite

Salaün, JM (2006). Web-médias – une synthèse http://blogues.ebsi.umontreal.ca/jms/index.php/2006/11/09/116-web-media-synthese

Session 4

Usages et Référentiels

Accessibilité des informations pertinentes des sites web accrue pour les personnes déficientes visuelles par extraction d'informations

Increased accessibility of informative content on web site for visually impaired people by information extraction

Sonia COLAS(1), Jérôme BULUCUA(1), Nicolas MONMARCHÉ(1), Mohamed SLIMANE(1)

(1)Université François-Rabelais de Tours,
Laboratoire d'informatique de l'Université de Tours,
Département Informatique de l'Ecole Polytechnique de l'Université de Tours
sonia.colas@univ-tours.fr
nicolas.monmarche@univ-tours.fr
mohamed.slimane@univ-tours.fr

Résumé. Les personnes handicapées visuelles utilisent des aides techniques telles que le clavier braille ou la synthèse vocale pour naviguer sur Internet. Bien que présentant une réelle avancée en termes d'autonomie pour ces personnes, ces aides techniques réalisent une lecture linéaire de la page. Cela implique d'une part la lecture d'informations générales (sommaire, publicité…) avant d'atteindre l'information proprement dite de la page web, et d'autre part la relecture des informations redondantes d'une page à l'autre d'un même site web. Pour pallier à ces problèmes, nous présentons dans cet article un outil permettant d'afficher uniquement l'information pertinente ou de réorganiser les divers éléments composant la page web.
Mots-clés. Handicapés visuels, web, contenu informationnel, extraction d'informations, réorganisation d'informations.

Abstract. Visually impaired people use assistive technologies such as Braille keyboards or voice synthesizers to browse the web. These technologies present a real advantage for the autonomy of these persons, but these technical helps are performing a linear reading of the web page. It implies, on the one hand, the reading of general information (contents, advertising…) before reaching the useful information itself of the web page, and on the other hand, the second reading of redundant information from a page to other one on the same Web site. To mitigate these problems, we present in this article a tool allowing to show only the relevant information or to reorganize the various elements in the web page.
Keywords. Visually impaired people, web, informative content, information extraction, information reorganization.

1 Introduction

Avec le développement d'Internet, de nombreux services ont été mis en place, tels des services commerciaux (Marin Lamellet et *al*, 2000), et des services administratifs. Outre le fait que ces e-services puissent être perçus comme un moyen plus rapide d'obtenir des documents, des renseignements ou même des objets, ils présentent surtout une réelle facilité pour bon nombre de personnes handicapées. Par exemple, des démarches administratives qui requièrent un déplacement en mairie — aussi banales ou anodines qu'elles puissent paraître — restent encore difficilement abordables pour les personnes atteintes de cécité. Dans l'idéal, l'utilisation d'Internet devrait s'avérer plus simple et plus rapide pour obtenir le même document. Or dans l'état actuel des choses, il n'en est rien. Effectivement, l'accès à des documents écrits (numériques ou non) n'est pas si simple pour des personnes handicapées visuelles. Même si l'apparition de l'écriture braille permît une réelle avancée des personnes atteintes de cécité en terme d'autonomie, son apprentissage est long et fastidieux. Avec l'avènement des nouvelles technologies numériques, des aides techniques apparurent, pour leur permettre d'utiliser seul un ordinateur. Ainsi grâce à la synthèse vocale, les personnes atteintes de cécité peuvent « lire », ou plutôt « écouter » des documents écrits, et ce sans connaissance du braille.

Les personnes handicapées, ayant accès à l'informatique grâce aux aides techniques, devraient par extension pouvoir naviguer sur Internet. Néanmoins pour pouvoir être interprétés correctement via les aides techniques, les sites Internet doivent respecter un certain nombre de normes. Précisons qu'un site est dit accessible s'il est consultable par tous quelque soit la technologie de consultation utilisée et quelque soit le système d'exploitation. Actuellement très peu de sites web sont accessibles (e-mediacité, 2006)(MeAC, 2007), c'est pourquoi de récentes lois en Europe (Loi D.D.A., 1995)(Loi n°2005-102, 2005) imposent aux sites web institutionnels, dans un premier temps, de respecter un niveau minimum d'accessibilité en suivant les recommandations établies (Chisholm et *al*, 1999) (BrailleNet, 2008) (WabCluster, 2007).

Toutefois, pour des personnes handicapées l'accès à l'information sur une page web, même accessible, reste difficile. Cette problématique sera présentée dans la section 2 afin d'introduire notre outil de réorganisation de l'information dans la section 3. Pour finir, la section 4 présente les résultats obtenus ainsi qu'un exemple d'application de notre outil.

2 L'accès à l'information sur une page web

Si l'utilisation des e-services administratifs est perçue comme une amélioration par les personnes handicapées, encore 40% ne s'en servent pas à cause de difficultés d'utilisation (Sandoz-Guermond et Bobillier-Chaumon, 2006). D'une part, peu de sites web publics sont effectivement accessibles. L'étude MeAC réalisée en 2007 précise que seulement 5,3% des sites web publics en Europe possède le niveau minimum d'accessibilité. D'autre part, une étude menée par T. Sullivan et R. Matson (2000) sur 50 sites très populaires a permis de montrer qu'il y avait une forte corrélation entre accessibilité et utilisabilité. Précisons que l'accessibilité est un sous ensemble de l'utilisabilité (traduction littérale du terme « usability ») (Stephanidis et Akoumianakis, 1999). L'utilisabilité représente le degré selon lequel un produit peut être utilisé, par des utilisateurs identifiés, pour atteindre des buts définis avec efficacité, efficience et satisfaction, dans un contexte d'utilisation spécifié. Communément, le terme « utilisabilité » est souvent utilisé pour désigner la capacité

d'un produit à être utilisé facilement. Un site web utilisable doit être efficace (le but doit être atteint) et efficient. Dès lors, pour mesurer l'efficience, il faut vérifier que le but de l'internaute a été atteint avec un minimum d'efforts, en un minimum de temps. Par conséquent un site web ergonomique minimise le temps de recherche d'un internaute par rapport au même site dans une version non ergonomique (Chevalier et *al*, 2004).

Il y a certaines règles d'ergonomie à respecter pour que l'information présente sur le site soit compréhensible et utilisable simplement. Les critères ergonomiques proviennent d'études empiriques ou de pratiques courantes. La version actuelle de ces critères a été publiée par Scapin et Bastien (1997). Ils se sont avérés utiles pour la classification de plus de trois cents recommandations ergonomiques pour la conception des sites web (Leulier et *al*, 1998). Ces recommandations sont répertoriées en catégories parmi lesquelles on trouve le guidage (incitation, lisibilité) et l'homogénéité/cohérence.

Certaines recommandations précisent par exemple qu'il est bon de fournir un moyen de navigation alternatif au sommaire comme un plan du site — ce qui permet à l'internaute de trouver rapidement la page qui l'intéresse. Lorsque ce plan de site n'existe pas sur le site web, il est alors possible d'utiliser un générateur de plan de site (Nation et *al*, 1997)(Nullpointer, 2006)(Colas et *al*, 2008).

D'autres recommandations s'attardent sur la lisibilité d'une page web. Elles préconisent d'utiliser une présentation homogène sur tout le site, ce qui facilite l'identification des différents éléments. Le visuel est fort utile pour structurer les propos et agrémenter la mise en page. Il est nécessaire de veiller à créer tout au long des pages d'un site, une identité et/ou une cohérence visuelle qui guidera l'internaute tout au long de sa visite. Le « visuel » — par lequel on entend également la position des éléments — est également nécessaire pour les malvoyants ou non-voyants qui trouveront sur chaque page les éléments qu'ils recherchent toujours au même endroit. Maurel et *al.* (2006) se sont intéressés à la représentation des éléments de structuration visuelle comme les énumérations lors de leur interprétation par les lecteurs d'écran.

Cependant les lecteurs d'écran lisent linéairement le contenu textuel d'une page. Pour naviguer au sein de la page les lecteurs d'écran proposent des fonctions permettant de passer au paragraphe suivant, d'aller de liens en liens… Sur une page web comportant de nombreux éléments (logo, titre, sommaire, publicité…) avant le contenu informationnel de la page, la lecture se trouve ralentie, même en utilisant les fonctions de navigation. Un internaute non-voyant peut donc rencontrer des difficultés pour accéder à l'information d'une page web, même si cette dernière respecte les critères d'accessibilité. Remarquons par exemple que des travaux (Oogane et Asakawa, 1998) ont été menés pour faciliter la lecture des tableaux de données via les lecteurs d'écran.

Dans cet article nous illustrons notre méthode en s'appuyant sur le site web de la mairie de Chambray-les-Tours. La figure 1 illustre une page de ce site en 2008. Ce site web représente typiquement les sites institutionnels de petite envergure (n'étant pas géré par des professionnels), pour lesquels nous avons réalisé, en collaboration avec le Conseil Général d'Indre-et-Loire, un outil proposant une démarche personnalisée vers l'accessibilité (Colas et *al*, 2007).

Sur cette illustration on identifie visuellement différents blocs d'informations. Le bloc (A) joue le rôle de menu principal et le bloc (B) celui de sous-menu. Cette figure représente la page qui s'affiche lorsque l'internaute clique dans le menu (A) sur le lien « La ville ». Les blocs (C) et (D) correspondent respectivement à un menu propre à la page (et à ses sous-pages) ainsi qu'à une zone d'information. Avant

d'accéder au contenu qui l'intéresse (zone (C) ou (D)), un internaute non-voyant devra passer par la lecture des menus (A) et (B). S'il est persuadé d'être sur la page internet qu'il cherche, cela représente une perte de temps. Après plusieurs pages visualisées sur un même site, la relecture d'informations redondantes d'une page à l'autre induit elle aussi un accès moins rapide à l'information pertinente.

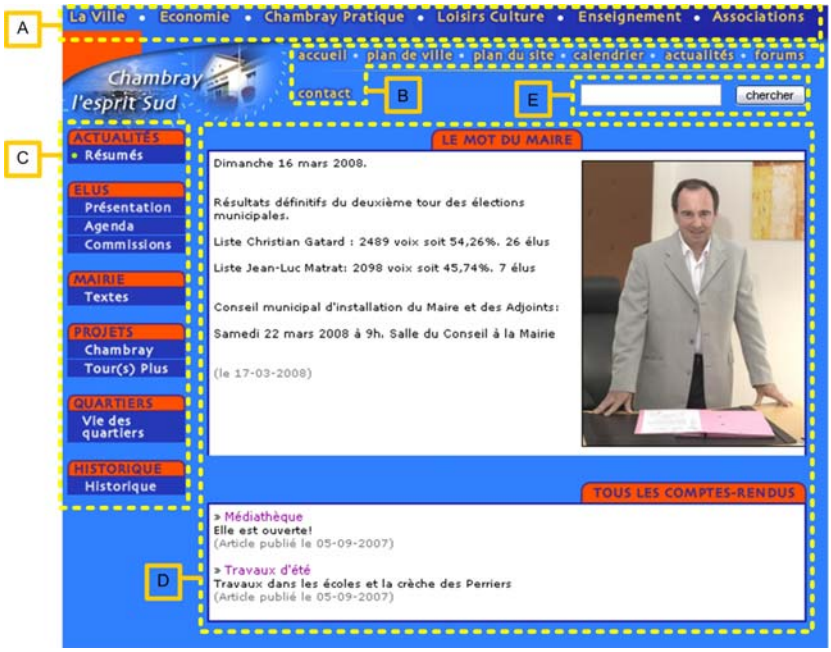


Figure 1. *Exemple de site web avec ses différentes zones*

3 Notre approche

3.1 Présentation générale de notre outil

Pour accélérer la vitesse de lecture via un lecteur d'écran, nous avons réalisé un outil capable de détecter dans une page web différentes zones d'information. Dans un premier temps, nous différencions les menus des autres zones d'information. Cette étape est détaillée dans la section 3.4. Grâce à cette distinction, lors de la visite d'une page d'un site web, l'internaute a la possibilité de n'afficher que les menus ou uniquement la page sans les menus. Dans un site web, tout ce qui n'est pas un menu n'est pas forcément de l'information utile propre à la page. Nous pouvons par exemple trouver dans un site web des bannières de publicité, le logo du site, un accès à un moteur de recherche… Tout ceci correspond généralement à des éléments qui se retrouvent sur toutes les pages du site mais qu'il n'est pas utile de relire à chaque nouvelle page. Nous proposons alors dans un deuxième temps de détecter toutes les zones redondantes sur les pages du site, et ce dès la deuxième page du site visitée. Notons qu'une zone est considérée comme redondante dès lors qu'une zone identique a été repérée sur la page précédemment affichée par le navigateur (la redondance n'est pas identifiée sur la totalité du site web). Cette étape de détection des zones redondantes fait l'objet de la section 3.3 de cet article.

Nous distinguons finalement 4 types de zones : les menus redondants, les menus spécifiques à la page, les zones redondantes qui ne sont pas des menus, et les zones d'information propre à la page qui ne sont pas des menus. L'utilisateur peut choisir d'afficher ou non chacune de ces zones. Notre outil propose également de réorganiser les éléments de la page pour lui permettre de personnaliser encore davantage son affichage. Ainsi l'internaute peut déterminer l'ordre d'apparition des zones qu'il a choisi d'afficher. Avant de décrire précisément les traitements de la page que nous venons d'exposer, nous devons préparer et nettoyer le code source de la page web.

3.2 Nettoyage du code source de la page et construction de l'arbre

Afin de préparer le code HTML pour la création de l'arbre, nous avons épuré ce code en fonction de nos besoins. Pour chaque page web que l'utilisateur souhaite visualiser, nous ne conservons que les structures de bloc et les liens, c'est-à-dire les balises DIV, SPAN et A, ainsi que les éléments de titres H1..H6 et les balises P représentant des paragraphes. En effet, un titre ne doit pas être assimilé à un paragraphe et deux paragraphes qui se suivent doivent rester distincts. Les tableaux de données, les formulaires et les listes sont considérés comme un bloc unique. Cela concerne les balises TABLE, FORM, UL, OL et DL. Quant aux balises dont le contenu est vide, elles sont bien évidemment supprimées.

Une fois le code source de la page nettoyé, chaque page du site web est représentée sous la forme d'un arbre. Chaque bloc présent dans la page HTML constituera un nœud de l'arbre. La figure 2 illustre deux exemples de pages HTML P_1 et P_2. On y distingue des zones $z_{i,k}$ qui sont considérées comme des blocs utiles des pages k={1,2}. La figure 3 illustre les deux arbres représentatifs correspondant aux pages P_1 et P_2.

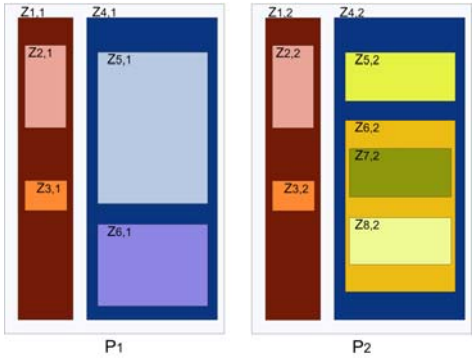


Figure 2. *Deux pages web d'un même site*

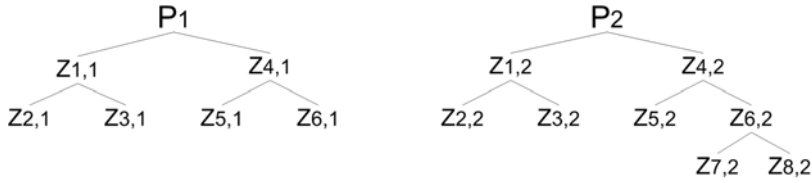


Figure 3. *Arbres représentatifs des deux pages*

3.3 Détection des parties redondantes sur des pages web

Lorsqu'une information est redondante sur plusieurs pages, nous considérons que c'est une information générale au site web. Ainsi lorsqu'un utilisateur navigue sur un site web, ces informations redondantes peuvent être supprimées dès lors qu'elles ont déjà été visualisées sur une page précédente. Pour détecter deux zones identiques, nous utilisons la méthode des *n*-grammes (un *n*-gramme est une sous-séquence de *n* éléments construite à partir d'une séquence donnée). Chaque zone peut alors être représentée par un vecteur contenant la fréquence d'apparition de chacun des *n*-grammes.

Soit $z_{i,k}$ et $z_{j,l}$ deux zones respectivement sur les pages P_k et P_l. Les équations suivantes permettent de définir deux zones strictement identiques et deux zones considérées comme identiques.

$$z_{i,k} = z_{j,l} \iff \cos(n\text{-gram}(z_{i,k}), n\text{-gram}(z_{j,l})) = 1$$
$$z_{i,k} \approx z_{j,l} \iff \cos(n\text{-gram}(z_{i,k}), n\text{-gram}(z_{j,l})) > \theta$$

avec θ un seuil fixé expérimentalement. Par définition le cosinus de l'angle formé entre les deux vecteurs *n*-gram($z_{i,k}$) et *n*-gram($z_{j,l}$) est le rapport du produit scalaire des vecteurs sur le produit des normes de chacun des vecteurs.

Lorsqu'une zone est considérée comme identique à une autre zone, elle est marquée en vue de sa suppression dans l'arbre représentatif de la page. Pour l'élimination des nœuds marqués, le parcours de l'arbre est effectué en largeur d'abord et le fils d'un nœud marqué est rattaché au père de ce dernier. Sur les deux pages de la figure 2, considérons que : $z_{1,1} = z_{1,2}$, $z_{2,1} = z_{2,2}$, $z_{3,1} = z_{3,2}$ et que $z_{4,1} \approx z_{4,2}$. La figure 4 reprend les arbres des pages P_1 et P_2 de la figure 3, dans lesquels les zones redondantes ont été supprimées.

Figure 4. *Arbres représentatifs des deux pages sans les zones redondantes*

3.4 Détection des menus

Que l'information soit redondante ou non, un internaute peut avoir besoin d'utiliser le menu d'un site web. Un menu est considéré comme une succession de liens. Dès que l'arbre contient une fratrie de nœuds correspondant à des balises « A », cet ensemble de nœuds est considéré comme un menu.

Sur les sites web non accessibles, il est courant de voir des menus composés uniquement d'images ou bien mis en forme à l'aide d'un tableau. Dans le premier cas, les images représentant des liens sont supprimées. Le lien est alors reconstruit en utilisant le texte alternatif de l'image, ou bien, s'il est inexistant, l'adresse de la cible du lien. Dans le second cas, ces tableaux n'ont pas lieu d'être puisque les recommandations du W3C précisent qu'ils ne doivent être utilisés que pour présenter des données et non pour faire de la mise en page. Si de tels tableaux sont détectés, ils sont automatiquement supprimés et leur contenu est restructuré à l'aide de balises « DIV ».

Finalement, une grande partie du travail consiste à être capable de distinguer les tableaux de données des tableaux de mise en page. Nous avons pour cela défini une règle permettant de détecter un tableau de données, et, tout tableau qui ne validera pas cette règle sera classé en tant que tableau de mise en page. Cette règle consiste à détecter si le tableau contient du code HTML dans ses pages. En effet la WAI

précise qu'un tableau de données ne doit contenir que des données, c'est-à-dire pas d'images, pas de vidéos… Cette règle consiste alors à rejeter (donc à les classer en tableau de mise en page) tous les tableaux contenant du code HTML dans leurs cases. Les tableaux de données sont conservés et restitués tels quels. Notons que des recherches ont été réalisées (Oogane et Asakawa, 1998) pour faciliter la lecture de ces tableaux par les lecteurs d'écran. Dans ces travaux chaque cellule devient un fichier HTML indépendant relié à ces cellules adjacentes ainsi qu'aux premières et dernières cellules de la ligne et de la colonne correspondante (contenant l'en-tête de la cellule).

Sur les deux pages représentées sur la figure 2, considérons que : $z_{1,k}$, $z_{2,k}$, $z_{3,k}$ pour k={1,2} sont des menus du site web et que $z_{5,2}$ correspond à un menu spécifique à la page P_2. La figure 5 reprend les arbres des pages P_1 et P_2 de la figure 3, dont tous les menus ont été supprimés.

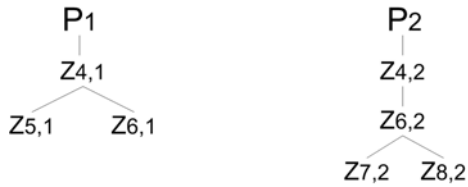


Figure 5. *Arbres représentatifs des deux pages sans les menus*

4 Application et résultats expérimentaux

4.1 Valeur de θ

Pour fixer la valeur θ représentant le seuil qui permet de considérer deux zones comme identiques, nous avons réalisé une série de tests sur des sites web des Ministères de l'Intérieur, de la Santé, de l'Education Nationale et enfin de la Santé (ces sites web ayant été utilisés lors de l'étude e-mediacité (2006) sur l'accessibilité des sites officiels). Nous avons comparé trois méthodes de mesure de similitude qui sont la méthode n-grammes, la distance de Levenshtein et celle de Jaro. Avec ces trois méthodes de similitude nous avons obtenu en moyenne respectivement les valeurs 0,01, 0,06 et 0,28 pour mesurer la similitude entre deux chaînes de caractères visuellement strictement différentes (ou la valeur 0 était attendue après observation visuelle). Pour s'approcher au mieux des résultats estimés par un observateur humain, nous avons conservé les n-grammes comme mesure de similarité. Afin de considérer comme redondants des nœuds très similaires visuellement mais ayant des codes très légèrement différents (comme par exemple une différence de saut de ligne ou d'espace), nous avons fixé par expérience θ=0,94. Ainsi, lorsqu'une différence inférieure à 6% entre deux nœuds est détectée, ces derniers sont considérés comme identiques. De ce fait, des menus différant uniquement par l'apparition d'un sous-menu sont considéré comme non redondant par notre programme avec la valeur θ=0,94. Or certains utilisateurs pourraient au contraire penser que ce sont deux menus identiques, avec une variante, et qu'ils devraient de ce fait être considéré comme redondant lors de la navigation. Dans ce cas la valeur de θ=0,94 est bien trop élevée. Cette appréciation de la ressemblance entre deux zones étant subjective, nous avons choisi de laisser l'utilisateur paramétrer lui-même le pourcentage au-delà duquel deux zones sont considérées comme identiques (la valeur par défaut étant fixée à 94%).

4.2 Détection des tableaux de mise en page

Nous avons testé notre méthode de distinction entre les tableaux de données et les tableaux de mise en page sur la page d'accueil des 73 sites web de l'étude e-mediacité (2006). En visualisant ces pages, nous avons observé qu'aucune ne comporte de tableaux de données.

Notre outil a permis de dénombrer 106 tableaux principaux (les sous-tableaux ne sont pas pris en compte ici) répartis sur 38 des 73 sites. D'après nos observations, ces derniers n'étant pas des tableaux de données, nous remarquons qu'encore 52% des sites web utilisent des tableaux de mise en page. Notre méthode de détection a permis d'en dénombrer 99 comme étant des tableaux de mise en page et 7 comme étant des tableaux de données (soit 6% d'erreur lors de la distinction). Ces erreurs de classement sont dues dans 3% des cas à des tableaux vides et dans 3% des cas à des tableaux ne possédant pas de balises fermantes. Dans notre application, une première étape consiste à nettoyer le code source. Les erreurs de détection concernant les tableaux vides n'apparaîtront plus étant donné que les balises vides sont supprimées dès le début.

4.3 Exemple d'application

Dans cette section nous présentons un exemple d'application sur le site web de Chambray-les-Tours. Précisons que dans cet exemple l'internaute désire n'avoir accès qu'aux informations propres à la page web et qui ne sont pas des menus. Toutes les zones redondantes (menu ou non) sont supprimées. Considérons que l'internaute décide de consulter la page présentée par la figure 1. Comme c'est la première page du site qu'il visite, aucune zone n'est considérée comme redondante. L'utilisateur a de plus précisé qu'il ne souhaitait pas afficher les menus ; par conséquent les zones (A), (B) et (C) sont supprimées. La figure 6 représente la page qui s'affiche dans son navigateur par défaut.

La figure 7 représente une autre page de ce site web vers laquelle l'internaute se dirige ensuite. On y trouve un menu (A'), un sous-menu (B') ainsi qu'une zone de recherche (E') identique respectivement aux zones (A), (B) et (E) de la page de la figure 1. Ces trois zones sont alors supprimées. La zone (C') est quant à elle détectée comme un menu (non redondant) et sera donc également supprimée.

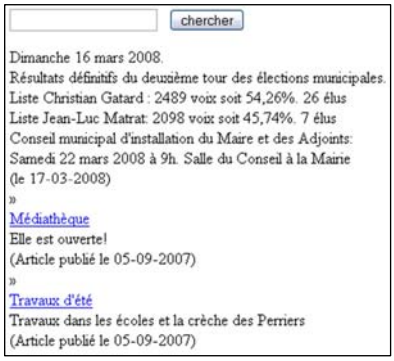


Figure 6. *Affichage du contenu propre à la page « La ville »*

Figure 7. *Page « Loisirs » du site web de Chambray-les-Tours*

Finalement l'internaute n'a accès qu'au texte de la zone (D') représentant le contenu propre à cette page et n'étant pas un menu (figure 8).

Politique culturelle de la ville
La ville de Chambray-Lès-Tours propose dans le cadre de sa politique culturelle une programmation offrant à la population des formes artistiques les plus variées. Parallèlement, la ville de Chambray-Lès-Tours développe des actions culturelles en collaboration avec les établissements scolaires et propose une politique de tarifs permettant l'accès au spectacle du public le plus large. Des partenariats nombreux se développent avec les institutions et les associations chambraisiennes ainsi qu'avec les acteurs culturels de l'agglomération tourangelle. Le service culturel de la ville a pour objectif d'offrir au public une programmation cohérente, en concertation avec l'école de musique et la bibliothèque.

Figure 8. *Affichage du contenu propre à la page « Loisirs »*

5 Conclusion et perspectives

Durant ces dernières années, l'accessibilité numérique a suscité beaucoup d'intérêt au sein de la communauté scientifique liée au handicap. L'accessibilité du web est particulièrement importante pour les personnes présentant des déficits de la perception. Ainsi les chercheurs apportent de nombreuses contributions permettant un espoir d'accroissement de l'autonomie des personnes handicapées à l'aide de cet outil qu'est internet.

Dans ce sens nous avons réalisé un outil d'extraction d'informations permettant à un internaute utilisant un lecteur d'écran d'accéder plus rapidement au contenu de la page qui l'intéresse. L'utilisateur peut également réorganiser les différents éléments de la page qu'il a décidé d'afficher. Actuellement nous distinguons quatre zones d'informations, mais il peut être envisagé d'en détecter d'autres comme par exemple celles correspondant à un moteur de recherche.

Références :

Loi n° 2005-102 du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées. JO n° 36 du 12 février 2005 page 2353. Disponible à : www.legifrance.gouv.fr/WAspad/ Visu?cid=719964 \&indice=1\&table=JORF\&ligneDeb=1

Loi D.D.A. (Disability Discrimination Act) en Grande-Bretagne, 1995. Disponible à : www.opsi.gov.uk/acts/acts1995/Ukpga_19950050_en_1.htm

Etude MeAC (2007). Measuring Progress of eAccessibility in Europe, octobre. Disponible à: ec.europa.eu/information_society/activities/einclusion/library/ studies/ meac_study/index_en.htm

BrailleNet (2008). Recommandations Accessiweb 1.1. Disponible à : www.accessiweb.org/fr/Label_Accessibilite/criteres_accessiweb/

Chevalier, A., Kicka, M., Cegarra, J. (2004). Quels sont les effets de la qualité ergonomique d'un site web et de l'expérience des utilisateurs sur la charge cognitive et le temps de navigation ?, in *10e Journée d'Étude sur le Traitement Cognitif des Systèmes d'Information Complexes*, JETCSIC'2004, Genève, Juin 2004.

Chisholm, W., Vanderheiden, G., Jacobs, I. (1999). Web Content Accessibility Guideline 1.0, WCAG 1.0, 1999. W3C Recommendation 5-May-1999. Disponible à : www.w3.org/TR/WAI-WEBCONTENT

Colas, S., Monmarché, N., Gaucher, P., Slimane, M. (2007). Accessibility of French publics web sites: a new tool and methodology to reach accessibility compliance, *International Conference on Human-Machine iNteraction* (HUMAN'07), Timimoun (Algerie), March 12-14 2007, pp.55-60.

Colas, S., Monmarché, N., Slimane, M. (2008). Génération de plan de site web pour les non-voyants par des fourmis artificielles. In *Revue d'Intelligence Artificielle, Numéro Spécial Métaheuristiques* (RIA), RSTI - RIA Métaheuristiques Vol. 22, numéro 2, Hermes. pp.137-159.

e-mediacité (2006). Etude portant sur l'accessibilité des sites Internet officiels. Audit réalisé du 30 janvier au 6 février 2006 sur un échantillon représentatif de 73 sites publics officiels.

Leulier, C., Bastien, J. M. C., Scapin, D. L. (1998). Compilation of ergonomic guidelines for the design and evaluation of Web sites, in *Commerce & Interaction Report*, Rocquencourt (France) : Institut National de Recherche en Informatique et en Automatique.

Marin Lamellet, C., Bruyas, M. P., Guyot, L. (2000). L'utilisabilité d'internet comme source d'information pour les voyageurs handicapés, in *Recherche Transports Sécurité*, no. 68.

Maurel, F., Mojahid, M., Vigouroux, N., Virbel, J. (2006). Documents numériques et transmodalité. Transposition automatique à l'oral des structures visuelles de texte. Document numérique, *Hermès* V. 09, N. 1, 25-42.

Nation, D., Plaisant, C., Marchionini, G., Komlodi, A. (1997). Visualizing Web Sites using a Hierarchical Table of Contents Browser: WebToc, in *Proceedings of Designing*

for the Web: Practices and Reflections (3rd Conference on Human factors and the Web), Denver.

Nullpointer (2006). WebTracer : générateur de plans de sites web. Disponible à : www.nullpointer.co.uk/-/webtracer.htm

Oogane, T., Asakawa, C. (1998). An interactive method for accessing tables in html. In *Proceedings of the third international ACM conference on Assistive technologies* (ASSETS '98), pages 126–128, New York, NY, USA. ACM.

Sandoz-Guermond, F., Bobillier-Chaumon, M.-E. (2006). L'accessibilité des e-services aux personnes non-voyantes : difficultés d'usage et recommandations, in *Actes du Congrès international IHM'2006*, Montréal, avril 2006, pp. 35-39.

Scapin, D. L., Bastien, J. M. C. (1997). Ergonomic criteria for evaluating the ergonomic quality of interactive systems, in *Behaviour & Information Technology*, vol. 17, 1997, pp. 220-231.

Stephanidis, C., Akoumianakis, S. (1999). Accessibility guidelines and scope of formative HCI design input : contrasting two perspectives, in *5th ERCIM Workshop on User Interfaces for All* .

Sullivan, T., Matson, R. (2000). Barriers to use : Usability and content accessibility on the web's most popular sites. *Proceedings of the ACM Conference on Universal Usability,* 139–144.

WAB Cluster, (2007). UWEM 1.2, 5/09/2007. Disponible à : www.wabcluster.org/uwem1_2/

Une approche *question-réponse* pour modéliser la recherche d'information

An Issue-Based Approach to Information Search Modelling

Alain LOISEL(1), Jean-Philippe KOTOWICZ(1), Nathalie CHAIGNAUD(1)

(1)INSA de Rouen – LITIS EA 4108, Mont-Saint-Aignan, France
alain.loisel@insa-rouen.fr
jean-philippe.kotowicz@insa-rouen.fr
nathalie.chaignaud@insa-rouen.fr

Résumé. Notre but est d'améliorer le moteur de recherche CISMeF en y intégrant un module de dialogue interagissant avec l'utilisateur. Pour étudier les processus cognitifs mis en œuvre pendant la recherche d'information, nous adoptons une démarche ascendante. Celle-ci consiste à mettre en place une expérimentation afin d'obtenir des dialogues humains durant la résolution de cette tâche. Cet article porte tout particulièrement sur l'analyse de ces dialogues par une approche fondée sur les « issues ».
Mots-clés. Dialogue humain-machine, analyse de corpus, recherche d'information.

Abstract. We aim at improving the health information search engine CISMeF, by including a conversational agent that interacts with the user in natural language. To study the cognitive processes involved during information search, a bottom-up methodology was adopted. An experiment has been set up to obtain human dialogs related to such searches. In this article, the emphasis lays on the analysis of these human dialogs through an issue-based approach.
Keywords. Human-computer dialog, corpus analysis, information search.

1 Introduction

La plupart des systèmes de recherche d'information ne cherchent pas à analyser les intentions de l'utilisateur. Pour cela, il faut donner à l'utilisateur la possibilité de décrire ce qu'il recherche en prenant en compte la façon dont il communique avec ses stratégies discursives, ses variations langagières et le contexte de recherche.

Notre but est de concevoir un agent conversationnel qui interagisse avec l'utilisateur en langage naturel. Cet agent est capable de construire une requête (formée de sous-requêtes) pas à pas. Chaque échange de ce dialogue humain-machine est traduit en items permettant la progression de la requête. Notre système est fondé sur des stratégies coopératives et constructives. La machine doit pour cela proposer à l'utilisateur des aides, des corrections ou des clarifications et présenter des choix en élargissant le but initial, si nécessaire.

Un tel système de dialogue permet d'expliciter le cheminement de l'utilisateur et propose les bons mots-clés à employer. De plus, la machine est capable de tester plusieurs requêtes en parallèle.

Notre cadre d'étude est l'annuaire de recherche de documents médicaux CISMeF (www.cismef.org) qui assiste les utilisateurs (médecins ou patients) dans leur recherche de documents médicaux sur le Web. Les termes sont issus du MeSH (medical subject headings) : mots-clés, qualificatifs (symptômes, traitements, etc.), méta-termes (spécialités médicales) et type de ressources. Nous avons mis au point une expérimentation pour obtenir des dialogues oraux entre un expert et des utilisateurs recherchant des informations médicales. Ces dialogues retranscrits (constituant un corpus) ont été analysés afin d'extraire leur structure discursive et leurs traits linguistiques, dans le but de concevoir un module de dialogue humain-machine.

2 Les systèmes de dialogue

2.1 La théorie des questions en discussion

Le modèle QUD (Ginzburg, 1996) est une théorie conventionnelle du dialogue fondée sur une sémantique formelle servant de modèle explicatif pour la résolution des ellipses et de certaines présuppositions. L'originalité de ce modèle, par rapport à ceux traitant des *questions* comme (Groenendijk et Stokhof, 1984), est de proposer une version structurée d'un tableau de conversation dirigé les questions en contexte. Le but de QUD est de déterminer très précisément les propriétés des couples *questions-réponses* (*issues*) et de montrer comment les questions et assertions en discussion enrichissent le tableau de conversation : que peut-on asserté ou demandé à un instant donné ? En reprenant l'approche par jeux de dialogue (Levin et Moore, 1980), Ginzburg conçoit le dialogue en termes de *coups* dans ces jeux.

2.2 Les dialogues fondés sur les issues et GoDIS

Fondé sur QUD, GoDIS (Larsson, 2002) est un modèle de dialogue implantable et ne garde qu'une sémantique simplifiée des questions. Il est possible de représenter trois types de question : interrogation totale *(?λ{}.P)*, l'interrogation partielle *(?λx.P(x))* ou l'interrogation parmi une liste de choix *(?set(P_1(x), P_2(y),…, P_n(z)))*. Les questions-réponses sont intégrées dans une structure de plans de dialogue (plans statiques composés d'une séquence d'actions abstraites appelées *actions de plan*). Les plans dans GoDIS représentent à la fois la tâche et le dialogue. GoDIS utilise la notion l'*état d'information* (IS), concept proche du tableau de conversation, qui se décompose en deux enregistrements :

Private
```
    Agenda : file d'Action
    Plan : pile de planConstruct
    Bel : ensemble de Propositions
    Nim : file d'actes de dialogue
```
Shared
```
    Com : ensemble de Propositions
    Qud : pile de Questions
    Issues : pile de Questions
    Actions : pile d'Action
    Previous moves : file d'actes de dialogue
    Last utterance : (énonciateur : participant)
                     (coups : ensemble d'actes de dialogue)
```

La partie privée représente les états mentaux de l'agent. La partie partagée permet de définir le tableau conversationnel mémorisant des informations partagées

par les deux interlocuteurs. Des règles de mise à jour et de sélection permettent de contrôler l'IS.

Le principe d'*accommodation* donne de la souplesse au dialogue. Il permet de mettre ou remettre des questions en discussion de manière non triviale. GoDIS propose des mécanismes d'accommodation s'inspirant du modèle de Ginzburg mais en l'étendant à d'autres champs de l'IS.

Le concept de *questions-réponses* peut être considéré comme une spécialisation des jeux de dialogue (Hulstijn, 2000). Comme le souligne (Beveridge et Milward, 2000), on peut définir des relations intentionnelles entre jeux de dialogue ainsi que des relations sémantiques. Or GoDIS ne définit que la seule relation de dominance entre questions-réponses et cette relation n'est pas récursive. D'autre part, aucune relation sémantique n'est définie et les différentes questions-réponses se succèdent par une relation de séquence implicite. Puisque aucune contrainte de *satisfaction-précédence* n'existe, GoDIS permet d'accommoder tout plan de dialogue à tout moment. Ce modèle bien que riche ne permet de traiter que des dialogues simples où chaque plan est défini comme une séquence simple d'actions ou de questions. D'autres part, sauf pour les phénomènes interactionnels de clarification ou d'établissement du *terrain commun*, le dialogue ne peut s'éloigner de la tâche pour introduire une digression, une explication ou pour proposer des suggestions à l'utilisateur. (Caelen, 2003) propose de distinguer au sein d'un dialogue la notion de stratégie de dialogue. Le changement de stratégie vise à choisir la meilleure direction d'ajustement des buts à un moment donné. Deux types de stratégies nous intéressent :

- La *stratégie coopérative* vise à modifier le but courant pour s'adapter au but de l'interlocuteur. Dans GoDIS, cela revient à modifier dynamiquement les actions de plan dans la pile d'actions de plan (par exemple, proposer des suggestions en fonction de l'IS) tout en gardant le même but.
- La *stratégie constructive* vise à abandonner provisoirement le but courant pour un nouveau but. Dans GoDIS, cela revient à abandonner la pile d'actions de plan pour un nouveau but. On peut ainsi introduire des digressions et ne pas suivre directement les séquences liées à la tâche.

Le principal avantage de GoDIS est de disposer d'un seul niveau de plan tout en étant très générique. Les actions de dialogue utilisées rendent la modélisation du dialogue proche de celle de la tâche. De plus, ce formalisme de plans permet de décrire des relations de dépendance entre questions et entre actions mais en assouplissant la notion de satisfaction-précédence. L'ordre des actions et questions n'est pas contraint au sein d'un même plan de dialogue.

3 Expérimentation et recueil du corpus

Comme nous l'avons déjà mentionné, nous avons mis au point une expérimentation pour obtenir des dialogues oraux entre un expert et des utilisateurs recherchant des informations médicales. Nous avons analysé à la main ces dialogues retranscrits et nous avons montré que les concepts de GoDIS permettent de décrire l'interaction humain-humain pour notre tâche.

Dans notre expérimentation, deux membres du projet, formés à CISMeF, jouent le rôle d'experts. Les questions émanent, sur la base du volontariat, de membres du laboratoire LITIS (secrétaires, doctorants, enseignants-chercheurs). Les deux experts permettent de contraster les démarches. Chaque expert se retrouve en

tête à tête avec son interlocuteur, face à une interface permettant d'utiliser la recherche avancée de CISMeF. L'expert mène la recherche et doit dans le même temps verbaliser au maximum tout ce qu'il est en train de faire. L'entretien se clôt lorsque la réponse satisfait l'utilisateur, ou qu'il semble bien qu'aucune réponse ne puisse être trouvée. Un corpus textuel est alors constitué des retranscriptions des 21 dialogues avec 21 sujets de cette expérimentation.

4 Analyse du corpus

Les différentes analyses présentées ont été réalisées à la main.

4.1 Décomposition des dialogues en sous-dialogues

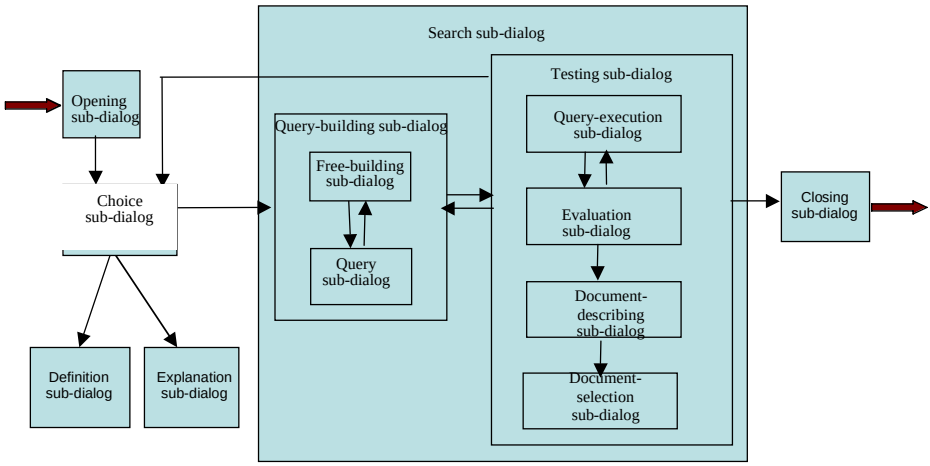


Figure 1. *Enchaînements possibles des sous-dialogues*

Puisque nos dialogues sont orientés vers une tâche bien précise, il est possible de décrire la structure globale d'une séquence idéale de recherche qui se dégage de l'analyse de notre corpus (voir Figure 1). Après une phase d'ouverture du dialogue (*opening sub-dialog*), il suit une phase de choix de sous-dialogue (*choice sub-dialog*) qui, généralement, conduit à une phase principale de recherche (*search sub-dialog*). Celle-ci se compose d'une ou plusieurs séquences de recherche qui aboutiront chacune à la résolution d'un problème posé par l'utilisateur. Enfin le dialogue se termine par une phase de clôture (*closing sub-dialog*). Une séquence de recherche se décompose en deux parties : une phase de formulation de requête (*query-building sub-dialog*) qui enrichit une requête courante et une phase de test (*testing sub-dialog*) de cette requête. La requête courante progresse tout au long de cette séquence de recherche.

La phase de formulation de requête se décompose elle-même en deux sous-dialogues : un premier sous-dialogue de formulation (*free-building sub-dialog*) a pour but de cerner le type de la demande et dans le second sous-dialogue de construction de la requête (*query sub-dialog*) un travail de reformulation se déploie. La requête est élaborée en coopération avec l'utilisateur. Des requêtes sont alors exécutées dans un sous-dialogue de lancement de la requête (*query execution sub-dialog*) et les résultats

sont présentés à l'utilisateur au cours d'un sous-dialogue d'évaluation de la liste de documents (*evaluation sub-dialog*) en fonction du nombre de documents trouvés. Cette liste est décrite à l'utilisateur dans un sous-dialogue de description de la liste de documents (*document-describing sub-dialog*) et des documents particuliers peuvent être lus plus en détail dans un dialogue de sélection de documents (*document selection sub-dialog*). A tout moment, cette phase de test de la requête peut être interrompue par de nouveaux sous-dialogues de demande de précision. Outre la recherche de documents, l'utilisateur peut demander une définition d'un terme médical (*definition sub-dialog*) ou des explications sur le système lui-même (*explanation sub-dialog*).

Nous pouvons relier ces sous-dialogues à des *Discourse Segment Purpose* (DSP) selon (Grosz et Sidner, 1986) : à chacun des segments linguistiques peuvent être associés des segments intentionnels ne formant pas forcément un bloc d'énoncés contigus. Il arrive souvent qu'un énoncé corresponde à deux segments de deux sous-dialogues. Nous situons les buts des DSP sur deux axes : les buts épistémiques où la résolution du segment dialogique permet d'apporter des connaissances dans le terrain commun (sous-dialogues de questions-réponses) et les buts actionnels permettant d'accomplir une action conjointe (sous-dialogues d'actions). Il existe alors deux relations entre segments dialogiques : *dominance* et *satisfaction-précédence*.

Dans la Figure 1, les dialogues emboîtés indiquent des relations de dominance entre sous-dialogues. Les flèches de transition entre sous-dialogues représentent les transitions naturellement induites par la tâche, qui n'enfreignent pas les relations de satisfaction-précédence. Cette structure de dialogue « idéale » se différencie de la structure linguistique des dialogues de notre corpus par un certain nombre de phénomènes que nous qualifierons d'opportunistes. En effet, le dialogue est constamment interrompu par un ensemble de phénomènes dialogiques à l'initiative de l'un des interlocuteurs. Ces phénomènes illustrent bien les difficultés d'une approche uniquement centrée sur la planification pour analyser notre corpus. En effet, parmi les sous-dialogues, certains sont facultatifs et peuvent être annulés explicitement ou implicitement ; des sous-dialogues incidents peuvent apparaître n'importe où, afin d'assurer le partage du terrain commun, et enfin, des sous-dialogues de choix peuvent apparaître pour relancer la recherche.

Après cette analyse orientée tâche, nous avons analysé à la main chaque énoncé. Une liste d'actes de dialogue a été construite à partir de traits linguistiques tirés du corpus. Cette taxinomie provient de (Weisser, 2003) et a été adaptée à nos besoins. Une description détaillée de cette taxinomie se trouve dans (Loisel et al., 2008). Dans la section suivante, nous décrivons les questions-réponses identifiées dans notre corpus.

4.2 Questions relatives à la tâche

Nous pouvons catégoriser les actes de dialogue par rapport aux phases dans la recherche de documents. A chaque phase du dialogue, est identifié un ensemble de questions-réponses qui fait progresser la tâche. Nous avons caractérisé chacune de ces questions-réponses en leur donnant une représentation formelle puis en étudiant des exemples issus du corpus. Une question-réponse peut se décomposer en plusieurs séquences d'actes de dialogue contigus ou non. Ces phases de la tâche font référence aux sous-dialogues mentionnés précédemment.

Par exemple, le dialogue suivant tiré de notre corpus correspond à une proposition de mots-clés dans le sous-dialogue de construction de requête. La traduction dans le langage du système est donnée ci-dessous :

- *Utilisateur : je serais plus sur l'arthroscopie moi*

```
Suggest(ajouterMotclé(arthroscopie.mc))
```

Expert: donc plus dans l'arthroscopie
`icm:sem*pos(ajouterMotclé(arthroscopie.mc))`

L'évaluation de `?intéressant(EnsD)` (avec `ensD` ensemble de documents retournés) permet principalement de répondre à la question implicite : « Est-ce que les résultats pris dans leur ensemble sont intéressants ? ».

- *Expert : ça par contre, c'est pas terrible*
  `Inform(¬intéressant(EnsD))`
- *Utilisateur : Voilà, c'est ce genre de documents qu'il faut chercher en fait*
  `Inform(intéressant(EnsD))`

4.3 Instanciation des questions-réponses

Nous avons ensuite défini des relations de cohérence entre ces questions-réponses et nous avons pu alors observer différentes instanciations de ces questions-réponses : une question suivie immédiatement d'une réponse n'est qu'un exemple parmi les différents types d'interactions possibles. Nous présentons ici des exemples tirés de notre corpus.

- *Expert : Vous voulez savoir l'évolution, le diagnostic, les traitements possibles ?*
  Utilisateur : euh ! plutôt le diagnostic.

Dans ce dialogue, lorsque la question est posée, elle est mise en discussion dans les champs `Qud` et `Issues`. Le fait d'obtenir une réponse à la question l'élimine du champ `Qud` (théorie QUD classique). Dans les ajouts proposés par GoDIS, répondre à une telle question la supprime également de `Qud` mais pas de `Issues` (Cooper et Larsson, 2003) (Larsson, 2002). Ainsi l'interlocuteur pourra corriger sa réponse dans la suite du dialogue par le phénomène de ré-accommodation. De plus, lorsque l'interlocuteur refuse de répondre à une question, elle est directement annulée de la pile des questions en discussion. Pour certaines questions, une ou plusieurs réponses sont possibles. La première réponse est donc résolvante, mais ne clôt pas la question, comme dans le dialogue suivant :

- *Expert : Est-ce qu'éventuellement vous pouvez préciser un petit peu ?*
  ```
  Com :
  Qud : ? λx.precisions(x)
  Issues : ? λx.precisions(x)
  LastMove : RequestInfo(?λx.precisions(x))
  ```
- *Utilisateur : Bah ! savoir les démarches à accomplir si on veut être donneur d'organes par exemple*
  Utilisateur : S'il y a des examens à passer
  ```
  Com : précisions("démarches à accomplir","donneur d'organes"),
        précisions("examens")
  Qud :
  Issues : ? λx.precisions(x)
  LastMove : Answer(("examens")
  ```

Il y a sur cet exemple deux réponses satisfaisantes à la question posée par l'expert. L'utilisation de connecteurs comme « aussi » et « encore » permet de préciser qu'il y a une nouvelle réponse à la question. Une question peut être aussi répondue par le locuteur qui a posé lui-même la question :

- *Expert : est-ce qu'un autre qualificatif va pouvoir nous intéresser ? On avait prévention et contrôle, on peut voir aussi le diagnostic des problèmes d'articulation*
  ```
  RequestInfo(? λm.ajouterQualificatif(m))
  Answer(ajouterQualificatif(prévention.qu))
  Suggest(ajouterQualificatif(diagnostic.qu))
  ```

Godis simplifie ce problème en considérant qu'une et une seule réponse est acceptable si elle est potentiellement résolvante. Selon cette modélisation, le comportement attendu serait que la deuxième réponse remplacerait la première car elle serait considérée comme une correction. Clairement ici, ce n'est pas le cas et les deux réponses sont acceptables et résolvantes.

Cependant, d'autres questions n'admettent qu'une seule réponse :

- *Utilisateur : c'est peut-être un truc osthéo*
  Expert : Oui c'est ça. Alors, donc ostéopathie. Merci. Euh ! Non, ils n'ont pas ça dans l'annuaire
  Utilisateur : rhumatologue. Rhumatologie dans la première
  Expert : D'accord. Euh ! Donc on va essayer ça, avec rhumatologie. Donc je lance la recherche comme ça.

Dans l'exemple ci-dessus, l'utilisateur propose une réponse à la question `? λm.ajouterMétaterme(m)` par un méta-terme « ostéo (pathie) ». Cette réponse est acceptée, mais le méta-terme n'existe pas dans la terminologie. L'utilisateur répond à nouveau par un autre méta-terme « rhumatologie » sans que la question ne lui soit reposée. Ici la première réponse est résolvante, mais l'utilisateur peut toujours proposer une seconde réponse. Ainsi, donner une deuxième réponse à la question doit être interprété comme une correction de la première réponse qui doit être annulée. Le mécanisme de ré-accommodation de GoDIS peut être utilisé pour modéliser cette séquence dialogique.

Il existe deux mécanismes différents (réponses multiples ou ré-accommodation) et le seul critère permettant de trancher est intrinsèque à la question-réponse. Dans notre corpus, de nombreux actes de dialogues d'assertion, d'information ou de suggestions sont présents. Nous avons associé ces énoncés à des questions-réponses même si aucune représentation n'existe sous la forme interrogative. Dans GoDIS, ces mécanismes ne sont pas du tout étudiés, mais dans QUD, le contenu propositionnel de ces actes de dialogue est représenté par des faits ajoutés dans l'IS partagé (champ `Com`). Lors de l'ajout d'un fait P(x), une question polaire ?P associée est aussi ajoutée dans `Qud`. L'interlocuteur peut alors approuver ou nier un fait. Il répond en fait à la question implicite associée.

- *Expert : (dans une phase de présentation des résultats) Ah ! on a trois choses : enregistrement <pligraphique> du sommeil, "syndrome du sommeil". Je ne sais pas si apparemment ç'est ça.*
  ```
  Com : ¬DocumentIntéressant(Document)
  Qud : ? DocumentIntéressant(Document), ? λx.precisions(x)
  Issues : ? DocumentIntéressant(Document)
  ```
- *Utilisateur : non, à mon avis non*
  ```
  Com : ¬DocumentIntéressant(Document)
  Qud :
  Issues : ? DocumentIntéressant(Document)
  ```

En reprenant le mécanisme d'accommodation de fait, nous pouvons identifier deux autres mécanismes d'accommodation analogues qui permettent d'interpréter certains actes de dialogue indirects : l'accommodation de questions polaires vers des questions ouvertes et les actes de dialogue indirects entre une question et une action associée.

Nous avons pu également observer dans notre corpus des exemples de stratégies coopératives, notamment lorsque l'expert répond lui-même à la question qu'il pose :

- *Expert : Comment traduire ça ? ce serait finalement des thérapeutiques*

Bien entendu, l'utilisateur a la liberté de contester cette réponse et d'en proposer une autre (phénomène de ré-accommodation).

De la même manière, l'expert peut suggérer une réponse sans même poser de question, ce qui permet de demander l'approbation de l'utilisateur. Une troisième forme de stratégies coopératives consiste à proposer des réponses à l'interlocuteur sans même lui demander son approbation (par un `Inform()`).

4.4 Relations entre questions-réponses.

Nous avons cherché dans notre corpus les relations qui existent entre les différentes questions-réponses en considérant les relations interactionnelles subordonnées (la deuxième question-réponse est une clarification, une correction, une reformulation ou une précision permettant de mettre la première question-réponse dans le terrain commun) ou les relations interactionnelles coordonnées (constituées de relations conventionnelles ritualisées comme les salutations et les remerciements) (Beveridge et Milward, 2000).

Comme nous l'avons déjà noté précédemment, nous avons également besoin de relations intentionnelles entre la tâche et les questions-réponses. Par exemple, l'action de haut niveau de construction de requête *action(construireRequête)* domine la question d'ajout de mot clés. De même, il y a une relation de satisfaction-précédence entre *action(construireRequête)* et *action(lancerRequête).*

Enfin les relations sémantiques pures n'établissent pas de lien intentionnel direct mais entraînent le lancement d'un sous-dialogue par stratégies constructives.

4.5 Transitions entre sous-dialogues

Ayant étudié les différents sous-dialogues et les relations entre questions-réponses, il nous reste à voir comment enchaîner ces sous-dialogues. Une approche par jeux de dialogue comme (Maudet, 2001) propose de décrire les transitions entre deux jeux de manière explicite en suivant quatre étapes : proposition d'ouverture d'un nouveau jeu, ouverture du jeu, proposition de fermeture du jeu courant, fermeture du jeu courant.

La proposition d'ouverture explicite d'un sous-dialogue permet de relancer un nouveau dialogue. Ce peut être un choix simple comme dans le dialogue ci-dessous, où la réponse négative permet de rester dans le sous-dialogue courant ou de passer dans le sous-dialogue qui suit le dialogue courant :

- *Expert : Est-ce que vous avez une autre requête?*
  Utilisateur : Non, c'était tout.

L'ouverture explicite peut se faire par l'acte de dialogue noté InformIntent().

- *Utilisateur: D'accord, je mets d'abord l'énoncé initial. Donc c'est un problème avec la nourriture ?*
  Expert: oui
  Utilisateur : D'accord ! Donc on va déjà aller dans les accès thématiques

Le passage dans un nouveau sous-dialogue est marqué par l'utilisation de connecteurs comme « donc » mais aussi « déjà ». Nous considérons que le dernier énoncé, noté `InformIntent()`, introduit lui aussi une question dans `Qud` par accommodation de fait, sur le désir de l'interlocuteur d'entrer dans ce jeu de dialogue. Si la réponse est négative, l'entrée dans le jeu peut être refusée.

Comme nous l'avons déjà vu, la tâche est représentée par des actions liées entre elles par des relations de séquences et/ou de satisfaction-précédence. Il existe également des notions de séquences représentant la succession « prototypique » des sous-dialogues. La transition entre deux sous-dialogues peut être alors implicite.

Les transitions entre sous-dialogues peuvent aussi ne pas suivre un ordre séquentiel. Ainsi l'accommodation de sous-dialogues (appelée accommodation de plans dans GoDIS) permet de court-circuiter une séquence de recherche :

- *Expert : Voilà ! Donc je commence la recherche avec ça (parasomnie). OK ! Bon. Alors il n'y a rien là-dedans*
  Utilisateur : Donc je crois qu'on va en rester là, non ?
  Expert : D'accord, ça marche

Ici, l'utilisateur propose directement la fin de l'entretien après un échec. C'est le passage d'un sous-dialogue de description de documents à un sous-dialogue de conclusion. Cette suite est non attendue par l'expert, mais elle peut être accommodée car les relations de satisfaction-précédence entre sous-dialogues sont respectées. Notons encore ici le connecteur « donc » indiquant le changement de sous-dialogue.

De plus, certains sous-dialogues définis sont peu informatifs et peuvent être sous-entendus dans le corpus. C'est le cas notamment des sous-dialogues d'ouverture et de lancement de requête. Ainsi, voici comment commence un dialogue :

- *Utilisateur : Donc je pose une question ? Je dis ce qui ne va pas et puis...*
  Expert : Voilà !
  `? PoserQuestion`

L'utilisateur peut ainsi initier le dialogue en interrogeant l'expert sur le but du dialogue. Cette étape permet de court-circuiter une phase initiale d'ouverture de dialogue. Or, il y a une relation de satisfaction-précédence entre le sous-dialogue d'ouverture et les autres sous-dialogues. Il faut donc considérer que l'ouverture du dialogue a bien eu lieu mais non verbalement.

Enfin, le passage dans un sous-dialogue peut se faire en suivant une stratégie constructive. Comme nous l'avons évoqué, cela intervient lorsqu'il y a des subordonnées coordonnées sémantiques mais sans liens directs avec la tâche. Le plus souvent ce sont des relations d'aide, de clarification, d'explication. En termes de jeux de dialogue, il s'agit de jeux emboîtés. Lorsque la séquence collaborative se termine, le contexte initial doit être retrouvé. Ainsi, nous rencontrons dans le corpus des relations rhétoriques d'explication :

- *Expert : alors insomnie, sinon il ne me trouve pas comme mot-clé, c'est bizarre !*

```
Suggest(?ajouterMotclés(Insomnie familiale.mc))
Inform(¬terminlogie(insomnie familiale.mc))
Com : (¬terminologie(insomnie.mc)
Qud : ? Cause(¬terminologie(insomnie.mc))
```

- *Utilisateur : peut-être que c'est uniquement inclus dans les troubles de sommeil ?*

```
Suggest(Cause (¬terminologie(insomnie.mc),
              (inclus(insomnie.mc,« trouble du sommeil »))))
Com : (¬terminologie(insomnie.mc),
        Cause (¬terminologie (insomnie.mc),
               (inclus(insomnie.mc,« trouble du sommeil »)))
```

Lorsqu'un fait est ajouté dans le terrain commun, une question implicite subordonnée portant sur la cause de ce fait, noté `? λc.cause(f,c)` avec f, c de type `Facts` est ajoutée dans `Qud`. C'est donc un mécanisme d'accommodation de fait un peu particulier puisque ce n'est pas la question polaire associée qui est ajoutée. L'utilisateur suggère une explication à ce fait et ainsi répond à la question implicite `? cause(¬terminologie(insomnie))`. La question sur la cause est alors résolue et retirée de `Qud` et les deux faits sont ajoutés dans `Com` :

- `inclus(insomnie.mc,« trouble du sommeil »)`
  `Cause(¬terminologie(insomnie.mc),`
  `(inclus(insomnie.mc,« trouble du sommeil »))))`

5 Discussion

Les applications dérivées de GoDIS sont restées relativement simples comme, la gestion des lumières d'un appartement par système vocal, l'utilisation d'un lecteur multimédia, ou la réservation de tickets d'avion. Cependant un système comme GoDIS peut être utilisé de façon plus complexes avec des modèles de la tâche et de l'interaction plus évolués.

Modéliser le dialogue de recherche d'information dans un annuaire électronique à l'aide de GoDIS nécessite alors l'introduction de nouvelles notions :

- La possibilité d'obtenir plusieurs réponses à une seule question posée : dans GoDIS, dès qu'une réponse a été donnée et que cette réponse est résolvante, la question sous-jacente est retirée du champs `Qud`. Or, dans le cas de la recherche de documents, une question de formulation de requête aboutit à la création d'une requête. Pour savoir si cette réponse est acceptable pour l'utilisateur, il ne suffit pas de lui proposer. Il faut aussi tester cette réponse à travers CISMeF, analyser les résultats de la recherche et les proposer à l'utilisateur. Si cela ne constitue pas une réponse valide, la question est toujours en discussion et ne doit pas être enlevée de la pile `Qud`.
- Les relations de satisfaction-précédence : dans GoDIS, le principe d'accommodation apporte de la flexibilité au dialogue. Cependant, celui-ci doit rester cohérent. Les relations de satisfaction-précédence interdisent l'accommodation si ses effets ne sont pas en accord pas avec la tâche.
- Les stratégies coopératives et constructives : en recherche de documents, les informations pertinentes retournées par le système permettent soit de progresser dans le plan courant (stratégie collaborative), soit de l'abandonner (stratégie constructive). Ainsi, le gestionnaire de dialogue (et non l'utilisateur) est capable d'accommoder de nouvelles questions-réponses.
- Les jeux de dialogue : l'accommodation est parfois trop souple et peut entrer en conflit avec les maximes de Grice (Grice, 1975). En utilisant la notion de contraintes sur les jeux de dialogue, il est possible de proposer des restrictions d'enchaînement de jeux. Les principes généraux d'accommodation et d'anticipation proposés par (Maudet, 2001) peuvent être incorporés dans GoDIS.

6 Conclusion

Dans le but de concevoir un système de dialogue humain-machine pour la recherche de documents médicaux, nous avons adopté une démarche expérimentale permettant d'obtenir des dialogues oraux de recherche d'information. Nous avons analysé ce corpus de dialogues retranscrits et nous avons pu mettre en évidence la

construction d'un terrain commun et des effets d'accommodation des dialogues. Le dialogue est constitué de sous-dialogues directement liés à la tâche. Le modèle de dialogue que nous proposons est fondé sur la notion de questions-réponses. Nous proposons une architecture d'agent dialogique modulaire qui comprend trois composantes essentielles :

- Le modèle de la langue qui réalise une analyse lexico-syntaxique (à partir du TreeTagger (Schmid, 1994)) ainsi qu'une analyse pragmatique (représentation en actes de dialogue) et une analyse sémantique (identification des termes CISMeF).
- Le modèle du dialogue qui comprend le gestionnaire du dialogue (à partir de Trindikit (Larsson et Traum, 2000)) et le générateur de phrases (utilisant des phrases à trous).
- Le modèle de la tâche qui englobe l'interface CISMeF avec la base de documents ainsi qu'un constructeur de requêtes à partir de termes proposés et un interpréteur de requêtes pour les raffiner.

Cette architecture d'agent dialogique est en cours d'implémentation (Loisel et al., 2008).

Références :

Beveridge, M., Milward, D. (2000). Ontologies and the Structure of Dialogue. *Workshop on the Semantics and Pragmatics of Dialogue*, Catalog'04, Barcelona, pages 69-76.

Caelen, J. (2003). Strategies of Dialogue. *Speech Technology and Human-Computer Dialogue Conference*, Bucarest, 27-42.

Cooper, R., Larsson, S. (2003). *Accommodation and reaccommodation in dialogue*. In Baüerle, R., Reyle, U., Zimmerman, E. (eds.), Presuppositions and Discourse, Elsevier.

Ginzburg, J. (1996). *Interrogatives: Questions, facts and dialogue.* The Handbook of Contemporary Semantic Theory, vol. 5(18), 359-423.

Grice, H.P., (1975). *Logic and conversation*, Cole (ed.), new york: academic press edition vol. 3, 41-58.

Groenendijk, J., Stokhof, M. (1984). *Studies on the semantics of questions and the pragmatics of answers*. PhD thesis, University of Amsterdam.

Grosz, B., Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, vol. 12(3), 175-204.

Hulstijn, J. (2000). *Dialogue Models for Inquiry and Transaction*. PhD thesis, University of Twente.

Larsson, S. (2002). *Issue-based Dialogue Management*. PhD thesis, University of Goteborg.

Larsson, S., Traum, D., (2000). *Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit*. Natural Language Engineering, Special Issue on Best Practice in Spoken Language, Dialogue Systems Engineering, Cambridge University Press, 323-340.

Levin, J. and Moore, J., "Dialogue-games: meta-communication structure for natural language interaction", *Cognitive Science*, vol. 4(1), pp. 395-420, 1980.

Loisel, A., Chaignaud, N., Kotowicz, J.-P. (2008). Modeling human interaction to design a human-computer dialog system. *International Conference on Enterprise Information Systems*, Barcelona, Spain.

Maudet, N., Modéliser les conventions des interactions langagières: la contribution des jeux de dialogue. *Thèse de doctorat*, Université de Toulouse, 2001.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing*, Manchester UK, 44-49

Weisser, M. (2003). SPAACy: A tool for Annotating Dialogue. *International Journal of Corpus Linguistics*, vol 8(1).

MPEG-21 : base normative pour les TICE du XXI siècle

MPEG-21: the core standard for the 21st century pedagogical resources

Françoise PRETEUX (1), Alain VAUCELLE (1), (2), Mokhtar BEN HENDA (3), Henri HUDRISIER (4)

(1) Institut TELECOM, TELECOM & Management SudParis, Département ARTEMIS, Evry, France
Francoise.Preteux@it-sudparis.eu

(2) Laboratoire Paragraphe Paris 8, Université Paris 8, Saint-Denis, France
Alain.Vaucelle@it-sudparis.eu

(3) CEM-GRESIC, Université Bordeaux III, Bordeaux, France
benhenda@yahoo.com

(4) Laboratoire Paragraphe Paris 8, Université Paris 8, Saint-Denis, France
henri.hudrisier@wanadoo.fr

Résumé. Face aux enjeux des TICE ainsi qu'aux différentes initiatives de normalisation, en particulier par l'ISO/IEC JTC1 SC36, les auteurs, dans un contexte de compréhension et description globale des TICE, tentent d'analyser comment la famille MPEG au travers des normes MPEG-4, 7 et 21, pourrait offrir une base normative pour les métadonnées associées aux TICE.
Mots-clés. TICE, EAD, normalisation ISO, SC36, SC29, normes MPEG-4, 7 et 2

Abstract. Facing the challenges of the ICT in education as well as various standard initiatives and specifically by ISO / IEC JTC1 SC36, authors attempt within a global and comprehensive description from work to analyze how the MPEG-4, 7 and 21 standards, could provide a normative basis for metadata associated with the ICT in education.
Keywords. ICT, e-learning, standardization process, ISO, SC36, SC29, MPEG-4, MPEG-7 and MPEG-21 standards.

1 Introduction

Dans un contexte de compréhension et description globale des TICE, cet article vise à analyser comment dans un environnement en profonde mutation, la famille MPEG pourrait offrir une base normative pour les métadonnées du monde de l'éducation et de la formation.

Les contextes et les enjeux des TICE sont tout d'abord rapidement brossés, puis l'histoire de la normalisation des TICE synthétisée. S'ensuit une brève analyse des normes MPEG. Les recommandations pour exploiter MPEG-21 au niveau des TICE sont finalement discutées.

2 TICE : contexte et enjeux

2.1 Un contexte en profonde mutation

Dans le domaine de l'éducation et de la formation, les défis sont liés à la mutation de la chaîne de production et diffusion de l'« analogique » vers le numérique. Ces changements affectent non seulement les outils, mais aussi la façon de s'approprier techniquement, institutionnellement, professionnellement, et cognitivement, ce nouvel environnement technique.

D'un point de vue technique, ces défis sont du même ordre que ceux déjà observés dans le monde de l'audiovisuel. Toutefois, comme les TICE concernent le monde de l'éducation au sens large du terme, *i.e.* ingénierie des connaissances, savoir et savoir-faire, de nouvelles spécificités sont à prendre en compte. En effet, les conditions socio-économiques et socio-techniques font que les TICE vont être déployées sur une grande échelle puisque intimement liées aux évolutions sociétales.

L'industrialisation de la connaissance, aussi bien dans la dimension *consumériste* de la production de biens et de services que dans la dimension *machinique* des industries de la connaissance, restent des questions ouvertes. Le processus mis en œuvre par les TICE se soldera-t-il par un progrès social ou un recul face à des technologies de transmission du savoir plus traditionnelles ?

Aujourd'hui, le grand défi des TICE concerne la mise à disposition des savoirs, et des savoir-faire, en même temps que leur accès (Rifkin, 2000), *i.e.* la manière de présenter, d'interagir et de structurer des contenus éducatifs. Cela n'est pas neutre d'un point de vue sociétal. Une transmission individualisée de masse des savoirs et savoir-faire peut conduire à une société de la connaissance essentiellement basée sur le service.

2.2 Normalisation des TICE : vers de nouvelles exigences

Pour répondre à ces enjeux et en fixer le périmètre, après l'émergence foisonnante de solutions visant à s'imposer comme des standards *de facto*, un contexte normatif global pour les TICE est en cours d'élaboration. Ces normes sont discutées dans le cadre mondial de l'ISO (Organisation internationale de normalisation) depuis novembre 1999, en accord avec l'IEEE (*Institute of Electrical and Electronics Engineers*) et le JTC1 (*Joint Technical Committee*). Le JTC1 est le comité de référence pour la normalisation des technologies de l'information. Les normes concernant les technologies de l'information pour l'éducation, la formation et l'apprentissage sont élaborées au sein du SC36.

Or, en presque 10 ans, le paysage des exigences culturelles, linguistiques, institutionnelles, disciplinaires s'est notablement diversifié. Les experts délégués exigent des normes acceptables pour toutes les langues, tout type d'organisation (de

la formation ou de l'éducation), tout style pédagogique (du plus magistral au plus collaboratif), tout type de gestion spatiale et temporelle (apprentissage en ligne « *e-learning* » synchrone ou asynchrone, présentiel électronique…), tout type de médiation (multimédia) et de modalités perceptives (simulation, « *mobile-learning* », « *television-learning* », réseaux collaboratifs, adaptation pour les déficients sensoriels…).

Aujourd'hui, la normalisation au sein des TICE couvre aussi bien les spécifications techniques communes aux futurs produits associés aux TICE, que les modèles interrelationnels entre les apprenants et les enseignants, ou la gestion et l'organisation des échanges. Le programme est vaste, riche et le contexte complexe et sensible.

Appréhender la normalisation au sein des technologies liées à l'éducation et à la formation revient aussi à imaginer différents scénarios d'échange des savoirs et des savoir-faire au sein des réseaux (Proulx, 2001). Un premier modèle pourrait reposer sur l'apprenant, seul devant son écran, qui accèderait à des bases de données structurées et indexées lui offrant le contenu dont il a besoin. Un deuxième scénario pourrait s'appuyer sur la constitution de « communautés apprenantes » permettant la mise en commun, l'échange de ressources et des expériences. Un troisième pourrait fonctionner sur la base du tutorat, où les ressources sont organisées autour des compétences et de l'intérêt des apprenants. Un quatrième pourrait s'organiser avec un enseignant-formateur, ce dernier devenant un chef d'orchestre chargé de rassembler les différents contenus mis à sa disposition, et de les utiliser de façon pédagogique. De tous ces scénarios, il ressort que interaction/interactivité est au cœur du dispositif d'enseignement aussi bien pour le passeur de savoir, que pour le créateur de contenu ou l'étudiant.

Ces nouvelles pratiques d'apprentissage induisent de profondes remises en cause du rôle de l'enseignant-formateur, de sa place dans le système éducatif et de sa position face à l'apprenant (Hudrisier, 2006).

Soulignons que ces outils favorisant la médiation induisent de forts enjeux financiers (matériel, réseau, support de cours…), une visite au BETT (*British Education and Training Technology*), le salon dédié aux technologies de l'éducation permet de s'en convaincre immédiatement.

La normalisation des TICE s'inscrit donc dans le cadre riche et complexe des interrelations entre la technologie et la société. L'historique des différentes étapes pour les spécifications d'un langage commun et d'espaces d'échanges normalisés est synthétisé dans le paragraphe suivant.

3 Sur le chemin de la normalisation des TICE

Deux périodes façonnent le paysage de la normalisation des TICE : la première avant 1999 et la seconde marquée par la création du SC36 en 1999. Une chronologie des faits marquants la normalisation de TICE ainsi qu'un panorama sont décrits dans les tableaux 1 et 2 (Blandin, 2003).

A ce jour 7 groupes de travail sont constitués au sein du JCT1-SC36 : WG1 - Vocabulaire (*Vocabulary*) ; WG2 -Technologie collaborative (*Collaborative technology*) ; WG3 - Information sur l'apprenant (*Learner Information*), WG4 - Gestion et livraison de l'apprentissage (*Management and Delivery of Learning Education and Training*). Les « Métadonnées pour les ressources d'apprentissage » sont incluses dans ce groupe de travail ; WG5 - Assurance qualité et architecture de support (*Quality Assurance and Descriptive Frameworks*) ; WG6 - Profils des normes internationales (*International*

Standardized Profiles : ISP) ; WG7 - Culture, langage, adaptabilités et accessibilités humaines (*Culture, Language, and Human Functioning Activities*)

Le JTC1-SC36 s'oriente vers l'exploitation de la norme ISO/CEI (Commission Electrotechnique Internationale) 24751 (ISO/IEC 24751-1 : 2008, 2008), en direction des autres normes issues du SC36. Cette norme a pour objectif « de répondre aux besoins des apprenants éprouvant une déficience et de toute personne en contexte de déficience ». C'est donc dans une démarche d'harmonisation de ses activités que le JCT1-SC36 est engagé : l'unification des normes et standards largement acceptés par les acteurs de l'éduction, de l'apprentissage et du marché de la formation (Arnaud, 2002).

1988 : AICC (*Aviation Industry* CBT (*Computer-Based Training*) *Committee*)/ Structuration des ressources pédagogiques à destination des personnels techniques
1994 : 2ème conférence du 3W / Première idée d'associer métadonnées sémantiques et ressource du web
1995 : Dublin Core / Cadre général des métadonnées pour la description des ressources électroniques
1996: - ARIADNE (*Alliance of Remote Instructional Authoring and Distribution Networks for Europe*)/Spécifications ET solutions pour la production de resources pédagogiques. - LOM (*Learning Object Metadata*) / Descriptions des ressources d'apprentissage. Plusieurs profils d'applications seront développés. Contient les éléments du Dublin Core
1997 : - LTMS (*Instructional Modeling System*) / Spécifications des métadonnées propres à l'enseignement - UIML (*User Interface Markup Language*) / Description des interfaces utilisateurs indépendamment de l'aspect graphique. - SCORM (*Sharable Content Object Reference Mode*) / Description des ressources pédagogiques en ligne (inclut une grande partie des spécifications déjà existantes) - CEN-ISSS (*Comité Européen de Normalisation-Information Society Standardization System*) est créée / Première prise ne compte du multilinguisme
1998 : XML est spécifié / S'appuie sur le langage normalisé SGML (*Standard Generalized Markup Language*)
1999 : Création du JTC1-SC36 (Sous-Comité 36 du *Joint Technical Committee* n°1 de *l'International Electrotechnical Commission* de l'ISO et de l'IEEE). IL est en charge de la normalisation pour les « Technologies pour l'éducation, la formation et l'apprentissage ». L'AFNOR (Association Française de Normalisation) prend part aux travaux de normalisation en cours dès 2000.
2001 : Propositions par le CEN-ISSS-LTW (*Learning Technologies Workshops*) de description de scénarios pédagogiques les EML (*Educational Modelling Languages*). Ce langage est la base d'une famille de normes ayant comme préfixe IMS.
2002 : L'AFNOR obtient le remplacement du numéro identifiant de l'apprenant (*Simple Human Identifier*), par le « *Participant Identifier* », afin d'assurer la protection et la sécurité des données personnelles. - Création au sein du JCT1-SC36 d'un groupe de travail sur les approches qualité des services.
2003 : Le MLR (*Metadata for Learning Ressources*) est proposé par le SC36 WG4/ Normalisation d'une interopérabilité de différents standards de métadonnées (recherche, acquisition, évaluation et utilisation).
2007 : Le MLO (*Metadata for Learning Opportunity*), fondée sur les travaux du CDM (*Courses Description Metadata*)-*Core elements* / Spécification des produits et services d'apprentissage pour l'éducation et la formation. Note : MLR et MLO sont présumées englober les normes de *l'e-learning*

Tableau 1. *Chronologie de la normalisation des TICE*

Le tableau 2 synthétise le panorama normatif en termes d'actions, de technologies et de champs d'application, principalement à partir de (Bourda, 2004), (Ben Henda, 2007), (Chartron *et al.*, 2004).

	Type	Structure	Champs d'application	Remarques
CANCORE 2002	Profil du standard IEEE-LOM et IMS	8 groupes, 61 éléments, tous optionnels	Créer et échanger des fichiers de métadonnées	Permet la recherche et la localisation des ressources pédagogiques.
DUBLIN CORE 1995	Norme ISO 15836:2003	15 éléments, tous optionnels	Métadonnées et interopérabilité	Reconnu par les acteurs de l'Internet. Nombre limité d'identificateurs car il n'a pas été spécifiquement conçu pour la formation en ligne
EML 1990				A la base de LTMS
IEEE/PAPI	Standard	8 spécifications sur l'information de l'apprenant : Informations personnelles, relation, sécurité, préférences, performances, portfolio.	Echange d'information avec l'apprenant	
IMS Version 1 2003	Spécifications		Description de la façon dont les objets pédagogiques doivent être conditionnés pour être échangés. C'est langage générique.	Pas de structuration des objets
LOM 2002	LTSC et IMS IEEE 1484.12.1-2002	9 groupes, 80 éléments, tous optionnels. Intègre les 16 champs du Dublin Core	Modèle international pour l'indexation des ressources pédagogiques.	Niveau de détails importants. Cela peut nuire à son implantation. Modèle orienté ou l'apprenant est face à sa machine.
LOM-fr 2005	AFNOR : NF Z76-040		Profil du standard IEEE-LOM et IMS Voir LOM	Communauté éducatives
MLO Relancé en 2007				Etablit la relation entre la description des cours et l'apprenant.
MLR	En cours de finalisation ISO : 19788		Mécanisme de conversion à partir du LOM.	Intègre des profils d'application Vient compléter le LOM et l'étend à l'usage du web. Compatible avec le Dublin Core
NORLOM 2005	Profil du LOM		Profil norvégien	
NORMETIC 2003	Profil du LOM	9 groupes, 62 éléments, 20 obligatoires, 12 recommandés, 30 facultatifs	Traitement des descripteurs du LOM sous 3 conditions : « Requis conditionnel », « Recommandé » ou « Facultatif »	Description des ressources d'enseignement et d'apprentissage
SCORM 1997	Profil de la spécification IMS	9 groupes, 61 éléments, 11, tous obligatoires	Création des objets pédagogiques structurés, agrégation des ressources, et suivi de l'activité de l'apprenant	Spécification technique en matière de conception de cours et de plateformes de e-learning. Exportation au format SCORM
SupLOMFR			Profil de la spécification du standard IEEE-LOM et IMS Voir LOM	Pour les institutions françaises de l'enseignement supérieur
UK LOM Core			Profil du Royaume Unis Voir LOM	
Vetadata			Profil australien du LOM	

Tableau 2. *Panorama des actions normatives dans le domaine des TICE*

4 TICE : normalisation et multimédia

4.1 Les normes du multimédia : la famille MPEG

Il apparaît évident à tout expert SC36 que de nombreux segments du SC36 sont aujourd'hui réinventés alors qu'ils sont déjà résolus ou en cours de développement normatif dans MPEG (Prêteux *et al.*, 2008). En s'appropriant et en adaptant des logiques normatives MPEG (notamment à travers MPEG-4 & 7), il devient possible de développer de façon beaucoup plus focalisée ce qui relève du métier de pédagogue proprement dit. Il est contreproductif de tenter de résoudre les problèmes de gestion et d'intégration multimédia au sein du SC36 alors qu'ils sont traités par les experts MPEG.

Si MPEG a de grandes chances de réussir cette fédération normative du multimédia, c'est bien parce qu'il renvoie à des leviers économiques qui ne peuvent être comparés qu'avec d'autres normes ou standards clefs comme l'ASCII (*American Standard Code for Information Interchange*), le TCP-IP (*Transmission Control Protocol-Internet Protocol*) ou HTML (*HyperText Markup Language*) et XML (*eXtensible Markup Language*).

Après MPEG-1 et MPEG-2, deux normes qui ont rendu possibles la vidéo sur DVD et, plus récemment, la télévision numérique, MPEG-4, MPEG-7 et MPEG-21 sont les nouvelles normes du SC29. Ces dernières s'articulent les unes par rapport aux autres, se prolongeant ou s'enrichissant de manière cohérentes de fonctionnalités supplémentaires (Figure 1).

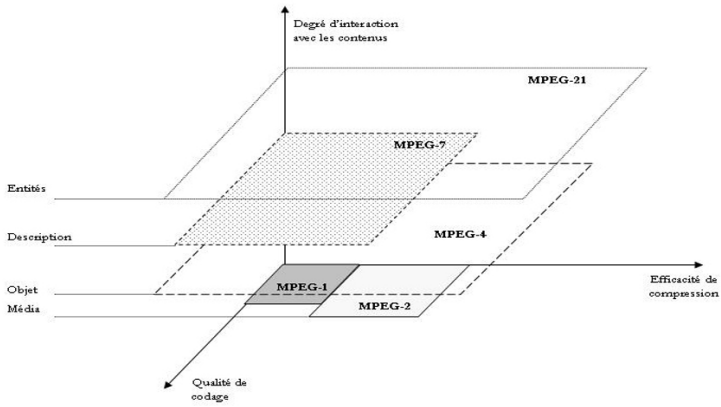


Figure 1. *Les différentes normes MPEG et leurs caractéristiques en termes de qualité de codage, d'efficacité de compression et d'interactivité*

4.2 MPEG-4

MPEG-4 (ISO/IEC 14496, 2000), norme depuis décembre 1999 pour sa version 1, traite des objets audiovisuels 2D/3D naturels et/ou synthétiques et décline des objectifs de codage sélectif et de composition de scènes. La norme offre donc un environnement d'outils génériques ainsi que des fonctionnalités nouvelles d'accès universel et d'interactivité.

Par le large éventail de fonctionnalités supportées, le standard ISO/IEC MPEG-4 révolutionne complètement le monde du multimédia numérique (Zaharia et Prêteux, 2007).

En effet, un flux MPEG-4 est un contenu vidéo enrichi de divers éléments d'information relatifs aux différents objets individuels considérés, comme durée de vie, régions support, emplacement dans une scène... Il vient tout naturellement à l'esprit la possibilité d'enrichir encore davantage cette représentation, en associant aux différents objets des descripteurs spécifiques débouchant sur des fonctionnalités nouvelles, comme par exemple l'accès automatique et les requêtes par le contenu. C'est l'objet de MPEG-7 (*Multimedia Content Description Interface* : MCDI).

4.3 MPEG-7

MPEG-7 (ISO/IEC 15938, 2002) spécifie une palette d'outils normalisés pour indexer et décrire syntaxiquement de façon automatique ou semi-automatique tout contenu multimédia. Une même information pourra donc être traitée en fonction des capacités communicationnelles recherchées, allant du spatio-temporel (audio et vidéo traités séparément) à une description sémantique du flux de données. MPEG-7 peut s'associer aux autres descripteurs spécifiant le format, les conditions d'accès, leurs classifications, les liens pertinents en relation avec l'information initiale, le contexte d'enregistrement ou de la diffusion du matériel : c'est la possibilité de naviguer, de chercher, de filtrer et de s'approprier l'information dans un corpus multimédia ouvert (Zaharia et Prêteux, *op. cit.*).

MPEG-7 a été développé pour s'harmoniser avec les autres normes utilisées dans les différents domaines d'application préconisés par le W3C. A ce titre, citons : XML, l'IETF (*Internet Engineering Task Force* qui propose les normes concernant Internet), la norme concernant les métadonnées du Dublin Core, celles concernant la terminologie et autres ressources linguistiques de l'ISO TC 37, les métadonnées garantissant les échanges entre les transactions (image, son, données alphanumériques), l'établissement de systèmes ouverts pour des applications de télévision interactive (TV Anytime), la norme ISO/IEC 11179 (ISO/IEC 11179, 2003) concernant les registres de métadonnées.

Cependant MPEG-7 n'inclut pas d'information particulière concernant l'utilisation d'objets multimédias dans le domaine de l'éducation. De ce fait, aujourd'hui MPEG-7 est dédié exclusivement aux descriptions de contenus multimédias et est complètement indépendant des canaux de transmission, des terminaux…

Toutefois, le monde des applications multimédias et des TICE en particulier ne peut ignorer la diversité des réseaux de communication et terminaux fixes ou mobiles disponibles aujourd'hui et doit proposer des services adaptés à chacun. Scalabilité, adaptation et convergence technologique deviennent les maîtres-mots du multimédia actuel. Comment assurer la diffusion des contenus et de leurs descriptions ainsi que les services proposées partout, tout en minimisant les coûts de production et en ré-utilisant au maximum les contenus existants ?

4.4 MPEG-21

MPEG-21 (ISO/IEC 21000. 2003), appelé *Multimedia Framework*, se propose notamment de lever ce verrou technologique en standardisant des descriptions non seulement des contenus, mais aussi de tous les éléments susceptibles d'intervenir

dans la chaîne de consommation, depuis la création, en passant par la diffusion et en allant jusqu'à l'utilisateur final.

Le concept central dans le contexte MPEG-21 est celui de DI - *Digital Item*, défini de façon générique et abstraite comme un produit numérique simple ou composite. Un exemple type est celui d'une page web, contenant différentes ressources multimédias comme du texte, des images, des vidéos, des éléments de mise en page (*e.g.* feuilles de style), des hyperliens, mais aussi des scripts de programmation qui conduisent à une apparence dynamique, en fonction de l'interaction de l'utilisateur. MPEG-21 fournit les mécanismes de description de tels produits numériques complexes. En particulier, les parties 2 – *Digital Item Declaration* et 3 – *Digital Item Identification* permettent respectivement la spécification complète et structurée des DI et leur identification/localisation.

Soulignons également la partie 7 du standard, dite *Digital Item Adaptation*, qui standardise des descripteurs et des schémas de description permettant l'adaptation des contenus vis-à-vis des utilisateurs, des réseaux, des terminaux ou encore de l'environnement d'utilisation.

Enfin, un important travail a été consacré aux aspects de propriété intellectuelle et de droits d'usage (partie 4 – *Intellectual Property Management and Protection Components*). Pour cela, MPEG-21 standardise un langage d'expression de droits (partie 5 – *Rights Expression Language - REL*) et un dictionnaire terminologique correspondant (partie 6 – *Rights Data Dictionary - RDD*).

REL décrit les droits et les permissions associés à un contenu multimédia. Ils protègent et garantissent les conditions d'utilisations. Ils fournissent les descripteurs d'accès et d'obtention des droits. C'est donc à travers l'expression de l'autorisation d'accès que l'interopérabilité est assurée. REL généralise l'échange de contenus et est garant de la bonne utilisation et de la protection de ceux-ci. Il définit les protocoles et autorise les droits d'accès aux contenus.

L'*Intellectual Property Management and Protection* (IPMP) décrit le système de gestion de droits associés aux objets multimédias. Il existait déjà un IPMP MPEG-4 mais celui-ci ne définissait pas les conditions de vérification des lieux de contrôle des droits. Cela permet donc de pouvoir accéder et interagir avec les outils IPMP, d'échanger des données entre les outils (décryptage…), et d'authentifier les outils.

Cette couche est nécessaire pour l'utilisation du REL et des RDD.

Les RDD se définissent comme : « Un ensemble de mots clairs, cohérents, structurés, intégrés, et identifiés de manière unique pour permettre la génération d'expressions de droits ». Ils permettent donc de décrire d'un point de vue sémantique les mots décrivant les droits. Ils facilitent les passerelles d'une terminologie à une autre dans le domaine des droits.

L'ensemble des travaux MPEG-21 s'inscrit donc en parfaite continuité avec ceux réalisés précédemment dans MPEG-7. Les descripteurs et schémas de description correspondants sont développés sous la responsabilité du même groupe MDS (*Multimedia Description Schemes*) et à l'aide du même langage de description fondé sur XLM. De façon synthétique, les interrelations possibles entre MPEG-21 et les normes (Lyon *et al.*, 2006) concernant les technologies d'apprentissage sont schématisées Figure 2.

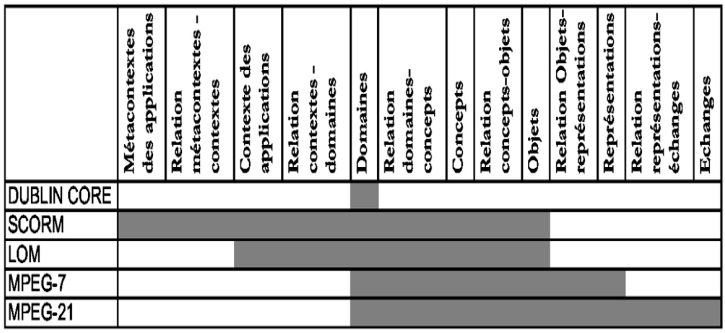


Figure 2. *Champs couverts par les principales normes des TICE*

5 Vers une convergence interdisciplinaire

Un aspect particulièrement intéressant dans l'essor des TICE est la façon dont les contenus pédagogiques se développent d'un point de vue technologique. Ce déploiement nécessite une instrumentation de tous les composants pédagogiques (image, son, texte, hyperlien…), et doit s'inscrire dans un cadre normatif afin de garantir interopérabilité et réutilisation par le plus grand nombre. Toutefois, il convient d'inclure dans les objets pédagogiques, toutes les ressources disponibles qui peuvent aider à la construction de la connaissance, comme les bibliothèques virtuelles, les sites Internet, les images fixes et animées…

Or, cette logique référentielle du corpus de documents numériques est l'un des postulats de base pour les concepteurs des systèmes d'information pour donner naissance à une normalisation de l'ingénierie linguistique (XML, MPEG-7, MPEG-21). Tous ces principes sont développés dans toutes leurs dimensions par le SC36.

5.1 SC36 : une orientation pluridisciplinaire

L'orientation des travaux du JTC1-SC36 (Burnett, *et al.*, 2003) repose essentiellement sur la portabilité, l'interopérabilité et l'adaptabilité culturelle des « technologies pour l'éducation, la formation et l'apprentissage ». Le SC36 n'a donc pas vocation à dupliquer les travaux effectués par d'autres comités techniques (le SC29 par exemple : codage du son, de l'image, de l'information multimédia et hypermédia).

Au sein du SC36, l'ADL qui a développé SCORM (et qui intervient comme Liaison A au SC36) veut agir en capitalisant sur les normes d'autres SC. En particulier, l'ADL cherche actuellement à faire adopter MPEG-21 partie 5 pour résoudre les questions de *copyright* et confie aux normes LOM (ou d'autres formats de métadonnées pédagogiques comme Dublin Core ou le futur MLR…) le soin de décrire les ressources pédagogiques dans leurs différentes facettes. Hors à l'évidence, cette description deviendrait concurrente de celles issues des normes multimédias de la famille MPEG.

Dans un premier temps, le travail passera inévitablement par une phase de spécification (*requirements*) propres à la pédagogie. Il serait alors souhaitable d'y associer pédagogues et experts de MPEG, voire de JPEG. L'ADL réussira-t-elle ce qui pourrait préfigurer une seconde phase d'intégration de MPEG, *i.e.* convaincra-t-

elle le groupe des experts du SC36 de repenser de façon unificatrice la description normative du document multimédia en tant que ressource pédagogique ?

Du côté de MPEG, deux scénarios contrastés sont envisageables. Le premier met en avant l'intérêt de MPEG pour devenir proactif non plus seulement dans le domaine des usages *broadcast*, mais aussi dans celui des usages convergents du monde des TICE. Dans le second scénario, la communauté MPEG attend d'avoir amorti ses usages *broadcast* pour attaquer un « second marché » vers les usages convergents. Bien sûr, ces scénarios extrêmes sont asymptotiques et la réalité sera plus hybridée et nuancée.

5.2 MPEG-21 : un cadre modulaire

Au-delà du strict recours par le SC36 au niveau MPEG-21 de la partie 5 (Bormans et Hill, 2002), la stratégie de certains experts tendrait à conserver au SC36 des fonctionnalités en cours de normalisation alors même qu'elles pourraient être empruntées au niveau des normes MPEG-21, 7 ou 4, puis adaptés aux nouveaux besoins de la transmission du savoir.

La gestion sera-t-elle réalisée avec des composants MPEG ou à partir des composants traditionnels des standards associés au SC36, voire des normes SC36 en cours de développement et qui prendront en compte ces fonctionnalités ?

La question reste cruciale et, en bonne logique ISO/IEC, devrait se traiter dans un consensus avec au minimum une liaison SC36-SC29 et un groupe de travail conjoint. Or, actuellement la liaison SC36-SC29 n'existe pas ! Voilà qui mérite, réflexion, voire action !

Côté SC29/WG11, l'intérêt est au déploiement le plus large possible des technologies normatives générales existantes. Il faudrait donc que les experts SC29/WG11 aient la conscience des enjeux du marché des TICE, et notamment de ceux à forte valeur ajoutée. L'objectif est d'optimiser ce qui s'effectue aujourd'hui par l'activité humaine et non *machinique*.

MPEG-21 est devenu le cadre modulaire de développement (*framework*), candidat normatif à l'intégration globale de tous les documents multimédias. Cette prétention peut paraître exorbitante, mais se justifie sous la forme d'un syllogisme très simplificateur : le multimédia n'est pas né de rien. C'est une conséquence directe de la normalisation des pratiques numériques : les concepteurs de téléphonie, d'audiovisuel, d'informatique textuelle se sont mis d'accord dans des instances de normalisation pour être interopérables et compatibles. Cela a conduit au multimédia numérique mondialement multilingue que nous connaissons. Le multimédia et la mondialisation numérique en réseaux se sont ainsi généralisés comme un effet direct de l'effort normatif. Dès lors, il est primordial de poursuivre cet effort même et surtout si cette normalisation achoppe de plus en plus sur des domaines de moins en moins triviaux et matériels, comme ceux d'applications sociales, de traitement du savoir, de sémantique, de disparité culturelle, linguistique ou disciplinaire.

Faute d'un cadre global d'intégration du « *commerce* entre les hommes » que nous pouvons synthétiser sous le terme général d'*e-procurement*, nous risquons une babélisation numérique. Cette globalisation se doit d'intégrer tous les aspects de diversité de l'information (mode de médiation, typologie d'usage, typologie d'accès, d'acheminement et d'échange, gestion référentielle, structurelle et sémantique …).

Cette normalisation a été bien sûr aux fondements mêmes de l'informatique et le code ASCII (ISO/IEC 646 : 1991, 1991), redéployé sur 4 octets dans la norme omni-écritures du monde ISO/IEC 10646 (Unicode) (ISO/IEC 10646 : 2003, 2003) en est un exemple emblématique. Ce qui frappe à l'évidence dans cette norme

fondamentale de l'informatique, c'est qu'en 30 ou 40 ans, les industriels et les usagers de l'informatique ont complètement assimilé la nécessité absolue d'intégrer la totalité des contraintes culturelles liées à l'écriture en tous lieux et à toutes les époques.

La voie est donc tracée. Les *New Work Item* (NWI) qui correspondent à tout nouveau champ proposé à normalisation après enquête des instances concernées s'attachent de plus en plus aux objets réseaux et composants. Ils prennent de plus en plus en compte les langages (tant humains qu'artificiels), les interactions sociales, la transmission ou la gestion du savoir… La normalisation des TIC se situe de plus en plus comme une pratique consistant à offrir des cadres de production comme dans TMF (*Terminological Markup Framework*) et MPEG-21.

Il est souhaitable que les experts du SC36 s'approprient et enrichissent les spécificités du monde des TICE, dans le cadre général et standardisé de la chaîne de production et de distribution de tout contenu numérique offert par MPEG-21. Pour faciliter ce travail collaboratif, nous recommandons de créer une liaison active entre le SC36 et le SC29-MPEG au bénéfice des métiers de l'éducation.

6 Conclusion

Aujourd'hui, sujet de débat entre chercheurs de cultures variées (sociologie, pédagogie, sciences exactes), les TICE sont largement acceptées et déployées au sein de toute structure à vocation éducative, formelle ou non : universités, écoles primaires, musées comme le Louvre, compagnies industrielles, associations pour l'intégration sociale des handicapés...

Ce contexte déjà favorable laisse entrevoir un futur technologique optimiste des TICE et du multimédia dont l'impact dépendra de la force de création et du niveau d'engagement de quelques communautés représentatives et moteur. En outre, susciter et coordonner efficacement l'action d'experts chercheurs, d'industriels et d'utilisateurs, tant au SC36 que dans MPEG, contribuera à promouvoir des solutions normatives à moyen et long terme.

Dans ce contexte, pour toute application multimédia, MPEG-21 constitue un cadre privilégié, naturellement accepté par l'utilisateur final, intensivement déployé par les industriels et essentiellement nourri par les académiques.

Références :

Arnaud, M. (2002). Normes et standards de l'enseignement à distance : enjeux et perspectives. In *Actes du colloque TICE 2002*, TICE 2002, Lyon, Novembre,

Ben Henda, M. (2007). SCORM specifications for an emerging world: The linguistic diversity at work. In *Open Forum Conference: Global Leadership & Governance of ICT standards for learning, education & training*. London, United Kingdom.

Blandin, B. (2003). *Les enjeux des normes sur les technologies de l'information pour l'éducation, la formation et l'apprentissage*. Colloque Synergie, Université de technologie de Troyes.

Bormans, J., Hill, K. (2002). MPEG-21 Overview v.5. ISO/IEC JTC1/SC29/WG11/N5231.

Bourda, Y. (2004). *Les évolutions du LOM. Compte rendu rédigé par l'ENSSIB.*

Burnett, I., Van de Walle, R., Hill, K., Bormans J., Pereira, F. (2003). *MPEG-21: goals and achievements*. IEEE Multimedia, vol. 10, Issue 4, 60 – 70.

Chartron, G., Gauthier G., Grandbastien, M., *et al.* (2004). Normes et standards. (2004). *Distances et savoirs*, vol. 2, num. 4.

Hudrisier, H. (2006). Société de la connaissance, le paradigme de l'appropriation. In *Hermès*, num. 45, 163 - 164.

ISO/IEC 646 : 1991. (1991).*Technologies de l'information - Jeu ISO de caractères codés à 7 éléments pour l'échange d'information.*

ISO/IEC 10646 : 2003. (2003). *Technologies de l'information - Jeu universel de caractères codés sur plusieurs octets (JUC).*

ISO/IEC 11179. (2003). *Information technology - Metadata registries.*

ISO/IEC 14496. (2000). *Information technology - Coding of audio-visual objects.*

ISO/IEC 15938. (2002). *Information technology - Multimedia content description Interface.*

ISO/IEC 21000. (2003). *Information technology - Multimedia framework (MPEG-21).*

ISO/IEC 24751-1 : 2008. (2008). *Technologies de l'information - Adaptabilité et accessibilité individualisées en e-apprentissage, en éducation et en formation -Partie 1: Cadre et modèle de référence.*

Lyon, L., Patel, M., Christodoulakis, S., *et al.* (2006). *Project num. 507618*.

Prêteux, F., Vaucelle, A., M. Ben Henda M., *et al.* (2008). *Normes MPEG : une base pour le e-procurement des TICE*. Rapport de recherche, N. 08009-ARTEMIS, TELECOM & Management SudParis, Evry, août.

Proulx, S. (2001). Usages de l'Internet : la « pensée-réseaux" et l'appropriation d'une culture numérique. In *Comprendre les usages de l'Internet,* Guichard, E. (éd.), Editions Rue d'Ulm, Presses de l'Ecole Normale Supérieure, Paris, 139-145.

Rifkin, J. (2000). *L'âge de l'accès – La révolution de la nouvelle économie.* Editions La Découverte & Syros, Paris.

Zaharia, T., Prêteux F. (2007). Normes de description des contenus multimédias. In *L'indexation multimédia - description et recherche automatique*, *Traité IC2 - Série Traitement du Signal et de l'Image*, Gros, P. (Ed.), Editions Hermès-Lavoisier, Paris, 163-185.

Conférences invitées

La communication scientifique et ses enjeux politiques : un regard transatlantique

Gérard BOISMENU(1)

(1)Professeur titulaire, Doyen de la Faculté des arts et des sciences, Université de Montréal, Président du Consortium Érudit

Que l'on me permette quelques remarques préalables pour situer ma réflexion sur les enjeux politiques de la communication scientifique. Dans une perspective historique, le rôle du politique dans la transformation des modes de communication scientifique ne peut passer inaperçu. Avec ou sans ce recul, la question se pose de plus en plus, et ce n'est pas l'effet du hasard.

Même si cela pêche par un égocentrisme historique, on peut voir là l'effet de l'« accélération de l'histoire », dopée par les changements prodigieux associés à l'électronique et au numérique. Le chamboulement des institutions et de la configuration des acteurs mène à redéfinir l'espace public et la place de l'autorité publique. Pourtant, un certain discours serait enclin à « naturaliser » ce cours fatal de l'histoire par l'emprunt d'un vocabulaire et de considérations essentiellement économiques. Si la communication scientifique peut être abordée sous l'angle du marché, elle ne peut l'être uniquement par le marché. Cette arène dissout en apparence les conflits, les intérêts, les enjeux, et ne laisse poindre que des acteurs anonymes mus par des lois souterraines et historiques. En prenant le contre-pied de cette vision, j'entends brosser brièvement un état des lieux, pour en arriver à examiner le rôle des pouvoirs publics, nationaux et supranationaux.

D'ailleurs, on observe que cette thématique est largement traitée, en Europe comme en Amérique du Nord, mais évidemment ailleurs également, dans les termes de l'accès à l'information scientifique. C'est le statut de la connaissance et son mode d'appropriation qui sont en cause. Les organismes de l'Union européenne mènent une réflexion intense et diversifiée sur cette question. Cela est d'autant plus intéressant qu'ils ne répugnent pas à appeler de leurs vœux une responsabilisation des pouvoirs publics, fut-ce à un niveau supranational.

Il ne s'agit pas là d'humeurs d'une bureaucratie en mal d'autojustification. Que ce soit en termes de rôle de la science pour la croissance économique, de financement public de l'activité scientifique ou encore de financement public du paiement de l'accès à la publication scientifique, l'accès à la science se révèle par sa haute pertinence politique (Dewatrinpont *et al.*, 2006). Par ailleurs, plusieurs acteurs œuvrent sur ce terrain. On pense aux éditeurs et plateformes numériques, aux grandes entreprises commerciales, aux bibliothécaires, aux institutions de financement de la recherche, aux chercheurs, ainsi qu'aux pouvoirs publics.

Pour tenter d'y voir clair, je procéderai à quatre coupes qui définissent autant de grandes thématiques. D'abord, je vais considérer le marché des revues ; ensuite, je discuterai l'enjeu de l'accès, pour continuer en traitant de la notion de cyberinfrastructure. Enfin, j'essaierai de montrer comment les pouvoirs publics sont interpellés.

1 Le marché des revues

La revue est le véhicule majeur de la diffusion de la connaissance dans de très nombreuses disciplines. Cette affirmation, qui n'invalide pas le rôle du livre et sa prépondérance dans plusieurs disciplines en sciences humaines, souligne seulement l'intérêt de rendre compte de la situation concernant l'accès à la connaissance et la structure socio-économique de la communication scientifique. Les revues jouent un rôle essentiel dans la communauté scientifique. Elles sont des institutions qui participent à la structuration de la communauté scientifique (nationale et internationale), elles assurent la validation, la légitimité, la reconnaissance, la diffusion et la conservation du patrimoine scientifique. Finalement, ce sont des vecteurs essentiels de la circulation de la connaissance de par le monde.

1.1 Le tracé des courbes

Le relevé de la tendance pour la croissance des prix pour les revues et les livres permet de lever le voile sur cette situation. Le graphique 1 (voir annexe) retrace l'évolution des prix par genre, telle qu'elle a pu être enregistrée par les bibliothèques universitaires aux États-Unis. C'est le point de départ pour l'appréciation du problème, car ce qui frappe c'est l'inflation vertigineuse pour le prix des revues savantes et ses effets. Parmi ces derniers, soulignons la position dominante des revues dans le budget des bibliothèques, qui a pour envers la détérioration de la position des ouvrages, malgré une stabilité des prix relatifs de ces derniers. Cela traduit les distorsions introduites par un marché « imparfait » et oligopolistique. Cette image globale doit conduire à une compréhension plus fine de la situation. Car se limiter à cette première appréciation conduit à une image simplifiée de la réalité et à des conclusions hâtives et inappropriées.

On pourrait conclure que la partie est jouée, au sens où les revues seraient pour l'essentiel sous l'emprise des grands groupes commerciaux. De ce fait, la bataille pour le maintien dans le secteur public de ces vecteurs de la diffusion de la connaissance serait perdue. Nous n'aurions plus qu'à s'en faire une raison ou à tenter de contourner ces « monstres » économiques en tablant sur des formes alternatives à la revue. La simplicité du diagnostique a pour corollaire la simplicité des voies de solution.

La réalité est rebelle et se conforme mal à cette image grossière. Les revues, qui échappent aux grands groupes commerciaux d'édition, occupent une place centrale. Et il devient assez évident que l'action publique, de même que celle des acteurs peut infléchir le cours de l'évolution des choses. Il importe de caractériser un peu mieux ce « milieu ».

1.2 Le marché imparfait des revues

Les revues, par leur fonctionnement et le rapport avec la communauté scientifique, tout autant les auteurs que les utilisateurs, présentent les caractéristiques d'un marché imparfait.

Il faut convenir qu'il repose sur des effets de réseau, ce qui apparaît lorsqu'on suit les modes de relations avec les scientifiques (Dewatrinpont *et al.*, 2006). Les auteurs publient le plus possible dans des revues reconnues dans leur secteur de spécialisation ou dans leur discipline. Puisqu'il faut sérier et sélectionner, les lecteurs privilégient les revues réputées pour leur « haute qualité » ou leur grande pertinence. De leur côté, les bibliothécaires s'abonnent de préférence aux revues lues, soit celles qui sont réputées répondre aux besoins des chercheurs. Quant à eux, les auteurs

citent les revues qu'ils ont lues. De ce fait, les articles qui n'apparaissent pas dans les index ou dans les revues de référence sont ignorés par les lecteurs. Dans la mesure où les auteurs et les lecteurs sont susceptibles de bouder les revues qui n'ont pas la cote, on peut affirmer que, dans ce « marché », il n'y a pas de valeur de substitution.

Une revue dans le même domaine n'est pas substituable à une autre parce qu'elle est moins chère ou plus attrayante, ou pour toute autre raison, si ce n'est qu'elle apparaisse majeure dans son domaine et qu'elle supplante la première par sa valeur intrinsèque ou sa valeur symbolique. Cet ensemble de pratiques constitue la trame de ce marché imparfait. Cela n'est pas fonction de la structure socio-économique qui prévaut dans le secteur, mais, inversement, cette structure socio-économique capitalise sur la caractéristique première du marché imparfait des revues.

1.3 Un marché oligopolistique

Depuis des décennies, mais avec une accélération phénoménale au cours des vingt dernières années, nous assistons à un processus de concentration du contrôle des revues dans un nombre très limité d'éditeurs commerciaux (Stanley, 2002). Les fusions « dopent » le mouvement, si bien quelques grands éditeurs occupent une place dominante dans l'édition et la commercialisation des revues savantes. On observe que chaque fusion est suivie d'une croissance prononcée des prix.

Quelques chiffres permettent de juger de l'ampleur du phénomène (Edlin et Rubinfield, 2004; McCabe, 2002; Bergstrom et Bergstrom, 2003). Dire que ce marché est oligopolistique ne relève pas de l'épithète ni de l'abus de langage. Les chiffres changent en quasi-permanence, mais il y a peu on pouvait estimer qu'Elsevier proposait un bouquet de 1800 titres, Blackwell, Wiley, Kluwer, 1350. Pour le secteur Sciences, techniques et médecine, Elsevier détenait 22,9 % des titres (de revues), Kluwer, 11,7 % et Thompson, 10,7 %. Par ailleurs, toujours dans ce secteur, les revenus encaissés sont allés à 70 % aux éditeurs commerciaux, 18 % aux éditeurs sans but lucratif et 12 % aux « agrégateurs ». On estimait aussi qu'Elsevier arrivait le 3e au monde pour les revenus Internet après OAL-Time Warner et Amazon. Voilà autant d'indices d'une très forte concentration et d'une position commerciale tout à fait « avantageuse », pour employer une litote.

Cette position oligopolistique permet aux grands éditeurs commerciaux de revues savantes de toucher une rente de situation. La forme privilégiée est celle de l'offre groupée et liée des titres d'un éditeur aux acheteurs institutionnels (bibliothèques, par exemple). La position dominante sur le marché s'est exprimée par une offre, connue sous le nom de *Big Deal*. Celle-ci a provoqué la constitution de consortiums de la part des acheteurs institutionnels. Cette dynamique est importante à étudier, mais auparavant revenons aux acteurs.

1.4 Les éditeurs commerciaux ne sont pas seuls.

Une récente étude (Dewatrinpont *et al.*, 2006) a montré que les revues publiées par les grands commerciaux sont vendues au moins trois fois le prix des revues éditées par des sociétés sans but lucratif (presses universitaires, sociétés savantes, ou autres institutions de mission publique). Et là, on compare la même chose : des revues dans la même discipline, avec le même indice de citation et avec la même durée de vie de la revue. Toutes ces variables étant contrôlées, l'aspect déterminant de cet écart semble devoir être la nature de l'éditeur.

On sait que de 1986 à 2002, la croissance des prix pour les périodiques achetés par les bibliothèques universitaires aux États-Unis a été de 226 %, alors qu'Elsevier augmentait ses prix de 642 % (Edlin et Rubinfield, 2004). Autre facette de cette même réalité : l'Université Cornell paie, en 2003, 1,5 million de dollars à Elsevier

pour son bouquet de revues. Ce dernier représente 2 % du nombre total de titres, mais 20 % du budget total alloué par la bibliothèque aux périodiques. Est-ce à dire que ces grands commerciaux occupent toute la place. Leur domination sur le marché et la commercialisation signifie-t-elle que ces éditeurs règnent sans partage sur la communication scientifique en tant que telle. ?

Nous avons récemment fait une étude pour distinguer ce « paysage des revues importantes » dans les disciplines. Cette étude retient les 25 revues les mieux classées au plan international (indices de citations ISI) dans dix disciplines (anthropologie, affaires et gestion, droit, sciences économiques, sciences de l'éducation, science politique et relations internationales, psychologie, travail social, sociologie, histoire et philosophie des sciences) en sciences sociales en 2003. Pour chaque revue de cet échantillon de 250 revues, les données suivantes ont été consignées : l'indice d'impact, l'éditeur et son statut et le prix de l'abonnement institutionnel (version imprimée). Les résultats donnent de bonnes indications sur l'état des lieux.

	Poids	Impact	Prix	Prix	Pointage moyen
Presses universitaires	30,5 %	1,59	149 $	122 $	13,4
Sociétés savantes	17,5 %	1,92	188 $	180 $	16,8
Éditeurs commerciaux	52 %	1,03	614 $	503 $	11,6

Tableau 1 : *Revues en sciences sociales, dix disciplines, 2003*[1]

Il en ressort que les revues sans but lucratif occupent un peu moins de la moitié de notre échantillon des revues les plus citées dans les différentes disciplines, qu'elles ont un indice d'impact nettement plus avantageux que pour celles éditées par les grands commerciaux, que leurs prix sont carrément moins élevés (3,5 fois moins chère) et qu'elles se distribuent avantageusement dans les revues de l'échantillon pour ce qui est du rayonnement. En d'autres termes, elles occupent la position la plus enviable, même si cela va à l'encontre de l'image reçue concernant la domination des revues des grands commerciaux comme vecteur de communication scientifique ; de ce fait, la domination est d'abord et surtout commerciale.

Il ne fait pas de doute que les revues savantes éditées par des organismes sans but lucratif représentent un meilleur investissement pour les bibliothèques et pour les universités. Même si un comportement « rationnel » devrait conduire à accorder une priorité à ces revues, il s'avère que « les premiers sont les derniers », car intervient toute une série de pratiques et de filtres qui nuisent au positionnement stratégique de ces revues dans la sphère marchande, où le combat est à armes inégales. En raison de leur mission et de leur morcellement (toute comparaison faite), ces revues ne prélèvent pas de rente de situation, ne sont pas en mesure de proposer une offre globale à fort volume et n'utilisent pas les stratégies marketing des grands éditeurs commerciaux.

Cela étant, le « marché des revues » se caractérise par l'existence de deux segments et de deux types de pratiques. D'un côté, avec les grands éditeurs commerciaux, nous avons des oligopoles qui occupent une place majeure dans la diffusion de la connaissance et dominante au plan commercial, ce qui leur permet

[1] Les données ont été compilées par nos soins à partir de Institute for Scientific Information, *Social Sciences Citation Index. Journal of Citation Reports*, 2003.

de pratiquer des prix fort élevés, de toucher des marges bénéficiaires extraordinairement avantageuses, de se doter des moyens d'un vaste rayonnement et de continuer à canaliser le plus possible de titres de qualité. De l'autre, les éditeurs sans but lucratif (ou apparentés que l'on identifie comme « éditeurs responsables ») occupent une place très avantageuse comme vecteur de communication scientifique, dominent dans les structures « nationales » de la communication scientifique, ont des pratiques qui les associent davantage à une politique de recouvrement de coût et ont du mal à adopter une démarche de groupe ou concertée, en raison de leur morcellement relatif.

1.5 Le big business et la communication scientifique

Les oligopoles tirent évidemment partie de leur puissance commerciale pour ériger des barrières stratégiques à l'entrée sur le marché des revues (Bergstrom et Bergstrom, 2003; Frazier, 2001; Hahn, 2006; Edlin et Rubinfield, 2004). Les *Big Deals* constituent des instruments privilégiés dans l'atteinte de cet objectif. Cette offre consiste pour les grands commerciaux à vendre l'accès à un panier fermé de revues (papier et numérique) à un prix d'ensemble souvent exorbitant et pour une durée significative, avec pénalité pour retrait. Les modalités varient, mais le principe est fondamentalement le même. L'idée est d'empêcher la liberté de choix du bibliothécaire et de l'engager dans une relation exigeante, contraignante et limitative. Cette pratique crée une fidélité mécanique, oblitère l'intermédiaire professionnel et change les règles du jeu ; au final, cette stratégie commerciale permet aux grands éditeurs de se ménager une immunité par rapport à la concurrence (Davis, 2003). Les *Big Deals* représentent une technique de vente au service des grands commerciaux qui a pour effet d'accroître la dépendance à leur égard et de laisser une portion congrue aux éditeurs sans but lucratif.

Une des réactions est venue du milieu des bibliothèques. En s'organisant en consortium, les bibliothèques ont voulu créer un rapport de force ; la création d'un interlocuteur unique au nom du plus grand nombre permet sans doute de changer la dynamique de négociation et de compter sur la possibilité de négocier de meilleures conditions de prix, d'utilisation, de service, etc. Quelques avantages peuvent être mis au crédit de cette option. Avec les consortiums, on peut espérer des ententes contractuelles permettant d'élargir et de diversifier l'accès à la documentation, d'établir un meilleur contrôle des coûts dans le cadre d'ententes pluriannuelles, de négocier des conditions acceptables pour les bibliothèques, en termes, par exemple, d'accès permanent, de prêt entre bibliothèques, etc.

La véritable question reste cependant la même sur le fond : s'agit-il d'une réelle contre-attaque ou d'une alliance objective. Les consortiums restent relativement modestes face à la concentration des éditeurs ; par exemple, le plus gros consortium représente de 2 à 3 % de l'achat, alors que le plus gros éditeur détient 20 % des ventes de revues. Mais, plus encore, les ententes pluriannuelles et les conséquences d'un éventuel non renouvellement ne font que poursuivre et accentuer la dépendance des bibliothèques envers les éditeurs commerciaux. En dépit de la volonté de changer le rapport de force, il s'avère que le coût des abonnements demeure globalement fort élevé. Cela est sans parler du fait que ce genre d'entente ne vise que marginalement les revues sans but lucratif ; elles ne seront touchées que si elles sont associées à un portail avec une collection significative, et encore.

1.6 Pactiser et faire pour le mieux

Cette démarche n'est pas sans alimenter des doutes sur son intérêt. Elle fait l'objet d'examens critiques (Dewatrinpont *et al.*, 2006). Les consortiums qui contractualisent les *Big Deals* sont réputés, non sans raison, créer de effets de

verrouillage, dans la mesure où cela perpétue la dépendance envers les commerciaux, exclut la concurrence et rigidifie les budgets des bibliothèques. La concurrence n'est plus entre les titres, mais entre de grandes collections de titres, et cela joue structurellement en défaveur des plus petites collections et des nouveaux arrivants. Ces ententes produisent un effet systémique discriminant à l'égard des revues des éditeurs sans but lucratif, des plus grandes aux plus modestes.

Qui plus est, il est possible d'estimer que, malgré les campagnes tapageuses qui ont accompagné leur constitution, les consortiums ont finalement eu relativement peu d'effets (en fonction de leur cible principale : contrer le marché oligopolistique). Les paramètres pour contrôler les prix permettent de programmer leur croissance et d'ériger des barrières pour la concurrence. Une intervention à ce seul niveau est nécessairement fragile, car on ne peut qu'occuper la position du consommateur dont l'éventail des choix est restreint. Il ne s'agit donc pas d'une opposition frontale mais d'une tentative de contournement. Avec ces expériences, reste posée la nécessité d'une action collective et publique pour enrayer la domination privée du marché.

Ces pratiques, en suivant leur cours, sans contrecarrer le diktat des grands commerciaux, peuvent en venir, de fait, à asphyxier les revues des éditeurs sans but lucratif. Or, ces revues — les plus importantes au plan international ou celles qui sont enracinées dans les structures nationales de la communication scientifique — sont des institutions qui participent pleinement à la structuration de la communauté scientifique (internationale et nationale), et elles contribuent à la constitution du patrimoine scientifique et à la diffusion des connaissances en plusieurs langues. Il ne faut pas s'y tromper, même avec un ancrage national, les revues sont des canaux ouverts sur le monde et le numérique vient donner corps à cette vocation première. Cela explique certainement que les pouvoirs publics aient été amenés à subventionner la mise en place de larges plateformes de revues « nationales » pour faciliter leur diffusion internationale.

L'enjeu majeur qui plane sur la communication scientifique, c'est celui de l'accès à la connaissance, que celle-ci soit diffusée par les revues ou grâce à d'autres véhicules.

2 La problématique de l'accès

L'une des thématiques majeures qui traversent la communication scientifique de par le monde, c'est celle d'avoir accès aux fichiers numériques sans frais et dans des délais courts pour les documents scientifiques. Plusieurs lettres, déclaration et appels au plan international ont été diffusés ces dernières années. Rappelons pour mémoire Budapest [2002], Berlin [2003], SMSI-ONU [2003], OCDE [2004].

Cet objectif suit deux directions : d'une part, les revues et plateformes en libre accès et, d'autre part, les dépôts (*archives*) en libre accès.

2.1 Plateformes ou revues en accès libre

Plusieurs plateformes en accès libre ont vu le jour ces dernières années. Il est inutile d'en faire l'inventaire. Mentionnons, à titre d'exemple, la plateforme *Persée*, qui diffuse la numérisation rétrospective des revues en sciences humaines et sociales depuis leur premier numéro jusqu'aux années récentes. Contrairement à *JSTOR*, alors que leur mandat est comparable, l'accès est libre et non soumis à des tarifs ou abonnements. De même, la plateforme *SciELO*, qui a été mise en place au Brésil avec financement public, met en réseau des plateformes nationales (portugais et espagnol) de revues dans plusieurs disciplines, dont des sciences et médecine.

Ces initiatives s'inscrivent dans un mouvement international et font l'objet d'une certaine attention (Kaufman-Wills Group, 2005; Tenopir, 2006; McCabe et Snyder, 2005; Getz, 2004). Très souvent, les revues en accès libre ont une assise institutionnelle, en tant qu'éditeur, qui peut être un département universitaire, un laboratoire de recherche, une association savante, etc. Ces revues, pour plusieurs, n'existent qu'en version numérique, mais ce n'est pas une règle. L'accès libre n'étant pas synonyme de gratuité impose un mode de financement par d'autres moyens que l'abonnement. Plusieurs cas de figure sont possibles et, le plus souvent, il y a hybridation. Se croisent subventions, volontariat, aide institutionnelle, en nature ou en service, etc. Une étude récente tend à montrer qu'en tout état de cause, on connaît une précarité financière. Il est possible de résumer en disant que le modèle économique est variable et non stabilisé (Kaufman-Wills Group, 2005; Tenopir, 2006).

L'accès libre pose un défi majeur qui peut se résumer en termes de viabilité économique des revues et éditeurs. Pour y faire face, la capacité de générer des revenus autonomes est à l'ordre du jour. Or, ce genre de modèle économique est très exigeant, incertain et produit des effets pervers, car la mise en vigueur de certaines solutions reste problématique, du moins si on pose la question en termes collectifs. Certaines des solutions (produits dérivés, valeur ajoutée sur certains services, etc.) supposent que l'on intériorise une logique marchande forte ou que l'on accroisse la dépendance aux commanditaires. Si les voies de solution restent à stabiliser au cours des prochaines années, on peut remarquer que l'idée d'un accès libre pour les publications au-delà d'une période de x mois, voire 24 mois en sciences humaines et sociales, semble se répandre. Et de toute façon, les solutions simples ne sont pas encore à la portée lorsqu'on songe au fait que plusieurs projets, cités comme une alternative à la publication commerciale oligopolistique, ne sont pas totalement en libre accès. Pensons à certains projets de SPARC, à *BioOne*, à *HighWire Press* ou encore à *Muse*.

2.2 Le dépôt de documents numérisés en accès libre

Au cours des dernières années, de nombreux dépôts institutionnels, interinstitutionnels, disciplinaires, nationaux (pré et post publications, revues, thèses, etc.) ont été créés. Le leitmotiv, c'est celui de l'accès le plus large possible et avec le moins d'obstacles aux résultats de la recherche. Il faut reconnaître l'engagement militant très soutenu de la communauté internationale des bibliothécaires, ainsi que des agences publiques dédiées à la production et à la diffusion de la connaissance scientifique.[2]

Au-delà de l'accès, d'autres motivations alimentent les discussions qui accompagnent ce foisonnement ((Dewatrinpont *et al.*, 2006). Les dépôts peuvent être également vus comme un moyen pour assurer la préservation institutionnelle de la production des membres d'une institution, ainsi que comme une vitrine témoignant de façon tangible de la qualité d'une institution de recherche. Concurremment, c'est aussi une façon de changer la structure de la communication scientifique en exerçant une pression en faveur de son contrôle par milieu universitaire et de la déconcentration de la structure de communication. Pour d'autres, les dépôts sont d'abord un moyen pour maximiser l'impact de la recherche par un accès sans entrave à des formes conventionnelles de publication.

---

[2] Par exemple : *Scientific Publication : Policy on Open Access*, European Research Advisory Board, Final Report, décembre 2005, 14 p.

La multiplication des dépôts, souvent accompagnée par des appels institutionnels (universités, organismes de financement, etc.) pour faire obligation aux chercheurs d'y déposer leurs publications, a un effet d'entraînement assez variable selon les disciplines, que ce soit pour le volume de documents déposés, la nature des documents, etc (Bergstrom et Lavaty, 2007). La réussite sera plus grande lorsqu'on prend les moyens pour que ce soit les employés du dépôt qui « déposent » les documents plutôt que les chercheurs qui agissent de leur propre initiative. Il est indéniable que ces dépôts augmentent l'impact des publications disponibles.

Certaines réserves doivent compléter cet aperçu (Warw, 2004; Xia et Sun, 2007). On remarque que l'accès au texte intégral n'est pas toujours assuré, que la qualité des documents est très variable et que leur statut n'est pas toujours clair. Force aussi est de constater que ces dépôts sont peu coûteux au départ, mais exigent d'importants moyens pour le service et la conservation, moyens qui sont loin d'être assurés actuellement.

2.3 Impacts potentiels collatéraux du libre accès

En raison de leur position dominante dans les circuits marchands de la diffusion des revues, les grands éditeurs commerciaux bénéficient d'une capacité réelle de répondre aux défis de l'accès libre. Au-delà de leur force relative, il ont démontré une certaine capacité d'adaptation et de récupération. Par contre, les éditeurs sans but lucratif se montrent préoccupés face aux revues en libre accès. Plusieurs anticipent un impact négatif sur la publication savante, d'autant plus que le modèle économique alternatif ne fait pas consensus.

À cela s'ajoutent les dépôts de documents qui sont source d'insécurité. Cela dit, il est vrai que l'impact sera à la mesure du développement réel de ces dépôts dans le proche avenir. Même si l'accès à la version finale chez l'éditeur devrait garder tout son intérêt de même que la fréquentation d'une plateforme de revues avec services de qualité, les éditeurs sont réticents à permettre le dépôt de la version finale des textes dans les dépôts. Évidemment, la crainte de la perte d'abonnements n'est pas étrangère à cette réaction.

De façon abstraite, on pourrait considérer que ce mouvement devrait idéalement nuire davantage aux éditeurs qui pratiquent des prix exorbitants, donc aux grands commerciaux, mais ce n'est pas le cas, car, encore faudrait-il que le modèle d'affaires des bibliothèques change concernant les achats documentaires et les droits d'accès.

C'est dans ce contexte que se manifeste le sentiment de vulnérabilité des petits éditeurs, qui explique la réaction et les débats, notamment dans *Nature*, concernant le libre accès. Pour ces petits éditeurs, la plupart sans but lucratif, le constat est simple : l'accès libre menace plus les revues au service de la communauté universitaire que les oligopoles. Cela amène l'éditeur de *HighWire Press* et de *Stanford University Press*, Michael Keller, parlant des diverses initiatives récentes (dont la gratuité), à déclarer que les plus touchés ce sont les « éditeurs responsables » : « Ironically, the combination of changes… reduce the competitiveness of responsible publishers, and reduce their utility to universities and scientific community. » (Keller, 2001) Si *HighWire Press* se sent menacé, que dire des revues ancrées dans structures nationales de la communication scientifique.

3 Vers une cyberinfrastructure

La très grande accumulation et la diversité des documents de recherche accumulés ou consignés sur support numérique, la configuration des institutions qui

y participent, et la capacité de mutualisation de ces documents conduisent à repenser le système d'information du document de recherche.

Les revues, les diverses publications plus ou moins arbitrées, les thèses, et autres sont dans des systèmes encore morcelés aux plans organisationnel, technique, socioéconomique, etc. Il est imaginable que les dépôts puissent agir comme levier d'intégration, mais leurs possibilités restent limitées : standard de publication, capacité de recherche, garanties de conservation à long terme, contenu limité et assez aléatoire, etc., toutes ces questions restent posées.

Par ailleurs, il est possible de penser la publication scientifique non seulement comme diffusion de la recherche mais aussi comme matériau de première main (sources premières) pour de nouvelles recherches.

De ce double constat découle l'idée de la cyberinfrastructure. Un récent rapport du *American Council of Learned Societies* insiste sur l'intérêt de considérer le système de publication universitaire comme un laboratoire de recherche, c'est-à-dire comme une infrastructure mise au service de protocoles de recherche particuliers. Particulièrement en sciences humaines et sociales, ce serait une contribution au développement d'un *Cultural Commonwealt* (*Our Cultural Commonwealth*, 2006).[3] Par ce terme, on entend qu'une communauté de personnes, avec un intérêt commun, se définit et que le bien-être public, et l'avantage ou le bien général, est le premier objectif du système d'information qui se construit.

Partant de cette notion de communauté, une cyberinfrastructure peut être comprise en termes de constitution d'un système d'informations, d'expertises, de normes, de politiques, d'outils et de services, mis largement à la disposition de communautés partageant les mêmes intérêts de recherche. La cyberinfrastructure est plus large qu'un outil ou un type d'information particulier à un projet, mais reste plus spécifique qu'un réseau. Il s'agit d'une perspective de travail, d'un projet en devenir, sans que les contours et l'architecture soient déjà déterminés, mais il y a un double préalable me semble-t-il. D'abord, il importe que la publication de recherche et l'information pertinente soient numérisées dans des formats de qualité et accessibles sans restrictions abusives (de prix ou autres). Ensuite, on doit pouvoir compter sur des réalisations, acquises dans le secteur sans but lucratif (aux USA, on s'appuie *Muse*, *JSTOR*, *ARTStor*, etc.). C'est sur cette base que la cyberinfrastructure prendra tout son sens, au-delà des équipements, des tuyaux et des outils logiciels.

4 Les pouvoirs publics interpellés

Les enjeux de la communication scientifique renvoient à l'économie, à la place occupée dans les réseaux scientifiques, aux systèmes de recherche nationaux ou supranationaux, etc. En un mot, ces enjeux interpellent le politique.[4] Aussi n'est-ce pas un hasard si, au cours des dernières années, plusieurs documents ont été produits à l'initiative des bibliothécaires, des chercheurs, mais aussi de comités conseils nommés par les pouvoirs publics.

On peut constater deux choses. Les questions soulevées touchent plusieurs milieux, que ce soit les professeurs-chercheurs, les sociétés savantes, les

---

[3] *Our Cultural Commonwealth*, Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences, American Council of Learned Societies, 2006, 44 p.

[4] Les réflexions qui suivent s'inspirent d'un engagement professionnel dans des projets de publication numérique de documents de recherche, d'abord et toujours avec Érudit (www.erudit.org), puis avec Synergies (www.synergiescanada.org), et dans le développement de plateformes multi-genres où se côtoient revues, livres, actes, pré- et post-publications.

bibliothécaires, les archivistes, la direction universitaire, les presses universitaires, le secteur commercial, le gouvernement et, éventuellement, les fondations privées. De plus, il ressort qu'une action collective s'impose à travers une intervention publique.

Il n'appartient pas ici de prescrire une politique spécifique ; au mieux, il est possible de dégager quelques principes. Partons d'un double constat. D'un côté, les intérêts sont diversifiés et les capacités d'action segmentées ou reposent sur la volonté de chacun. D'un autre côté, les intérêts en présence, non seulement corporatistes, mais aussi d'affaires, sont puissamment ancrés. Cette réalité nous conduit à retenir la nécessité d'une action collective, car elle s'appuierait sur une capacité réelle de décider, sur des ressources disponibles conséquentes, et sur la disposition de moyens pour voir à l'application des décisions. Autrement, il faudrait compter sur la concordance spontanée des initiatives dispersées, ce qui apparaît plutôt illusoire.

4.1 Penser en termes de bien public

Le système de communication scientifique est intimement lié au développement de la recherche, à l'avancement de la connaissance et à l'érudition dans une société. Par la place occupée dans la société, la communication scientifique doit être vue comme un bien public plutôt qu'un produit commercial ; en cela, on peut faire une analogie avec la défense, la santé publique, le système de navigation GPS, les barrages, l'éducation, etc. Il est curieux que ce raisonnement ne semble pas s'appliquer à la connaissance.

Il est symptomatique que l'*American Council of Learned Societies* insiste sur l'importance d'agir collectivement pour soutenir un système de communication scientifique, compris comme bien public ; sur cette lancée, on souligne que l'action publique doit viser le plus grand nombre. Pour ce faire, il est nécessaire de développer une vision globale et systémique du problème, avec une perspective de service public. De plus, il est impératif de considérer que l'investissement dans l'infrastructure de la communication scientifique est une priorité stratégique.

Parler de service public ne signifie pas que l'on cède à une perspective du tout à l'État. Les pouvoirs publics devraient opter pour s'engager dans ce champ sur la base des acteurs (chercheurs, éditeurs responsables, bibliothécaires, divers experts, universités et centres, plateformes numériques, etc.) et non en essayant de redéfinir arbitrairement ou par décret de terrain d'action. On peut très bien considérer l'intérêt de travailler en partenariat, avec des conventions ou cahiers des charges, ou, sous conditions, avec les acteurs privés. Il est primordial que l'engagement se fasse dans une perspective de long terme et s'inscrive dans un pacte social, afin d'acquérir une certaine prévisibilité des règles du jeu. Il reste que tout cela implique qu'il faille faire des choix.

Parallèlement à la démarche relative aux acteurs, l'approche devrait être inclusive, en ce qui concerne les matières et les matériaux participant à la communication scientifique.

Le secteur des revues sans but lucratif occupe une place capitale et stratégique dans la restructuration de la communication scientifique. Mais il est également fort utile d'associer, tant que possible, les autres genres de la communication scientifique. D'où l'idée de travailler dans la perspective de la complémentarité. Les dépôts de documents sont un bon exemple à ce propos.

Mais cela ne devrait pas faire perdre de vue la *revue*, comme genre « noble » de la communication scientifique. Promouvoir la concurrence (commerciale) dans le milieu des revues, notamment en luttant contre les barrières stratégiques à l'entrée, devient un élément primordial. C'est en ce sens que l'on s'interroge de plus en plus sur la possibilité de casser la mécanique des *Big Deals*, en encadrant et soutenant les

bibliothécaires pour éliminer les conditions abusives (prix publics et individuels, paniers ouverts, rejet des pénalités, fourchette pour prix juste, etc.). Les réflexions avancées à ce propos en Europe sont d'un grand intérêt (Dewatrinpont *et al.*, 2006). D'ailleurs, que ce soit en Europe ou aux États-Unis, la tentation est grande de procéder à un examen attentif des prochaines fusions eu égard au bien public (en fonction des législations anti-trust) et de considérer les *Big Deals* comme une entrave inadmissible au commerce. Sur un versant plus positif, l'intervention publique peut chercher à offrir une alternative professionnelle et crédible au modèle commercial oligopolistique pour la production, la diffusion et la préservation à long terme de la documentation scientifique et à financer de larges plateformes donnant un accès libre aux revues publiées dans le (les) pays. Il s'agit de grandes avenues qui demandent à être discutées et opérationnalisées, ce que je ne ferai pas ici.

4.2 Affirmation de la diversité

Le secteur des revues sans but lucratif occupe une place capitale et stratégique dans la restructuration de la communication scientifique pour autant qu'elles en aient les moyens. De plus, pour les revues qui participent à des structures d'abord « nationales » de communication scientifique ou à des sous-ensembles linguistiques de la communication scientifique, leur présence structurée et structurante contribue, notamment, à confirmer le caractère polyglotte du Web et de la communication scientifique. Nous sommes face à la nécessité de relever le défi de la diffusion mondiale du document universitaire de ces sous-ensembles et, pour cela, l'exploitation de la logique et des possibilités du numérique et de la mise en réseau est d'une grande aide.

Pour prendre le sous-ensemble francophone, l'objectif est de diffuser bien au-delà de la francophonie. Il s'agit d'imposer sa présence dans la communication scientifique, en milieu francophone, mais aussi dans l'anglophonie, en ayant un positionnement stratégique dans les principaux outils au plan international : index, répertoires, moteurs de recherche, agrégateurs. Il faut pouvoir compter sur un effet de masse, soit disposer de collections de quelques centaines de revues ; cela facilitera d'autant la présence dans les bases de données et systèmes d'information. Établir des partenariats avec des pôles importants (*Synergies, Muse* et *HighWire*) des réseaux anglophones serait aussi précieux. En somme, l'objectif c'est d'avoir droit de cité dans l'anglophonie, mais aussi… d'être cité.

Ce raisonnement ne se limite pas à la francophonie évidemment. On peut noter le grand intérêt que représentent des projets, tels SciELO pour les pays latino-américains et à J-STAGE pour le Japon. Les objectifs sous-jacents sont tout à fait apparentés.

5 Accès public et pérenne aux résultats de la recherche

Dans le monde numérique, cette question se présente sous un nouveau jour (Schonfeld *et al.*, 2004). Face à l'hétérogénéité des dispositions des éditeurs, on peut ressentir la nécessité de créer une plateforme pour les éditeurs sans but lucratif qui aurait pour mandat la préservation à long terme des documents. On peut s'étonner et déplorer la lenteur avec laquelle on en vient à introduire le dépôt légal obligatoire et une norme d'identifiant permanent dans le domaine public.

La mise en place d'une plateforme qui fournisse un accès central aux diverses revues n'est sans doute qu'une méthode qui n'est pas incontournable. L'objectif, qu'il ne faut pas perdre de vue, c'est bien plutôt de garantir l'accès public aux résultats de la recherche financée avec des fonds publics, dans des conditions qui permettent la viabilité de tous les acteurs dans un environnement où les règles du

jeu permettent une prévisibilité et une stabilité. Il va de soi que l'on puisse aussi utiliser et créer au besoin des dépôts qui mettent à disposition (institutionnels, interinstitutionnels, spécialisés) les résultats de la recherche. Dans tout cela, on doit s'assurer que l'on applique des normes compatibles assurant l'interopérabilité, l'accessibilité et la diffusion des plateformes participant à ces réseaux de la communication scientifique.

C'est dans ce contexte que se pose le défi de la conception d'un système d'information intégré. L'idée de base consiste à aller au-delà de la division du travail actuelle, en créant un système d'information multigenre, qui permette la recherche intégrée de l'ensemble des corpus issus de la recherche, tout en reconnaissant et identifiant chacun des genres mis à contribution (conférence, article, chapitre de livre, chapitre de thèse, note de recherche, etc.). On peut aussi aller plus loin. C'est d'ailleurs tout l'intérêt de la notion de cyberinfrastructure, considérée sous l'angle de son contenu. La conception de ce système d'information multigenre peut ne plus se limiter à la dimension de la communication, et mettre l'accent sur la dimension infrastructure et laboratoire de recherche, adaptable à des protocoles de recherche distincts. Il s'agit là d'une condition pour s'engager dans de nouvelles recherches avec de nouveaux protocoles et de nouveaux outils.

5.1 Un bien public face aux forces du marché

On ne peut plus ignorer la question de l'action publique sur la configuration et les conditions de la communication scientifique. Les pouvoirs publics ont toujours, à un titre ou à un autre, joué un rôle. Il faut dès lors s'interroger sur les contours de cette action publique dans le monde numérique et compte tenu de la force relative des acteurs.

Cet article a voulu présenter une évaluation des principaux enjeux, fondée sur l'état des travaux, sur l'expérience, mais aussi sur des rapports récents venant des États-Unis et de l'Union européenne. Il est significatif qu'aux États-Unis on parle d'action collective alors que tous les principaux acquis ont pour origine des fondations privées. Il est aussi significatif que la Commission européenne suscite et reçoive des propositions plaidant pour une action nationale et supranationale en la matière afin d'assurer la circulation des résultats de recherche dans un environnement non-oligopolistique et le plus grand accès au savoir. L'idée d'une action publique fondée sur un bien public s'exprime de façon plus insistante des deux côtés de l'Atlantique, sachant bien que ce qui est en cause ce n'est pas un réflexe étatiste. Plutôt, la réflexion sur l'accès à la connaissance met en lumière, plus généralement que la culture, et la culture scientifique en particulier, ne peut être laissée aux seules forces du marché.

Références :

Bergstrom, Carl T., Bergstrom, Theodore C. (2003). The Costs and Benefits of Library Site Licenses to Academic Journals. Proceedings of the National Academy of Sciences, vol 1001, num 3, 897-902.

Bergstrom, Ted C., Lavaty, Rosemarie (2007). How Often Do Economists Self-Archive ? Scholarship Repository, University of California, Paper 2007a, Disponible à :
http:repositories.cdlib.org/ucsbecon/bergstrom/2007a

Davis, Philip M. (2003). Tragedy of the Commons Revisited : Librarians, Publishers, Faculty and the Demise of a Public Resource. Libraries and the Academy, vol. 3, num 4, 547-562.

Dewatrinpont, Mathias et al. (2006). Study on the Economic and Technical Evolution of the Scientific Publication Markets in Europe, Rapport final déposé à la DG-Rercherche, Commission européenne, janvier.

Edlin Aaron S. et Daniel L. Rubinfeld (2004). Exclusion or Efficient Pricing: The «Big Deal» Bundling of Academic Journals, Disponible à :

http://works.bepress.com/aaron edlin/37

Frazier, Kenneth (2001). The Librarians' Dilemma. D-Lib Magazine, vol. 7, num 3, mars.

Getz, Malcolm (2004) Open-Access Scholarly Publishing in Economic Perspective, Working Paper n° 04-W14, Department of Economics Vanderbilt University, Nashvill, juin, 58 p.

Hahn, Karla (2006). The State of the Large Publisher Bindle : Findings from an ARL Member Survey. ARL, num 245, avril, 1-6.

Kaufman-Wills Group, LLC (2005). The Facts About Open Access. Rapport de recherche pour The Association of Learned and Professional Society Publishers, p. 1-25.

Keller, Michael (2001). Innovation and Service in Scientific Publishing Requires More, Not Less, Competition. Nature, web debates, 6 septembre.

McCabe, Mark J. (2002). Journal Pricing and Mergers : A portfolio Approach. The American Economic Review, vol 92, num 1, mars 2002, p. 219-269.

McCabe Mark J., Snyder, Christopher M. (2005). Open Access and Academic Journal Quality, AEA Papers and Proceeding, mai.

Our Cultural Commonwealth, Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences, American Council of Learned Societies, 2006, 44 p.

Schonfeld Roger C. et al. (2004). The Nonsubscription Side of Periodicals : Changes in Labrary Operations and Costs between Print and Electronic Formats, Rapport de recherche, Council on Library and Information Resources, Washington, juin, 58 p.

Scientific Publication : Policy on Open Access, European Research Advisory Board, Final Report, décembre 2005, 14 p.

Stanley, Morgan (2002). « Scientific Publishing: Knowledge is Power», Industry Overview, Equity Research, 30 septembre.

Tenopir, Carol (2006). Not-for-Profit Scholarly Societies and Open Access Journal Publishing (diapo), School of Information Sciences, The University of Yennessee.

Warw, Mark. (2004). Institutional Repositories and Scholarly Publishing. Learned Publishing, vol. 17, num 2, avril, 115-124.

Xia, Jingfeng, Sun, Li (2007). Assessment of Self-Archiving in Institutional Repositories : Depositorship and Full-Text Availability. Serials Review, vol. 33, num 1, 14-21.

Lecture(s) et genre(s) du document numérique

Reading(s) and Genre(s) of the Digital Document

Ioannis KANELLOS(1)

(1)Département Informatique, TELECOM-Bretagne, Brest, France
ioannis.kanellos@telecom-bretagne.eu

Résumé. L'article propose quelques éléments de réflexion sur le genre du document numérique (DN) qu'il cherche à situer dans un cadre interprétatif. Dans un premier temps, il visite certains travaux relatifs dans l'objectif d'en distinguer et d'unifier d'éventuelles tendances. Il résume les visions sur la notion en deux. Suivant la première, le genre serait le produit d'une analyse logique et sa détermination relèverait d'une activité classificatoire. Suivant la seconde, il serait de la nature d'un prototype et sa détermination procèderait de l'explicitation d'une structure fondée sur la similitude. À ces deux, on proposerait volontiers une vision du genre du DN en termes de projet d'interprétation : le genre serait alors une donnée permettant de déclencher des lectures efficaces, respectant les normes d'une communauté. L'article conclut sur une discrète critique du projet du Web Sémantique qui l'interroge, précisément, sur cet « oubli » inexplicable de la notion de genre du DN.
Mots-clés. Document numérique (électronique), genre, interprétation, stratégie de lecture, schème, normes, communauté de pratiques, action, énaction, classification, prototype.

Abstract. The paper presents some thoughts concerning the genre of the digital (electronic) document from an interpretative point of view. In the first part, it deals with the state of the art, aiming at distinguishing and unifying already developed conceptions about the notion. It suggests to divide them into two main categories. According to the first one, the genre becomes the product of a logical analysis and its determination comes out from a classification objective. According to the second one, it is of the nature of a prototype and its determination gives generally rise to some similarity founded structure. The paper gives clues for an additional vision of the digital document genre in terms of interpretation project: the genre may be seen as the framework that allows efficient readings respecting the norms of a community. This idea is, in the conclusion, the basis of a sober critique of the Semantic Web project, questioned closely about its silence concerning the genre of the digital document.
Keywords. Electronic (or digital) document, genre, interpretation, reading strategy, schema, norms, community of practices, action, enaction, classification, prototype.

1 Introduction

La notion de genre désigne généralement un des aspects de normativité dans une écologie de formes de communication. Il concerne certes la production mais, surtout, la consommation sémiotique. Ces lignes proposent une réflexion sur le genre du document numérique (DN), entendu ici plus que comme un objet et un véhicule d'information numérique (doxa), plus que comme la représentation d'une vérité partagée au-delà du chaos (le silence et le bruit), de la cacophonie (la confusion et le sensible) et de l'oubli (l'intime et l'éphémère) (Pédauque, 2006), plus même que comme une mémoire collective, externalisée certes par la technologie, et souvent en constante évolution (Bachimont, 2007) : comme quelque chose qui, dans une pratique de communication normée, peut être convoqué dans un projet d'interprétation.

Revendiquée, cette place au sein d'un horizon interprétatif serait même sa cause efficiente. En effet, pour assurer ce nouveau type de communication, désormais possible à travers les réseaux, le document numérique doit répondre des normes de production et de réception novatrices. Un tel argument demeure dans l'attendu et l'évident et reste généralement recevable : une communication ne peut promouvoir de nouvelles formes de sociabilité que si elle est efficace, i.e. que si elle participe effectivement dans une économie sémantique réglée qui offre, à ceux qui en usent, des moyens d'expression suffisants. Mais cet argument stagne dans l'ombre de sa propre évidence. L'évidence n'est d'ailleurs pas la seule cause du peu d'intérêt que l'on crédite généralement à la notion de genre. En vérité, on ne sait pas quoi en faire « dans la pratique ». Si, comme on le dit, en théorie, il n'y a pas de différence entre théorie et pratique, on sait pertinemment que, en pratique, il y en a ! Il n'est pas très difficile de parler de normativité et de genre de DN ; mais essayer d'en expliciter des aspects opératoires, avoir l'ambition de les décrire minutieusement, de les formaliser même, voire d'en tirer quelque profit dans le cadre d'une application, c'est une autre histoire. Et un autre défi. Ainsi, même si la valeur d'une réflexion sur le genre du DN n'est pas à contester, son intérêt subit nécessairement des variations inflationnistes suivant les marchés et les modes. À la rigueur, de la pression applicative.

Penser le genre du DN ne comporte cependant pas de litige épistémologique : le genre concerne prioritairement les degrés de production sémiotique les plus élevés, et un DN constitue effectivement une entité d'un tel niveau.

Mais comment penser le genre du DN ? Comment le penser en s'éloignant de la leçon peu exploitable des dictionnaires et des traditions théoriques, toujours d'un emprunt possible et facile ? Serait-il un objet ou un concept ? Serait-il la forme que prendraient des conditions ou des agrégats de propriétés ? Serait-il, tout simplement, une classe, un ensemble ou une catégorie ? Peut-être un type ? Un prototype ? Un système de filiations ou des airs de famille ? Serait-il, éventuellement, un schéma de communication attendue ou encore un moyen pour scénariser des attentes partagées en matière de communication ? On conviendra, certes, que les genres sont des constructions culturelles relativement stables, des formes globales de régulation intersubjective dans le commerce sémiotique (tant en production qu'en réception) et encore des catégories englobantes de la tolérance culturelle. Mais plus précisément ?

Thème peu passionnel mais fortement passionnant, embêtant ou utile, quelque fois même dérangeant voire agaçant, le genre nous transporte violemment vers des considérations relevant de ces procédures continues mais fugaces d'encodage des

échanges sociaux que réalisent nos codes sémiotiques, forcément normés puisque obligés d'être partagés.

2 Des deux visions « génériques » attestées sur le genre du DN

Il y a déjà pas mal de travaux sur le genre du DN. Leur courte histoire atteste sans surprise de cette marche, somme toute habituelle, d'une envie de connaissance qui va de l'objet vers le sujet, par un progrès successif de notre conscience sur la complexité de tout thème d'étude. En effet, dans un premier temps, on a cherché à décrire, d'une manière logique la notion de genre du DN, à l'objectiver en quelque sorte, en lui décernant, un peu précipitamment peut-être, le statut d'un objet intégralement saisissable sous forme de propriétés et de valeurs. Puis, le prétendu « objet », en se révélant plus fluide et plus instable, ou seulement plus complexe, en faisant exploser le cadre que souhaitait pour lui une clarté de facture formelle, en repoussant ses limites hors de la capacité de nos calculs, a fini par établir de nouvelles exigences d'approche, souvent empruntées à des notions plus souples, plus ouvertes, plus « poétiques » peut-être, comme la « similitude » et la « distance ».

2.1 Le genre du DN serait une classe

L'abord logique repose sur la confiance de la pensée humaine en ses capacités d'accéder à la totalité de la réalité d'un objet. À l'idée, aussi, que la connaissance peut être indépendante du sujet qui questionne ((Arendt, 1972), ch. « Le concept de l'histoire », p. 68). L'autorité des mathématiques qui l'accompagne, plus précisément de cette part des mathématiques que constitue le calcul, devient un héritage de droit, qui l'amène même souvent à des postures d'arrogance en matière de connaissance. Une connaissance qui dérive, d'ailleurs, vers la forme d'une tutelle concernant la vérité même.

On y discerne encore la leçon, toute vivante, de Descartes (*Discours de la Méthode*, Deuxième partie, 141) : « Ne recevoir jamais aucune chose pour vraie que je ne la connusse évidemment être telle ; c'est-à-dire, d'éviter soigneusement la précipitation et la prévention, et de ne comprendre rien de plus en mes jugements que ce qui se présenterait si clairement et si distinctement à mon esprit, que je n'eusse aucune occasion de le mettre en doute. » Propos qui évincent d'emblée toute interrogation sur le rôle du sujet dans la constitution même des connaissances.

Les travaux de Biber ((Biber, 1988, 1990, 1993a, 1993b, 1995), (Biber et al., 1996), (Biber and Conrad, 2001), (Biber and Kurjian, 2007) entre autres) en sont probablement emblématiques. Ils témoignent de cette pensée logique, foncièrement classificatoire, qui se sert des calculs pour établir des statistiques nécessaires à ses projets de classification. Le DN se représente sous forme d'un tableau truffé de 0 et de 1, valeurs qui disent si un prédicat est ou non vérifié. (Kessler et al., 1997) et leur genre comme faisceau de facettes, (Crowston and Williams, 1997), et la classification des DN du Web, (Schmid-Isler, 2000) et son « langage » des genres numériques, (Shepherd & Polanyi, 2000) et leur tentative de modéliser les genres du DN tout en s'ouvrant à des concepts moins « objectivisants », (Malrieu et Rastier, 2001) et la recherche des corrélations entre genres et variations morphosyntaxiques, (Shepherd and Watters, 2004) et leur conception des genres du DN comme une entité mouvante et évolutive… tous, ne font que reprendre l'entreprise classificatoire en suivant des critères choisis propres à des objectifs applicatifs différents.

Par la description, les critères et les classes, par les conditions nécessaires et suffisantes sans cesse recherchées et immanquablement étendues et affinées, la

définition du genre devient celle d'un objet, d'un objet obtenu par objectivation dans une entreprise angoissante d'objectivité, l'équivalent d'une classe de propriétés logiques. Définir c'est bien, mais le prix est celui de l'essence : on ne saurait avoir essence et définition rappelleraient de vieilles leçons de philosophie (Aristote, *Analytiques Seconds*, *passim*).

2.2 Le genre du DN serait un (proto)type

Or, l'abord logique a ses limites. Peut-être nos propres limites, pas si différents que celle de notre pensée discursive. En tout cas, le calcul, on le sait, ne peut garantir la vérité, quelle qu'elle soit (et pour autant qu'il y en ait une), par le biais des classes qu'il permet de construire. La critique de la confiance logique en sciences cognitives, plutôt diffuse ou cantonnée aux critiques philosophiques, a trouvé son représentant le plus médiatisé dans la mouvance de la typicalité. Sans surprise, ce mouvement s'est répété : la théorisation de la notion de genre du DN a, semble-t-il, également expérimenté, sous plusieurs variantes, sciemment ou non, cette voie. Il s'agit de tentatives qu'on pourrait comprendre comme « génétiques ». Suivant ce point de vue, les genres seraient des entités présentant certaines ressemblances entre elles, des sortes de « catégories » envisagées en termes de typicalité et de représentativité, de similitude et de graduation, suivant une vision plutôt radiale de l'organisation de la classe. Mais la notion de fond, celle de classe, reste immuable.

Cette vision reconnaît d'illustres prédécesseurs, même si peu formalistes (et pour cause). (Bakhtine, 1953), par exemple, et ses genres premiers et seconds, (Todorov, 1978), aussi, et la distinction entre les genres théoriques et historiques et, bien sûr, l'indépassable (Genette, 1979) et ses tentatives de dérivation des genres par complexification des genres premiers d'une certaine tradition. Pour (Adam, 1992), il s'agirait de catégories « superordonnées » tandis que dans (Compagnion, 2001) on lit explicitement que les genres feraient appel à des relations de type « air de famille ». Enfin, (Schaeffer, 1989, 1995), entre autres, parle explicitement de régimes de généricité, induits par une relation de répétition (ou de duplication) et une relation de transformation (ou d'écart).

Ces deux approches couvrent pratiquement l'ensemble de la question du genre du DN (et pas seulement). On remarquera cependant que, sous la pression applicative, même dans ce cadre génétique (et, plus avant, transformationnel et dérivationnel), on rejoint, à terme, une perspective plus ou moins classificatoire. Seuls les principes de classification et la vision de l'organisation de la classe diffèrent : mais on calcule le prototype de façon classique, on fait des mesures et on établit aussi des statistiques pour en déduire l'appartenance à la classe-genre.

L'écho de l'opposition classique entre génotype et phénotype n'est pas difficile à entendre dans cette vision du genre où la filiation qui le fonderait reposerait, à son tour, sur la notion de similitude. Le genre dérive ainsi, comme une structure ouverte, au moyen d'une autre notion-pivot, celle de distance, mobilisée métaphoriquement pour offrir quelque sol ferme aux calculs.

3 Le genre serait plutôt une dimension essentielle de la lecture

Mais l'expérience de la lecture semble nous livrer, à toute occasion, une toute autre leçon. En effet, en s'interrogeant sur notre façon de faire pendant que nous lisons, en considérant notre lecture comme cet acte qui réalise notre intention de comprendre, on s'aperçoit que la problématique du genre doit se déplacer de façon radicale, définitivement radicale, vers le sujet et ses stratégies de lecture. De telles stratégies sont moins « subjectives » qu'on ne le croit et qu'on ne le dit

habituellement, puisqu'elles sont normées par le besoin d'être reprises et validées par les membres d'une communauté engagée dans les mêmes projets de compréhension.

Le DN étant de la nature des construits, son genre ne ferait pas exception : il est en réalité actualisé par la forme de la lecture, i.e. par une intention de comprendre, intention toujours placée dans le cadre d'un usage. Le genre devient de cette manière quelque chose d'intimement lié à la lecture et à ses conditions de réalisation. Autrement dit, chaque nouvelle condition de lecture est susceptible de reconfigurer l'identité générique d'un document, d'un DN en particulier. Le régime de la lecture impose, par conséquent, au genre un régime d'insécurité soutenue : un genre peut subir une catastrophe au profit d'un autre genre. Il est clair que la pluralité des codes sémiotiques qu'un DN met à disposition augmente de manière significative ce risque.

Mais, précisément, la lecture d'un DN inclut une dimension interactionnelle qui fait de lui, également, un document à manipuler. On lit certes, on manipule aussi. Ce qui veut dire, qu'un DN est le produit d'une lecture qui engage une dimension de maniement. Cette manipulation procède, bien entendu, d'un projet de lecture, tout en le validant et/ou en le rectifiant en permanence. Probablement, le DN est le premier document qui dévoile avec autant de probité combien les rôles entre le lecteur et l'auteur sont interdépendants mais fondamentalement asymétriques ; combien, aussi, le lecteur est un auteur. Car, évidemment, le lecteur-acteur construit son propre document à partir des opportunités d'interaction que le DN met à sa disposition. Le DN est, en somme, et même devient, l'élaboration que son lecteur lui fait subir.

Le repérage du genre du DN serait cet élément indispensable à tout projet de communication active pour déclencher les stratégies de lecture (choix et intégration des informations) et d'action (maniements de configuration et de reconfiguration) qui correspondent à un projet d'interprétation. En d'autres termes, le genre du DN est comme le maître de Delphes : ne dit rien, il n'affirme rien, il n'infirme rien ; seulement il indique. Il situe et il guide. Il suggère surtout et, avant même le déroulement des opérations propres à la lecture, il offre des moyens pour évaluer la dynamique entre la situation de lecture et son objectif d'un côté et, de l'autre, le type de lecture et les scénarios des manipulations du DN qui l'accompagneraient au mieux. Il apparaît, ainsi, comme un authentique *schème* sémiotique qui encode des normes issues d'une histoire de lectures partagées. Le genre du DN serait donc un fonds permanent de légitimation d'une lecture (et d'une écriture, d'ailleurs) dans ce nouveau média porté par les technologies de la communication et de l'information. Il n'y a pas de genre sans consensus. Et ce consensus serait aussi à l'origine d'un lien de cohérence du DN, lien nécessaire pour fonder une lecture.

Cette vision de la notion de genre de DN, qu'on appellerait volontiers « schématique » nous amène sans médiation aux problématiques centrales de l'*énaction*. Le genre, en réalité, n'est ni objet ni concept : seulement un « programme » d'action inhérent à la forme de la lecture. Il commande non seulement ce qu'il faut faire pour lire correctement, i.e. en respectant les normes d'usage. Mais, aussi, ce qu'il ne faut pas faire. Il tient le rôle d'un catalyseur pour l'exploitation du DN en orientant la lecture et en simplifiant la complexité sémiotique. Il offre, disons, des « heuristiques » pour rendre la lecture conforme aux pratiques d'une communauté qui est bâtie et se reconnaît dans la même économie sémiotique en mettant, précisément, en place des référentiels (par exemple, des conventions constituantes, régulatrices et traditionnelles) pour évaluer et légitimer un projet de lecture. Il offre,

également, des éléments pour gérer l'inconnu. De plus, il autorise des anticipations et des projections, en circonscrivant des horizons d'attentes ((Jauss, 1978), (Rastier et Pincemin, 1999)) sous forme de présomptions qui pilotent la suite, à court et à long terme, du processus de lecture (cf., par exemple, (Rastier, 1987, 1989)).

De façon sommaire, on pourrait représenter la dialectique entre genre et lecture de la manière suivante :

> Détection du genre → choix d'une stratégie de lecture → présomptions, collecte des informations et recherche d'interprétants → mise en épreuve, validation → confirmation du genre → nouvelles présomptions, collecte d'informations et recherche d'interprétants → nouvelle mise en épreuve etc.

Si, dans la suite précédente, le genre n'est pas confirmé, le scénario ne change pas de manière significative : seule la nature des présomptions nouvelles, de la nouvelle collecte d'informations etc. se voit altérer. Pour revenir au début, avec, éventuellement, une nouvelle stratégie de lecture. Le genre, par conséquent, aurait une double vocation, tant prospective que rétrospective, et dont l'objectif final serait l'apaisement de la tension entre l'originalité individuelle, qui semble sans limites dans le cas du DN, et certaines formes de conservatisme collectif, nécessaire pour la transmission des cultures et des savoirs. Il est là pour opérer le modelage mutuel d'un espace communément accepté et en permanente construction par une action conjuguée. Il codifierait, donc, cette part de la production et de l'interprétation qui relèvent d'un niveau sociolectal (et non du système fonctionnel de la compétence sémiotique convoquée pour les besoins de lecture d'un DN).

4 Conclusion

Ce papier ne visait pas à remettre à la mode l'affaire du genre du document numérique qui rebondit, semble-t-il, de toute façon, presque par une nécessité interne, dans le cadre d'une ambition de poser correctement la question d'un authentique Web Sémantique (WS). En effet, voici comment le *W3C* consortium décrit une telle entreprise (http://www.w3.org/2001/sw/) :

> "The *Semantic Web* provides a common framework that allows *data* to be shared and reused across application, enterprise, and community boundaries."

Et plus avant, il précise :

> "The Semantic Web is a web of data. There is lots of data we all use every day, and it is not part of the web. I can see my bank statements on the web, and my photographs, and I can see my appointments in a calendar. But can I see my photos in a calendar to see what I was doing when I took them? Can I see bank statement lines in a calendar?
>
> Why not? Because we don't have a web of data. Because data is controlled by applications, and each application keeps it to itself.
>
> The Semantic Web is about two things. It is about common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents. It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then

> move through an unending set of databases which are connected not by wires but by being about the same thing."

On pardonnera, bien sûr, le style désespéramment utilitariste sinon naïf de la formulation de la question. On se gardera, aussi, par simple prudence, de dresser la n-ième critique sur la vision latente de la sémantique sur laquelle ce projet repose, définitivement objectiviste. On retiendra, cependant, le type des objectifs du WS, qui sont essentiellement pratiques. L'organisation de l'information, censée représenter des connaissances, suit globalement une visée descriptive, suivant des protocoles qui tendent à rejoindre les préoccupations sur les standards (cf. le *Resource Description Framework* (RDF), les *Gleaning Resource Descriptions from Dialects of Languages* (GRDDL), le *SPARQL Query Language for RDF*, le célèbre *Web Ontology Language* (OWL) entre autres). Dans toutes ces initiatives, partielles mais indispensables et constitutives, le genre du DN est un absent silencieusement dissonant. On ne le trouve ni au niveau de la norme ni à celui du standard, il ne fait même pas partie des préoccupations descriptives. Comme si les connaissances étaient « atomisables », indépendantes de la forme des documents et sans aucun rapport avec les lectures qui leurs sont adressées et qui constituent leur contenu.

Le projet du WS devient, en réalité, aujourd'hui, un projet de partage des ressources. D'inter-utilisabilité et d'inter-opérabilité, de fusion ou de capitalisation de données. D'« inter-cohérence » partagée entre structures de connaissances cumulées. D'« inter-pratique », aussi, si l'on se permet le terme, puisqu'il ambitionne d'établir des passerelles entre contenus fondateurs des pratiques de diverses communautés. On imagine, et non sans fondement, que cette entreprise ne saurait voir le jour qu'en cherchant des appuis sur une dimension pivot, commune, nommément le sens. Plus précisément, le sens qui se trouve dans des « contenus », généralement multimédia. Donc, un sens porté par des DN. Mais si un tel horizon ne saurait faire l'économie de la question des pratiques, dans la mesure où les ressources n'émergent pas *ex nihilo* mais au sein même des pratiques, et même, des pratiques qui sont codifiées et normées socialement, il ne saurait non plus éviter la question des genres, puisque toute pratique se fonde sur des genres sémiotiques, qu'elle déploie et organise pour les besoins de communication des personnes qu'elle socialise.

L'entreprise du WS n'est en réalité qu'une reprise des thèmes qui ont largement dominé les orientations et les ambitions de l'IA d'antan, formulés certes plus prudemment et plus modestement aujourd'hui, mais qui restent, de toute évidence, toujours vivaces. D'une certaine manière, on pourrait même résumer le projet de WS en la tentative d'étendre et de généraliser, sur la dimension des ressources cette fois, le paradigme des réseaux sémantiques d'autrefois. En tout cas, s'agit-il encore d'une démarche clairement attachée à une épistémologie où le sens est uniformément considéré de façon discrète, atomique, logique, enfin : comme un objet détachable de la pratique sémiotique qui l'engage.

Cet élan de normativité des connaissances risque de vite s'estomper sans quelque concept susceptible de ramener, aux principes fondateurs du WS, une notion de normativité documentaire. Comment le WS pourra passer outre la problématique du genre du DN ? Le DN, vecteur de cette sémiose nécessaire à la communication à travers les réseaux et qu'on cherche par ailleurs à formaliser comme ressource sémantique, amène irrémédiablement la question du genre, tout simplement parce que le genre encapsule des éléments cruciaux pour la production et la réception d'un document. Et donc, pour son exploitation en tant que ressource. C'est ainsi que nous interprétons, d'ailleurs, le rapprochement récent du

WS et du Web 2.0, un Web, en d'autres termes, des sociétés émergentes par et dans l'usage des DN, pour en faire sa nouvelle transfiguration, le « Web 3.0 ». L'histoire, pour encore une fois, semble nous précéder.

Pour résumer et pour conclure : une réflexion sur le genre du DN au sein du projet de WS ne semble ni fortuite ni gratuite. Probablement même, n'y aura-t-il pas de WS sans notion de genre du DN.

On peut considérer cette dernière affirmation comme un pari.

Références :

Adam, J.-M. (1992). Les textes, types et prototypes : récit, description, argumentation, explication et dialogue, Nathan.

Arendt, H. (1972). La crise de la culture. Gallimard (1972 pour la trad. française).

Bachimont, B. (2007). « Bibliothèques numériques audiovisuelles. Des enjeux scientifiques et techniques. » http://www.utc.fr/~bachimon/Publications attachments/Bachimont-Biblios-AV.pdf

Bakhtin M. (1953). « Les genres du discours ». *In Esthétique de la création verbale*, NRF, Gallimard (nouvelle édition 1984).

Biber, D., Conrad, S. and Reppen, R. (1996). "Corpus-based investigations of language use". *Annual Review of Applied Linguistics*, v. 16, pp. 115-136.

Biber, D., and Conrad, S. (2001). "Quantitative corpus-based research: Much more than bean counting". *TESOL Quarterly* 35, pp. 331-336.

Biber, D., and Kurjian, J. (2007). "Towards a taxonomy of web registers and text types: A multi-dimensional analysis". *In* M. Hundt, N. Nesselhauf, and C. Biewer (eds.), *Corpus linguistics and the web*, pp. 109-132. Amsterdam, Rodopi.

Biber, D. (1988). *Variation across speech and writing*, Cambridge, Cambridge University Press.

Biber, D. (1990). "Methodological issues regarding corpus-based analyses of linguistic variation". *Literary and Linguistic Computing*, 5(4), pp. 257-269.

Biber, D. (1993a). "The multi-dimensional approach to linguistic analysis of genre variation: an overview of methodology and findings". *Computers and the Humanities*, 26 (5-6), pp. 331-345.

Biber, D. (1993b). "Using register-diversified corpora for general language studies", *Computational Linguistics*, 19(2), special issue on Using Large Corpora II, pp. 219-241.

Biber, D. (1995). *Dimensions of register variation: a cross-linguistic comparison*, Cambridge University Press.

Compagnon, A. (2001). "Théorie de la littérature : la notion de genre ». http://www.fabula.org/compagnon/genre.php.

Crowston, K. and Williams, M. (1997). "Reproduced and emergent genres of communication on the World Wide Web". *Proceedings of the 30th Annual Hawaii International Conference on System Sciences (HICSS '97)*. Maui, Hawaii, vol. VI, 1997, pp. 30-39. (aussi : http://crowston.syr.edu/papers/genres-journal.html).

Genette, G. (1979). *Introduction à l'architexte*. Seuil.

Jauss, H. R. (1978). *Pour une esthétique de la réception*, Gallimard (coll. « Tel ») (nouvelle édition 1990).

Kessler B., Nunberg G. and Schütze H. (1997). *Automatic detection of genre*. Palo Alto Research Centre.

Pédauque, R. (2006). « Comprendre et maîtriser la re-documentarisation ». Proposition de projet de recherche (déposé auprès de l'ANR). http://rtp-doc.enssib.fr/rubrique.php3?id rubrique=63.

Rastier, F. (1987). *Sémantique interprétative*. P.U.F.

Rastier, F. (1989). *Sens et textualité*. Hachette.

Rastier, F. (2001). *Arts et sciences du texte*. PUF (coll. « Formes sémiotiques »)

Rastier, F. & Pincemin, B. (1999). « Des genres à l'intertexte », *Cahiers de Praxématique* 33, pp. 12-43.

Schaeffer, J.-M. (1989). *Qu'est-ce qu'un genre littéraire ?* Seuil.

Schaeffer J.M. (1995). *Les célibataires de l'art. Pour une esthétique sans mythes*. Gallimard.

Todorov, T. (1978). *Les genres du discours*. Seuil.

Le document : à la frontière du modélisable

Jacques LABICHE(1)

(1)Litis, Université de Rouen, Rouen, France
Jacques.Labiche@univ-rouen.fr

Résumé. Une analyse de l'évolution de l'interprétation automatisée d'images de documents permet de dégager de nouvelles pistes de recherche tenant compte d'une théorie récemment apparue en sciences cognitives et en particulier de ses implications dans le domaine de l'interprétation textuelle. Ces nouvelles pistes vont à l'encontre des méthodes habituelles de développement informatique en assignant au système informatique un rôle d'acteur à part entière lors de chaque utilisation considérée comme une des étapes d'une démarche expérimentale.
Mots-clés. système informatique, traitement d'images, cognition, interprétation.

1 Introduction

Le traitement d'images construit sur les méthodes et outils du traitement du signal, et la reconnaissance de formes issue de l'Intelligence Artificielle, ont bénéficié des avancées de l'ingénierie informatique pour proposer les systèmes informatisés actuels d'interprétation de documents textuels ou graphiques conçus par les entreprises spécialisées pour les grands comptes. Cette communication tente d'analyser l'évolution de ces systèmes pour dégager des perspectives de recherche, tenant compte des attentes sociétales et industrielles, pour les équipes de recherche investies dans ce champ particulier de la gestion électronique du document qu'est l'ingénierie du document numérisé.

La collaboration nécessaire avec les experts « métier » engagés dans la recherche et développement pour la conception de systèmes dédiés amène en effet à s'interroger sur la relation entre la formalisation des connaissances dans les dispositifs informatiques et les compétences cognitives des concepteurs et utilisateurs.

Les nouvelles interactions apparues entre systèmes et acteurs, amènent des glissements entre les rôles traditionnels. Cette évolution amène à s'interroger sur la nature des compétences cognitives des acteurs mises en jeu au cours de ces interactions qui pourraient permettre de mieux contextualiser les procédures et algorithmes des systèmes informatisés.

C'est ce dernier angle d'attaque qui a été choisi par le « groupe NU » qui regroupe des informaticiens, des linguistes et des « traiteurs d'images » qui s'interrogent sur l'apport de l'informatique pour des systèmes d'aide à l'interprétation et ont décidé d'analyser dans ce contexte les conséquences de la théorie de l'enaction qui propose en particulier de reconsidérer les liens entre perception et action ; liens essentiels lors de la conception des systèmes si on souhaite qu'il puisse y avoir émergence de nouveaux usages.

2 Historique du traitement d'images de documents

Les applications visées ici appartiennent au domaine de l'interprétation d'images de documents. Une image de document (figure 1 : Exemples de documents) n'est pas n'importe quelle image… Pour un lecteur humain, tout document, qu'il soit textuel, graphique ou mixte est un support direct d'information. Il peut en être extrait immédiatement des informations : lecture du texte si la langue en est connue, extraction de données si le document graphique (ou plan) respecte les conventions élémentaires (Bertin, 1967). L'image, elle, n'est qu'un tableau de valeurs (pixels) que l'ordinateur ne peut décrypter immédiatement. La difficulté provient de ce que l'apprentissage de la lecture est prégnant pour le sujet humain, et qu'il lui est donc impossible d'expliciter simplement comment il procède (Labiche et al., 1993).

Figure 1. *exemples de documents*

2.1 Des débuts prometteurs

Dans les années 90, l'Ia quoique en retrait par rapport aux espoirs fous (traduction automatique, diagnostic médical, jeux, …) de ses débuts, reste un domaine en expansion dont on attend beaucoup. L'informaticien, « maître » des algorithmes, doit pouvoir résoudre des problèmes aussi simples que la reconnaissance optique de caractères et de symboles cartographiques connus, ou bien la reconstitution de graphes de traits (type plans EDF ou France Télécom) à partir d'algorithmes de suivi de contours, de croissance de zones, de recherche de formes connexes etc., suivis d'une reconstruction à base de graphes. On dispose de tous les atouts pour développer des systèmes industriels de rétroconversion de plans techniques ou de documents textuels de bonne qualité en utilisant des systèmes à base de règles. Les grandes entreprises pressentant le futur marché des données numériques, mettent en chantier d'ambitieux plans de rétroconversion de documents (BNF, France Télécom, EDF, IGN, DGI,…) dont le but est d'extraire automatiquement les données présentes sur les documents et de les stocker dans de grandes bases de données qui peuvent être topologiques (plans de masse, cadastre, …). L'objectif est bien d'extraire ces données et leurs relations spatiales de manière à pouvoir construire les systèmes d'information géographique (SIG) correspondants. Les algorithmes de traitement d'images sont chaînés de manière à rechercher toutes les informations présentes, la stratégie est la même que celle qui est utilisée dans le cadre de l'analyse de scène (robotique, analyse de flux optique, …), c'est celle que les informaticiens attribuent à la vision humaine : par analogie,

l'information véhiculée par l'image est contenue dans ses pixels comme dans les activités ioniques des cônes et bâtonnets de la rétine pour la vision. Les algorithmes qui imitent la vision tentent de prendre en compte la définition variable du capteur (vision fovéale, vision périphérique) ce qui permet de faire collaborer traitements globaux et locaux, comme dans les rétines artificielles de Zavidovique (Nguyen, 1993) ou bien permet l'analyse séparée des différents plans « couleur » (Mariani, 1997). Avec ce choix du « tout automatique » l'utilisateur n'aurait qu'à configurer le système pour mettre en œuvre une application qui reconnaît, puis stocke toutes les données contenues dans le document. Cette démarche du « tout automatique » perdure et est nécessaire dans beaucoup d'applications industrielles ; elle correspond à des types de problèmes suffisamment contraints et stables pour que les connaissances métier nécessaires puissent être codées « en dur » dans le dispositif, comme lors du traitement automatique d'un type donné de facture pour lequel le taux de correction manuelle pourra être extrêmement faible, ou bien dans le cas de documents issus d'imageurs médicaux pour lesquels les connaissances anatomiques spatiales sont suffisamment génériques pour être formalisées par des relations spatiales et des mesures de distances floues (Bloch, 2003). Cette démarche nécessite une numérisation avec une bonne définition lors de l'acquisition car la reconnaissance nécessite la meilleure précision possible compatible avec le coût et la durée du traitement.

2.2 Interactions plus nombreuses

Les systèmes précédents sont imparfaits, ils doivent être améliorés et pouvoir être réutilisés par modifications ou ajouts de procédures, tout en restant robustes. Par conséquent la maintenance et les développements pour des marchés autres (banques ou postes étrangères par exemple) sont très onéreux. Il apparaît une nouvelle catégorie de projets industriels pour lesquels l'intervention de l'utilisateur devient indispensable pour corriger les résultats ou faire un choix lorsque le système en est incapable.

L'exemple caractéristique de cette approche concerne le traitement des formulaires du recensement de 1999 (Gilloux, 2001) pour lequel l'opérateur valide, ou non, toutes les propositions du système. Pour d'autres dispositifs dotés d'algorithmes d'évaluation des résultats, seuls ceux qui posent problème sont présentés à l'opérateur. Ces nouveaux dispositifs nécessitent une intervention humaine dans la boucle : ce sont des outils logiciels reconfigurables pouvant être chaînés par l'utilisateur final ou le fournisseurs d'outils et de services. Dans ce cadre, l'intervention humaine est celle du concepteur de la chaîne de traitement. Ce concepteur industriel dispose d'outils robustes et paramétrables qu'il doit chaîner pour traiter un problème particulier. Il dispose également d'outils de test et d'interfaces permettant de visualiser et corriger si besoin les résultats intermédiaires. Dans l'entreprise, l'organisation du travail n'est pas fondamentalement remise en cause, les procédures métier restent identiques en leur principe, les utilisateurs finaux doivent accepter l'aide informatique et le maniement d'outils informatisés de plus en plus sophistiqués. Le modèle sous-jacent du système logiciel est celui de la perception. En effet, alors que la vision permet d'extraire des données d'une image projetée sur la rétine, comme l'homoncule en extrait de son théâtre d'ombres chinoises, la perception met clairement en jeu les connaissances de l'utilisateur ou de l'expert qui seules lui permettent de désambiguïser les données.

Le système cognitif humain apporte sa compétence lors de la validation (ou de l'invalidation) des données estimées peu fiables par les algorithmes de classification

comme il le fait lors de la reconnaissance d'images ambiguës. L'intervention de l'utilisateur dans la boucle permet de corriger le système, mais elle doit avoir été prévue dès la conception du système et elle nécessite de pouvoir évaluer les résultats de manière à ne faire intervenir l'utilisateur (expert) qu'à bon escient. Cette problématique met l'accent sur le rôle des utilisateurs selon leur expertise et amène à s'interroger sur les capacités d'interaction des systèmes avec les utilisateurs.

2.3 Perception

Une première évolution des théories de la perception visuelle a consisté à tenir compte de l'aspect dynamique de la perception. En effet une image n'est pas perçue de manière figée, elle succède à d'autres images, comme dans un film; la perception se situe dans le flux d'informations visuelles. Ensuite l'approche perceptive confrontée à la phénoménologie (Husserl, Merleau Ponty, ...) devient la perception « active » ; le retour proprioceptif est essentiel dans le « cycle perceptif », la perception est reliée à la commande du récepteur (les degrés de liberté du système oculaire). La démarche cognitive humaine d'analyse de scène se conclut par une action ou une décision en un cycle répétitif hétérarchique. Les « requêtes d'analyse » qui représentent les intentions du sujet interagissent avec le système perceptif pour sélectionner les attributs utiles et reconstruire (action) les objets correspondants.

De nombreux chercheurs en traitement d'image ont tenté d'imiter cette démarche issue des sciences cognitives pour construire des systèmes logiciels qui opérationnalisent ces cycles perceptifs (Ogier, 1994) et (Ramel, 1998) . Cette opérationnalisation pouvant utiliser un modèle en couches, inspiré du modèle OSI, qui sépare les niveaux conceptuels correspondants aux connaissances mises en œuvre, qu'elles soient applicatives, descriptives ou « traitement d'images ». Ce cycle perceptif permet de réaliser une recherche d'information dans une image de document selon l'intention d'un utilisateur (démarche descendante) exprimée automatiquement (démarche ascendante) en séquencement d'algorithmes qui extraient des données recombinées ensuite automatiquement pour reconstruire les objets recherchés (Ogier, 2000). Les connaissances descriptives ou opératoires sont exprimées sous forme de règles dont l'élaboration nécessite un important investissement de la part des experts.

En conclusion, selon (Trupin, 2003), parce qu'elle n'intègre réellement ni les usages ni les besoins des utilisateurs des documents analysés, cette problématique ne met pas en œuvre une réelle interaction avec les utilisateurs. Il faut doter le système d'information d'une intelligence artificielle (ou d'une cognition) suffisante pour qu'il puisse accéder, à la demande d'un utilisateur, aux connaissances utiles portées par un document. Il faut poser le problème de l'interaction avec l'utilisateur durant le processus de modélisation des connaissances et d'interrogation des systèmes documentaires.

3 Une évolution qui rencontre des limites

La réponse à la nécessité de proposer des solutions informatiques susceptibles de prendre en compte des connaissances, une contextualisation, une mise en situation, poussées pour automatiser au mieux des services dédiés a été une évolution des logiciels informatiques grand public vers des systèmes généralistes. Systèmes que l'utilisateur doit s'approprier pour pouvoir articuler les traitements informatique proposés de manière à automatiser au mieux son système dédié à une tâche totalement définie. Cette approche laisse ouverts deux types de problèmes :

ceux qui concernent la modélisation des connaissances expertes et ceux qui concernent leur utilisation.

3.1 Scénario

Proposer à des utilisateurs non compétents en informatique les plate formes mises au point par les services informatiques pour le développement de produits spécifiques, c'est à dire remplacer les chefs de projet des sociétés de développement par des utilisateurs non professionnels, suppose qu'un apprentissage par l'exemple de conception de scénarios de traitements est possible. Dans ce cadre, d'intéressants travaux ont été menés pour le développement de la plate-forme ACTI VA (Baudouin, 2003), pour laquelle les référentiels de termes retenus pour l'IHM ont été construits a partir des « primitives intentionnelles » recueillies lors d'interviews tutorés d'experts selon une méthodologie basée sur des travaux de Vandervecken sur les actes de dialogue (Vandervecken 97) et inspirée des expérimentations en deuxième personne (Petitmengin, 2001), (Vermersch, 2000). Ce travail propose de modéliser les connaissances, non plus avec des ontologies, mais à l'aide de Bases de Connaissances Terminologiques BCT qui sont plus aptes à modéliser de façon légère des connaissances partagées par une communauté.

On est confronté directement aux problèmes cognitifs, le concepteur du système et l'utilisateur final sont souvent une seule et même personne et c'est à eux de modéliser le problème posé et d'utiliser les modèles sousjacents des outils logiciels.

3.2 Modèles

Améliorer les plate formes pour qu'elles deviennent plus génériques et plus performantes. On se heurte là aussi au problème de modélisation : modélisation de scénarios de traitement, modélisation des documents, modélisation des utilisateurs finaux et de leurs usages... Créer des modèles génériques aisés à instancier pour chaque problème formalisé reste une difficulté majeure lié à celui de la modélisation des connaissances nécessaires, donc à ce que l'on appelle "ontologie" dans la communauté des informaticiens. C'est une problématique difficile qui demande un fort investissement des experts et des concepteurs, et doit être remise en chantier périodiquement.

L'action Spécifique « Document et Organisation » du RTP 33 du CNRS a permis de préciser la notion de « document métier » en privilégiant un point de vue pluridisciplinaire (Sciences de l'information, Sciences de l'ingénieur, Socio terminologie, Sociologie des organisations et des nouvelles technologies). Cette recherche a montré que l'analyse des documents professionnels « en situation » était de fait un problème d'interprétation et qu'il était vain de tenter de modéliser l'ensemble du faisceau d'interactions sociales et culturelles qui interviennent « en temps réel » lors de l'analyse de documents même dans un cadre apparemment fermé (Blanc-Merigot et al., 2004).

Mais une piste intéressante est la perspective interprétative dans laquelle il faut prendre en compte le contexte en accord avec la sémantique différentielle qui permet de concilier la description statique du lexique avec son comportement dynamique dans le cadre de l'interaction (Rastier, 2004).

4 Propositions et travaux

Les apports de l'ingénierie informatique ont été extraordinaires, pour s'en convaincre il suffit de se replacer aux débuts des OCR, des traitements de texte et

aux balbutiements des systèmes de GED... Les apports des approches Markoviennes et des réseaux Bayésiens comme les apports des outils de vectorisation, ont permis de concevoir et réaliser des outils très efficaces, mais la difficulté reste de concevoir les systèmes qui permettent d'utiliser ces outils dans de bonnes conditions, d'autant plus que l'évolution a clairement montré qu'il apparaît un glissement entre les rôles d'utilisateur et de concepteur ; la conception des systèmes dédiés est dorénavant centrale. Les compétences cognitives, que les anciens regroupaient sous le terme d'intellect, doivent être revisitées afin de pouvoir dégager les nouvelles pistes de recherche.

4.1 L'intellect, l'âme, la raison et la substantialité

La rationalité a été considérée comme un des modes possibles de la pensée depuis l'antiquité. Si l'intellect est ce qui permet de concevoir, donc de modéliser, d'interpréter, de créer... , pour nous, il est d'abord constitué de la pensée rationnelle qui correspond à la pensée computationnelle érigée en dogme par les sciences cognitives des années 90. Pour Aristote, l'intellect, capable d'intuition, est supérieur à la raison procédant par le discours.

S'il y a effectivement une différentiation entre intellect et pensée rationnelle, il faut s'intéresser à leur relation pour améliorer l'interaction entre le concepteur-usager muni d'un intellect et le système informatisé disposant d'une forme de rationalité.

Il faut noter qu'au XIIIième siècle, le décentrement « averroïste » du sujet a été interprété comme une séparation totale de la pensée et du sujet ! L'homme ne pense pas quelque chose – l'intellect – se sert de lui pour penser, cité par (De Libera, 1998) et critiqué par St Thomas et l'église.

Pour Aristote, dans le traité De l'âme, l'intellect est réalité substantielle ; vu sa destination à tout connaître, il est de nature « séparée » du corps, c'est-à-dire supérieur au niveau sensible, et donc incorruptible.

Cette question restée ouverte de la dualité corps esprit (et corps social ?), trouvera, après Spinoza (Spinoza, 1667) qui refuse de séparer l'âme du corps, une réponse encore plus radicale, basée sur la corporeïté de la cognition, avec la théorie de l'enaction qui pose également la question de la modélisation de cette cognition ; question qui peut être reliée à l'interrogation d'Aristote : « l'intellect est-il intelligible comme les autres intelligibles ? " cité par (De Libera, 1998).

4.2 Enaction

Le terme "enaction" a été proposé par Francisco Varela (Varela 89) pour désigner un nouveau paradigme[1] en sciences cognitives, basé non pas sur la métaphore de l'ordinateur comme dans le cognitivisme classique, mais sur celle des organismes vivants. La proposition initiale de Maturana et Varela (Cognition = Vie = Autopoïèse) constitue une réponse originale. Les deux autres questions fondamentales abordées par le paradigme de l'enaction sont celles de la relation sujet-objet (objectivisme versus constructivisme) et la relation entre l'expérience vécue à la première personne et la connaissance à la troisième personne. De fait, ce paradigme entretient des relations de parenté notamment avec la phénoménologie, la philosophie de la vie et de l'individuation (Jonas, Simondon), le constructivisme

[1] Tout paradigme en Sciences Cognitives comporte deux éléments majeurs : a/ le noyau théorique qui doit permettre de résoudre le problème de la relation entre matière et esprit et b/ une réelle articulation transdisciplinaire, notamment entre les domaines de la Philosophie, la Psychologie, la Linguistique, les Neurosciences et l'Informatique. (Olivier Gapenne dans "proposition d'ARP STC")

(Piaget, Latour), la Gestalt psychologie, l'approche écologique (Gibson), la robotique autonome…

L'enaction, établie à partir de l'observation du vivant par deux biologistes, Maturana et Varela, repose sur la propriété d'autopoïèse (auto constitution) propre au vivant, et sur l'affirmation que la connaissance est incarnée, donc indissociable du vivant et de l'histoire du sujet pensant. Cette théorie permet d'envisager d'un point de vue formel le couplage structurel entre l'utilisateur et le système via le contexte.

La théorie enactive impose de développer des systèmes laissant la charge de production de sens à l'utilisateur, et de privilégier l'inscription du logiciel dans l'action cognitive de l'utilisateur, plutôt que d'introduire l'intentionnalité de ce dernier dans la machine. Les solutions passeraient donc par la mise en œuvre du couplage structurel entre les processus logiciels et leur environnement, dans lequel ils trouvent la source des perturbations leur permettant de s'auto-organiser (Dionisi, 2006).

Si l'on suit cette approche il n'y a plus de modèle a priori, mais l'utilisation d'un système informatique participe à une expérimentation qui est l'acte de modélisation lui-même.

4.3 Un exemple d'approche enactive (projet AIDé)

Porté par le (groupe NU, 2008) ce travail vise à améliorer les interactions d'utilisateurs avec un système d'information dédié au droit du transport qui repose sur un corpus de textes réglementaires et de compte rendus de jurisprudence. Une perspective ouverte par la théorie de l'enaction est donc la réalisation de systèmes informatiques qui seront des ateliers logiciels permettant aux utilisateurs de mener des expérimentations dans leur domaine d'intérêt et de réaliser ainsi eux-mêmes un logiciel éphémère qui apportera une réponse au problème qu'ils auront ainsi implicitement traité.

Inventer de nouveaux usages des outils de TAL

L'approche herméneutique et énactive dans la conception et l'intégration d'outils de TAL marque une différence de point de vue avec les méthodes classiques en favorisant une démarche scientifique expérimentale. L'expérimentation est mise en avant comme une boucle de conception où la modélisation n'est pas une étape initiale, pas plus que les évaluations (non nécessairement comparatives) ne sont des étapes finales. L'objectif ici est d'inventer de nouvelles façons d'utiliser des outils informatisés dans des recherches sur le langage.

L'interprétation comme "énaction de"

D'un point de vue expérimental, il s'agit de savoir comment un environnement numérique de travail, et le couplage qu'il induit, permettent l'émergence par énaction d'une perception sémantique du corpus et ainsi un meilleur accès aux documents juridiques. Si on considère que le sens provient de la démarche outillée de l'interprétant face à un texte et à son intertexte, alors, une expérimentation mettant en œuvre un environnement numérique de travail qui permet d'effectuer des traitements sur des documents électroniques participe à la co-production de sens pour l'expérimentateur. Dans le couplage personne-système les interprétations des utilisateurs et les traitements des machines ne sont pas en concurrence. Ils doivent être pensés comme complémentaires dans la mesure où l'activité d'une machine a pour objectif de produire dans l'interaction des traces (Laflaquières,

2008), (Cram, 2007) qui vont participer aux interprétations du ou des utilisateurs. Cette démarche est conforme à l'idée de (Dionisi et al, 2006) qui consiste à caractériser des « processus logiciels » impliqués dans des « processus expérientiels », eux-mêmes impliquant des « processus cognitifs ».

D'un point de vue énactif, la cognition résulte de l'histoire et du couplage des diverses actions qu'accomplit un être dans le monde. Il s'agit là, comme le remarque François Rastier (Rastier, 2005) d'un courant de pensée qui, comme l'herméneutique matérielle, ne se présente pas comme une théorie globale mais comme une voie de recherche conduisant à un questionnement permanent des textes et de la place qu'il convient de réserver aux outils informatiques dans le traitement des données.

Pour une démarche terminologique "centrée utilisateur"

Là où le Web Sémantique cherche à rendre le plus possible partagées de vastes ontologies qui synthétisent une connaissance pensée comme objective, nous préférons manipuler des ressources termino-ontologiques (bases de données terminologiques, représentations du contenu lexical etc.) propres à un utilisateur ou un petit groupe d'utilisateurs et liées à leur tâche, leurs besoins et de leurs centres d'intérêt. Il en découle une certaine légèreté sémantique de ces ressources, au sens de (Perlerin, 2004), dans la mesure où elles ne représentent que ce qui est important du point de vue de l'utilisateur et restent ainsi de taille raisonnable.

La tradition logico-grammaticale et plus précisément la sémantique formelle et computationnelle cherchent à représenter et à produire, automatiquement ou pas, des formes le plus possible objectivées des significations et du sens. Dans la démarche centrée utilisateur, on considère que les traitements sémantiques appliqués à l'accès aux contenus des documents ont tout à gagner à être le plus possible subjectivés, tant du point de vue des ressources que du point de vue des résultats opératoires.

Il est remarquable que la posture énactive rejoigne néanmoins une méthode du domaine du web sémantique « centrée communauté d'utilisateurs », celle qui est apparue dans le cadre des folksonomies (Pedauque, 2006), (Mathes, 2006) ; il sera intéressant de les relier plus avant. La théorie de l'enaction apparaît alors comme une théorisation possible de l'approche cognitive sous-jacente.

4.4 Approche autopoïétique ?

Le développement du paragraphe précédent montre une des directions possibles de recherche basée sur la théorie de l'enaction. En effet la charge cognitive durant l'expérimentation est laissée à l'utilisateur (qui est aussi en partie concepteur du système), le système informatique ne faisant que proposer, si possible à bon escient, des outils , des paramètres, des modes de présentation de données, des séquencements d'outils, ... Une étape supplémentaire sera franchie lorsque le système pourra évoluer de lui-même en générant de nouveaux outils, paramètres, séquencements, présentations au cours d'un véritable dialogue avec l'usager-concepteur-expérimentateur. Il est déjà possible que le système fasse lui même des propositions, pourquoi pas issues de librairies d'outils disponibles dans la communauté des chercheurs du domaine ou sur le web ? mais il reste la difficulté de l'interopérabilité !

Quant au système capable de s'auto adapter, s'auto configurer, et pourquoi pas de s'auto générer, il reste à développer...

Tout en notant la remarquable progression des jeux interactifs, d'abord en synthèse et animation d'images dans les années 80, puis plus récemment dans

l'adaptation de scénario ludiques puis enfin dans l'auto-génération de jeux (ou ce qui semble être de l'auto génération).

Si le problème de la confidentialité pouvait être résolu, les concepteurs de jeux auraient certainement beaucoup à dire...

5 Perspectives et conclusion

Le seul modèle acceptable de la réalité, disait T.D. Ross, le créateur de SADT, c'est la réalité elle-même...

Mettre un accent particulier sur le caractère situé de la connaissance est nécessaire, cela ne peut résulter que de l'amélioration de l'interactivité des systèmes comme on a pu les constater depuis de nombreuses années. Une piste intéressante semble être le paradigme de l'énaction.

Le « tournant énactif » que l'on peut replacer au sein d'une démarche pragmatique et expérimentale donne un cadre général pour théoriser ces approches nouvelles. En remettant en cause les analyses classiques du fonctionnement cognitif humain il permet de reconsidérer les aspects fondamentaux de la perception en rendant indissociables perception et action en cours d'expérience.

Le domaine de l'interprétation d'images de documents se prête particulièrement bien aux expérimentations à mener car l'interprétation ne peut se faire qu'en situation, le contexte y est toujours prégnant et elle est liée à la « perception-action».

Une posture qui pourrait se révéler efficiente consiste alors à construire des ponts :

- entre le modélisable et le non modélisable, c'est-à-dire entre le computationnel et le non computationnel ; modélisons ce qui peut l'être en respectant la possibilité d'interaction avec ce qui ne peut pas l'être. Construisons des modèles avec lesquels on peut interagir, c'est à dire modèles qui peuvent être modifiés aisément par l'utilisateur-concepteur (cf proxidoc), par le système informatique lui-même, ou qui peuvent s'auto-modifier ;

- entre des temporalités, celle du présent immédiat et celle des ruptures entre ces micromondes (petits domaines de connaissance selon F. Varela) ; donc temporalités des apprentissages acquis (actions incarnées- énaction d'un monde) et temporalités des analyses rationnelles expertes. Les outils algorithmiques (rationnels), doivent interagir avec l'utilisateur-concepteur (cf le projet AIDé). La rationalité concerne aussi bien le système informatique lorsqu'il met en œuvre des outils bien maîtrisés, que l'utilisateur-concepteur lorsqu'il raisonne froidement. La non rationalité concerne le système informatique lorsqu'il met en œuvre des outils nouveaux
mal maîtrisés (ou apparaissant comme non maîtrisés par exemple dans le cas d'un pilotage par système multi agents), ou lorsqu'il s'auto génère comme elle concerne l'utilisateur-concepteur lorsqu'il utilise son expertise ou son imagination.

Ces ponts doivent permettre qu'un véritable dialogue s'institue entre les différentes entités :

- chacune peut interrompre l'autre ;
- chacune peut refuser l'interruption de l'autre ;
- chacune peut progresser seule.

La progression de l'expérimentation doit pouvoir échapper à tout contrôle ! Le système doit pouvoir échapper aux utilisateurs qui doivent également pouvoir échapper au système.

Il n'y a pas de réussite programmée lors d'une expérimentation, mais seulement une progression par essai-erreur.

La mémoire du comportement de l'ensemble système-utilisateurs sera alors constituée par les outils de traçage mis en place par le concepteur ; traces entre actes de dialogue, entre sessions concepteur et session utilisateur, traces formalisées (Cram, 2007), (Laflaquières 2008) dans lesquelles on doit pouvoir naviguer si besoin et seulement si besoin.

Les théories utiles pour le développement des dispositifs expérimentaux seront certainement l'enaction, mais aussi l'herméneutique matérielle ; les technologies seront celles qui sous tendent les outils du domaine, mais aussi l'ingénierie des traces et l'ingénierie des folksonomies, la technologie des jeux, ...

Références ;

Baudouin, N., Holzem, M., Saidali, Y., Labiche, J. (2003). Acquisition itérative de connaissances en traitement d'images : consultation d'un collège d'experts; *Plate-forme AFIA, Conférence Ingénierie des Connaissances* IC2003, Laval Fr, pp 101-116, juillet.

Bertin, J. (1967). *Sémiologie Graphique. Les diagrammes, les réseaux, les cartes.* La Haye, Mouton, Gauthier-Villars,Paris. 2e édition : 1973, 3e édition : 1999 (complétée de la « Théorie matricielle de la graphique », EHESS, Paris).

Blanc-Merigot, M., et al. (2004). *Document et Organisation*, sous la direction de Maryvonne Holzem et Jacques Labiche. Europia Editions Paris, ISBN 2-909285-29-4; 101 p.

Bloch, I., Géraud, T., Maître, H. (2003). Representation and fusion of heterogeneous fuzzy information in the 3D space for model-based structural recognition – application to 3D brain imaging. *Artificial Intelligence*, 148:141-175, Août 2003.

De Libera A. (1998) *Averroès, L'intelligence de la pensée ; Sur le De anima*, Présentation et traduction de Alain de Libera, p 21, édition GF Flammarion, Paris. et réf. p 36 à Aristote De anima Livre III chapitre 4, 429b22-430a9.

Dionisi, D. (2006). *Proposition d'une méthodologie d'opérationnalisation informatique de l'approche enactive de la cognition* Thèse de doctorat de l'INSA de Rouen.

Dionisi, D., Labiche J. (2006). "Enaction et informatique : les enjeux de l'opérationnalisation technologique d'une théorie de la cognition » ; In *COGNITICA Actes du Colloque de l'Association pour la Recherche Cognitive* ARCo, p 97-110, Bordeaux, décembre.

Gilloux M. (2001) Une application industrielle de numérisation et de lecture automatique de documents : la saisie des questionnaires du recensement de la population de 1999. , *4e Colloque International sur le Document Électronique*, CIDE 2001, Toulouse, 24-26 octobre.

Groupe NU. (2008) Conception et usages d'un environnement numérique de travail pour une aide à l'interprétation de documents juridiques, *Colloque International sur le document électronique*, CIDE 2008, Rouen, Octobre.

Labiche, J., Ogier, JM., Balan B., Caston J. (1993). Stroop effect : an example of conflict analysis. *IEEE-SMC'93 Vol3, Le Touquet France, 458-462,* Octobre.

Laflaquière, J., Prié, Y., Mille, A. (2008) Ingénierie des traces numériques d'interaction comme inscriptions de connaissances. to appear in *Ingénierie des Connaissances 2008*, june.

Mariani, R. (1997). *Contribution à la lecture automatique de cartes*. Thèse de doctorat de l'Université de Rouen.

Mathes, A. *Folksonomies. Cooperative Classification and Communication Through Shared Metadata.*
http://www.adammathes.com/academic/computer-media

Cram, D., Jouvin, A., Mille A. (2007) Visualisation interactive de traces et réflexivité : application à l'EIAH collaboratif synchrone eMédiathèque. *STICEF*, (Numéro spécial Analyse des traces d'interactions dans les EIAH) 14().

Nguyen, P., Bernard, T., Zavidovique, B. (1993). Cisc Rétina vs Risc Retina, Camp' 93, USA (Nouvelles Orléans), Décembre.

Ogier, JM. (1994). *Contribution à l'analyse automatique de documents cartographiques : interprétation de données cadastrales*. Thèse de Doctorat Université de Rouen

Ogier, JM., Mullot, R., Labiche, J., Lecourtier, Y. (2000) The semantic coherency : the basis of an image interpretation device – application to the french cadastral map interpretation; *IEEE Trans on System Man and cybernetics part B*, 30(2); pp 322-338.

Pédauque, RT. (2006). *Le document à la lumière du numérique*. C&F Editions. Caen

Perlerin, V. (2004) *Sémantique légère pour le document* Thèse d'informatique Université de Caen

Petitmengin C. (2001) *L'expérience intuitive*, l'Harmattan, Paris

Ramel JY., Vincent, N., Emptoz, H. (1998) Interprétation de documents techniques par « cycles perceptifs » à partir d'une perception globale du document. *Traitement du Signal*. Presse universitaire de Grenoble. Vol. 15. No. 2. p. 1-20.

Rastier F. (2004) *Enjeux épistémologiques de la linguistique de corpus* ; Texto.disponible à
www.revue-texto.net//Inedits/Rastier/Rastier Enjeux.html

Rastier, F (2005) *Sémiotique du cognitivisme et sémantique cognitive : questions d'histoire et d'épistémologie*. Texto, [en ligne] mars 2005 (consulté le 12 févier 2008)

Spinoza, B. (1677) *L'éthique*, trad. de A. Guerinot, Ivrea, 1993

Trupin E. (2003) *De la reconnaissance automatique d'images de documents*. Habilitation à Diriger des Recherches, Université de Rouen.

Vanderveken D. (1997) Formal Pragmatics of Non Literal Meaning - in *Linguistiche Berichte*, volume 8: Pragmatics.

Varela F.J. (1989) *Autonomie et connaissance*, traduction Paul Dumouchel et Paul Bourgine Seuil, Paris.

Vermersch P. (2000), « Conscience directe et conscience réfléchie », *Intellectica* 2000/2: 31, pp. 269-311

Des « données » aux documents

François Rastier (1)

(1) CNRS-ERTIM, INALCO, Paris

Résumé : Le programme du Web sémantique entend remplacer le « Web des documents » par le « Web des données » et prolonge ainsi le programme classique de la représentation des connaissances. En revanche, pour une sémantique du Web inspirée par la linguistique de corpus, les connaissances résident dans les textes et les documents qui les véhiculent, et ne peuvent en être abstraites sans perdre leur valeur contextuelle et leur pertinence. Cela conduit à recontextualiser la notion même de donnée, ainsi qu'à problématiser le rapport entre données et métadonnées Cette étude invite à un remembrement de la tripartition de fait entre discours, texte, et document. Elle s'appuie pour cela sur la sémantique de corpus et la philologie numérique. Il s'agit en effet, de revenir des « données » aux documents et d'exploiter pour la recherche d'informations leur irremplaçable complexité.
Mots-clés : texte, document, passage, complexité, donnée, métadonnée, connaissance, pertinence.

Abstract: From « data » to documents. The Semantic Web program aims to replace the "Web of Documents" by the "Web of Data", thus prolonging the classical programme of knowledge representation. In contrast, a corpus-linguistic inspired Web Semantics/ situates knowledge within texts and the documents that convey them. Data cannot therefore be abstracted without losing their contextual valeur and pertinence. This leads to a recontextualisation of the notion of "data" and a rethinking of the relationship between data and metadata.
This study seeks to realign the triangular division which exists between discourse, text and documents. To further this end, it relies on corpus semantics and digital philology. In contrast to the Semantic Web programme, the goal here is to move beyond "data" back to documents, and to make use of their irreplaceable complexity in order to find specific information.`
Keywords: semantics, data, metadata, knowledge, keyness.

1 Ambitions et crédibilité du Web sémantique

On sait que le Web fonctionne d'après trois standards : le protocole HTTP, l'adressage par les URL et le langage HTML. Tim Berners-Lee, qui dirige le W3C, instance qui préside aux destinées du Web mondial, a présenté depuis 1994 le Web sémantique comme une extension du Web qui le transformerait en un espace d'échange de documents permettant d'accéder à leurs *contenus* et à effectuer des *raisonnements*. Cela exigerait une représentation du contenu des documents par des ontologies pourvues d'une sémantique dénotationnelle (le Web sémantique n'en reconnaît pas d'autre) ; l'ensemble des contributeurs au Web sémantique, et bientôt l'ensemble de ceux qui mettent en ligne des contenus, doivent donc respecter une

infrastructure commune figurée d'abord par le fameux « cake » de Tim Berners-Lee, présenté à la conférence XML 2001. Cette infrastructure est aujourd'hui jugée effectivement normalisée jusqu'au niveau des ontologies : elles fournissent notamment le vocabulaire de ces métadonnées pour représenter le contenu des documents de la même manière qu'un thésaurus, composé de termes (ou concepts) et non de mots. Notons bien qu'au-dessus du deuxième niveau, on perd ainsi tout contact avec les textes et les langues, en passant à des langages (« formels ») de représentation. En promouvant le *Web sémantique*, le W3C entend remplacer le « Web des documents » par le « Web des données » (cf. Tim Berners-Lee, 2007). En utilisant des ontologies, il s'agit de s'affranchir de la complexité des documents et de leur diversité linguistique et sémiotique.

En accord avec l'objectivisme de la philosophie du langage issue du positivisme logique, la donnée est alors conçue comme une simple chaîne de caractères (ex. la donnée *pêche*, qui peut être reliée soit à *poisson*, soit à *fruit* ; Berners-Lee, *loc. cit.*). Il serait discourtois d'insister sur l'indigence banale de cette conception des données.

Rassurantes, car présentées comme purement pratiques, les recommandations du W3C ont vocation à devenir des standards. Or, l'adoption de standards « de bas niveau » comme HTML, ou Unicode voire XML n'entraîne aucunement que l'on doive ériger en standard des langages de représentation comme RDF ou OWL, sauf à céder benoîtement à la tentative de coup de force du W3C en faveur du « Web sémantique ».

Il serait plus discourtois encore d'insister sur les enjeux économiques du Web sémantique : on comprend parfaitement que le Département US du Commerce soutienne le Web sémantique, car la normalisation des contenus accessibles sur le Web se concrétise par des ontologies généralement faites de mots anglais écrits en majuscules, et réputées toutefois représenter des « métadonnées » permettant d'accéder aux documents.

Les standards de métadonnées sont l'un des trois éléments clés de la Stratégie données en *réseau centré* (*Net-Centric Data Strategy*) du Département de la Défense des États-Unis, arrêtée en décembre 2001 et rendue publique en mai 2003 (cf. Stenbit éd., 2003). Rendant obligatoires certains types de métadonnées, cette stratégie consiste à contrôler un réseau définitoirement non centré : « to ensure that all data are visible, available, and usable », pour en finir décisoirement avec le Web caché qui échappe au contrôle. Par ailleurs, « all posted data will have associated metadata » (Stenbit éd., 2003, Avant-propos, p. 4), de manière que le volume des données privées soit drastiquement réduit (divisé par deux, alors que les données communes seraient multipliées par trois ; cf. p. 10). La centralisation du réseau ainsi réalisée (pour réaliser la *Net-Centricity*) permet alors « a completely different approach to *warfighting and business operations* » (Appendice A, p. 2, mes italiques).

On comprend ainsi que les soutiens militaires n'aient pas manqué depuis septembre 2001 : le Web sémantique se veut en effet collaboratif, chaque fournisseur de contenu devant mettre en ligne ses bases de données selon un format unique qui les rendra interopérables et permettra par là-même d'y accéder — par exemple, pour découvrir de nouveaux médicaments (selon Tim Berners-Lee, 2007). L'intelligence, au sens économique et militaire du terme, a tout à gagner à cette transparence coopérative.

Les enjeux politiques et économiques ne doivent pas faire oublier les conséquences épistémologiques de ce programme. On peut s'interroger sur la cohérence du « cake », qui graphiquement évoque les délices étagés de la tranche

napolitaine : il s'agit vraisemblablement d'une simple juxtaposition éclectique, mais cet éclectisme est orienté par les objectifs dont le statut scientifique reste douteux. Quiconque est un peu frotté de sémiotique visuelle aura en effet reconnu dans les gradins du « cake » les marches d'un *gradus ad Parnassum*, qui nous conduit d'*Unicode* à *Trust* (au sens de *confiance*, comme dans *In God we trust*, plutôt que de *monopole*). Bref, on édicte des standards, puis on les érige au rang de modèles théoriques, ce qui est caractéristique de la technoscience, non seulement instrumentaliste, mais instrumentalisée. Sir Tim Berners-Lee est ingénieur ; en proclamant opportunément en 2007 la formation d'une *Web Science*, Tim Berners-Lee évite cependant que les problèmes scientifiques soient posés et débattus hors de la communauté du Web sémantique, qui s'est auto-engendrée et se doit aussi de s'auto-évaluer.

Laissons les analystes futurs s'interroger sur l'unité de pensée entre le W3C et le *Department of Defense*. Internet est né d'une contradiction toute militaire entre sécurité du réseau et contrôle des informations. Le réseau devait être distribué pour pouvoir résister à toute tentative de destruction ; mais son succès même et son extension à l'économie et aux données privées l'a rendu difficile à contrôler. Or tout Appareil, économique ou militaire, et le *Department of Defense* n'est qu'un exemple éminent, se doit de maintenir une hiérarchie pour exercer son pouvoir et constituer sa légitimité : il projette donc nécessairement sa structure sur le monde qui l'environne, et nous avons déjà souligné par exemple la parenté théorique, au sens le plus métaphysique, entre les ontologies et les organigrammes (l'auteur, 2004b). Le « cake » du W3C restitue à sa manière une hiérarchie de hiérarchies (les ontologies, les DTD XML, etc.) et permet de subsumer par différents niveaux de métadonnées puis de « données » la diversité incontrôlable des documents, comme des langues et des systèmes de signes qu'ils mettent en jeu.

Aussi le titre de cette étude met-il en scène de manière trop simple deux conceptions différentes qui ne s'opposent pas directement et ne sont pas commensurables. Ne nous attendons pas à un combat de David contre Goliath : le Web sémantique est un programme politico-technique, alors que la sémantique du web est un projet méthodologique et un domaine d'applications fondées sur une sémiotique des corpus.

S'il bénéficie de soutiens influents, le Web sémantique rencontre aussi un facile assentiment, car il concrétise un ensemble de conceptions reçues qui appartiennent à la tradition de l'Intelligence artificielle classique et que nous allons questionner.

2 Redéfinir la concept de donnée

Rien ne nous est donné. La notion de *donnée* prend toutefois un relief particulier si l'on s'avise qu'en promouvant le *Web sémantique*, le W3C, instance qui préside aux destinées du Web mondial, entend remplacer le « Web des documents » par le « Web des données » (cf. Tim Berners-Lee, 2007). En utilisant des ontologies, il s'agit de s'affranchir de la complexité des documents et de leur diversité linguistique et sémiotique. En accord avec l'objectivisme de la philosophie du langage issue du positivisme logique, la donnée est alors conçue comme une simple chaîne de caractères (ex. la donnée *pêche*, qui peut être reliée soit à *poisson*, soit à *fruit* ; Berners-Lee, *loc. cit.*). Nous proposerons ici un modèle moins sommaire de la donnée, qui tienne compte de la dualité sémiotique irréductible entre expression et contenu, ou plus généralement entre *phore* et *valeur*. Cela s'étend à toute chaîne de caractères, du signe de ponctuation au chapitre, sans égard pour le modèle

apocryphe du signe prêté à Saussure par les rédacteurs du *Cours de linguistique générale* et contredit par les écrits autographes.

La dualité phore/valeur, qui constitue le corps sémiotique de la donnée, se trouve sous la rection d'une dualité de rang supérieur entre le *point de vue* et le *garant*. Le point de vue n'est pas un simple point d'observation : il est déterminé par une pratique et un agent individuel ou collectif ; dans un traitement de données, il dépend donc de l'application. Le garant est l'instance de validation qui fonde l'évaluation de la donnée : cette instance est une norme sociale qui peut être juridique, scientifique, religieuse ou simplement endoxale. En linguistique de corpus, le garant est l'autorité qui a présidé à la constitution du corpus ; certaines métadonnées documentaires, comme l'auteur ou l'éditeur, relèvent de cette instance.

Le point de vue est « subjectif » dans la mesure où il est occasionnel ; le garant, « objectif » dans la mesure où il est constitutionnel ou du moins constituant. La dualité du point de vue et du garant définit deux régimes de pertinence, saillance pour le point de vue et prégnance pour le garant. Puisque les données sont bien ce qu'on se donne, elles sont ainsi les résultats initiaux d'un processus d'élaboration — et leur traitement produit des résultats ultérieurs, dans un cycle susceptible de récursivité.

Dans les termes de la sémiotique des zones anthropiques (cf. l'auteur, 1996, 2001a, 2002), le corps (phore+valeur) de la donnée, en tant qu'elle est objectivée, relève de la zone proximale de l'environnement ; le point de vue, de la zone identitaire ; enfin, le garant, de la zone distale où se situent les instances de normativité. L'axe sur lequel se répartissent ces zones est celui de la *médiation symbolique*, alors que l'axe subordonné qui relie le phore et la valeur relève de la *médiation sémiotique* (cf. l'auteur, 2001a). Soit en bref :

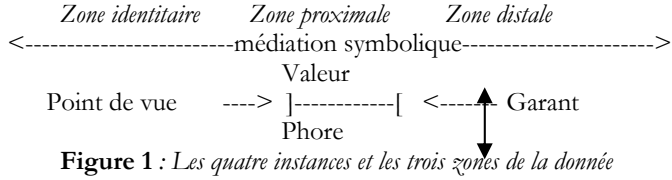


Figure 1 *: Les quatre instances et les trois zones de la donnée*

En ne percevant pas le caractère instituant de la valeur, du point de vue et du garant, en réduisant la donnée à la seule instance du phore, le positivisme ordinaire élude toute dimension critique et épistémologique. Recueil de données ainsi appauvries, un « corpus » sans point de vue ni garant n'est pas véritablement un objet scientifique mais un amas numérique inexploitable en tant que tel ; ainsi des pseudo-corpus recueillis par aspiration aléatoire de sites.

3 Une conception du texte issue de la sémantique interprétative

Les modèles issus de la linguistique textuelle dépendent largement de la tradition logico-grammaticale, qu'il s'agisse de théories macrosyntaxiques, de modèles propositionnels arborescents (van Dijk et Kintsch) ou simplement consécutifs (Kamp, Asher), voire de séquences discrètes successives (Adam)[1]. À ces

[1] Pour compenser le laconisme de cette section, nécessairement elliptique, nous nous permettrons de renvoyer à des publications antérieures ; sur les modèles du texte, cf. 2008 ; sur l'interrelation des composantes textuelles, l'auteur, 1989, 2001 b ; sur les passages, 2003, 2007 ; sur les variations du lexique selon les genres et discours, 2004.

modèles hiérarchiques ou séquentiels nous préférons des représentations hétérarchiques pour la structure architectonique et rhapsodique pour la structure compositionnelle.

On confond souvent aujourd'hui les modèles théoriques et les formats de représentation, voire les modes d'implémentation. En raison de la complexité des textes, il reste indispensable d'élaborer des conceptions spécifiques du texte et du corpus issues de la linguistique et de la philologie qui soient fondées sur une théorie des performances sémiotiques. La question des modèles devient alors subsidiaire, si l'on s'avise que les « modèles du texte » restent des schématisations partielles qui ne permettent pas de rendre compte de la complexité constituante des textes.

Hétérarchie des composantes textuelles. — À la différence de la linguistique textuelle qui privilégie les modèles hiérarchiques, la sémantique interprétative présente le contenu du texte comme une hétérarchie de composantes sémantiques (thématique, dialectique, dialogique, tactique). Le plan de l'expression est décrit par d'autres composantes (médiatique, etc.). Le genre se définit par un type d'interaction entre composantes au sein des deux plans du contenu et de l'expression, ainsi qu'entre ces deux plans : elles norment ainsi la *sémiosis textuelle.* Alors que la sémiosis au palier du mot reste trivialement problématique (en raison des faux problèmes induits par la polysémie, la synonymie, etc.), la sémiosis des paliers supérieurs comme le paragraphe dépend de la sémiosis textuelle telle qu'elle est normée notamment par le genre ; par exemple, *amour* n'a pas le même sens en poésie et dans le roman et n'a pas de cooccurrents communs dans ces deux corpus.

Passages et problématisation des unités. — Pour rendre opératoire le concept de *forme sémiotique* et permettre l'extraction assistée de telles formes, il convient de revenir sur la notion d'unité textuelle. Dans la perspective néo-saussurienne qui est la nôtre, les grandeurs ne sont pas des unités empiriquement constatables ou des « données » d'évidence. Elles ne sont pas des unités simplement isolables car discrètes ni déterminables à un seul plan d'analyse (du contenu ou de l'expression). Les signes et plus généralement les grandeurs sémiotiques sont construits dans l'interprétation qui leur assigne leur valeur. Si l'ontologie logico-grammaticale attribue aux grandeurs textuelles la discrétion et la présence, l'identité à soi et l'isonomie, sans doute à l'image naïve des objets physiques, la conception rhétorique / herméneutique dont nous nous inspirons admet en revanche que les grandeurs qu'elle construit soient continues, parfois implicites, varient dans le temps et selon leurs occurrences et leurs contextes, connaissent entre elles des inégalités qualitatives et ne relèvent pas uniformément des mêmes règles.

Le texte échappe au modèle du signe : les grandeurs sémantiques textuelles n'ont pas de signifiants uniformément isolables comme des parties du discours ; elles sont constituées par des connexions de signifiés et d'expressions des paliers inférieurs de la période, du syntagme, de la sémie. L'articulation entre l'expression qui détermine l'identité de la grandeur et le contenu qui détermine sa valeur s'opère au sein du *passage*, lieu de la sémiosis locale. Dans la perspective interprétative, cette grandeur locale correspond indifféremment à un signe, à une phrase, ou par exemple à un paragraphe. Au plan du signifiant, le passage est un *extrait*, entre deux blancs s'il s'agit d'une chaîne minimale de caractères ; entre deux pauses ou ponctuations, s'il s'agit par exemple d'une période. Au plan du signifié, le passage est un *fragment* qui pointe vers ses contextes gauche et droit, proches et lointains.

Le passage renvoie aux étendues contiguës ou plus lointaines. L'extrait peut renvoyer aux étendues connexes, par exemple par des règles d'isophonie ou de concordance de morphèmes : ce sont des *cooccurrents* expressifs. Le fragment se relie

à d'autres par des phénomènes d'isotopie : ils ont le statut de *corrélats* sémantiques. Pour ce qui concerne leur connectivité externe, on distinguera l'*incidence* de l'extrait et la *portée* du fragment. Un extrait peut être conventionnellement isolé, car les structures de l'expression relèvent pour l'essentiel de la mésolinguistique ; en revanche, un fragment ne peut l'être sans perte, car les structures du contenu sont macrolinguistiques.

La sélection d'un passage et *a fortiori* l'isolation d'un " signe" exigent deux opérations : faire l'hypothèse qu'à un extrait minimal correspond un fragment, de façon à pouvoir les isoler ; puis, en les décontextualisant, leur assigner un rapport terme à terme entre signification et expression qui littéralise la première et fixe la seconde.

Chaque interprétation isole, construit, analyse et hiérarchise des passages, sur le modèle du commentaire. Elle les recontextualise en elle, tout en permettant d'accéder à sa source : dans la dualité où le texte commenté revêt la fonction de garant, et où le commentaire concrétise le point de vue qui préside à la description, la donnée construite qu'est le passage sert de médiation objectivante entre le texte et sa description.

Quand il est mal choisi, le passage remplace la complexité par l'indétermination (ainsi celle du mot isolé) ; en revanche, le « bon » passage témoigne de la complexité locale et permet de pointer vers la complexité globale.

Des méthodes statistiques[2] permettent à présent de proposer, pour chaque passage, des *cooccurrents* expressifs de l'extrait qui restent à qualifier comme des *corrélats* sémantiques du fragment.

Plan du contenu

⊃ frag. corrélat $_1$⊂ ⊃ *fragment* ⊂ ⊃ frag. corrélat $_n$⊂

⊃ ext. cooccurrent $_1$⊂ ⊃ *extrait* ⊂ ⊃ ext. cooccurrent $_n$⊂

Plan de l'expression

Figure 2 *: Le passage et ses contextes*

Les corrélats d'un fragment sont d'autres fragments ; les cooccurrents d'un extrait, d'autres extraits ; les relations entre passages intéressent ainsi tant la textualité que l'intertextualité.

Notons que le passage n'a pas de bornes fixes et son empan dépend évidemment du point de vue qui a déterminé sa sélection. Sa définition s'écarte donc de l'objectivisme traditionnel de la tradition logico-grammaticale. Redéfinir le signe comme un passage conduit à s'éloigner de la logique des « idées » et des représentations, pour en élaborer une conception purement relationnelle et donc contextuelle. La relative clôture organisationnelle du passage se traduit par le fait que les relations au sein du passage sont plus denses et sémiotiquement plus fortes que les relations entre passages. Le rapport entre global et local va du texte au

[2] Quand il s'appuie sur des corpus de textes appartenant au même genre et au même discours que le texte analysé, le test de l'écart réduit permet de repérer des groupements de cooccurrents qui sont de bons candidats pour la constitution de passages (cf. la fonction Thème du logiciel Hyperbase, obligeamment ménagée par Étienne Brunet). Enfin, la thèse de Mauceri (2007) ouvre des perspectives fort intéressantes.

passage : le passage est une *zone de localité*, définie par une sémiosis propre (mode d'appariement entre contenu et expression) et, sur chacun de ses plans (fragment et extrait) par des relations contextuelles internes fortes.

Ainsi les « données textuelles » peuvent-elles être qualifiées comme des passages, fussent-ils de petite taille, comme les lexies. La recherche de passages en fonction de régimes de pertinence objective (liée au genre et au discours) ou occasionnelle (liée à un type de requête ou d'application) devient à présent pour la linguistique de corpus une question empirique [3]. Le passage relève du modèle général de la donnée textuelle, définie par quatre postes (cf. l'auteur, 2008), un signifiant, l'extrait ; un signifié, le fragment ; un point de vue, celui qui préside à la description ou à l'application ; une garantie (permise par l'établissement du texte et la constitution du corpus. Soit :

point de vue (applicatif) > ---- fragment PASSAGE extrait ---- < garantie (philologique)

pertinence subjective > saillance < pertinence objective

Figure 3 *: Le passage comme donnée qualifiée*

Formes sémiotiques et métamorphismes [4] .— Dans l'hypothèse de la *perception sémantique*, l'opposition entre fond et forme qui détermine le traitement de l'expression vaut aussi pour le plan du contenu. Les fonds sémantiques sont des isotopies et faisceaux d'isotopies. Les formes sémantiques, comme les thèmes, les acteurs, les foyers énonciatifs, s'apparient avec des formes expressives (périodes) pour constituer des *formes sémiotiques*.

Cet appariement ne va pas de soi et suppose une interprétation : quand on utilise des méthodes lexicométriques, les groupes de *cooccurrents*, qui, en tant que chaînes de caractères, relèvent du plan de l'expression, doivent être qualifiés comme des *corrélats* sémantiques. L'interprétation locale est ainsi constitutive des signes et dépend du régime herméneutique global propre au genre et au texte.

Dans la conception morphosémantique du texte, les transformations se spécifient en *métamorphismes* (changements de forme), *métatopies* (changements de fond) et *transpositions* (changements des rapports entre forme et fond : par exemple, une forme peut se diffuser dans un fond). Les transformations des formes sémantiques et des formes expressives se manifestent par des changements de contexte, comme par des modifications corrélatives du rapport entre contenu et expression.

Le principe des transformations est analogue pour le contenu et pour l'expression. Ces deux plans étant solidaires en raison du principe même de la sémiosis, qui s'étend à des relations de contextualité entre plans, toute transformation sur un plan s'accompagne d'une transformation sur l'autre.

Rompant avec l'ontologie prégnante dans la tradition logico-grammaticale, la conception du texte comme série de transformations relève d'une *praxéologie* : non seulement tout texte est inscrit dans une pratique, mais cette pratique s'inscrit en lui, car il est produit par une activité constante de réécriture qui assure sa cohésion. La

[3] Cf. l'auteur, 2008.

[4] Dans les paragraphes qui suivent nous reprenons des éléments de l'auteur (2001b, 2007).

textualisation se définit alors comme un *cours d'action* que les parcours interprétatifs peuvent avoir ultimement l'ambition de restituer.

4 Des ontologies aux connaissances textuelles

Les débats. — Dans le domaine de la représentation des connaissances, on a élaboré des formalismes de représentation considérés comme adéquats, dans la mesure où ils conviennent à des applications peu ambitieuses. L'adoption de standards comme XML ou RDF permet une interopérativité de principe, mais ne résout pas le problème de la production, de l'identification et de l'évolution des connaissances.

Les débats portent en amont sur le problème de la réification des connaissances hors des contextes d'utilisation, ou complémentairement sur l'adéquation de leurs modes de représentation à leur utilisation effective.

Les tenants de la position réifiante s'appuient sur l'essor des ontologies, qui radicalisent la préconception objectiviste des connaissances. Les ontologies restent des *thésaurus* – celui de Roget a d'ailleurs explicitement servi de modèle à Miller et à ses collaborateurs pour WordNet. Elles en gardent les inconvénients notoires : une généralité qui ne leur permet pas de s'adapter aux points de vue sélectifs exigés par les tâches et un manque d'évolutivité qui exige une maintenance manuelle. Elles réduisent la langue à une nomenclature, qui ne rend compte ni des structures textuelles, ni des variations considérables de genres et de discours.

Même dans le domaine du Web sémantique, pourtant très lié aux ontologies, la perspective centrée sur les utilisateurs conduit à des constats résignés comme celui-ci : « Semantic Web researchers accept that paradoxes and unanswerable questions are a price that must be paid to achieve versatility » (Berners-Lee et coll., 2001). La variété des points de vue des utilisateurs et des régimes de pertinence propres à leurs tâches leur interdit de se satisfaire d'une norme unique au demeurant arbitraire : mais l'absence de contradiction reste un postulat absolu des ontologies, conformément aux lois d'identité, de non-contradiction et de tiers exclu qui fondent leur conception logiciste du monde.

Plus radicalement, les tenants de la cognition située et les ergonomes spécialisés en recherche d'information insistent sur la diversité imprévisible des applications et sur le fait que les formalismes ne sont que des supports à des parcours d'interprétations. Dès lors, les facettes définitoires d'un objet, quel qu'il soit, ne peuvent être fixées *a priori* : en d'autres termes, ce sont les pratiques qui définissent les propriétés pertinentes des objets.

Cette divergence peut aujourd'hui être tranchée empiriquement. En effet, l'étude de grands corpus, y compris techniques, a montré que les relations sémantiques qui organisent les ontologies différaient selon les discours et les domaines, au point que certaines relations sémantiques de base sont tout bonnement absentes de certains corpus pourtant étendus (cf. projet Safir conduit par un consortium Crim-Lip6-Edf).

Par ailleurs, l'expérience de Wordnet et EuroWordnet est instructive : ces ontologies se sont révélés inutilement complexes. Fondées sur les postulats psychologiques datés de Miller et Johnson-Laird (1976), elles ignorent des savoirs linguistiques élémentaires comme la notion cruciale de *morphème*, ce qui conduit à créer des sous-réseaux distincts pour les noms, les verbes et les adjectifs. Malgré des

coûts sans précédent, les ontologies se révèlent peu utiles et sont ordinairement consultées comme des dictionnaires ou des aides à la traduction. Enfin, à l'échelle du Web, la fusion des connaissances provenant de différentes ontologies reste problématique, du fait que, même au sein d'une même discipline, elles ne sont pas interopérables entre elles, malgré les consignes de standardisation.

Ontologies et Web sémantique. — Dans les sciences de la communication et dans le domaine des traitements automatiques du langage, la séparation entre cognition et communication s'est classiquement traduite par le privilège donné à la représentation des connaissances, sans préoccupations particulières pour leur production, leur sélection et leur transmission. On extrait l'information, puis on la communique, la seule condition mise à la communication se limitant à *l'information packaging*, conçue comme simple emballage des connaissances.

Les ontologies sont l'aboutissement de cette conception héritée du positivisme logique : semblant faites par personne pour personne et donc indépendantes de tout point de vue, elles sont censées représenter un monde objectif, indépendant de toute langue et de tout système de signes, comme d'ailleurs de toute tâche. La nomenclature des objets du « monde » n'est évidemment pas problématisée, puisqu'elle repose sur l'évidence partagée ; dans le cas des ontologies « locales » ou spécialisées, l'inventaire des entités dépend simplement de l'état de l'art tel qu'il est admis.

Dans ce type de représentation, la différence entre les langues s'efface, de même que la diversité des discours et des genres : le format des connaissances dans les hiérarchies ontologiques reste celui des réseaux sémantiques : un thesaurus en *basic english*, dopé par des relations sémantico-logiques stéréotypées et d'ailleurs hétérogènes, comme l'hypéronymie, la méronymie, etc.

Tout entier dépendant de cette problématique, le « Web sémantique » reste tributaire d'un petit nombre de relations sémantiques universelles et pauvres. Comment peut-on supposer que la pertinence d'un mot soit liée à la position de son référent dans une hiérarchie ontologique ? Les inégalités qualitatives dans un texte sont sans rapport déterminable avec la position hiérarchique des entités : en général, comme les entités superordonnées sont triviales, plus un concept est superordonné moins il est discuté et donc moins pertinent. La pertinence, si on la définit comme comme un principe d'économie cognitive (selon Sperber & Wilson), ne privilégie alors que les concepts les plus triviaux, mais non ceux sur lesquels portent effectivement les débats.

En outre, la richesse sémiotique des documents numériques n'est pas ou peu prise en compte, car elle est inconciliable avec la problématique référentielle et ne contribue pas à la dénotation : or les indices de l'expression (typographie, codes de couleur, etc.) peuvent se révéler hautement discriminants.

Enfin, la variété des tâches d'application impose de pouvoir définir et faire varier des régimes de pertinence : aucune connaissance n'est indépendante d'une tâche. Comme toute pratique définit son régime de pertinence, c'est à une *praxéologie* (et non à une ontologie) de déterminer quelles sont les « informations-clé » dans les textes et les corpus.

Exigences pour la linguistique. — Concernant le Web, l'enjeu majeur est évidemment l'amélioration des moteurs de recherche, l'adaptation des stratégies en fonction des tâches d'une part, de la nature des documents d'autre part.

Cela demande le recours à une linguistique *applicable* qui puisse traiter des textes, analyser leur sémantique et refléter leur diversité linguistique et sémiotique. La linguistique de corpus se voit ainsi dans la nécessité d'innover. Elle est issue

d'une part de la linguistique computationnelle, qui pose des problèmes dérivés du cognitivisme chomskyen (génération de phrases, construction d'arbres syntaxiques, etc.) et de la lexicométrie (issue de la linguistique mathématique et des statistiques).

La linguistique computationnelle se heurte à des obstacles issus de la philosophie du positivisme logique (notamment par la séparation entre syntaxe, sémantique et pragmatique). En revanche, la lexicométrie, en tant que « simple » méthodologie, ne défend pas de préconception du langage, ce qui la rend plus adaptable. Ces deux problématiques ont en commun de ne pas avoir de conception théorique du texte : pour la linguistique computationnelle, c'est une suite de phrases ; pour la lexicométrie, un ensemble de mots. Aussi ces disciplines se trouvent-elles dépourvues quand elle se trouvent affrontées à la fois à des corpus, massifs, multilingues, polysémiotiques, dont l'abord dépend de multiples demandes sociales et culturelles.

C'est donc à une linguistique conçue comme science des textes et consciente de son appartenance aux sciences de la culture qu'il revient de faire des propositions d'unification et de remembrement. En tenant compte des demandes sociales et non simplement en appliquant des théories : la linguistique ne peut être véritablement appliquée que si elle est également *impliquée*. Elle se doit d'intervenir, même de façon auxiliaire, à diverses étapes : création des logiciels, constitution des corpus, balisages, expérimentations avec outils sur corpus (balisés), interprétation et discussion des résultats. À toutes ces étapes de la chaîne de traitement, des connaissances linguistiques, et plus largement sémiotiques, se révèlent indispensables.

Pour mettre fin à l'oubli des textes. — La problématique ontologique de la représentation des connaissances reste sans doute tributaire d'un état de l'art obsolète, celui d'un temps où l'on *n'avait pas accès* au plein texte. Les thésaurus et autres classifications formalisées servaient alors à indexer les textes à partir d'une représentation statique de leur contenu présumé. Les inconvénients sont connus : coûts de construction et d'entretien considérables, pertinence insuffisante et non modulée en fonction de la tâche qui préside à la recherche d'information.

Le point de vue normatif repose sur des oublis méthodologiques voire épistémologiques qui affectent : (i) les contextes locaux et globaux des informations au sein des textes ; (ii) le contexte des corpus où les textes (et donc les informations) prennent sens ; (iii) les points de vue dont les informations dépendent et qui les ont configurées ; (iv) les collectivités auxquelles elles sont destinées. En bref, la soustraction des contextes est aussi une soustraction des usages dont dépend la notion même de pertinence.

Ces obstacles sont inévitables si l'on réduit les textes à des « ensembles de mots » sans tenir compte des structures, des genres, etc. En revanche, l'accès au plein texte permet désormais des réponses plus adaptées, dès lors qu'il est guidé par les propositions de la linguistique de corpus. En effet, les métadonnées que l'on accumule à présent n'ont tout de même pas pour fonction de permettre d'oublier les données !

Élaboration dynamique. — Nous formulons la proposition méthodologique de fonder toute représentation des connaissances sur l'analyse sémantique et sémiotique des corpus effectifs qui les manifestent : *les connaissances et les ontologies qui les "normalisent" doivent et peuvent être élaborées dynamiquement, en fonction des applications et de leurs corpus.* En effet, les « connaissances » sont des interprétations objectivées de textes et d'autres performances sémiotiques.

Chaque application définit dans son corpus un régime de pertinence propre. Aucun concept n'est pertinent en toute application. Par ailleurs, un des grands

problèmes des ontologies est la définition de leur « nomenclature » : comment distinguer les concepts qui doivent y figurer, alors que tous les mots du lexique sont des candidats potentiels, sans parler des syntagmes phraséologiques. La pratique de George Miller montre qu'il n'a pas d'autre critère que le « bon sens », c'est-à-dire le préjugé du créateur d'ontologies[5].

Si l'on admet que le lexique n'est pas organisé en une arborescence unique, car chaque discours et chaque genre a son lexique, on doit substituer à l'image totalisante du réseau unifié des zones locales organisées par des rapports de *profilage* plutôt que des rapports de subsomption : chaque concept est une *forme sémantique* qui se profile sur un fond. Certains termes lexicalisent des formes ou des parties de formes, d'autres des fonds. Par exemple, le mot *texte* en critique littéraire est un élément de fond, et non un concept : il sert de base compositionnelle à des expressions comme *texte balzacien*, mais il ne se trouve jamais dans le contexte de termes comme *notion* ou *concept*.

Par ailleurs, les formes sémantiques sont *valuées*, alors que les concepts d'une ontologie ne le sont pas : par exemple, dans un réseau comme WordNet, *carré pané* pourrait fort bien être le plus proche voisin de *caviar*. Or il est évident, et l'exploration des corpus le confirme, qu'on ne les rencontre aucunement dans les mêmes contextes (cf. Rastier et Valette, à paraître). Aussi la hiérarchie évaluative prime-t-elle la hiérarchie ontologique construite sans tenir compte des évaluations.

Les concepts peuvent être décrits comme des formes sémantiques propres aux textes théoriques : leurs lexicalisations diffuses ou synthétiques, leurs évolutions, de leur constitution à leur disparition (par extinction ou banalisation désémantisée), leurs corrélats sémantiques, leurs cooccurrents expressifs, tout cela dessine un champ de recherche qui commence à peine à être exploré.

L'alternative que nous proposons est celle de moteurs de recherche en plein texte qui tiennent compte des avancées de la sémantique textuelle, notamment : (i) la définition d'unités textuelles non strictement bornées et séquentielles (les passages) ; (ii) l'extension du principe différentiel de la sémantique au contraste de corpus, entre discours, genres, et sections de textes ; (iii) l'analyse des genres textuels en zones de pertinence différenciées.

L'enjeu est non pas la représentation mais la *production* de connaissances à partir données massives non structurées issues du Web — ou, de préférence, de banques documentaires.

Enfin, la problématique de la représentation des connaissances doit être conçue dans le cadre d'une sémiotique. En effet, les textes ne sont pas de simples chaînes de caractères. Leur découpage, leur structure « logique », leur typographie, voire leurs balises, font partie de leur sémiotique. Par exemple, en philosophie classique, l'usage des majuscules désignait les concepts principaux. Au-delà, les textes scientifiques et techniques intègrent ce qu'on appelle improprement des *hors-textes* : figures, tableaux, diagrammes, photographies participent à la textualisation des connaissances et appellent pour leur traitement une sémiotique multimédia.

Les connaissances textuelles. — Une connaissance est un *ensemble de passages* de textes (éventuellement multimédia) : dans leurs récurrences, le contenu de ces passages (les fragments) et leurs expressions (les extraits) sont en relation de transformation, ne serait-ce que par changement de position. Résultant de figements

[5] En 2002, il fait ainsi sortir de l'ontologie le franc, la lire et le mark, puisque ces monnaies n'avaient plus cours, et il y fait entrer l'*intifada* et le *bacillus anthracis* (cf. l'auteur, 2004b).

et de réductions de syntagmes, les mots sont une sorte très particulière de ces passages, et comme les autres passages, ils restent impossibles à interpréter sans recontextualisation.

En somme, la connaissance est issue d'une décontextualisation de certaines formes sémantiques saillantes et des expressions qui leur correspondent, qu'elles soient compactes (comme les lexicalisations) ou diffuses (comme les définitions). Les formes donnent l'illusion de l'indépendance, voire de leur idéalité, parce que les formes sont par définition éminemment transposables.

Toutefois, aucun mot ni aucun passage ne peut prétendre résumer un texte. Certes, définir une saillance, comme on le fait en faisant figurer en tête d'un article une liste de mots-clés, c'est donner une « instruction » interprétative : la clé n'est cependant pas une clé qui ouvre la serrure du sens, car il reste à construire dans l'interprétation. Aussi, les métadonnées doivent-elles garder trace du texte et du contexte et permettre d'y accéder – sans jamais pouvoir s'y substituer.

Puisque la problématique logico-grammaticale ne peut penser la textualité, les métadonnées utiles n'ont pas de statut logico-grammatical déterminable. En revanche, dans la problématique que nous adoptons, elles revêtent un statut philologique (pour documenter le texte) et herméneutique (pour permettre de l'interpréter). Les informations ne sont plus alors simplement assimilées à des connaissances : on n'appellera *connaissances* que les informations sélectionnées pour une pertinence interprétative. Il reste à les *comprendre*, c'est-à-dire à les relier entre elles en fonction de la structure du texte dont elles sont extraites et de l'objectif de la tâche en cours.

5 Propositions

Typologie des formes de pertinence. — La communication scientifique n'est pas plus directe et pas plus claire que les autres types de communication. De toute façon, la prétention à la clarté n'exclut pas la nécessité de l'interprétation, même si l'herméneutique des textes scientifiques et techniques reste peu développée.

Ces textes se caractérisent par un usage notoire de l'indexation et une structure hiérarchique particulière. Concrétisant une inégalité qualitative, la pertinence résulte d'une valorisation : tel point du texte sera primé, et servira de point d'accès à d'autres, considérés alors comme secondaires. La pertinence affichée doit alors régir les parcours interprétatifs. Alors que les discours scientifiques se limitent en principe à des faits, la pertinence y introduit des valeurs qui concernent tant les faits eux-mêmes que le mode d'accès à ces faits. En effet, les connaissances sont des objets culturels et, à ce titre, elles ne peuvent être dissociées des valeurs.

La pertinence « objective ». — Selon les parties du texte, on peut distinguer plusieurs types de pertinence qui introduisent des indices d'inégalité qualitative et donnent ainsi des indications de valeur.

A/ Le péritexte. — Tant par sa fonction que par sa structure sémiotique (capitales, corps, graisse) il établit des inégalités qualitatives : par exemple les titres ne se limitent pas à des résumés, mais sont des indications interprétatives.

Partie du péritexte, les mots-clé explicites placés en début de texte sont également des indications interprétatives qui pointent des formes sémantiques saillantes.

B/ L'intratexte (ou corps du texte).— Dans cette partie du texte, les unités sont moins normalisées. Les passages clé peuvent être des mots, des syntagmes, des

phrases, des paragraphes, etc. Leur caractérisation suppose des contrastes par des méthodes de linguistique de corpus, quantitatives notamment.

On relève traditionnellement la pertinence des mots singuliers : ils peuvent être isolés par un test probabiliste comme caractéristiques d'un passage ou du texte (cf. la fonction *thème* du logiciel Hyperbase).

La pertinence des passages reste plus importante mais moins étudiée reste: en raison des phénomènes de diffusion sémantique, les passages réunissent des faisceaux de corrélats (lexicalisations partielles d'une même forme sémantique) que l'on peut appeler des *paratopies*. Il faut alors définir des techniques de *zonage* minimal : l'unité textuelle retenue n'est plus le mot, mais le *passage*, si bien qu'un mot clé n'est utile que s'il conduit à un passage clé.

Dans tous les cas, la pertinence intrinsèque est construite par trois types de contrastes : entre passages du texte ; avec des passages d'autres textes du corpus ; avec le corpus choisi considéré dans son ensemble.

C/ L'infratexte. — Conventionnellement, le contenu de l'*infratexte* (les notes, ou la bibliographie, par exemple) est considéré comme faiblement pertinent. Mais c'est « l'inconscient du texte » et la lecture experte peut y déceler des indices cruciaux, comme de simples références bibliographiques, qui situent l'ensemble du texte ou permettent d'en reconsidérer des passages.

La pertinence « subjective ». — La paresse voudrait que l'on se satisfasse de la pertinence proposée : il est vrai que la complexité des rapports entre le péritexte et l'intratexte appelle des recherches propres. Mais l'on doit conserver une attitude critique, car la communication scientifique est aussi « indirecte ». Au-delà de la pertinence « affichée », il peut exister une pertinence cachée : le double langage existe aussi dans les domaines scientifique et technique.

Les textes sont inclus dans des pratiques sociales et leur production comme leur lecture dépend de tâches et de stratégies différenciées. Outre la pertinence objective, un autre régime de pertinence dépend de la lecture et de la tâche qu'elle concrétise : on peut la nommer pertinence « subjective ».

Pour une pertinence dynamique. — La distinction entre pertinence objective et subjective n'est que temporaire. L'auteur propose, le lecteur dispose : parmi les indications proposées par l'auteur, il ne retient que les mots ou passages-clé qui correspondent à sa tâche, en soulignant des mots ou passages-clé qu'il désigne comme tels en fonction de sa tâche. Ni subjective, ni objective, la pertinence doit ainsi être construite dynamiquement en fonction : (i) de la structure du document, (ii) de ses spécificités telles qu'elles peuvent être déterminées par contraste avec son corpus de référence, (iii) enfin, de la pratique en cours.

Incidences sur la redéfinition textuelle des concepts. — La caractérisation textuelle des concepts s'appuie sur les contextes locaux et globaux.

1/ Les indices contextuels locaux comprennent : (i) Les lexèmes cooccurrents, les entités nommées adjacentes (noms d'auteurs notamment), les morphèmes, les ponctèmes. (ii) Les indices d'expression : typographie, balises.

2/ Les indices contextuels globaux comprennent : (i) La position des concepts dans le texte. (ii) La spécificité des concepts et de leurs contextes immédiats, pour caractériser un texte. (iii) La spécificité du texte dans son corpus de référence (de manière experte, on peut caractériser aussi un texte par les concepts absents).

3/ La position temporelle des concepts : l'évolutivité des concepts impose des études en diachronie (exemple : les travaux de Mathieu Valette sur un corpus de Gustave Guillaume étendu sur 40 ans).

Pour une sémantique du Web. —Développée à partir de la sémantique des textes et de la philologie numérique, une sémantique du Web peut les mettre à profit pour adapter à la diversité des requêtes la diversité des réponses, qui seront pertinentes si elles reflètent la diversité des textes. Elle n'est à son tour qu'une étape de médiation pour constituer une *sémiotique comparée* des documents numériques.

On doit tenir compte tant au plan épistémologique que méthodologique des sources de diversité que la problématique actuelle du Web sémantique ne permet pas de traiter de manière satisfaisante.

1/ *La diversité des langues* . — Le Web est multilingue et le sera de plus en plus. L'hégémonie initiale de l'anglais a été renversée par la montée en puissance d'autres grandes langues. Les moteurs de recherche doivent donc gérer un multilinguisme croissant, ce qu'ils ne font pas encore de façon satisfaisante.

Par ailleurs, les représentations des connaissances devraient varier avec les langues : il ne s'agit pas simplement de découpages différents des mêmes champs de réalité, mais encore de définitions différentes de ces champs comme l'attestent par exemple les contrastes « ontologiques » entre le chinois et l'anglais.

2/ *La diversité des discours et des genres.* — Les « concepts » qui peuplent les « ontologies » dépendent étroitement des discours et des genres. L'existence de communautés internationalement structurées a favorisé la constitution de discours disciplinaires plurilingues et la diffusion de genres comparables malgré les différences linguistiques : cela peut se traduire par des calques terminologiques, mais aussi par des modes de structuration textuelle, tant au plan du contenu que de l'expression. Toutefois, l'adoption de normes internationales limite la diversité linguistique, mais sans pouvoir l'annuler.

3/ *La diversité des styles.* — La formation et l'évolution des concepts sont l'objet d'importantes différences non seulement selon les disciplines, mais encore selon les auteurs. Le style philosophique de Deleuze définit par exemple un régime de transformations conceptuelles tout à fait différent de celui de Bourdieu. Les méthodes de la linguistique de corpus ont permis sur ce point des résultats qui confirment le bien-fondé d'un programme comparatif (cf. Loiseau, 2006).

4/ *Les inégalités qualitatives au sein des documents.* — Chaque genre, chaque texte singulier définit un régime de pertinence qui prime certaines formes sémiotiques relativement à d'autres. Cela impose de définir, ici encore avec les méthodes de linguistique de corpus, des techniques de détection des inégalités qualitatives.

5/ *La diversité sémiotique intrinsèque des documents.* — La distinction entre textes (multimédia) et documents doit être réduite voire supprimée, car le texte n'a pas un contenu indépendant de son expression, et le document ne peut être véritablement décrit en faisant abstraction de son contenu. Corrélativement, au plan épistémologique, les divergences de fait entre linguistique et philologie doivent être reconsidérées au sein d'une sémiotique générale de la communication. La sémantique du Web relève en effet d'une *sémiotique comparée* des documents numériques.

6/ *La diversité des tâches.* — Malgré le problème lancinant mais faussé de la réutilisabilité, les représentations des connaissances qui ne sont pas établies en fonction d'une application déterminée sont en général peu utilisables et guère réutilisables. Construire de telles représentations avec une ambition de généralité

reste une tâche indéfinie sinon infinie, car les tâches déterminent en effet un régime de pertinence.

En revanche, la rencontre entre *l'horizon de pertinence* déterminé par la tâche et *les formes sémiotiquement saillantes* détectées par l'analyse contrastive du corpus de travail permet de qualifier les passages essentiels et de restreindre drastiquement les réponses en recherche d'information.

7/ *La diversité des statuts de fiabilité* . — Au plan pratique comme au plan éthique, la question de la fiabilité des documents ne doit pas être négligée, car le Web fourmille d'écrits apocryphes, de faux, sans parler de textes diversement négationnistes. Un écrit non authentique ne peut évidemment jouir que d'une autorité usurpée.

Cette question doit être traitée dans le cadre d'une réflexion sur les types de communication, les dimensions de la destination et l'adresse, comme enfin de l'autorité et de l'authenticité. Le nombre de liens et le *page-ranking* ne définissent qu'une métrique conformiste de l'autorité. On ne peut véritablement parvenir à une recherche d'information fiable si l'on ne tient pas compte du degré de confiance que l'on peut attribuer aux documents : c'est là un point faible du Web 2, quand il fait de l'anonymat un principe — comme on le voit avec Wikipedia.

6 Du texte au document

Clarifier le rapport entre texte et document engage à concilier linguistique et philologie.

La reconquête de l'expression. — Consommée de fait depuis un demi-siècle, la séparation entre linguistique et philologie a beau être récente, ses raisons mêmes furent oubliées avec l'oubli des textes par les linguistiques universelles. Cependant, depuis sa formation disciplinaire à Alexandrie, la grammaire avait toujours été considérée comme une discipline auxiliaire pour la lecture et l'analyse critique des textes. La sémantique (Semasiologie), lors de sa création par Reisig dans les années 1820, était une sorte de lexicologie des textes classiques. Alors qu'il était ordinaire de combiner l'analyse de textes et la linguistique historique et comparée (Steinthal, Bréal, Saussure en sont des exemples bien connus), à partir des années 1950, l'adoption unilatérale de perspectives synchroniques, le privilège exclusif donné à la modélisation de la morphosyntaxe, l'image prégnante des langages formels, et complémentairement les études sur l'oral privilégiées par une pragmatique du hic et nunc, ont accompagné voire causé la quasi-disparition de la philologie en linguistique générale.

Les grammaires de texte ont confirmé ou consommé l'abstraction du concept de texte. Il fut généralement réduit à son plan sémantique : la Sémantique structurale de Greimas (1966) proposait ainsi sa formalisation en une série de propositions dans un format inspiré explicitement de la logique de Reichenbach. Il reste la source principale du modèle de Van Dijk (dont Greimas dirigea les premières recherches), puis des divers modèles propositionnels du cognitivisme orthodoxe (la Forme Logique chomskyenne des années 1980, par exemple). On est revenu ainsi au programme des grammaires générales, justement dites philosophiques, qui à l'âge classique, avant donc la formation de la linguistique, représentaient les phrases comme des propositions logiques et le discours comme un raisonnement (sur le modèle de la logique des classes). L'appauvrissement qui

accompagnait cette réduction « sémantico-logique » conduisit à faire du texte une série de chaînes de caractères considérées comme des données. Si en revanche, à la suite de Saussure, on rapatrie le signifié dans les langues et dans les textes, on doit donner corrélativement toute sa place au plan de l'expression, puisque la corrélation entre expression et contenu détermine la sémiosis textuelle.

Les indications philologiques élémentaires, quand elles sont retenues, sont aujourd'hui classées comme des « métadonnées ». Cette conception prévaut aujourd'hui avec le Web sémantique. La plus grande confusion règne dans ce domaine, puisqu'on classe dans les métadonnées toutes sortes de données incompatibles avec la théorie appauvrie du texte qui prévaut généralement : on juxtapose des indications simplement bibliographiques, comme l'auteur, l'éditeur, l'ISBN, le lieu d'édition ; des indications documentaires, comme le résumé ou les mots-clé ; des caractérisations textuelles globales, comme le genre. Les théories linguistiques du péritexte, qui limitent le texte à l'intratexte, en séparent les titres, voire les notes, etc. n'ont fait qu'ajouter à la confusion. En règle générale, les données relèvent de la linguistique interne, les métadonnées de la linguistique externe et, faute de réfléchir leur dualité, on ne peut théoriser le rapport entre données et métadonnées. Les problèmes négligés reviennent alors, réifiés, sous la forme de métadonnées. Par exemple, dans le domaine du multimédia, les textes eux-mêmes deviennent les métadonnées des images.

Sans trop croire à l'efficacité d'un moratoire sur les métadonnées, retenons que les métadonnées sont des critères globaux et les données des grandeurs locales qui en dépendent : au lieu de les séparer a priori, c'est à une théorie élaborée de la textualité qu'il revient d'établir systématiquement les corrélations entre métadonnées et données, pour restituer la complexité des textes.

La notion de métadonnée doit ainsi être critiquée et refondue. Par exemple, le succès de Google s'explique par l'introduction d'un nouveau type de métadonnées (les liens qui pointent vers le document) et par une perspective praxéologique implicite qui représente le document en fonction d'un point de vue (de qui sélectionne les liens) et d'un garant (celui qui pose les liens et apporte ainsi une évaluation).

Extension du domaine des données. — Formations historiques, les langues sont des artéfacts : le nombre et la nature de leurs niveaux de sémiotisation ne sont pas fixés a priori, comme on le sait depuis l'invention de l'écriture[6]. Les codes des formats numériques poursuivent cette évolution, même s'ils ne sont pas réservés exclusivement aux textes. Par exemple, le HTML et le XML qui codent un texte ou des éléments de ce texte peuvent être considérés comme des niveaux sémiotiques supplémentaires. En effet, le texte est en quelque sorte une idéalisation linguistique, et son support documentaire introduit inévitablement d'autres sémiotiques : le rouge n'est pas réservé aux textes, mais, signe d'excellence dans l'Antiquité, il a donné lieu au signalement des rubriques, si bien nommées, dans les manuscrits, puis aux titrages en rouge dans l'imprimerie renaissance, etc.

Il faut cependant dans la notion de document distinguer deux choses : l'expression du texte et la configuration matérielle du support. La structure typographique de la mise en forme relève d'un niveau de l'expression textuelle ; en revanche, la mise en page relève de la norme du document — même si elle n'est pas

[6] Cela vaut d'ailleurs sur les deux plans, puisque les « recodages » sémantiques sont attestés par différentes herméneutiques, en général ésotériques.

sans effet sur l'appréhension du texte. Ainsi les paginations comme les titres courants font-ils partie du document, non du texte. À la philologie numérique répond ainsi une diplomatique numérique, qui ne traite point du texte, mais seulement de caractères spécifiques au document qui le véhicule.

Enfin, la stratification même du langage doit encore beaucoup à des simplifications théoriques : ainsi la séparation entre contenu et expression a-t-elle été creusée, sinon imposée, par le dualisme métaphysique qui oppose matière et pensée ; mais aussi à des simplifications méthodologiques : on a maintenu séparés les niveaux linguistiques de manière à réduire la complexité et pouvoir formuler des règles dont l'application ne soit pas conditionnée par l'incidence de phénomènes situés à d'autres niveaux.

De fait, la linguistique de corpus a permis de produire de nouveaux observables qui associent des éléments de niveaux d'analyse ordinairement séparés (ainsi de la corrélation entre des noms de sentiments ou des temps verbaux avec certaines ponctuations, cf. l'auteur, 2005) : elle témoigne ainsi de la solidarité sémiotique entre plans du langage. Plus impressionnante encore, la solidarité entre paliers de complexité permet de caractériser des textes en corrélant des indices d'expression locaux (comme la ponctuation ou la longueur moyenne des mots) à des « métadonnées » sémantiques globales comme le genre et le discours : les résultats présentés dans Malrieu et Rastier (2001) établissent ainsi la concordance complète entre la classification « manuelle » en genres et discours sur un corpus de 2600 textes et la classification automatique à partir de moyennes établies pour chaque texte sur 251 variables locales d'expression.

Des critères ordinairement réputés non textuels, comme le code de la police de caractères peuvent se révéler fort discriminants : par exemple, dans une application de détection de sites racistes, le code des polices gothiques obligeamment téléchargeables sur les sites néo-nazis peuvent suffire à caractériser un texte à la volée, sans grand risque d'erreur. Bref, à partir de la sémantique des textes, la « reconquête » de l'expression a d'autant plus de conséquences que les systèmes informatiques ne traitent aisément que les indices de cette strate linguistique. Cela engage une reconception sémiotique du texte, plus précise que celle qui a été élaborée naguère dans les études littéraires. Pour les applications, cela permet de développer des caractérisations multicritériales, les coalitions d'indices permettant de former des conjectures éprouvées, dans des systèmes multi-agents par exemple. Ces coalitions sont extraites par « des patrons complexes d'expressions régulières multiniveaux combinant différents items de natures formelles différentes (chaînes de caractères, POS, annotations, positions, etc.) » (Valette et Slodzian, à paraître).

Enjeux actuels. — On sait que le programme du Web Sémantique consiste à passer du « Web des documents » au « Web des données » sur le mode de ce que l'on nommait naguère la représentation des connaissances, en extrayant des textes des données, et en les organisant en ontologies. Les enjeux politiques et économiques ne peuvent cacher les conséquences épistémologiques de ce programme. En effet, l'adoption de standards « de bas niveau » comme HTML, ou Unicode voire XML n'entraîne aucunement que l'on doive ériger en standard des langages de représentation comme RDF ou OWL, sauf à céder benoîtement à la tentative de coup de force du W3C en faveur du « Web sémantique ».

Un texte n'est pas un réservoir de connaissances qui pourraient être extraites par indexation et condensées en données résumant son contenu informationnel ; l'indexation donc n'a qu'une relative valeur de recherche et de classement. Prenons

un exemple : à l'heure actuelle, dans les services de renseignement militaire d'un grand pays européen, des personnels extraient de documents Word des mots et expressions qu'ils transfèrent dans des feuilles Excel où ils sont classés en « ontologies ». Ces feuilles sont ensuite transmises à des analystes qui en font la synthèse sous la forme de Powerpoints présentés à l'état-major [7]. L'éloquence militaire prise certes le laconisme, mais toute modification systématique d'un texte en change le genre et donc l'interprétation. La sélection de ces passages minimaux que sont les mots et expressions reste de fait incontrôlable, puisque le recouvrement de deux indexations du même texte par la même personne s'établit en moyenne à 40%. La délinéarisation, la « compression » augmentent l'équivoque et créent l'ambiguïté.

Au « Web sémantique », il faudra inévitablement substituer une sémantique du Web (cf. l'auteur, à paraître), car les besoins sociaux pour la recherche d'information, l'amélioration des moteurs de recherche, le data-mining, ne pourront être satisfaits que par une linguistique et une sémiotique de corpus permettant l'analyse des données textuelles.

N.B. : J'ai plaisir à remercier Évelyne Bourion, Carmela Chateau et Bénédicte Pincemin.

Références :

Adam, J.-M. (1992) *Les textes : types et prototypes*. Bruxelles, Mardaga.

Adam, J.-M. (2005) *La linguistique textuelle. Introduction à l'analyse textuelle des discours*. Paris, Nathan.

Adam, J.-M. (2006) Autour du concept de *texte*. Pour un dialogue des disciplines de l'analyse des données textuelles. *Actes des 8emes JADT*, Besançon.

Berners-Lee, T. (1998) *Weaving the Web*, Harper, San Francisco.

Berners-Lee, T. (2007) Le web va changer de dimension. *La Recherche*, 413, pp. 34-38. Propos recueillis par Marie-Laure Théodule.

Bird S. & Liberman M. (2001). A formal framework for linguistic annotation, *Speech Communication*, 1/2-33, pp. 23–60.

Bourion, E. (2001) *L'aide à l'interprétation des textes électroniques*, Thèse de doctorat, Université de Nancy II, http://www.texto-revue.net.

Eco, U. (1974) *Trattato di semiotica generale*. Milan, Bompiani.

Greimas, A.-J. (1966) *Sémantique structurale*. Paris, Larousse.

Halliday M.A.K et Hasan, R. (1976) *Cohesion in English*. Londres, Longman.

Loiseau, S. (2006) *Sémantique du discours philosophique : du corpus aux normes. Autour de G. Deleuze et des années 60*. Thèse de doctorat, Université Paris X-Nanterre.

Loiseau, S. (2007) CorpusReader. un dispositif de codage pour articuler plusieurs interprétations. *Corpus*, 6.

[7] Je n'invente rien : partir de Word, passer par Excel pour arriver à Powerpoint, tel est aujourd'hui le cycle d'extraction et de d'exploitation des « connaissances ».

Louw, B. (2007) Collocation as the determinant of verbal art. In *Language and Verbal Art Revisited*, Donna R. Miller et Monica Turci (eds). Londres, Equinox, pp. 149-180.

Malrieu, D. et Rastier F. (2001) Genres et variations morphosyntaxiques. *Traitements automatiques du langage*, 42, 2, pp. 547-577.

Mauceri, C. (2007) *Indexation et isotopie : vers une analyse interprétative des données textuelles*. Thèse de doctorat, ENST-Bretagne (Telecom Bretagne) et Université de Bretagne Sud. Rééd. : http://www.texto-revue.net

Pédauque, R. T. (2006) *Le document à la lumière du numérique*. Caen, C&F éditions.

Poudat, C. (2006) *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*. Thèse de doctorat, Université d'Orléans, http://www.texto-revue.net.

Rastier, F. éd. (1995) *L'analyse thématique des données textuelles*. Paris, Didier.

Rastier F. (1996) Représentation ou interprétation ? — Une perspective herméneutique sur la médiation sémiotique. In V. Rialle et D. Fisette (dir.), *Penser l'esprit : des sciences de la cognition à une philosophie de l'esprit*, Grenoble, Presses Universitaires de Grenoble, pp. 219-239

Rastier F. (2001a) L'action et le sens. — Pour une sémiotique des cultures, *Journal des Anthropologues,* 85-86, pp. 183-219.

Rastier F. (2001b) *Arts et sciences du texte*. Paris, PUF.

Rastier, F. (2002) Anthropologie linguistique et sémiotique des cultures. In *Une introduction aux sciences de la culture*, ch. 14, pp. 243-267.

Rastier F. (2004). Doxa et lexique en corpus - Pour une sémantique des « idéologies ». *Actes des Journées scientifiques en linguistique 2002-2003 du CIRLLEP*, Reims : Presses Universitaires de Reims.

Rastier, F. (2005) Enjeux épistémologiques de la linguistique de corpus. In G. Williams (éd.). *La Linguistique de corpus*, Rennes : Presses Universitaires de Rennes, 31-46.

Rastier, F. (2006) Sémiotique des sites racistes. *Mots*, 80, pp. 73-85.

Rastier, F. (2007a) Indices et parcours interprétatifs, in Denis Thouard, éd. *L'interprétation des indices*, Lille, Presses du Septentrion, pp. 123-152.

Rastier, F. (2007b) Passages. *Corpus*, 6, pp. 127-162.

Rastier, F. (à paraître) Web semantics v. Semantic Web. *International Journal of Corpus Linguistics*.

Rastier, F., Cavazza, M., Abeillé, A. (1994) *Sémantique pour l'analyse : de la linguistique à l'informatique*, Paris, Masson.

Valette, M. (2004) Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet. *Approches Sémantiques du Document Numérique, Actes du 7e Colloque International sur le Document Electronique*, 22-25 juin 2004, Patrice Enjalbert et Mauro Gaio, eds, 2004, pp. 215-230.

Valette, M. et Slodzian, M. (à paraître) Sémantique des textes et recherche d'information. *Revue française de linguistique appliquée* (soumis).

Conférences CIDE-CIFED

Progrès récents en interprétation automatique des images

Frédéric Jurie (1)

(1)Université de Caen - UFR des Sciences-Greyc

Résumé *:* Bien qu'il s'agisse d'un domaine de recherche ancien, l'interprétation automatique des images est restée relativement longtemps à un niveau de développement tel qu'il n'était pas possible d'envisager de l'utiliser dans des situations opérationnelles. L'augmentation de puissance des calculateurs, la mise au point de nouveaux modèles pour la représentation des images, les avancées récentes des techniques d'apprentissage automatique ainsi que la focalisation des travaux sur des tâches pertinentes ont permis, depuis peu, de faire d'utiliser certaines de ces technologies dans des produits industriels (appareils photos qui détectent les visages, vidéo surveillance, biométrie, etc.). Cette présentation aura pour objet de présenter les difficultés que soulève l'interprétation automatique des images, la nature et l'impact des avancées que nous avons mentionnées, la présentation de résultats récents ainsi que les défis qui restent posés à ce jour.

La Théorie Cinématique des Mouvements Humains Rapides : Développements Récents.

Réjean Plamondon (1)

(1) Laboratoire Scribens, Département de Génie Électrique,
École Polytechnique de Montréal
rejean.plamondon@polymtl.ca

Résumé : Lorsqu'employée en reconnaissance des formes, la modélisation des mouvements humains vise, entre autres, à procurer certaines assises théoriques au traitement en ligne de l'écriture manuscrite et à fournir des connaissances fondamentales pouvant servir de balises dans la conception de systèmes automatiques. À ce jour, plusieurs approches ont été proposées pour modéliser la production des mouvements en général et de l'écriture en particulier: des modèles à base de réseaux de neurones, des modèles dynamiques, des modèles psychophysiques, des modèles cinématiques, des modèles reposant sur des principes de minimisation. Parmi les modèles dits analytiques, la Théorie Cinématique et son modèle delta-lognormal se sont avérés des plus prometteurs. Mais, bien qu'il ait été démontré que ce paradigme permettait de prendre en compte la majorité des phénomènes couramment observés en motricité fine, plusieurs problèmes théoriques et techniques ont retardé son intégration directe ou indirecte dans la conception de systèmes.
Dans cet exposé, nous ferons le point sur ces différentes difficultés et nous dévoilerons les résultats de récents travaux que notre équipe a réalisés pour les surmonter. Dans un premier temps, dans une perspective de généralisation, nous présenterons toute la famille des modèles de types log-rnormaux. Ensuite, du point de vue pratique, nous décrirons deux nouveaux algorithmes d'extraction de paramètres. Nous montrerons également comment la nouvelle représentation qui en résulte peut être employée pour caractériser des signatures, des scripteurs et pour étudier l'effet de différents facteurs, par exemple le vieillissement, sur le contrôle neuromoteur. Nous proposerons également une méthodologie pour générer automatiquement des banques de données manuscrites dépendantes ou indépendantes du scripteur. Du point de vue théorique, nous montrerons comment, à l'aide de nouvelles expériences psychophysiques, nous avons pu valider les hypothèses de base de la Théorie Cinématique de même que ses prédictions les plus distinctives. Nous terminerons en expliquant comment l'écriture manuscrite peut-être employée pour améliorer le traitement des signaux électro-myographiques et électro-encéphalographiques, ouvrant ainsi la porte à de nouvelles applications en génie biomédical et en neurosciences.

Recherche d'information : de la RI orientée système vers la RI orientée contexte

Mohand Boughanem(1)

(1)IRIT – Université Paul Sabatier France
Mohand.Boughanemrit.fr

Résumé : La recherche d'Information (RI) a une longue histoire au cœur de la science informatique. Dés les années 60 on s'est interrogé sur les possibilités de traitement informatique dans l'accès aux bibliothèques, de même que dans la gestion automatisée des documents. Le domaine de la RI n'a pas cessé d'évoluer dans le but de rationnaliser le processus d'identification, au sein de collections de documents, ceux qui sont potentiellement pertinents pour l'utilisateur. Dés ces années là, la RI a développé des techniques d'indexation de documents et de requêtes, des modèles de mesure de pertinence document-requête des techniques de stockage ainsi que des méthodes d'évaluation qui ont été affinées au fil des ans. Le début des années 90 a marqué une période de profusion des travaux de recherche dans le domaine de la RI de manière générale. Cet essor est le résultat de plusieurs facteurs dont les plus importants sont l'accroissement considérable des volumes d'information à gérer et l'apparition du web qui a lui seul bouleversé le domaine de la RI. Ces facteurs ont posé de nouvelles problématiques qui ont conduit à la définition de nouvelles thématiques, comme la RI personnalisée-contextuelle, la RI mobile. Une simple comparaison entre l'appel à communication de la conférence SIGIR (Special Interest Group of Information Retrieval) des deux années 1990 et 2008 fait apparaître plus d'une dizaine de nouveaux sujets.
L'objectif de cet exposé est de faire un tour d'horizon sur les avancées effectuées ces dernières années dans le domaine de la RI. Un accent particulier sera mis sur les modèles contextuels. Ces modèles tentent de revisiter l'approche classique de la RI, communément qualifiée d' «approche orientée système », vers une approche, qualifiée d' «approche orientée utilisateur ou contexte » qui intègre l'utilisateur comme paramètre du modèle d'accès à l'information. Plus précisément, dans l'approche classique, un besoin en information est uniquement traduit à travers la requête soumise par l'utilisateur, or l'approche contextuelle tente, quant à elle, d'intégrer le contexte de la recherche comme paramètre dans le modèle d'appariement requête-document. L'interprétation du besoin est donc subordonnée au contexte de l'utilisateur qui l'a exprimé. Celui-ci est représenté entre autres, par ces centres d'intérêt, son expertise, ses objectifs, la tâche qu'il effectue et la situation de recherche correspondante.

Atelier Fracture Numérique

L'usage de l'internet à l'université de Ouagadougou

Usage of the internet in the University of Ouagadougou

Pascal RENAUD (1)

(1) Unité de recherche « savoirs & développement » (R105), Institut de recherche pour le développement, Bondy, France,
Pascal.Renaud@ird.fr

Résumé. L'université de Ouagadougou est une des plus engagées de la sous-région dans l'adhésion au processus de Bologne. Adhérente au «Réseau pour l'Excellence de l'Enseignement Supérieur en Afrique de l'Ouest» (REESAO, 2005) depuis sa création, elle s'est fixée pour objectif de terminer la mise en place du LMD en 2010. Nous avons mené une enquête sur l'usage de l'internet par les étudiants durant l'année universitaire 2006-2007. Elle porte sur un échantillon représentatif de 1000 étudiants et un certain nombre d'entretiens d'enseignants. Ce travail met en évidence l'importance que les étudiants accordent aux TIC dans l'amélioration de l'enseignement et leur désir de se placer dans la perspective d'un marché mondial des diplômes.
Mots-clés. TIC, société des savoirs, enseignement supérieur, pays en développement, internet, processus de Bologne, mondialisation.

Abstract. The university of Ouagadougou is, in its sub-region, one of more involved establishment in the EU Bologna university reform process. Member of the " Network for the Excellence of the Higher Education in Western Africa" since its creation, its objective is to finalise the implementation the reform process by 2010. We carried out a survey on the usage of the internet by the students during the 2006-2007 academic year. It is based on the analyse of a representative panel of 1000 students and several interviews of trainers. This work points out the importance that the students grant in ICTs for the improvement of the education and brings to light their hope to do play their part in the perspective of a world market of diplomas.
Keywords. ICTs, Knowledge society, Higher Education, developping countries, internet, Bologna Process, globalisation

1 Le burkina Faso

Rappelons que le Burkina Faso (13 millions d'habitants) fait parti du groupe des pays les moins avancés (PMA), il se situait au 174ème rang sur 177 pour son indice de développement humain dans le rapport du PNUD publié en 2006. Il a peu de ressource naturelle, son économie est très dépendante du coton, vulnérable aux catastrophes naturelles et à l'instabilité régionale. Le revenu national ne progresse pas de plus de 5% par an depuis 1994 tandis que sa croissance démographique est

proche de 4%. Le revenu par habitant reste inférieur à 400 US$ ce qui le situe au dessous de la moyenne des PMA (590 US$).

L'éducation est cependant une forte priorité budgétaire du Burkina qui y consacre plus de 20% de ses ressources. Le taux d'étudiants pour 100 000 habitants reste faible (124) mais est en forte progression. L'effectif de l'enseignement supérieur public devrait dépasser les 32000 étudiants à l'horizon 2015 (Brossard, Foko, 2006). Le pays dispose actuellement de trois universités, outre celle de Ouagadougou sur laquelle nous avons travaillé, l'université polytechnique de Bobo-Dioulasso créé en 1995 accueille environ 1500 étudiants. Enfin, l'université de Koudougou créée en 2005 pour décentraliser et désengorger certaines UFR pléthoriques de Ouagadougou, accueille entre 3000 et 4000 étudiants.

Dans le domaine des technologies de l'information le Burkina se situe dans la moyenne des PME. Sa télédensité (nombre d'abonnements téléphoniques pour 100 habitants) était de 8,1 en 2006. Un point fort, la mise en œuvre précoce d'une offre l'ADSL. Elle a permis d'améliorer très sensiblement la vitesse d'accès de nombreux cybercafés sans augmenter leurs prix de revient. Les étudiants en ont été les premiers bénéficiaires.

2 Université de Ouagadougou

Créée en 1974, avec seulement 374 étudiants, l'université de Ouagadougou a connu une évolution quantitative et qualitative très rapide qui a conduit les autorité à engager une déconcentration vers la nouvelle université de Koudougou. En 2006-2007, on recensait près de 24 000 étudiants répartis en sept Unités de Formation et de Recherche (UFR) et un Institut (Arts et Métiers).

A l'instar des universités de la sous-région, l'université de Ouagadougou, à connaît d'importantes difficultés (amphithéâtres surchargés, manque d'équipement et problèmes d'infrastructures, insuffisance du nombre d'enseignants...). Ces difficultés génèrent des conflits et expliquent en partie la fréquence des grèves qui paralysent le campus.

3 Les TIC à l'Université

Le réseau informatique universitaire s'appuie sur une infrastructure en fibre optique qui dessert presque tous les bâtiments du campus. Il comprend environ 600 postes de travail répartis entre les bureaux des enseignants - chercheurs, de l'administration et les trois *centres de ressources*, placés dans les trois principales facultés (lettres, sciences, droit). L'ensemble est relié à l'internet par l'Onatel à travers une ligne dont le débit, qui doit être prochainement porté à 2Mbs, reste notoirement insuffisant par rapport à la charge du réseau.

L'enquête

L'enquête a été réalisé par une équipe de recherche pilotée par l'IRD (projet OUIE pour observation des usages de l'internet par les étudiants) et porte sur la place de l'internet dans l'acquisition des savoirs, laissant volontairement de côté les autres usages de l'internet par les étudiants. Cet objectif a été martelé par les enquêteurs et rappelée dans l'intitulé des questions que nous analysons ici.

Elle a été réalisée entre le 19 mars et 12 avril 2007, sur l'ensemble des UFR de l'Université de Ouagadougou. L'échantillon a été défini pour comporter environ 1000 étudiants représentant la population de l'université par sex et UFR.

74% d'internautes

Moyenne	Femmes	Homme
73,78%	76,96%	72,47%

74% des étudiants déclarent utiliser régulièrement l'internet. Ce chiffre est très élevé si on prend en compte les conditions dans lesquelles les étudiants accèdent au réseau. Une étude réalisée en France en novembre 2005 par Médiametrie pour la Délégation aux usages de l'Internet indique que 31 % des étudiants de 19 à 24 ans, déclare ne *jamais* utiliser l'ordinateur *pour faire leurs travail* (Médiametrie, 2005), chiffre confirmé par un rapport du CREDOC publié en 2006 sur la diffusion des technologies de l'information dans la société française. Celui-ci met en évidence que 21% des 18/24 ans n'utilisent jamais ni ordinateur, ni internet même s'ils ne sont plus que 12% des élèves et étudiants (Bigot, 2006).

Une première constatation s'impose, le taux d'utilisation de l'internet par étudiants burkinabé est comparable à celui de leurs collègues du Nord. Il s'agit cependant d'un indicateur brut qui n'indique rien sur la durée ce cet utilisation. Si celle-ci se compte en heures par semaine à Ouagadougou elle s'exprime en heures par jour à Paris… Ce qui rend comparable ces 74% d'utilisateurs réguliers au Sud avec un taux de 70% d'utilisateurs régulier au Nord, c'est la proximité des représentations comme des objectifs poursuivis.

On notera aussi que dans l'université de Ouagadougou, les hommes et les femmes ont un comportement similaire face aux TIC, la différence de 4% dans l'échantillon (même redressé) ne semble pas assez significative.

« Je ne maîtrise pas mais je vais m'y mettre »

230 étudiants de l'échantillon, représentant environs 25% de la population totale de l'université de Ouagadougou, n'utilisent pas régulièrement l'internet. Plus précisément la question retenue pour sélectionner ces *non-utilisateurs* est celle qui concerne le temps passé sur l'ordinateur : « quel temps passez-vous sur l'internet dans la semaine ? ». Plusieurs réponses étaient proposées : « *plus de 10 h, entre 5 et 10 h, entre 2 et 5 h, moins de 2h* » et finalement « *je n'utilise pas l'internet* ». Les étudiants qui ont coché cette case, semblent, dans leur grande majorité, connaître le Net et même pour nombre d'entres eux, l'avoir déjà utilisé. Leurs réponses précises à la question de savoir si le Net est utile à l'enseignement supérieur indique que s'ils connaissent l'internet et savent ce qu'il pourrait leur apporter, ils ne l'utilisent pas régulièrement, notamment dans le cadre de leurs études.

Lorsque les étudiants expliquent leur *non-usage* ils mettent en avant le coût de l'accès et le manque de formation[1]. Celle-ci fait cruellement défaut. Parmi les 74% d'utilisateurs réguliers, près de 80% d'entre eux, déclarent s'être formé seul ou à l'aide d'un ami. Seuls 21% des étudiants indiquent avoir reçu une formation et ils sont moins de 4% a avoir bénéficié d'une formation à l'université.

Plus ils avancent dans leurs études, plus ils utilisent l'internet

1 La question posée est « si vous n'utilisez pas l'internet, dites pourquoi »

Et, sans aucun doute, ils s'y mettent. Plus ils avancent dans leurs études, plus ils utilisent l'internet. A tel point qu'en fin d'étude le taux d'utilisateurs régulier avoisine les 100 %. Il est vrai qu'en première année, ils n'ont pas une claire connaissance des services proposés et se rendent presque exclusivement dans les cybercafés. Et force est de constater que plus ils utilisent l'internet plus ils considèrent le Net essentiel au processus d'acquisition de savoirs. Cette expérience plus longue ou plus dense de l'internet (ils sont en deuxième ou en troisième année et se connectent plus de 2h par semaine) les conduit d'une part à affirmer un point de vue plus critique à l'égard de leurs enseignants et d'autre part à étayer leur vision mondialiste.

Mieux comprendre, compléter, se mettre au niveau mondial

Le Net permet de « mieux comprendre les cours », de « bien en voir l'intégralité » et de « compléter » ou « d'approfondir les connaissances ». Ils « permet de connaître beaucoup de choses que l'enseignant n'a pas fait, les cours en faculté représentent 30% de nos besoins intellectuels ».

30% des étudiants passent plus de 5 heures par semaine sur le Net et c'est parmis eux que s'affirme nettement la demande « d'ouverture sur le monde », « d'enrichissement de la culture générale », de « contacts avec le reste du monde » afin de rester « au parfum des dernières découvertes », « de confronter ce qu'on reçoit avec d'autres cours de différents pays », « de se mettre au niveau de nos homologues du monde entier » car « la connaissance est universelle », « le système LMD nous permettra de suivre le même programme des mêmes facultés dans les autres universités ».

Le Net se présente comme une double promesse. Celle d'un accès au cours dactylographié[2] et celle de l'accès à une bibliothèque presque inépuisable. Deux pièces indispensables à un enseignement de qualité. Avec cette bibliothèque globale, les jeunes burkinabé vont enfin pouvoir faire jeux égal avec leurs homologues du Nord. Ils y puiseront les supports de cours et les ressources documentaires qui s'inscrivent dans leur cursus pédagogiques, des documents à jour – au parfum – pour préparer un dossier, un exposé, un mémoire de fin d'année.

L'enseignement à distance qui mobilise d'important financement et dispose des faveurs des médias et des agences de coopération retient assez peu leur attention. Font-il preuve d'un simple bon sens ? Le e-learning, outre son caractère parachuté, off shore, exigent généralement des moyens supérieurs à l'enseignement classique dit présentiel. Ils sont conscients des contraintes budgétaires d'un pays très endetté et plus ou moins sous la tutelle des institutions financières internationales. Ils ne peuvent espérer d'augmentation importante du taux d'encadrement.

4 Conclusion

Si l'objectif des étudiants est clair, entrer dans le marché mondial du savoir. Leur université est-elle prête à relever ce défit ? Les premières réunions de restitution de cette enquête ont permis de mettre en évidence plusieurs éléments. Si tous étaient conscients de l'enthousiasme de la jeunesse pour le Net, ils pensaient que la relation des étudiants avec le Net était essentiellement ludique et que ceux-ci s'intéressaient avant tout aux sites de rencontres, de musique, voire aux sites

2 Les autorités de l'université avec l'appui d'agences de coopération, souhaitent généraliser la mise en ligne des cours des enseignants

pornographiques. Et, même si nombre d'enseignants proposent des références bibliographiques sous forme d'adresse internet (URL), ils ne soupçonnaient pas l'importance que les étudiants attribuaient à l'internet pour leur formation. Cette capacité des étudiant ouagalais à se saisir de l'internet comme outils professionnel est une véritable découverte qui, très probablement, va influencer les décideurs lorsqu'ils auront à faire des arbitrages.

Références :

Bigot R. (2006), "La diffusion des technologies de l'information dans la société française." in *Enquête « Conditions de vie et Aspirations des Français »*. Paris: CREDOCBrossard Mathieu, Foko Borel, 2006, "Coûts et financement de l'enseignement supérieur dans les pays d'Afrique francophone." Pôle de Dakar (UNESCO-BREDA)

Charlier J.-É. (2006), "Savants et sorciers. Les universités africaines francophones face à la prétendue universalité des critères de qualité", *Education et sociétés* 2006/2 (18), 93-108

Charlier J.-É. , Croché S. (2003), "Le processus de Bologne, ses acteurs et leurs complices", *Education et sociétés* 2003/2

Fall B. (2007), "Survey of ICT and education in Africa: Burkina Faso Country Report." edited by InfoDev. DC: The World Bank

Hachicha S., Ouerfelli T. (2005), "Les stratégies d'apprentissage des technologies de l'information et de la communication : Le cas des étudiants de l'institut supérieur de documentation de Tunis", *Revue maghrébine de documentation*

Karsenti T., Ngamo S. T. (2007). " Qualité De L'éducation En Afrique: Le Rôle Potentiel Des Tic", *International Review of Education* 53 (5-6), 665-686

Médiametrie (2005). "Extrait «Education nationale» du premier baromètre des usages de l'Internet." Délégation aux usages de l'Internet, http://media.education.gouv.fr/file/52/6/526.pdf

Reesao (2005). "Réseau pour l'Excellence de l'Enseignement Supérieur en Afrique de l'Ouest", http://www.ub.tg/reesao/index.htm

Renaud P., Guyot B., (2007). "Projet OUIE, Enquête au Centre d'information sur la recherche pour le développement." Ouagadougou: IRD, http://www.tic.ird.fr/article.php?id_article=252

Traoré D. (2007). "Intégration des TIC dans l'éducation au Mali, Etat des lieux, enjeux et évaluation", *Distances et savoirs* 5 2007/1, 67 à 82

Unesco-Bangkok (2008). "ICT in Education." edited by Website. Bangkok, http://www.unescobkk.org/index.php?id=787

Unesco (2005). "Capacity Building of Teacher Training Institutions in Sub-Saharan Africa." http://unesdoc.unesco.org/images/0014/001406/140665M.pdf

Worldbank (2003). "Higher Education Development for Ethiopia: Pursuing the Vision." edited by A.S. STUDY. Washington: The World Bank, Africa Region, Education sector

Contribution des projets d'informatisation des ressources documentaires à la production scientifique dans les pays de zone TACIS

Contribution of the computerization projects of documentary resources to the scientific production in the countries of zone (TACIS)

Omar LAROUK

ELICO (Equipe de Lyon en Science de l'Information et de la Communication)
École Nationale Supérieure des Sciences de l'Information et des Bibliothèques
Lyon-Villeurbanne Cedex. France
omar.larouk@enssib.fr

Résumé. L'objectif des projets TEMPUS1 consiste à réduire le fossé numérique dans le cadre des orientations des programmes pour la réforme économique et sociale. Ces projets participent au développement des systèmes d'enseignement supérieur dans les pays bénéficiaires par une coopération avec les États membres de l'Union européenne. Nous allons nous intéresser uniquement aux pays de la zone TACIS (Europe de l'est et d'Asie) et notamment aux projets de management de l'université et d'informatisation de bibliothèques qui réduisent la fracture numérique, via l'accès à la documentation et le transfert de connaissances vers ces pays. Ces projets ont favorisé l'introduction de nouvelles pratiques documentaires dans les facultés des pays partenaires. Des résultats tangibles ont été obtenus grâce à la pertinence des projets, qui offrent un réel accès aux ressources documentaires multilingues, à travers la mise en place de centres de formation en IST (Internet et documentation en ligne) destinés aux étudiants, aux chercheurs et aux enseignants des pays bénéficiaires. Nous avons constaté, via notre participation à cinq projets européens TEMPUS, et aux indicateurs bibliométriques tirés de la base documentaire pluridisciplinaire Scopus®, une amélioration au niveau des productions scientifiques de certains pays de la zone TACIS.
Mots-clés. Ressources documentaires multilingues. Production scientifique. Transfert de connaissances. TEMPUS-TACIS. Information Scientifique et Technique (IST). Fossé numérique.

Abstract. The objective of TEMPUS projects consists in reducing the numeric ditch within the framework of the orientations of programs for the economic and

[1] TEMPUS (Trans-European Mobility scheme Program for University Studies) : *programme de mobilité transeuropéenne pour l'enseignement supérieur*

social reform. These projects participate in the development of the systems of higher education in the profitable countries by a cooperation with States members of the European Union. We are going to be interested only in the countries of the zone TACIS (Eastern Europe and of Asia) and notably in the projects of management of the university and the computerization of libraries which reduce the digital fracture, by way of the access to the documentation and the transfer of knowledge towards these countries. un résumé en anglais.
Keywords. Multilingual documentary resources. Scientific production. Transfer of knowledge. TEMPUS-TACIS. Scientific and technical information (IST). digital ditch.

1 Spécificités des projets TEMPUS : Objectifs et pays bénéficiaires

Le programme Tempus[2] I a débuté en 1990 avec tous les mouvements de démocratisation en Europe de l'Est. Ce programme, couverait la période1990-1994, suivi par un autre programme Tempus II dont la durée a commencé en 1994 pour une période de 6 ans (1994-2000). Ce programme concernait les pays comme *la Pologne, la Hongrie, la Bulgarie, la Roumanie, les Républiques Tchèque et Slovaque, l'Albanie, l'Estonie, la Lettonie, la Lituanie et la Slovénie* y compris les républiques de l'ancienne Union soviétique. Le programme TEMPUS III (2000-2006), destiné initialement aux seuls pays d'Europe centrale et orientale et à la Mongolie, a été étendu aux bénéficiaires des programmes TACIS précédents.

L'Union Européenne, à travers les projets TEMPUS, vise à mettre en place une politique informationnelle et des actions pour promouvoir l'infrastructure des TIC pour réduire ce fossé numérique. Les projets servent d'instruments à la modernisation du fonctionnement des services des universités, et notamment la modernisation des bibliothèques, la structuration des facultés, la réorganisation des programmes de formation, l'utilisation des TIC, et le renforcement des formations des enseignants. Cette coopération a été lancée, après la chute du mur de Berlin, en priorité entre les États indépendants du bloc communiste (ex. URSS) et l'Union européenne pour l'enseignement supérieur.

Le programme TEMPUS concerne plusieurs zones en contribuant efficacement, dans les pays bénéficiaires, à la diversification de l'offre d'enseignement et à la coopération inter-universités. Il crée, ainsi, des conditions favorables au développement de la coopération scientifique. Culturelle, économique et sociale. Les pays TACIS qui bénéficient de ces programmes sont : *Arménie, Azerbaïdjan, Belarusse, Fédération de Russie, Géorgie, Kazakhstan, Kirghizistan, Moldavie, Mongolie, Ouzbékistan, Tadjikistan, Turkménistan et Ukraine,*

Nous parlerons surtout des projets qui participent à la réduction de la fracture documentaire, via les programmes dans les domaines culturels et scientifiques, comme l'informatisation des bibliothèques universitaires et la mise en place de centre de documentation en IST.

[2] http://www.tempus.gov.mk/EN/tempus1.htm; http://europa.eu/scadplus/leg/fr/cha/c11020b.htm; http://europa.eu/scadplus/leg/fr/cha/c11020c.htm

2 Ressources documentaires multilingues : Apport des projets TEMPUS d'informatisation des bibliothèques

Le programme TEMPUS[3] a pour ambition générale de consolider la qualité de l'enseignement en encourageant la coopération avec des pays tiers afin d'améliorer la valorisation des ressources humaines et de promouvoir le dialogue entre les peuples et les cultures. L'objectif des projets TEMPUS est aussi de moderniser les structures universitaires et notamment l'informatisation des bibliothèques universitaires. En appliquant les méthodes modernes de gestion documentaire, les bibliothèques permettent une transmission des savoirs. En outre, elles favorisent la collaboration dans l'enseignement supérieur (conférences, ateliers, etc.).

Les projets TEMPUS donnent aux partenaires la possibilité d'acquérir des expériences dans l'Union européenne en favorisant les mobilités des bibliothécaires, des universitaires et des institutionnels entre les établissements partenaires. Cette coopération améliore, certes, l'image de l'Union européenne par la diffusion d'informations sur les pays membres, mais elle permet aux bénéficiaires de s'informer des nouvelles techniques utilisées pour valoriser l'information documentaire.

Nous limitons notre analyse aux bibliothèques pour étudier le degré de leur implication dans la diffusion de la culture et la production scientifique numérique. Notre participation et/ou l'expertise de plusieurs bibliothèques universitaires comme la bibliothèque de l'université de Volgograd[5], la bibliothèque de l'université de Chisinau[4], la Bibliothèque de l'université d'État de La Crimée, Yalta[5], la bibliothèque de l'université d'État d'Achkhabad[6], la bibliothèque de l'université polytechnique de Bichkek[7].

2.1 Etat des lieux de la documentation dans les universités évaluées : quelques points faibles

Nous avons participé à l'expertise des bibliothèques des universités précédentes, et nous pouvons dresser l'état des lieux suivant :

- Que les publications de documents scientifiques (articles, actes de colloques, thèses, mémoires, etc.) sont très faibles ;

---

[3] L'ENSSIB-Lyon est l'un des partenaires européens des projets TEMPUS portant sur la modernisation de la gestion des universités de Bakou (Azerbaïdjan), Bichkek (Kirghizistan), Achkhabad (Turkménistan) avec notamment l'université de Nice. En tant qu'expert-européen, j'ai participé activement aux formations des bibliothécaires et des universitaires (Yalta, Volgograd, Achkhabad, Bichkek, Bakou) depuis 1998.

[4] Expertise du centre des ressources documentaires de la bibliothèque de l'Université Internationale de Chisinau (ULIM) en 2007.

[5] Rapport d'expertise à la bibliothèque de l'université de Yalta du 27 février au 7 mars 2005, Tempus JEP 22040-2001. TEMPUS-TACIS PROJECT N°IB-JEP-22045-2001/UKR; Land Reform And Land Market Development In Ukraine; Yalta, 14 April 2005.

[6] En tant que responsable de la documentation électronique et Centre de Formation de l'Information Scientifique et Technique à l'Université d'Achkhabad, Turkménistan ; TEMPUS TACIS; PROJET EUROPEEN COMMUN UM-JEP/TME 22040-2001, 15 April 2003- 31 May 2006.

[7] Modernisation de la bibliothèque et Mise en place d'un Centre de Formation et de Promotion de l'Information Scientifique et Technique à l'Université Technique du Kirghizistan (KTY) ; Projet TEMPUS-TACIS n° UM-JEP-22042-2001/KYR

- Que l'accès à la production scientifique (BBD spécialisées, portails, Internet, ..) était très limitée ;
- Que les fonds documentaires gérés par ces bibliothèques sont pauvres et obsolètes ;

Etc..

2.2 Une documentation électronique inexistante et une production scientifique irrégulière

Depuis les années 1990, on a constaté une absence d'informations sur les périodiques, et les nouveaux éditeurs de ressources électroniques. En outre, les personnels universitaires (enseignants, chercheurs, bibliothécaires) n'ont pas été formés aux nouvelles techniques bibliothéconomiques et, surtout, aux nouvelles technologies de l'information et de la communication. Les services rendus à la communauté universitaire se limitent uniquement aux documents primaires. Les principales lacunes que nous avons rencontrées dans ces bibliothèques universitaires sont :

- Un manque important sur la documentation électronique, sur les bases de données spécialisées comme (*DIALOG, MEDLINE, SCIRUS, Ingenta, Science Direct, ...)*, sur les bibliothèques numériques *(OPAC-web, SFQC, SBIG, Bases de signets, ...), etc.* ;
- Les productions scientifiques en langues étrangères (anglais, français. etc.) sont inexistantes ;
- Les publications ou communications scientifiques en russe et/ou en langue nationale sont très faibles ;
- L'absence de pratiques de normes internationales, comme les formats *(MARC, Z39'50, ISO2780, etc.)*, ne favorise pas l'interopérabilité et la visibilité des catalogues locaux ;
- Le retard pris dans la mise en place de système de gestion des bibliothèques, des réservoirs de documents électroniques, des sites de publications des conférences, des archives de la littérature grise, etc.

3 Transfert des ressources documentaires aux universités et aux usagers, via les centres de documentation (IST)

3.1 Mise en place de Centres de Formation et de Promotion de l'Information Scientifique et Technique (IST) dans les universités

Les projets TEMPUS favorisent la création de centres de formation et de promotion de l'information Scientifique et Techniques dans les universités. Tous les projets de coopération doivent veiller à la pérennité des résultats obtenus, et au transfert des expertises. Ce transfert doit comprendre le développement de centre de formation en IST afin de renforcer la production de l'information scientifique.

L'objectif final de ces centres est d'insérer dans le cursus de tous les étudiants un module de méthodologie de recherche documentaire pour optimiser l'utilisation des outils de la documentation électronique. mais la priorité à la formation de formateurs est donnée aux bibliothécaires et enseignants. qui joueront le rôle de relais de transmission.

Les ressources documentaires dédiées spécifiquement aux langues et aux corpus linguistiques progressent de manière importante sur le réseau Internet. Cette disposition offre à l'étudiant. à l'enseignant et au chercheur. un ensemble de ressources sélectionnées par des bibliothécaires ou des experts du domaine.

De même, la consultation des base de signalement d'articles scientifiques et bases bibliographiques gratuites comme :- OAIster (www.oaister.com); SCIRUS (www.scirus.com); SUDOC (www.sudoc.fr); GoogleScholar (http://scholar.google.com) -, permettent aux chercheurs de constituer des bibliographiques actualisées de leurs travaux.

Un accès à la documentions numérique multilingue (gratuite et/ou payantes) a permis de consulter la masse des ressources existantes, qui est complétée par les revues électroniques et des bases de données payantes. La documentation en ligne est une information actualisée et disponible à partir de plusieurs points d'accès, en plusieurs langues, que les lecteurs de ces pays peuvent consultés.

4 Production scientifique de certains pays TACIS : Synthèse sur les observations bibliométriques

Développée par Elsevier, (http://www.scopus.com/), la base de données SCOPUS est pluridisciplinaire et multilingue en sciences techniques et sociales (STM & SHS). Elle contient *-plus de 15 000 revues scientifiques concernées, 500 journaux universitaires de libre d'accès, 700 compte-rendu de conférences, 600 publications commerciales-, etc).* Plus variée que le WoS [8] (web of Science -ISI), elle permet de faire des recherches [9] par pays, via le moteur *Scimagojr (SJR)* qui montre la visibilité des journaux et des productions scientifiques des pays depuis 1996.

On peut constater que l'Ukraine, qui a bénéficie de beaucoup de projets TEMPUS, présente une nette augmentation des productions documentaires liée à l'organisation de colloques, de séminaires et/ou de journées de formation depuis 1996. Rappelons que la communauté européenne exige la production de document de dissémination pour valoriser les projets TEMPUS. Néanmoins, on observe que la production de l'Ukraine (64985 documents) est largement dépassée par la Russie (367560 documents), qui est elle-même, est surpassée par la production française (729133 documents). Tous ces pays sont dominés par le principal producteur de la science (*les Etats-Unis)* avec presque 4 millions de documents *(cf. tableaux 1, Annexes).*

On relève la domination écrasante des pays du Nord (USA, UK, Japon, Allemagne, France, Canada, Italie, Espagne, etc.) et des pays émergents comme la Chine en 5ème position avec 960 669 documents, l'Inde en 12ème position 334512 documents, la Corée du Sud en 14ème position avec 263401 documents produits, le Brésil en 17ème position avec 195541 documents, et enfin Taiwan en 18ème position avec 195522 documents produits.

---

[8] Il est difficile de tirer des enseignements totalement fiables dans la mesure où le web scientifique est avant tout dominé par la langue anglaise. Le contenu de la base WoS (web of Science) contient plus de 98% d'articles en anglais, 0.23% en français, 0.20 en chinois, etc.) dans son index.

[9] A l'index de la base SCOPUS, il faut ajouter tous les documents répertoriés par le moteur gratuit SCIRUS : http://www.scirus.com/. C'est un moteur de recherche spécialisé sur le web scientifique, il intègre outre les ressources publiées par le groupe Elsevier, des documents issus de bases de données, de serveurs de prépublications et d'Open Archives.

Au delà des aspects liés aux algorithmes de Scopus et Scimago, qui favorisent les publications en langue anglaise et notamment le contenu des revues et des actes de congrès produits par les pays du Nord, néanmoins, on observe une légère percée de certaines langues comme le chinois, l'espagnol, l'allemand, le russe, etc..

Il faut relever également la très grande faiblesse des productions de documents numériques des pays comme le Turkménistan, le Kirghizstan, la Moldavie, et l'Azerbaïdjan par rapport à l'Ukraine et la Russie (cf. tableau 2 & 3, Annexes). Il faut signaler que toute la documentation *(fonds documentaires des bibliothèques évaluées)* dans ces pays est en langue russe (plus 95%) avec un très faible pourcentage des langues nationales et de l'anglais.

Il faut noter la part des publications scientifiques des pays comme le Turkménistan et le Kirghizstan est presque nulle (0.3% de la production mondiale). Le potentiel des ces pays est certes faible, mais il faut croiser ces résultats avec les données répertoriées par d'autres bases comme le WoS (Web of Science) et les réservoirs d'open access.

Dans le tableau 4 *(cf. Annexes),* on peut constater que le pourcentage de collaboration internationale dans la production scientifique de documents des pays suivants : *Turkménistan ; Kirghizstan ; Moldavie ; Azerbaïdjan* reste très élevé. Ceci s'explique par l'apport de projets européens de coopération du type TEMPUS.

Les projets TEMPUS-TACIS contribuent ainsi, à réduire les disparités existantes à l'ère du numérique par la réforme des institutions d'enseignement supérieur, par le renforcement de la coopération, et notamment par la diffusion des résultats issus des programmes au niveau national et international. Ils devront déboucher sur des résultats tels que des programmes de mobilités et/ou stages de formation qui présentent un impact durable pour plusieurs universités.

5 Conclusion

Les projets TEMPUS-TACIS sont destinés à faciliter l'adaptation de l'enseignement supérieur aux nouveaux impératifs socio-économiques et culturels dans les pays bénéficiaires pour réduire la fracture numérique.

Ces projets visent à moderniser la gestion des structures universitaires pour les aider à s'adapter à un environnement en pleine évolution. L'introduction de nouvelles méthodes de management : *-modernisation de la bibliothèque et des services pour les chercheurs, renforcement de la publication scientifique via les actes de colloques, renforcement des liens entre universités,* permettent de réduire le fossé numérique entre universités des pays développés et celles des pays de l'Europe de l'Est ou d'Asie.

L'université et sa bibliothèque doivent accomplir les missions de formation, mais aussi de lancer des opérations de vulgarisation des connaissances comme la constitution des bases de signets validées par les experts du domaine (enseignants de l'université). La constitution de sites fédérateurs spécialisés compensera le manque d'ouvrages, de documentations techniques et de références actualisées dans le domaine scientifique et technique. L'introduction de ressources documentaires électroniques locales et la généralisation de l'accès à la documentation électronique mondiale (gratuite ou payante) constituent les principaux facteurs pour réduire le fossé numérique, via l'ingénierie du document. L'encouragement des échanges, la production et la diffusion d'informations scientifiques spécifiques, via les outils documentaires et technologiques, répondent à certains besoins stratégiques pour réduire la fracture documentaire.

6 Annexes : Données statistiques issues la base *Scopus* Scimago

Pays	Productions de Documents	Classement
United States	3 872 452	**1**
……	……	……
China	960 669	5
France	729 133	**6**
Russian	367 560	**10**
……	……	
Ukraine	64 985	**33**
……..	……..	
Azerbaijan	3 294	**86**
……. ;	……..	
Moldova. Republic	2 639	**91**
……		
Kyrgyzstan	500	**139**
……	……..	
Turkmenistan	100	**184**

Tableau 1 : *Classement 'mondial' des pays TACIS sur 233 pays. tiré à partir de la base documentaire Scopus, via la base de SCImago Research Group. Scopus® pour la période 1996-2007*

	Ukraine	Russie	France	USA
1996	5 099	29 728	50 493	309 452
1997	5 535	30 152	53 669	303 655
1998	5 498	30 867	54 625	299 429
1999	5 274	29 672	55 218	293 685
2000	5 205	29 974	55 072	296 490
2001	*5 888*	30 615	52 978	290 187
2002	*5 216*	30 052	52 959	292 723
2003	*5 144*	31 019	59 728	310 961
2004	*5 830*	30 839	61 220	292 962
2005	*6 311*	34 356	66 724	313 478
2006	*5 031*	29 743	66 292	323 049
2007	4 603	28 236	63 122	313 441

Ordre de la production scientifique : **Ukraine < Russie < France < USA**

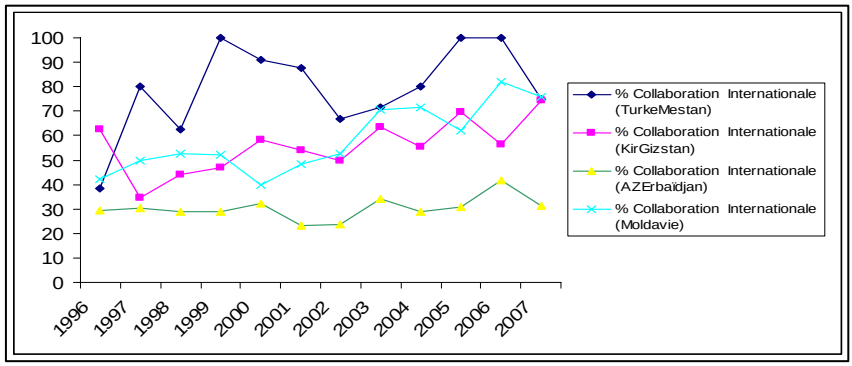
Tableau 2 *: Comparaison des productions de documents scientifiques dans les pays comme l'Ukraine et la Russie par rapport aux publications scientifiques de la France et les Etats-Unis entre 1996-2007*

	Turkmenistan	*Kirghizstan*	*Moldavie*	*Azerbaijan*	**Ukraine**	**Russie**
1996	12	30	219	247	5 099	29 728
1997	10	23	201	226	5 535	30 152
1998	8	34	191	196	5 498	30 867
1999	3	34	201	185	5 274	29 672
2000	11	36	232	183	5 205	29 974
2001	8	39	184	168	5 888	30 615
2002	9	36	193	257	5 216	30 052
2003	6	30	229	322	5 144	31 019
2004	5	38	191	393	5 830	30 839
2005	8	71	314	394	6 311	34 356
2006	10	64	255	340	5 031	29 743
2007	8	61	211	363	4 603	28 236

Tableau 3 *: Comparaison de la production scientifique des 6 pays TACIS :*
Turkmenistan < Kirghizstan < Moldavie < Azerbaijan < Ukraine < Russie

%Collaboration Internationale	% (Turkménistan)	% (Kirghizstan)	%(Azerbaïdjan)	% (Moldavie)
1996	38.46	62.50	29.15	42.01
1997	**80.00**	34.78	30.53	49.75
1998	62.50	44.12	28.93	52.60
1999	**100.00**	47.06	29.03	52.24
2000	**90.91**	58.33	32.24	39.91
2001	**87.50**	53.85	23.21	48.37
2002	66.67	50.00	23.74	52.82
2003	71.43	63.33	33.95	**70.56**
2004	**80.00**	55.26	28.89	**71.65**
2005	**100.00**	**69.44**	30.83	62.03
2006	**100.00**	56.25	**41.52**	**82.03**
2007	75.00	**74.19**	31.06	**76.04**

Tableau 4 : *Pourcentage des collaborations internationales dans la production scientifique de documents des pays : Turkménistan; Kirghizstan ; Moldavie; Azerbaïdjan.*



Les boutiques de communication à Château-Rouge (Paris) : une contribution privée à la réduction de la fracture numérique ?

Phone cabins and cybercafes in the Chateau-Rouge commercial zone in Paris : a migrants' contribution to reduce the digital divide.

Claire SCOPSI

CRIS, Paris 10-Nanterre, Paris, France
claire.scopsi@wanadoo.fr

Résumé. De 2000 à 2005, le nombre de commerces consacrés à la communication numérique (téléphonie, cybercafés) a connu une importante croissance. Dans le quartier de commerce ethnique de Château-Rouge (Paris 18) ces commerces, créés par des migrants, prennent modèle sur les télécentres des pays du Sud et proposent une gamme de services adaptée à la clientèle immigrée. Toujours aussi nombreuses en 2008, ces boutiques qui ont accompagné le développement de l'internet haut débit et de la voix sur IP font partie du paysage urbain . Elles ont contribué à rendre les technologies de communication familières aux migrants comme aux autochtones.
Mots-clés. Fracture numérique – commerce ethnique – cyber café – télé boutique-immigration.

Abstract. From 2000 to 2005, numerous cybercafes and phone cabins were created by migrants in the Chateau-Rouge commercial zone in Paris (18). Imitating the southern countries' telecenters, these stores offer services adapted to migrants' needs, and contribute to accustom both migrants and french people to Information and Communication Technologies.
Keywords. Digital divide – telecenters – ethnic trade.

1 Château-Rouge : le « triangle d'or » de la communication numérique

Château-Rouge est un territoire moderne, marchand et pluriethnique, et surtout étonnamment contrasté. Ce quartier « sensible », à l'habitat dégradé, concentre depuis quelques années la vente et la consommation de crack de la capitale. Le XVIIIème arrondissement parisien, comptait en 1999, 184 581 habitants dont 35 213 étrangers (19,1%), parmi lesquels 28 365 (15,4%) n'étaient pas issus de la communauté européenne. Les populations Algériennes, Tunisiennes, Marocaines, Africaines (hors Maghreb), Indiennes (sous continent indien) et européennes (hors Communauté), y sont sur-représentées par rapport à la moyenne parisienne et le taux de chômage y était à cette date de 16,9 % (le taux de chômage parisien se situant à 12 %). Le quartier administratif de la Goutte d'Or, où est localisé le secteur « Château-Rouge », affichait un taux de chômage de 23,1 %. 18,1 % de la population de l'arrondissement n'avait aucun diplôme (13 % à Paris, 28,6 % à la Goutte d'Or).

Ces chiffres suggèrent un lieu déshérité, un ghetto. Mais impossible d'appliquer ce qualificatif à un quartier de commerces prospères, abondamment desservi par les transports en commun et qui reçoit chaque semaine la visite de plusieurs milliers de visiteurs attirés de toute l'Ile de France par son « marché africain » : on achète la nourriture à Château-Rouge, parce qu'on est missionné pour cela par un groupe d'amis, parce qu'on manque de certains ingrédients et que "ce soir c'est africain", par nostalgie, pour retrouver les marques, les emballages de là-bas, glaner des nouvelles du pays dans les boutiques... et puis l'on file à Barbès-Rochechouart, chez Tati, parce qu'on y trouve une gamme de prix imbattable et au passage on fait un saut chez Toto où le tissu vendu au poids est si avantageux.

Château-Rouge, qu'une lettre spécialisée en nouvelles technologies désignait comme le « triangle d'or de la voix sur IP »[1], est aussi, depuis la fin 2000 et de plus en plus ouvertement un territoire de télécommunications. En hébergeant, à l'instar des autres quartiers ethniques des villes occidentales, un nombre important de commerces et services consacrés aux télécommunications, sur un mode adapté à la clientèle des migrants, ce quartier réputé difficile a, à sa façon, contribué à minimiser la fracture numérique auprès de ce public a priori « éloigné », et apporté un coup de pouce aux développement des usages des télécommunications dans leurs pays d'origine. Au delà, les boutiques de communication des migrants ont maintenu en France une offre d'accès public à l'internet volontairement négligée par les services publics.

2 Château-Rouge et la communication de 2001 à 2003

Dans les ruelles étroites et encombrées de Château-Rouge et le long des grands axes routiers (Boulevard Ornano et Boulevard Barbès) nous avons dénombré en 2004 pas moins de 80 boutiques consacrées sous une forme ou une autre aux télécommunications : cabines de téléphonie internationale, consultation internet, accessoires et forfaits de téléphonie mobile- dont de mystérieux commerces de « GSM Afrique », pour beaucoup nées entre la fin de l'année 2000 et 2004. Encore a-

[1] La voix sur IP fait ses premières armes dans les téléboutiques. Christophe Guilemin .-Zdnet France, 04 février 2004. http://www.zdnet.fr/actualites/technologie/0,39020809,39140424,00.htm

t-il fallu renoncer à identifier les commerces de cartes téléphoniques prépayées en direction de tous les pays du monde car tous les commerces de Château-Rouge, épiceries, salons de coiffure les proposent à leurs clients.

Le nombre de téléboutiques a doublé entre 2001 et 1003, (il s'est créé presque une téléboutique par mois en 2003). Le nombre de points d'accès à l'internet a également doublé pendant cette période montrant que l'introduction de cette technologie dans le quartier n'était pas un avatar sans lendemain de la netéconomie, mais que ces implantations pouvaient survivre à la crise des start-up.

Mais au delà de leur densité, la logique d'exploitation de ces commerces intrigue.

- L'implantation des boutiques de communication recouvre exactement la zone de commerce ethnique :

- L'aspect des boutiques, petites et sommairement meublées aux vitrines occultées par de nombreuses affichettes, les enseignes (Super Kin City, Ganesa.com, World Communication Afro), l'association même des produits et services proposés (téléboutique-salon de coiffure, marchands de wax et de cartes prépayées) relève d'un modèle non occidental. Les commerçants de Château-Rouge confirment s'inspirer en cela des télécentres nés dès le début 2000, dans les capitales Africaines (à Dakar notamment). Il s'agit là, et c'est un phénomène assez rare, d'une dynamique Sud/Nord.

- La combinaison des services proposés, soit dans une même boutique, soit par la juxtaposition dans une même zone de plusieurs boutiques complémentaire, offre une réponse à la fois originale et adaptée aux problématiques des migrants. Nombre de boutiques offrent des « pôles » d'accès aux technologies utilisés par les commerçants et les petits entrepreneurs locaux : reprographie, photographie, fax et e-mail en émission et réception, téléphonie et consultation internet. Les boutiques de téléphonie s'installent à proximité des agences de voyages spécialisées sur la zone Afrique, et des services de fret et d'import export. A la pratique de l'internet sont associés des services d'initiation à la micro-informatique, de traduction d'actes administratifs, d'assistance à la rédaction des mails, d' « aide aux familles[2] », ou de « communication avec image »[3] .

- Le ressort de cette activité, comme celle marché exotique tout entier (car salons de coiffure, restaurants, épiceries sont autant de lieux de retrouvailles et d'échanges d'information entre compatriotes), réside dans le souci de conserver le lien avec le pays quitté, pourtant cette relation n'apparaît que très rarement sous une forme nostalgique.

- Si les grands groupes de communication usent de l'argument nostalgique en illustrant les télécartes de paysages africains ou de visages d'enfants ou de personnes âgées, le « discours » des téléboutiques, celui choisi par les commerçants migrants, révèle un grand pragmatisme : les prix sont tirés vers le bas, car le migrant prend le plus souvent en charge le coût des appels téléphoniques et y consacre un important budget ou offre des téléphones portables aux membres de la famille, lors de ses séjours au pays, le téléphone portable n'est pas représenté comme un bijou technologique, sophistiqué et performant, mais vendu dans les bazars, parmi les ustensiles ménagers, comme un objet d'usage courant.

[2] C'est-à-dire : envoi d'argent.

[3] C'est-à-dire : visiophonie.

3 Château-Rouge et la communication en 2008

En 2008 la situation n'a pas radicalement changé : dans la quartier en cours de réhabilitation, certaines de ces boutiques ont fermé ou changé de propriétaire, mais d'autres se sont créées à un rythme cependant moins spectaculaires qu'en 2001-2003, à l'exception des boutiques de téléphones portables dont le nombre a plus que doublé dans la zone sud du boulevard Barbès où elles constituent une quasi mono activité, les boutiques de communication demeurent aussi nombreuses à Château-Rouge, elle n'ont pas connu le sort des cybercafés des quartiers branchés de Paris de la fin des années 90 qui après une croissance spectaculaire ont quasiment disparu en quelques années.

Cependant la forme de ces commerces a évolué et s'est, en tous cas en apparence, « professionnalisée ». Les boutiques mixtes associant deux ou trois cabines téléphoniques à un restaurant, un commerce de tissus ou un salon de coiffure ont disparu, les boutiques suivent désormais deux modèles : les boutiques de téléphones mobiles, Gsm, et accessoires, souvent franchisées, et les boutiques de services multiples : cabines téléphoniques internationales, postes internet, fax, photocopies, scanners complétées par la vente de cartes internationales prépayées et d'accessoires pour téléphones mobiles. Les propositions de traduction ou « d'aide aux familles » ont disparu, mais des agences spécialisées dans le transfert d'argent Money Gram et Western Union sont venues à partir de 2006 rejoindre la gamme des services spécialisés aux migrants (agence de voyages, agence de fret, pompes funèbres islamiques...) qui, avec les commerces alimentaires exotiques, ont construit l'identité de Château Rouge.

Les usages de ces services ont également évolué : en 2003 les accès publics à l'internet étaient généralement moins utilisés que les cabines téléphoniques internationales, et les usagers étaient plutôt des français ou des européens (alors que les cabines téléphoniques étaient beaucoup utilisées par les clients maghrébins, indiens ou africains). En 2008 les postes internet sont beaucoup plus utilisés, par tout type de public, et les boutiques ont généralement augmenté le nombre postes. La visiophonie est désormais une pratique courante.

4 Une contribution privée à la réduction de la fracture numérique

Indirectement, les boutiques de communication des quartiers migrants ont marqué le développement des TIC dans les pays du Sud auxquels elles ont emprunté leur modèle. Les « Gsm Afrique » vendus en « gros, demi-gros et détail » ont alimenté le marché africain, revendus par les migrants ou donnés, car ils constituent un présent apprécié, lors du retour dans la famille. Les pratiques développées dans les territoires bien équipés du Nord, le téléchargement, la visio-conférence, créent une demande d'infrastructure dans les pays du Sud. Au delà des accès publics, les migrants ont joué un rôle de médiation constant dans le développement des TIC au Sud. En initiant, depuis l'occident, des portails d'actualité, de promotion de leur pays et de leur culture, ou des sites destinés à rassembler la diaspora d'un village ou d'une région, ils ont assuré, même modestement, la présence de leur pays et de leur culture d'origine sur le web.

En proposant, dans les quartiers dits « difficiles », des accès à l'internet à des tarifs bas et dans un environnement de proximité facilement appropriable par les immigrés, en accompagnant les premiers pas de leurs clients sur le net, en proposant des services de rédaction de mails et de traduction, les téléboutiques ont contribué à la réduction de la fracture numérique « sociale ». En 1997, dans le cadre du Programme d'Action Gouvernementale pour la Société de l'Information (PAGSI),le

gouvernement Jospin avait confié ce rôle d'initiation des « publics éloignés » aux espaces publics numériques, labellisés et organisés en divers réseaux subventionnés. La suppression du statut d'emploi jeune (statut de nombre des animateurs socio culturels de ces Espaces publiques) par le gouvernement Raffarin, puis sa décision de privilégier l'équipement individuel pour la lutte contre la fracture numérique, portèrent un coup fatal aux EPN. Si certains survécurent (il existe un Espace Publique Numérique à Château Rouge), leur développement fut limité à partir de 2001 et 2008 a vu l'annonce de la suppression du label Espace Culture Multimédia (ECM) l'un des principaux réseaux d'accès public à internet. Cette situation abandonne de fait ce rôle d'initiation et d'animation au secteur privé, sans accompagnement ni contrôle.

On assiste donc, au Nord comme au Sud, à une prise en charge des accès publics aux télécommunications par le secteur privé en raison de la faiblesse du secteur public au Sud, de l'indifférence des pouvoirs publics à la question des accès collectifs au Nord. Les boutiques de communication apportent leurs services aux touristes, étudiants, riverains non équipés ou momentanément privés de leur équipement. Les téléboutiques, toujours créées par des migrants, sont présentes désormais (mais de façon moins dense), dans les quartiers non ethniques: comme l'épicerie de proximité, les services de télécommunications sont devenus dans tout Paris une spécialité de l'entrepreneuriat migrant.

Mais c'est peut dire que les pouvoirs publics manifestent peu d'intérêt pour ces entreprises : au contraire, depuis les attentats terroristes d'Al Qaida et notamment la tentative infructueuse de Richard Reid[4], qui a séjourné à Château-Rouge et y a fréquenté une téléboutique, l'image des téléboutiques des quartiers ethniques est associée au risque terroriste. Comme le souligne, en novembre 2005, le rapport à l'Assemblée Nationale sur le projet de loi relatif à la lutte contre le terrorisme : « [les criminels]...*privilégient des accès au réseau qui protègent leur anonymat, par exemple les connexions « Wi-Fi », ou l'utilisation d'ordinateurs publics, notamment dans les cybercafés, comme l'avait montré l'affaire Richard Reid.*[5] *».*

L'image officielle des téléboutiques, même des années après les attentats d'Al Qaida, reste donc diabolisée. Trop marchands pour les politiques de gauche, trop modestes pour intéresser les politiques de droite, ces commerces n'auront jamais pu voir leur rôle souligné. Si leur présence témoigne encore de la contribution des migrants au développement des TIC au Sud, cette trace urbaine est vouée à s'effacer lorsque les franchises des grands groupes de télécommunication auront achevé dans quelques années de remplacer les premières boutiques indépendantes.

Références :

Arnaud, M., Perriault, J., (2002). Les Espaces Publics d'accès à internet. PUF, Paris.

Castells, M. (1998). La Société en réseaux : l'ère de l'information. Fayard, Paris.

[4] Connu comme « l'homme à la chaussure piégée », Richard Reid est l'auteur d'une tentative d'attentat à l'explosif dans le vol Paris/Miami en décembre 2001. Les passagers du vol sont parvenus à le neutraliser.

[5] Rapport fait au nom de la commission des lois constitutionnelles, de la législation et de l'administration générale de la République, sur le projet de loi (N° 2615), après déclaration d'urgence, relatif à la lutte contre le terrorisme et portant dispositions diverses relatives à la sécurité et aux contrôles frontaliers, Par M. Alain Marsaud, Député.-Assemblée Nationale n°2681.

Chéneau-Loquay, A. (dir.). (2004). Mondialisation et Technologies de la communication en Afrique». Karthala, Paris.

Institut Panos Paris, (2001). D'un Voyage à l'autre : des voix de l'immigration pour un développement pluriel. Karthala, Paris.

Kiyindou, A., (2003). La Place des savoirs africains sur internet ou penser la "fracture numérique" par le contenu disponible à : http://www.ticom.info/indexarti.htm.

Sayad, A. (1985). Du Message oral au message sur cassette : la communication avec l'absent. In Actes de la Recherche en Sciences Sociales, n°59, 61-73.

Scopsi, C., (2004). Sortir du modèle de la fracture numérique : l'apport de la diaspora. in « Société de l'information, Société du contrôle » CREIS/Terminal.

De l'oral à l'écran : des administrations numérisées. Réduire la fracture, créer la fracture. Exemples africains.

From oral to screen : digital administrations. Reduce digital divide, create divide. African examples.

Michel LESOURD

UMR Cnrs 6266 IDEES (Equipe LEDRA), Université de Rouen, Mont-Saint-Aignan, France
Michel.Lesourd@univ-rouen.fr

Résumé. Les pays du Sud développent leurs propres innovations numériques : la fracture numérique doit être appréciée de manière nuancée. L'un de ses aspects les plus marquants en Afrique est l'implication de l'Etat comme maître d'oeuvre de la création d'une société de l'information. Outre la réalisation d'infrastructures, une véritable volonté politique a permis de mettre en place des appareils juridico-administratifs et de planification, avec une filiale de grand groupe international placée en situation de monopole commercial, avec une mission de service public.
Si la remise en cause actuelle des monopoles « historiques » semble se traduire par le triomphe de l'oralité, avec l'explosion des ventes de téléphones cellulaires, les nouvelles pratiques innovantes sont plutôt mises au service du document numérique : e-gouvernement pour une bonne gouvernance politique, avec priorité à la e-administration au service du citoyen. L'innovation « par le bas », ou les usages populaires du document numérique, est active : les innovations, « réinterprêtées » (le fax comme « lettre de crédit ») sont mises au service des pêcheurs et des paysans.
Les inégalités socio-spatiales sont plus que jamais à l'ordre du jour : entre pays, entre espace rural et villes, entre régions, entre info-riches et info-pauvres. La disponibilité en énergie électrique joue un rôle décisif. Le « modèle fractures numériques » se caractérise par le renforcement du pouvoir des e-élites, informées et communicantes, et un nombre considérable de pauvres et « d'oubliés du numérique ».
Mots-clés. Révolution numérique, fracture numérique, e-gouvernance, e-élites, info-riche, info-pauvre

Abstract. Developing countries have their own digital innovations : digital divide must be estimate with nuances. In Africa, one of the main aspects is the State participation as a command work to create an information society. With the infrastructures, a high political voluntary permits to realise legal and administrative structures with monopolistic international companies subsidiary, with public service.

The end of historic monopoly seems to be the orality triumph, with the cellular using explosion.But innovating practices are oriented to digital document : e-government for best governancy, with priority to e-administration for best citizen services. Popular use of digital document is active : reinterpreted (the Fax as a credit card), they are now for fishermen an peasants.
Socio-spatial inegalities are topicality : between countries, rural land and cities, between regions, info-rich and info-poor people. Electricity has a decisive constrainst rôle. Characteristics of the digital divide model are the reinforcement of e-elites power, informated, and a great number of poor and « digital forgotten people ».
Keywords. Digital revolution, digital divide, e-governancy, e-elite, rich digital people , poor digital people

Même si c'est d'abord dans les pays dits « du Nord » que les divers aspects de la *révolution numérique* se développent, l'informatisation des sociétés des pays en développement est aussi en marche. Loin de se tenir à l'écart de cette révolution, la plupart des Etats s'y intéressent depuis au moins une dizaine d'années, particulièrement du point de vue de la gouvernance. Mais, dans les Suds, les politiques, les capacités d'innovation et les rythmes des avancées de la « société de l'information » sont différents de ceux des pays du Nord.

On s'intéressera ici au cas des pays du continent africain, afin de montrer que la *fracture numérique* doit être non seulement appréciée de manière nuancée entre Nord(s) et Suds, mais qu'elle révèle des situations et des dimensions variables à différentes échelles, L'un des paradoxes de cette fracture est que, dans les Suds, *l'intelligence numérique* peut aussi être innovante, en s'adaptant à des conditions locales caractérisées par des ressources financières limitées, des qualifications humaines rares et des infrastructures spatialement déficientes.

L'un des aspects les plus marquants de la révolution numérique en Afrique est l'implication des gouvernements des états comme maître d'oeuvre de la création d'une société de l'information dans leur pays, notamment en matière de réalisation d'infrastructures, de politique d'offre de services, de législation et de régulation. Il n'est donc pas surprenant de constater que nombreux sont les pays à avoir donné la priorité à l'informatisation de leur administration d'Etat, tant dans son fonctionnement interne que dans ses rapports avec le citoyen administré. Beaucoup ont aussi mis en place une politique numérique dans des secteurs où le document écrit symbolise à la fois la puissance de l'Etat et celle de la maîtrise du savoir contemporain, comme l'éducation-formation, les archives, les bibliothèques et aussi de nombreux autres services d'Etat, comme la surveillance de l'environnement et l'aménagement du territoire. De plus, quelques uns, aux marges des structures politico-administratives, mais à travers l'encadrement ministériel ou consulaire, favorisent des initiatives privées utilisant l'écrit et les technologies de l'information pour améliorer la gestion d'activités économiques vitales pour le pays (pêche, agriculture...) et rechercher l'indispensable mise à niveau technologique, seule capable de permettre l'insertion dans l'économie mondialisée.

On rappellera que la révolution numérique (de l'oral et du papier à l'écran et à l'oral) est ambiguë : elle est, certes, progression et généralisation de l'écran et de l'écrit comme moyen de communication et de gestion sociétale ou privée, mais elle est d'abord, à ce jour, en Afrique, triomphe de l'oralité, avec la formidable expansion de l'usage du télephone cellulaire (en 2007, le taux d'équipement de l'Afrique est de 47%)

Rappelons qu'en matière de TIC, l'écrit et l'oral vont de pair, avec une utilisation croissante de l'écran-écrit pour le cellulaire (SMS, texto, Web) et de l'écran-oral pour l'ordinateur (Skype...).

1 De la volonté numérique à la e-gouvernance : politiques et outils au service du document numérique

En Afrique, le développement du numérique s'est fait « par le haut » : Etat, collectivités territoriales, entreprises de service public sont le « moteur » de l'informatisation des sociétés et des territoires. Même si « le bas » se révèle porteur d'initiatives, celles-ci sont toujours « en retrait » par rapport aux « machineries numériques » lancées par les gouvernements. Mais les conditions de la dynamique des politiques numériques sont d'abord au service de l'écrit.

1.1 Une volonté politique au service du document numérique

Les objectifs stratégiques des politiques de développement des TIC en Afrique sont en partie liées à la volonté de pallier l'indigence des infrastructures anciennes (le téléphone filaire) dans les territoires. Elles se caractérisent, dans les années 1995-2000, par la mise en place et la promotion d'outils de communication nouveaux, quoique déjà opérationnels dans les pays du Nord, comme le fax et l'internet. Une autre ligne de force de ces politiques est la dépendance des Etats vis à vis des promoteurs et installateurs d'infrastructures de circulation des données : les « backbone » et les réseaux internationaux de câbles sont presque tous entre les mains de consortiums dominés par les financements et des technologies des pays du Nord.

Enfin, les objectifs majeurs sont prioritairement tournés vers la satisfaction des besoins des Etats en administration, et la promotion de l'outil écrit et de l'image. De très nombreux pays ont élaboré des plan de développement : ainsi, du Cap-Vert, qui a lancé, en novembre 2005, deux plans stratégiques : le PESI, Programme Stratégique pour la Société de l'Information, et le PAGE, Plan d'Action pour la Gouvernance Electronique qui définissent les options de développement de la société de l'information cap-verdienne, et où la promotion du document numérique joue un rôle essentiel.

1.2 L'organisation institutionnelle, juridique et technologique au service du numérique

Ces politiques, conçues et suivies par l'Etat, souvent au plus haut niveau (la Primature ou la Présidence), ont oeuvré plus particulièrement dans trois domaines : juridico-réglementaire, commercial-financier, et l'ingénierie informatique.

Le domaine juridique et réglementaire a fait l'objet de l'attention de l'Etat dans la mesure où il conditionne la pérennisation de la société de l'information, crée des lois et décrets adaptés à une situation inédite, organise la structuration administrative de la société de l'information, favorise l'innovation, et lutte contre la criminalité informatique.

Le domaine commercial a été confié le plus souvent, dans les années 1995-2000, à un opérateur unique, placé en situation de monopole. Cette situation résulte

d'une bataille d'influence entre de très grands groupes d'investisseurs étrangers, comme Vivendi, France Télécom ou Marconi, qui ont mis en place une filiale locale contrôlée à la fois par l'entreprise-mère principale actionnaire et par l'Etat, parfois majoritaire. Au Sénégal, Sonatel, filiale de France Télécom, a, comme Cabo Verde Telecom (filiale de Portugal Telecom) joué ce rôle d'opérateur « historique », avant d'être mis, surtout à partir des années 2005, en situation de concurrence. Le monopole historique avait ses avantages : l'Etat confiait à une unique entreprise une mission de service public. Il avait aussi beaucoup d'inconvénients : cherté des accès aux services et des abonnements, qui ont considérablement ralenti la généralisation des usages, notamment de l'internet.

L'ingénierie d'accompagnement est fréquemment sous la tutelle directe du gouvernement, et étroitement associée, comme maître d'ouvrage, à l'élaboration de plans stratégiques et à la réalisation des équipements, comme par exemple les réseaux interconnectant les bureaux d'une structure administrative, ou reliant plusieurs structures entre elles. Elle est également chargée de développer des logiciels adaptés aux besoins des administrations, ainsi que des sites internet, souvent « de service public ». Au Cap-Vert, c'est le NOSI (Nucleo Operacional para a Sociedade da Informaçao) qui joue ce rôle : agence de conception et d'exécution financièrement autonome rattachée au Ministère des Finances, NOSI est, de facto, une PME para-étatique d'une cinquantaine de cadres de haut niveau par qui passent la plupart des offres de marché public.

1.3 La remise en cause des monopoles « historiques » : le triomphe de l'oralité?

Ce n'est que très récemment, depuis 2005, que les monopoles des opérateurs historiques commerciaux ont été remis en cause, avec la politique d'ouverture et de concurrence. Celle-ci apparaît en effet comme la seule susceptible de résoudre le difficile problème de la cherté des services, génératrice de blocage aux équipements et à la généralisation des usages des outils. Au Cap-Vert, c'est l'Etat en personne qui a décidé de mettre fin, en 2009, au monopole de CVT, pourtant garanti (par lui) pour une période de 25 ans (1995-2020).

C'est aussi très récemment que l'équipement individuel ou des ménages en téléphone cellulaire a littéralement « explosé » en Afrique. Les raisons de cet engouement sont multiples : rareté du téléphone filaire; équipement en antennes-relais, qui ont progressivement couvert la plus grande partie des territoires nationaux; abaissement des coûts d'équipement et d'abonnement en raison de la mise en concurrence d'opérateurs rivaux (au Cap-Vert, T+ et CVT, au Sénégal : Sonatel, Sentel, . Mais surtout, le cellulaire est particulièrement bien adapté aux besoins d'une société où la communication orale demeure fondamentale : nul besoin d'être alphabétisé pour utiliser cet appareil! Toutefois, la fin des interdictions d'utiliser la téléphonie fixe par Internet Protocol (IP) a, d'un coup, permis la création, à Praia et Mindelo (Cap-Vert), de dizaines de télé(cyber)centres : il est moins coûteux de téléphoner en IP depuis un téléphone dans une boutique que d'utiliser sa ligne téléphonique domestique : revanche du fixe sur le cellulaire et du téléphone « public » sur le « privé »!

C'est donc dans un contexte souvent très dynamique, tant du côté de l'Etat que de celui des opérateurs, que les sociétés de l'information africaines se développent. Mais, pour autant, comment innovent-elles? Et l'Etat est-il le seul acteur de l'innovation?

2 Les nouvelles pratiques innovantes au service du document numérique

La dynamique numérique voulue par l'Etat concerne particulièrement l'administration publique et dans une certaine mesure celle des collectivités territoriales. Mais elle établit aussi un lien fort avec les structures privées (les banques par exemple), ayant une mission de service public ou d'intérêt public.

2.1 Un e-gouvernement pour une bonne gouvernance politique

Les politiques des Etats ont presque toutes donné la priorité à l'informatisation des structures de gouvernement et à l'information-communication de l'organisation et des activités gouvernementales. Cette philosophie de la transparence renvoie à la résolution prise par de nombreux pays d'Afrique de mettre en pratique le slogan politique de la « bonne gouvernance ». Ainsi, privilégiant l'image et le texte écrit plutôt que d'autres médias, les gouvernements ont-ils crée des sites internet « du gouvernement » et des principaux organes de la démocratie (assemblée nationale, sénat). Plus rarement, certains gouvernements, comme le Cap-Vert (www.governo.cv) ont pris la décision de mettre en ligne les débats ou les actes des décisions prises à l'occasion des séances des assemblées. Jusqu'alors limitée à des retransmissions télévisuelles et radiophoniques combinant oral et image, le débat politique est donc désormais en image et écrit. Par rapport à la presse écrite, cette innovation numérique présente l'avantage d'être instantanément accessible dans les villes et villages les plus éloignés du pays : la démocratie ne peut qu'en sortir gagnante, pour autant que les maîtres de ce nouveau e-pouvoir fassent preuve de la plus grande honnêteté politique.

Dans le champ très politique de promotion de la démocratie, l'Etat cap-verdien, comme d'autres pays d'Afrique (Sénégal, Afrique du Sud) a informatisé tout le processus électoral, de l'observatoire des élections à la proclamation des résultats, ainsi que leur archivage. Les commentaires politiques, officiels, des partis et des citoyens se sont développés sur Internet, tant dans les médias en ligne que dans des sites et sur des forums de discussion.

Dans un domaine moins politique mais davantage tourné vers l'amélioration de la qualité de la gestion des administrations centrales, priorité a été donnée à l'informatisation des services financiers et des ressources humaines, ainsi qu'à la maintenance. Toutefois, la réalisation de ces plans sectoriels de modernisation est très inégale : par exemple, le Cap-Vert et le Sénégal sont beaucoup plus avancés que leurs voisins Malien et Guinéen. Le Cap-Vert, avec le SIGOF (Sistema Integrado de Gestao Orçamental e Financeira) a significativement amélioré la capacité de gestion budgétaire et comptable des administrations.

2.2 Une e-administration au service de la démocratie, pour une e-gouvernance citoyenne

Toujours dans le cadre de la « bonne gouvernance », les politiques d'Etat privilégient l'équipement informatique pour la vie politique et la participation citoyenne. L'exemple du Cap-Vert est ici particulièrement éclairant. Ce pays a mis en place en novembre 2007 une « Casa do Cidadao » (Maison du Citoyen), dont la

fonction est de faciliter les relations entre l'Administration d'Etat et le citoyen. Réalisée par le NOSI, la Casa, accessible par le site Internet du gouvernement (www.governo.cv) propose plusieurs services : document unique automobile, des certifications d'état-civil, les paiements électroniques à l'Etat, et la prestation « Empresa no dia » (Une entreprise créée en un jour). De nouveaux services seront progressivement installées sur le site. Même si elle existe encore au guichet ou dans les bureaux de l'Administration, la relation (orale) directe citoyen-administration a désormais fait une place au « tout-écrit » numérique.

Une autre réalisation technique cap-verdienne témoignant du même état d'esprit est l'équipement des trois plus grandes places publiques de la capitale, Praia, en WiFi. Sur l'une d'entre elles, la place de la Cruz do Papa (La Croix du Pape), les sorties familiales permettent aux enfants de jouer sur des portiques installés pendant que les parents surfent sur Internet et lisent leurs e-mails!

La politique des Etats a également favorisé l'informatisation des entreprises para-publiques ou privées ayant une obligation de service public : télécommunications, services bancaires, services de gestion des structures de transport etc... La gouvernance administrative est souvent conçue d'une manière large et l'Etat voyait un grand intérêt à soutenir et même inciter les entreprises à se moderniser. C'est ainsi que, dès la fin des années 90, le gouvernement du Cap-Vert a appuyé des programmes d'informatisation bancaire, qui ont abouti, dans un premier temps, à créer des prestations offrant au client monnaie électronique, distributeurs et accès aux comptes *on line,* avant d'aboutir à l'interconnexion permettant un usage multi-systèmes des cartes de crédit.

2.3 L'Administration n'est pas seule: TIC et innovation « par le bas » ou les usages populaires du document numérique

Les initiatives de la société civile pour s'approprier les TIC, au besoin en inventant de nouvelles utilisations, ne sont pas récentes : téléphone filaire, fax, mobile, ordinateur ont suscité depuis une dizaine d'années, et compte tenu des contraintes locales, des utilisations souvent bien peu développées dans les pays du Nord. On en examinera ici trois : le fax comme transport de fonds, l'ordinateur comme service communautaire villageois, le cellulaire comme base d'information pour des usages professionnels.

Le citoyen innove : depuis de nombreuses années, le fax est utilisé comme lettre de crédit et comme bon de commande pour transférer de l'argent entre émigrés et village d'origine. De nombreux exemples de ce que les services de transfert de fonds comme La Poste et Western Union considèrent comme un « détournement » d'usage ont été montrés par des études sur les régions d'émigration au Sénégal et dans le Nord-Ouest du Mali. Le système est simple : un chef de famille, au village, présente à un commerçant local une liste de produits qu'il souhaite acquérir. Ne disposant pas de la somme nécessaire pour les acheter, il se tourne vers son parent, émigré, par exemple en France. Le commerçant contacte ce parent, par Fax, lui demandant s'il consent à payer la commande proposée. Si ce dernier répond positivement, le commerçant livre la marchandise. L'émigré fait alors un virement bancaire au commerçant, qui a généralement un compte en France dans la même banque...Les frais prélevés par le commerçant pour l'opération sont réduits au seul coût d'envoi d'un Fax : le document écrit s'impose et vaut « lettre de change », même au coeur de populations très inégalement alphabétisées.

L'Internet villageois communautaire s'est peu à peu imposé comme une solution à l'impossibilité matérielle et financière de se connecter en milieu rural. La

distance à la ville, le coût des équipements, la rareté ou l'absence d'énergie électrique conditionne fortement l'usage des TIC dans les campagnes africaines. Les organisations non gouvernementales, sur projet, ont, localement, dans plusieurs pays d'Afrique de l'ouest, comme le très pauvre Burkina Faso, résolu le problème en installant d'abord une source d'énergie électrique, solaire mais le plus souvent un générateur diesel, et en proposant une connection web avec tout le matériel nécessaire. L'écran de l'ordinateur permet collectivement aux villageois de découvrir les richesses des sites Internet, lesquels sont abondamment commentés : la différentes facettes de la modernité, les autres pays, le monde sont accessibles en image, et en texte, lu par celles et ceux qui sont alphabétisés. Au sein de l'oralité, le document numérique occupe toute sa place.

Plus proche des pratiques professionnelles en usage dans les pays du Nord, l'écran du téléphone cellulaire est désormais utilisé comme source d'information et comme appui aux activités économiques. Au Sénégal, l'expérience de Manobi est éclairante. Manobi, petite PME installée à Montpellier (France), a crée en 2003 avec le groupe Sonatel, Manobi-Sénégal, sa filiale sénégalaise. Manobi-Sénégal est un opérateur de services à valeur ajoutée sur téléphonie, assurant la fourniture de services métiers sur GSM. La société exploite les possibilités offertes par les services SMS et surtout WAP (Wireless Application Protocol), qui permet d'adapter les formats d'Internet aux contraintes des téléphones cellulaires. Elle transmet donc des informations techniques, météorologiques, de positionnement géographique et de cours de produits halieutiques et maraîchers sur les marchés urbains sous forme de courts messages et par l'accès mobile à Internet. L'utilité de ces services est remarquable : sécurité des pêcheurs en mer, amélioration de la lisibilité de la commercialisation pour les producteurs, en sus de l'utilisation banale de téléphonie. Dans cet exemple, l'oralité est encore une fois associée au document numérique, mais l'utilité de ce dernier se révèle bien supérieure à l'utilisation classique du cellulaire. Les opérateurs économiques se sont désormais complètement approprié cette innovation.

Au total, les innovations, « réinterprêtées », qu'elles viennent de l'Etat ou de la société civile, contribuent à réduire la fracture avec les pays du Nord pour toute une série de services et de comportements « modernes » d'utilisation des services relationnels, ainsi que des moyens d'information et de communication.

3 Les nouvelles fractures : le document numérique en partage ?

Loin d'être uniformément développées dans tous les pays, ces processus et actions innovantes suivent des dynamiques très diverses. Les raisons en sont variées, mais relèvent globalement d'un complexe de facteurs « institutionnels » qui en rendent (ou non) l'extension malaisée.

3.1 Les systèmes institutionnels incomplets ralentissent l'essor du numérique

La dynamique de l'innovation numérique dépend d'un ensemble de facteurs interdépendants et interagissants. Ce système peut être caractérisé de la manière suivante :

- un e-gouvernement qui définit une politique généralement orientée vers la promotion d'une e-administration et éventuellement de services divers,

- des collectivités territoriales inégalement volontaristes en matière de politique de TIC,
- un ou plusieurs opérateurs économiques contribuant à la promotion d'outils numériques,
- un ou plusieurs opérateurs responsables des coûts d'accès publics et privés aux produits et services numériques,
- des territoires variés, régionaux et locaux, inégalement desservis par des infrastructures d'accès,
- la disponibilité en ressources énergétiques, et plus particulièrement en électricité, avec une desserte généralement socio-spatialement inéquitable.
- le degré d'implication des populations dans la mondialisation économique, et par la mobilité migratoire.

Le fonctionnement optimal de ce système conditionne l'essor et la réussite de la société de l'information dans un pays. Or, dans la plupart des pays d'Afrique, un ou plusieurs éléments du système sont déficients (l'énergie électrique notamment) : tout en réduisant -un peu- la fracture entre Nors et Suds, la révolution numérique en Afrique génère ses propres fractures.

3.2 Des inégalités socio-spatiales plus que jamais à l'ordre du jour

La dynamique de la fracture numérique se déploie à plusieurs échelles et elle est de nature variée. Même si les pays d'Afrique sont désormais presque tous desservis par des infrastructures d'accès (câbles sous-marins et terrestres internationaux en fibre optique, couverture multi-satellitaire), les fractures entre pays sont profondes, chacun avançant, du point de vue de l'utilisation et la rentabilisation de ces infrastructures, à un rythme très inégal. De multiples indicateurs d'usage confirment ces écarts : parmi les plus significatifs, les pays d'Afrique qui ne disposaient pas encore du haut débit en 2007 sont le Congo, la Rép. Démocratique du Congo, la Guinée, le Soudan, le Tchad, le Burundi. Au contraire, on comptait parmi les bénéficiaires de la plus lage bande passante (bandwidth) tous les pays du Maghreb, l'Egypte, l'Afrique du Sud, Maurice et les Seychelles, le Sénégal et la Gabon.

La question de la disponibilité en énergie électrique est décisive pour la promotion de l'usage du document numérique : Si l'Afrique du Nord a, selon le dernier rapport (2008) de l'Union Internationale des Télécommunications, le taux d'équipement le plus élevé (95% des ménages avaient accès, en 2005, à l'énergie électrique) de l'ensemble des pays en développement, ce taux n'est, pour l'Afrique subsaharienne, que de 28%, et pour l'Afrique subsaharienne rurale, seulement 8%. L'ensemble de l'Afrique affiche pour sa part un taux de 38% (pays du Nord : 100%).

Le document numérique est un enjeu géographique. Dans les régions, entre Etat et collectivités territoriales, les différences se creusent. Le SIM (Sistema de Informaçao Municipal) crée en 2003 par l'Etat du Cap-Vert n'est pas encore opérationnel dans tous les Municipes (régions) du pays. Praia est informatisé depuis 2003, alors que les municipalités de Boa Vista, Brava, Sao Nicolau espèrent encore, en 2008, dans un plan qui tarde à être mis en chantier.

L'inégalité des équipements entre espace rural et la ville s'accroît. Même si les interactions rural-urbain se développent, les villes sont généralement mieux équipées en infrastructures d'accès que les campagnes : ces dernières demeurent davantage à l'écart du document numérique, de l'écrit, de l'image. Il en est de même entre la métropole et le reste du pays. Les logiques d'aménagement du territoire

privilégient certains centres sélectionnés, comme au Maroc, qui promeut depuis 2005 un petit nombre de ses capitales comme e-centre administratif, universitaire et/ou d'affaires comme CasaNearShore à Casablanca. En Afrique, les plans volontaristes d'aménagement d'accès et de services TIC fondés sur le principe d'équité socio-spatiale ne sont pas des plus répandus : pour un Maroc qui, depuis 2005, multiplie les EPN -espaces publics numériques- fixes (4 kiosques publics par ville, dans les gares, gares routières, ports, centre ville) et mobiles (7 EPN ruraux par région), combien de pays dénués de vision et de politique territoriale?

D'anciennes fractures sociales sont renforcées par la capacité financière d'accès au numérique : le pouvoir de celles et ceux qui contrôlent ou maîtrisent l'information et la communication écrite et documentaire est singulièrement renforcée par les TIC. Cette fracture est territoriale, mais aussi sociale : entre riches et pauvres, les politiques d'accès n'ont que très inégalement pris en compte les grandes différences de pouvoir d'achat : les pauvres n'ont toujours pas, ou peu, accès au document numérique. On n'insistera pas, dans cette courte présentation, sur des approches identitaires et culturelles du rapport aux TIC et leurs usages. Mais on retrouve, semble-t-il, des clivages variés, qui questionnent : au Sénégal, l'appropriation du document numérique par les différentes confréries musulmanes est très inégale. Alors que l'oralité demeure privilégiée dans les *daaras* (écoles coraniques) come mode d'acquisition des connaissances fondamentales de l'islam, certains cyber-marabouts figurent aujourd'hui parmi les plus grands utilisateurs de l'image numérique, tandis que les débats entre *taalibé* sur des forums modérés font plutôt un large usage de l'écrit!

4 Conclusion

Avec retard, les Suds, et notamment l'Afrique, ont reçu un transfert technologique du Nord. Mais en s'appropriant ces outils importés, les sociétés inventent de nouvelles normes d'utilisation, que ce soit avec le fax (banque), le téléphone (télécentre et cybercentres communautaires), ou le téléphone cellulaire et l'ordinateur (communautaire). Partout, le document numérique, triomphe de l'écrit et de l'image, progresse, sans pourtant, semble-t-il, remettre en cause l'oralité comme mode relationnel. Cependant, la cyberutilisation à marche forcée laisse de côté beaucoup de monde. De nouvelles fractures se dessinent, d'autres sont renforcées, d'autres s'atténuent : le développement des cybersociétés est à deux ou plusieurs vitesses. Actuellement, le *« modèle fracture**s** numérique**s** »* se caractérise plutôt par le renforcement du pouvoir des e-élites, informées et communicantes, avec des e-administrations contribuant au renforcement de l'Etat, la montée de classes moyennes pour qui le téléphone cellulaire est avant tout un moyen commode de communication dans une oralité renforcée mais où l'écrit numérique est secondaire, et un nombre considérable de pauvres et *« d'oubliés du numérique »,* laissés pour compte de la société de l'information, qui n'ignorent pas ces nouveaux outils, mais ne les utilisent que très faiblement.

Références

Bad-Ocde, (2007) *Afrique. Vue d'ensemble.* Coll. Perspectives économiques, Paris, OCDE, 93 p.

Chéneau-Loquay, A. (ss. la dir. de), 2004 *Mondialisation et technologies de la communication en Afrique.* Paris, Karthala-MSHA, 322 p.

Elie, M., (2001) Le fossé numérique. L'Internet, facteur de nouvelles inégalités? *Problèmes politiques et sociaux,* n° 861, 3-82.

ITU, 2008 *African Telecom / ICT. Indicators 2008 : At a crossroads.* Genève, 2008, 140 p.

Lesourd, M. (2003) Nouvelles technologies, nouvelles inégalités ? Les NTIC et les fractures socio-spatiales. Exemples au Sénégal et aux îles du Cap-Vert, in *L'Afrique. Vulnérabilité et défis* (M. Lesourd, Coord.), Nantes, Ed. du Temps, 421-447.

Lesourd, M. (2004) Les NTIC au Cap-Vert. Des médias à l'avènement d'une société de l'information ? in *Lusotopie « Médias, pouvoir et identités »,* 337-361.

Royaume du Maroc, (2007) *Stratégie e-Maroc 2010. Réalisations, Orientations et Plans d'action* Premier Ministre, Ministère des Affaires Economiques et Générales, Rabat, 119 p.

Pnud, (2004) *Rapport National sur le développement humain : Nouvelles technologies de l'Information et de la Communication et transformation du Cap Vert.* PNUD, Praia, 120 p.