ISBN 2-909285-31-6

15, avenue de Ségur,
75007 Paris, France.
Téléphone (Fr) 01 45 51 26 07 - (Int.) +31 1 45 51 26 07
Télecopie (Fr) 01 45 51 26 32- (Int.) +31 1 45 51 26 32
e-mail: production@europia.org
http://europia.org

Le Document Electronique

Actes du neuvième colloque international
sur le document électronique

Fribourg Suisse 18 - 20 septembre 2006

TABLE DES MATIÈRES

Session 03
Production collaborative de documents et partage de connaissances

Session 04
Gestion et accès à des collections de documents

PRÉFACE

Initié en 1998, à Rabat, au Maroc, le cycle de colloques CIDE (Colloque International sur le Document Electronique) vise à rassembler selon une base annuelle des communautés de chercheurs pour lesquelles le document électronique constitue un objet fondamental d'étude. Les propriétés intrinsèques du document électronique, que ne possède pas son homologue papier - telles la dimension multimédia, la navigation hypertextuelle et l'usage de ses diverses modalités - lui confèrent un rôle de plus en plus important dans la construction et la capitalisation de savoirs dans un contexte de globalisation de l'information. Ne se limitant volontairement pas aux aspects purement liés aux technologies, CIDE intègre dans son champ d'investigation scientifique des contributions qui mettent en exergue la dimension sociale liée à l'évolution des pratiques et usages du document.

S'inscrivant dans une mouvance qui dénote un intérêt croissant pour le document électronique et draine des communautés de chercheurs adressant différents aspects tels la production, la reconnaissance, l'interprétation, l'extraction d'information, la diffusion ou l'usage collaboratif, CIDE s'est associé en 2004 à la première Semaine du Document Numérique qui s'est tenue à La Rochelle, en France, et participe à la seconde édition en 2006, à Fribourg, en Suisse.

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

Le rôle communicationnel du document est en pleine mutation ; il n'est clairement plus limité à la simple diffusion d'un contenu mais devient un objet de support à une communication plus complexe entre les différents acteurs des systèmes d'informations distribués. Reliés aux pratiques où ils sont produits et interprétés, les documents numériques sont à la source de nouveaux scénarios d'utilisation mais également confrontés au problème de leur interopérabilité suite à la prolifération d'une multitude de normes adressant divers aspects, depuis le plus bas niveau (tel le codage des informations) jusqu'à l'expression de leur sémantique (tels la définition de méta données et l'usage d'ontologies).

Les communications retenues pour la 9ème édition du colloque CIDE, s'articulent autour de cette problématique et sont réparties en quatre sessions. La première session intitulée " Modèles de documents, transformations et

modes d'accès " adresse des aspects fondamentaux inhérents au document électronique dans son rôle de vecteur de communication : la modélisation (en particulier, la difficulté de tenir compte de la pluralité de modèles s'appliquant à un même document), les mécanismes de transformations nécessaires pour exploiter le caractère numérique du document en vue d'en assurer la diffusion sur une variété non limitée de supports et enfin, l'exploitation des diverses modalités du document.
La seconde session intitulée "Analyse et interprétation des documents" s'intéresse à la découverte de structures logiques et sémantiques de documents en vue d'en faciliter l'exploitation - le partage, la réutilisation, l'accès - dans divers contextes d'utilisation.

La troisième session intitulée " Production collaborative de documents et partage de connaissances " rassemble des contributions qui mettent en évidence le rôle joué par le document en tant que support de communication et de partage de connaissances dans le cadre de collectivités.

Enfin, la quatrième session intitulée "Gestion et accès à des collections de documents " se focalise sur des travaux qui s'intéressent non pas" au document " en tant que tel mais à des ensembles de documents ; visant ainsi à exploiter des interdépendances de nature sémantique et ancrer des processus de communication ; notamment par le biais de mécanismes d'annotations.

Khaldoun ZREIK (GREYC, Université de Caen), *Co-Président*
Christine VANOIRBEEK (EPFL, Suisse), *Co-Présidente*

REMERCIEMENTS

Nous tenons à exprimer nos remerciements aux auteurs des contributions et surtout aux membres du comité du programme CiDE.9

Patrick Andries *(Cooptel, Canada)*
Abdel Belaïd *(Université Nancy 2 - LORIA UMR 7503, France)*
Jean Caelen *(CLIPS-CNRS, France)*
Jacques Ducloy *(INISIT, France)*
Dominique Dutoit *(Memodata, France)*
Claudie Faure *(ENST-Paris, France)*
Jean-Daniel Fekete (*INRIA, France)*
Christian Fluhr *(CEA, France)*
Patrick Gallinari *(LIP6 - Université Pierre et Marie Curie, France)*
Jean-Gabriel Ganascia *(LIP6 - Université Pierre et Marie Curie, France)*
Joël Gardes *(Francetelecom, France)*
Franck Ghitalla *(UTC, France)*
Ioannis Kanellos *(ENST-bretagne, Brest, France)*
Peter King *(Université de Manitoba, Canada)*
Jacques Labiche *(PSI, Université de Rouen, France)*
Omar Larouk *(ENSSIB-Lyon, France)*
Jacques Madelaine *(GREYC, Université de Caen, France)*
Mustapha Mojahid *(IRIT, Toulouse, France)*
Anne Nicolle *(GREYC, Université de Caen, France)*
Philippe Palanque *(IRIT, Toulouse, France)*
Yannick Prié *(LIRIS - CNRS, France)*
Imad Saleh *(Université de Paris 8, France)*
Chantal Soulé-Dupuy *(Université de Toulouse 1, France)*
Saîd Tazi *(LAAS, France)*
Eric Trupin *(PSI, Université de Rouen, France)*
Jean Vanderdonckt *(Université catholique de Louvain, Belgique)*
Jean Vivier *(Modescos, Université de Caen)*
Manuel Zacklad *(Université de Technologie de Troyes, France)*

HASLERSTIFTUNG

Session 01

Modèles de documents, transformations et modes d'accès

Un modèle et un schéma pour représenter des documents textuels multistructurés

Emmanuel BRUNO
Elisabeth MURISASCO

LSIS - UMR CNRS 6168
Université du Sud Toulon-Var
Avenue de l'Université, BP 20132
F-83957 La Garde cedex, France
{bruno,murisasco}@univ-tln.fr

RÉSUMÉ

Un document XML est organisé essentiellement par une structure hiérarchique, modélisée par une représentation arborescente. Cette structuration correspond à un niveau d'analyse (par exemple logique) des données contenues dans le document. Mais, certaines applications, en particulier dans un contexte documentaire, nécessitent la description d'un texte selon différents niveaux d'analyse qui correspondent à des usages différents de ce texte. Cet article présente dans une première partie la problématique de la représentation de ce type de documents sur lesquels plusieurs structures peuvent être superposées ainsi qu'un court état de l'art. Dans une seconde partie, nous proposons un modèle et un langage de schéma pour documents textuels multi structurés.

MOTS-CLES :
Documents textuels, structures multiples, XML, modèles et langages.

1. INTRODUCTION

L'usage du langage XML [BRA98] et de l'ensemble des modèles, outils et langages qui lui sont associés pour modéliser [FER03] et manipuler ces documents numériques (édition, interrogation XPath [CLA99b], XQUERY [BOA03], transformation XSLT [CLA99a]) sont devenus maintenant incontournables. Dans un document XML, des marques, appelées balises, sont ajoutées au sein même du texte: la structure du document est donc mêlée à son contenu. La structuration induite par les balises correspond à un niveau d'analyse (par exemple logique) des données contenues dans le document. Un document ainsi structuré peut alors être vu comme un arbre dont les nœuds non terminaux sont les balises et les feuilles les fragments de texte [GON87].

Dans le domaine documentaire, les chercheurs ont éprouvé le besoin d'associer **simultanément** au même texte plusieurs structures. Ce besoin a été identifié dès SGML [GOL90] avec l'option CONCUR. Lors de l'émergence de XML, CONCUR n'a pas été adapté mais ces dernières années, de nombreux travaux à propos du marquage concurrent de données ont été publiés [SPE01, TEN02, WIT02, ABA03, JAG04, IAC04, JAG04, DEK05]. Il s'agit de pouvoir représenter ces différentes hiérarchies qui se superposent sur le même document en impliquant en particulier des recouvrements/ chevauchements [DUR01]. C'est le contexte dans lequel ce travail se situe. Nous étudions comment représenter simultanément plusieurs structures hiérarchiques associées à un document textuel, chaque structure concernant une segmentation a priori différente du texte. De plus, nous nous plaçons dans un cadre où les structures sont le plus souvent autonomes car développées indépendamment les unes des autres. Enfin, des chevauchements étant possibles, elles sont sans relations hiérarchiques ; elles ne peuvent donc pas être fusionnées toutes en une unique hiérarchie grâce par exemple aux espaces de noms [BRA04].

Notre objectif dans cet article est double. Dans une première partie, nous présentons la problématique de la représentation de ce type de documents (que nous appelons documents textuels multistructurés) ainsi qu'un court état de l'art (pour un état de l'art en version étendue, voir [BRU06] ou encore [DER04, WIT04]). Pour cela, nous avons sélectionné une application en philologie (science dédiée à l'étude de l'histoire des textes) en collaboration avec Sofia Bozzi-Corradini [COR90], Philologue au Dipartimento di lingue e letterature romanze de l'Université de Pise en Italie. Il s'agit d'un fragment de manuscrit ancien écrit en occitan. Dans une seconde partie, nous présentons un modèle et un schéma adaptés à la représentation des documents textuels multistructurés. Notre modèle s'appuie sur la notion de hedge [MUR00] (fondation de RelaxNG [VAN04]). Il est associé à une algèbre définie sur les structures d'un document pour spécifier des contraintes entre elles au moyen des relations de Allen [ALL91]. Notre schéma permet de positionner ainsi certains éléments de structures par rapport à d'autres appartenant à d'autres structures. L'aspect interrogation n'est pas abordé dans cet article.

L'article est organisé de la façon suivante, la section 2 décrit les difficultés de la multistructuration sur notre exemple du manuscrit ancien, la section 3 présente l'état de l'art, la section 4 définit respectivement notre modèle et un schéma associé pour les documents multistructurés. La section 5 discute de l'implantation de notre modèle, appelé MSXD. La section 6 établit un bilan et ouvre des perspectives.

[1] Ce travail a été financé par le projet ACI SemWeb (Interrogation du Web sémantique avec XQuery). Notre partenaire sur ce sujet est Sylvie Calabretto du LIRIS Lyon

```
<?xml version="1.0" encoding="utf-8"?>
<manuscrit titre="Princeton, Garrett 80">
 <page>
  <colonne>
    <ligne>Per recobrar maniar</ligne>
    <ligne>Ad home cant a perdut</ligne>
    <ligne>lo maniar prin de l erba</ligne>
    <ligne>blanca so es l'eixens et un</ligne>
    ...
  </colonne>
  <colonne>...</colonne>
 </page>
</manuscrit>

<?xml version="1.0" encoding="utf-8"?>
<manuscrit titre="Princeton,Garrett 80">
 <syntaxe>
  <phrase><m>Per</m> <m>recobrar</m> <m>maniar</m></phrase>
  <phrase><m>Ad</m> <m>home</m> <m>cant</m> <m>a</m>
   <m>perdut</m>
   <m>lo</m> <m>maniar</m> <m>prin</m> <m>de</m> <m>l</m>
   <m>erba</m>
   <m>blanca</m> <m>so</m> <m>es</m> <m>l</m>'<m>eixens</m>
   <m>et</m><m>un</m> ...
  </phrase>...
 </syntaxe>
</manuscrit>

<?xml version="1.0" encoding="utf-8"?>
<manuscrit titre="Princeton, Garrett 80">
 <prescriptions>
  <prescription>
    Per recobrar maniar <indication>Ad home cant a perdut
    lo maniar</indication> prin de l
    <ingredient><plante>erba blanca</plante>
    so es l'<plante>eixens</plante></ingredient> et un...
  </prescription>
  <prescription>...</prescription>
 </prescriptions>
</manuscrit>
```

Figure 1: Structure physique (S1), syntaxique (S2) et sémantique (S3)

2. PROBLÉMATIQUE ET PRÉSENTATION DE L'EXEMPLE

La multistructuralité des textes littéraires ou traitant d'histoire a été mise en évidence très tôt en particulier dans le cadre des travaux de la TEI . C'est pourquoi nous avons choisi d'illustrer notre problématique sur un extrait d'un manuscrit médiéval traitant de recettes médico-pharmaceutiques et écrit en occitan.

Ce type de texte est étudié par les philologues de l'Université de Pise au Dipartimento di lingue e letterature romanze. A partir d'un tel document au format image, nous avons extrait le contenu textuel suivant : " ... Per recobrar maniar Ad home cant a perdut lo maniar prin de l erba blanca so es l'eixens et un... " (voir [BRU06]).

Trois structures correspondant à trois usages/points de vue différents ont été dégagées sur ce même contenu textuel (voir figure 1) : une structure physique S1 (le manuscrit est organisé en pages, sur plusieurs colonnes composées de lignes), une structure syntaxique S2 simplifiée (le manuscrit est structuré en phrases et en mots), une structure sémantique S3 (le manuscrit décrit des prescriptions médicales qui ont une indication, des ingrédients, un mode d'administration et des effets). Chaque structure segmente le texte différemment d'une autre. Par exemple le segment de texte " Ad home cant a perdut lo maniar " est marqué par indication dans la structure sémantique, alors que dans la structure physique ce même texte est segmenté autrement : " Ad home cant a perdut " est marqué par ligne tandis que " lo maniar " débute un autre segment de texte marqué également par ligne. On notera d'ailleurs qu'une indication peut donc chevaucher deux lignes. Chaque structure existe indépendamment d'une autre, elle peut être décrite par une DTD XML, un schéma XML [BIR01] ou un schéma relaxNG [CLA01] comme dans la figure 2.

Notre problématique s'intéresse à la prise en compte simultanée de ces trois structures afin de pouvoir les modéliser et les manipuler conjointement pour en étudier les corrélations. En particulier, nous aimerions qu'un utilisateur puisse à terme exprimer des requêtes qui combinent plusieurs de ces structures simultanément, comme par exemple : Quelle est la prescription qui comporte le plus de lignes ? (cette requête fait référence simultanément aux structures physique et sémantique); Quels sont les mots qui sont coupés en fin de ligne ? (ici, ce sont les structures syntaxique et physique qui sont concernées); Combien y-a-t-il en moyenne de prescriptions par colonne ? (les structures physique et sémantique sont concernées).

2 Text Encoding Initiative. http://www.tei-c.org/

3. ETAT DE L'ART

Pour répondre à cette problématique, on peut distinguer deux types d'approches. La première génère un unique document multistructuré (DMS) à partir des différentes structures. Cependant, pour pouvoir l'exploiter au mieux avec les outils XML classiques (interrogation, transformation), le modèle du DMS doit alors être hiérarchique. Toute la difficulté de cette approche réside dans la conception d'un tel

modèle. En effet, le problème du chevauchement possible de deux structures se pose : dans notre exemple, les éléments ligne et prescription ne sont pas hiérarchiquement imbriquables.

Pour cette approche, on peut citer deux propositions apportant des solutions d'ordre syntaxique et un langage alternatif à XML. CONCUR [GOL90], travail précurseur, permet d'intégrer au sein du même document des balises extraites de DTDs différentes pour représenter dans un texte plusieurs structures hiérarchiques. Les travaux de la TEI [SPE01] proposent un ensemble de guides de bonnes pratiques pour prendre en compte la multistructuralité dans un document XML. L'idée est d'utiliser une partie de la structure du document pour positionner les différentes autres structures. Ce qui nécessite soit (1) de choisir une structure hiérarchique principale et d'utiliser les références (ID/IDREF) pour décrire les autres structures soit (2) de choisir une représentation plate du DMS; ainsi l'avantage de la structure hiérarchique est perdu, en particulier pour l'aspect interrogation. De plus l'ajout d'une nouvelle structure peut impliquer une mise à jour importante du DMS liée à une nouvelle segmentation du texte. Une autre proposition, LNML (Layered Markup Annotation Language) [TEN02] a opté pour un nouveau langage de balisage - non XML (même si l'import export existe) - pour exprimer des chevauchements lors de la structuration d'un document textuel et pour pouvoir associer à des fragments de texte des annotations structurées. Comme précédemment, un seul DMS est produit, cependant les chevauchements doivent être traduits, ainsi les mêmes difficultés que celles posées par les codages proposés par la TEI sont à prévoir.

La deuxième approche s'appuie sur des documents différents, un par structure. La difficulté se situe alors dans l'interrogation de ces structures concurrentes. Comment interroger simultanément plusieurs hiérarchies ? Comment naviguer d'une hiérarchie à l'autre ? Comment les positionner l'une par rapport à l'autre ?

Pour cette approche, les propositions s'appuient en les étendant sur les modèles de données associés à XML [FER03, HOR04]. On peut citer les arbres colorés [JAG04] dont le titre " One hierarchy is not enough " établit

```
(S1)
start =
          element manuscrit {
    attribute titre { text } ,
          element page {
     element colonne {
      element ligne { text }+
     }+
    }+
   }
(S2)
start =
          element manuscrit {
           attribute titre { text } ,
           element syntaxe {
            element phrase {
      element m { text }+
     }+
    }
   }
(S3)
start =
          element manuscrit {
           attribute titre { text } ,
           element prescriptions {
            element prescription {
      ( text | dosage | effet | ingredient |
        modeAdministration
       | element indication { text }
       | element preparation { ( text
       | effet
       | modeAdministration
       | element action { ( text | ingredient)+ })+ }
       )+
    }+
   }
  }

effet = element effet { text }
ingredient = element ingredient { ( text | dosage
          | element mineral { text } | element plante { text })+}
modeAdministration = element modeAdministration { text }
dosage = element dosage { text }
```

Figure 2 : Grammaires des trois structures considérées

que l'unique hiérarchie avec laquelle le monde XML a l'habitude de travailler ne suffit pas. En réalité, cette proposition n'est pas adaptée à notre problématique ; elle est " orientée donnée " et partage d'information déjà segmentée. Il s'agit de construire plusieurs structures arborescentes au dessus du même ensemble de valeurs (des nœuds textes). Cette contrainte ne convient pas à notre contexte de travail. Cependant, nous citons cette proposition pour deux raisons : son titre révélateur et ses contributions compatibles XML en particulier en ce qui concerne l'interrogation. Le modèle étend celui de XML, il est composé d'un ensemble fini de couleurs. Chaque nœud peut avoir plusieurs couleurs, l'ensemble des nœuds de même couleur forment une hiérarchie. Pour naviguer dans ces hiérarchies de couleurs, les auteurs ont choisi XPath [CLA99b] en étendant la notion de " pas ". Un pas devient le choix d'une couleur (une structure) puis classiquement d'un axe et d'un test de nœud. Une seconde proposition définie dans [DEK05] est, selon notre point de vue, la plus complète car le système proposé est proche d'un système de gestion de documents. Elle s'appuie sur le modèle logique des Goddag (General Ordered-descendant directed acyclic graph) [SPE00]. Plusieurs arbres sont définis sur le même texte en partageant leurs feuilles c'est-à-dire des fragments du texte. Ainsi, une partie importante des propriétés des arbres sont conservées mais les nœuds peuvent également avoir plusieurs parents. Une instance de DMS est générée sous la forme d'une extension du modèle DOM pour la représentation de documents XML multihiérarchiques. La proposition décrit aussi une extension du langage XPath par de nouveaux axes qui permettent d'atteindre des nœuds en fonction de leurs relations indépendamment des structures auxquelles ils appartiennent : il est par exemple possible d'atteindre tous les ancêtres (axe xancestor d'un nœud dans toutes les hiérarchies dans lesquelles il se trouve (obtenus éventuellement à partir de leurs multiples parents).

Finalement, quelle que soit l'approche nous notons qu'aucune de ces propositions n'offre une solution au problème de la définition d'une classe de documents multistructurés. Or définir les contraintes (éventuellement faibles) existant entre les éléments des structures nous semble très utile, en particulier pour la vérification de la cohérence entre structures et leur interrogation conjointe. Dans notre exemple, une page (structure S1) débute toujours par une prescription (structure S3), les prescriptions sont toujours décrites par (contiennent toujours) des phrases (structure S2). Dans la suite de cet article nous proposons une solution à ce manque en définissant un modèle et un schéma pour DMS textuels. Ce modèle est appelé MSXD.

4. MSXD : UN MODÈLE POUR DOCUMENTS TEXTUELS MULTISTRUCTURÉS

Notre objectif est de définir un modèle formel qui (1) propose plusieurs segmentations du même texte (2) permet de définir une structure hiérarchique pour chacune de ces segmentations (produisant ainsi autant

de documents XML) (3) permet la définition d'un schéma de documents textuels multistructurés afin de décrire des contraintes (éventuellement faibles) sur des structures concurrentes (au moyen des relations de Allen). Le document multistructuré n'est jamais instancié : nous souhaitons conserver intacte chaque structure hiérarchique pour faciliter sa construction et son interrogation avec les langages et outils disponibles en XML.

4.1. Le modèle

DÉFINITION 1 : Un document multistructuré M est un triplet (T,V,S) où T est une valeur textuelle, V un ensemble de segmentations de T et S un ensemble de structures associées aux segmentations de V .

Pour définir la notion de structure, on utilise le concept de fragment. Les fragments permettent la structuration d'une segmentation. Leurs positions au sein de la valeur textuelle sont utiles pour calculer leur position relative.

DÉFINITION 2 : Un fragment f est défini sur une segmentation XV d'une valeur textuelle V et un alphabet SV :

1. f = e (le fragment vide), début(f) = fin(f) = 0
2. f = vi avec vi ? XV , début(f) = début(XV [i]), fin(f) = fin(XV [i])
3. f = v<x> avec v ? SV et x un fragment, début(f) = début(x), fin(f) = fin(x) (f est un arbre)
4. f = xy avec x et y deux fragments et début(y) = fin(x) + 1, début(f) = début(x), fin(f) = fin(y)

DÉFINITION 3 : Une structure est un arbre f (un fragment étiqueté) défini sur une segmentation XV de la valeur textuelle V, avec début(f)=0 et fin(f) = |XV |-1. Pour notre exemple, V est le texte. Nous avons défini trois segmentations , et sur V puis trois structures S1, S2 et S3 sont ensuite définies sur, et (et représentées dans des documents XML comme le montre la figure 1) :

Pour l'analyse physique, S = {manuscrit, page, colonne, ligne, titre},
= x1 ? ... ? x5 avec x1 = " Princeton, Garrett 80 " et x5 = " blanca so es l'eixens et un ".

Le fragment XML <ligne>blanca so es l'eixens et un</ligne> construit sur, est représenté dans notre modèle par ligne<"blanca so es l'eixens et un">.

Pour l'analyse syntaxique, S = { manuscrit, syntaxe, phrase, m, titre},
= x1 ? ... ? x21 avec x1 = " Princeton, Garrett 80 ", x2 = " Per ", x3 = " recobrar ", x4 = " maniar " et x21 = " un ".

Le fragment XML <phrase><m>Per</m><m>recobrar</m><m>maniar</m></phrase> construit sur , est représenté dans notre modèle par phrase<m<"Per">m<"recobrar">m<"maniar">>.

Pour l'analyse sémantique, S = {manuscrit, prescriptions, prescription, indication, ingrédient, plante, titre}, = x1 ? ... ? x8 avec x1 = " Princeton, Garrett 80 ", x2 = " Per recobrar maniar " et x8 = " et un... ".

Le fragment XML " Per recobrar maniar " construit sur , est représenté dans notre modèle par "Per recobrar maniar".

De plus, notre modèle est construit de façon à pouvoir utiliser les relations de Allen [ALL91] pour calculer la position relative des fragments dans une segmentation ou entre deux segmentations.

DÉFINITION 4 : Des prédicats sur deux fragments f1 et f2 sont définis sur une ou deux segmentations construites sur la même valeur textuelle de la façon suivante :

avant(f1, f2)	*= après(f2, f1)*	*= fin(f1) < début(f2)*
avant(f1, f2, n)	*= après(f2, f1, n)*	*= début(f2) - fin(f1) = n*
touche(f1, f2)	*= touché-par(f2, f1)*	*= fin(f1)=début(f2)*
pendant(f1, f2)	*= contient(f2, f1)*	
	= début(f1) > début(f2) et fin(f1) < fin(f2)	
chevauche(f1, f2)	*= est-chevauché(f2, f1)*	
	= début(f1) < début(f2) et fin(f1) > début(f2) et fin(f1)<fin(f2)	
commence(f1, f2)	*= commencé-par(f2, f1)*	
	= début(f1) = début(f2) et fin(f1) < fin(f2)	
finit(f1, f2)	*= fini-par(f2, f1)*	
	= fin(f1) = fin(f2) et début(f1) > début(f2)	
égal(f1, f2)	*= début(f1) = début(f2) et fin(f1) = fin(f2)*	

Si f1 et f2 sont définis sur la même segmentation, les prédicats touche et chevauche sont toujours faux. Enfin, pour capturer les relations parent/enfant entre fragments d'une structure, nous avons besoin de calculer le niveau d'un fragment dans une structure.

DÉFINITION 5 : Soit F(s) (s est une structure) un ensemble de fragments f tels que f = s ou ?x ? F(s), ?a ? SV |x = a < f >. La fonction niveau(s, f) retourne le niveau du fragment f dans la structure s, il est calculé avec l'algorithme suivant :
- niveau(s, s) = 0
- niveau(s, y) = niveau(s, x) + 1 avec x = a < y > (x et y ? F(s)).

Nous donnons des exemples de prédicats et de calculs de niveau pour l'extrait du manuscrit :
- si fi = page<...> (utilisé dans S1) et fj = prescription<...> (utilisé dans S3), commence(fi, fj) est vrai,
- si fj = manuscrit<...> (utilisé dans S1), fj = manuscrit<...> (utilisé dans S2) et fk = manuscrit<...> (utilisé dans S3), égal(fi, fj) et égal(fj, fk) sont vrais,
- niveau(S1, manuscrit<...>) = 0, niveau(S1, page<...>) = 1, niveau(S1, colonne<...>) = 2, ..., niveau (S1, ligne<...>) = 3.

- niveau(S3, manuscrit<...>) = 0, niveau(S3, prescriptions<...>) = 1, niveau(S3, prescription<...>) = 2, niveau(S3, indication<...>) = 3, niveau (S3, ingrédient<...>) = 3, ..., niveau(S3, plante<...>) = 4.

4.2. Définir un schéma de documents multistructurés

Nous proposons une syntaxe XML pour représenter un DMS (voir figure 3). Les segmentations sont implicites.

Nous définissons un schéma pour documents multistructurés comme un ensemble de règles (vs une définition de modèle de contenu) car nos structures sont faiblement couplées et que le document multistructuré n'est pas hiérarchique. Nous utilisons les relations de Allen (commence, chevauche, égal, ...) pour contraindre les positions relatives des fragments appartenant aux différentes structures. Une contrainte est une expression booléenne construite à partir de prédicats de Allen qui prennent en paramètres deux ensembles de fragments. Chaque ensemble de fragments est décrit par un modèle de chemin XPath dans une structure. Les contraintes portent donc sur des fragments appartenant à deux structures.

DÉFINITION 6 : Un schéma de documents multistructurés est défini comme un couple (GS, C) où GS est un ensemble de grammaires définissant des structures valides et C = {ci|ci = c(p1 in s1, p2 in s2)} est un ensemble de contraintes où c est nom d'un prédicat de Allen et p1, p2 sont des expressions XPath appliquées aux structures s1 et s2. La contrainte est vraie si pour chaque fragment f1 de val(p1), il existe un fragment f2 dans val(p2) tel que c(f1, f2) est vrai. Un document est valide par rapport à un schéma si et seulement si toutes les contraintes de C sont vraies.

La figure 4 présente une syntaxe XML pour définir un schéma pour notre exemple (noter les commentaires dans la figure) et illustre la définition de quelques contraintes entre les fragments de nos trois structures. Chaque contrainte peut être lue de la façon suivante, par exemple :
- Règle 1 : Les fragments racine des structures physique et syntaxique sont égaux. Tout fragment vérifiant /manuscrit dans tout document valide par rapport à la structure (dont l'alias est) " mPhys " doit être égal à au moins un fragment qui vérifie /manuscrit dans tout document valide par rapport à la structure (dont l'alias est) " mSynt " ;
- Règle 5 : Une page commence par une prescription. Tout fragment vérifiant page dans tout document valide par rapport à la structure (dont l'alias est) " mPhys " commence au moins un fragment qui vérifie prescription dans tout document valide par rapport à la structure (dont l'alias est) " mSem " ;
- Règle 6 : Une prescription contient des phrases. Tout fragment vérifiant prescription dans tout document valide par rapport à la structure (dont l'alias est) "mSem" contient au moins un fragment qui vérifie phrase dans tout document valide par rapport à la structure (dont l'alias est) " mSynt ".

```
<MsXml>
<ValeurTextuelle
uri="http://sis.univ-tln.fr/msxd/valeur/manuscrit"/>
<Structure
type="http://sis.univ-tln.fr/msxd/structure/manuscrit/
physique"
        uri="http://sis.univ-tln.fr/msxd/instance/S1.xml">
<Structure
type="http://sis.univ-tln.fr/msxd/structure/manuscrit/ syntaxique"
        uri="http://sis.univ-tln.fr/msxd/instance/S2.xml">
<Structure
type="http://sis.univ-tln.fr/msxd/structure/manuscrit/
semantique"
        uri="http://sis.univ-tln.fr/msxd/instance/S3.xml">
</MsXml>
```

Figure 3 : Syntaxe XML pour le DMS exemple

5. PROTOTYPE

Pour valider pratiquement nos propositions nous développons un prototype dédié à la manipulation de documents multistructurés. Celui-ci est constitué de deux composants : (i) une implantation proprement dite du modèle présenté dans cet article et (ii) une implantation du langage XQuery étendu à l'interrogation de documents multistructurés. La spécification de cette extension est un travail en cours. Ce prototype est développé en Java et des informations sur son avancement sont disponibles à l'adresse :
http://sis.univ-tln.fr/msxd/, une version de test sera bientôt disponible.

Notre objectif en ce qui concerne la représentation et la manipulation des instances du modèle est de pouvoir les considérer comme des documents conformes au modèle DOM [HOR04] (plus précisément à une extension de celui-ci). La description XML d'un document multistructuré (cf. section 4.2) est analysée pour en construire une indexation (dynamiquement à la demande de l'interprète de requêtes). L'analyse de la première structure permet de déduire la valeur textuelle qui est représentée de façon centralisée dans un composant spécifique, ce composant a en charge l'indexation du texte et l'accès aux valeurs des fragments. Lors de l'analyse des structures suivantes, la cohérence du texte est vérifiée et un alignement sur les séparateurs (espaces, tabulations et fins de ligne) est réalisé automatiquement. L'analyse des différentes structures permet aussi de construire une représentation des structures et des fragments dans une base de données relationnelle. Cette représentation est indépendante des valeurs textuelles, elle utilise les positions de début et de fin et calcule le niveau des fragments dans chaque structure.

```
<MsXmlSchema >
 <!-- IDENTIFICATION DES STRUCTURES -->
  <Structures>
   <Structure type="http://sis.univ-tln.fr/msxd/structure/
    manuscrit/physique" alias="mPhys"
    grammar="mPhys.rnc"/>
   <Structure type="http://sis.univ-tln.fr/msxd/structure/
    manuscrit/syntaxique" alias="mSynt"
    grammar="mSynt.rnc"/>
   <Structure type="http://sis.univ-tln.fr/msxd/structure/
    manuscrit/semantique" alias="mSem"
    grammar="mSem.rnc"/>
  <Structures>
  <Contraintes >
  <!-- CONTRAINTES RELATIVES ENTRE LES STRUCTURES -->
  <!-- Règles 1 et 2 : Les fragments manuscrit  sont égaux
   dans toutes les structures -->
  <Egal>
   <Fragments nom="mPhys" select="/manuscrit"/>
   <Fragments nom="mSynt" select="/manuscrit"/>
  </Egal>
  <Egal>
   <Fragments nom="mPhys" select="/manuscrit"/>
   <Fragments nom=" mSem" select="/prescriptions"/>
  </Egal>
  <!-- Règles 3 et 4 : Les attributs titre des 3 structures
   sont égaux -->
  <Egal>
   <Fragments nom="mPhys" select="manuscrit/@title"/>
   <Fragments nom="mSynt" select="manuscrit/@title"/>
  </Egal>
  <Egal>
   <Fragments nom="mPhys" select="manuscrit/@title"/>
   <Fragments nom="mSem" select="manuscrit/@title"/>
  </Egal>
  <!-- Règle 5 : Une page commence par une prescription -->
  <Commence>
   <Fragments nom="mPhys" select="page"/>
   <Fragments nom="mSem" select="prescription"/>
  </Commence>
  <!-- Règle 6 : Une prescription contient des phrases -->
  <Contient>
   <Fragments nom="mSem" select="prescription"/>
   <Fragments nom="mSynt" select="phrase"/>
  </Contient>
 </Contraintes >
</MsXmlSchema >
```

Figure 4 : Une grammaire pour notre DMS

L'implantation de l'API DOM est réalisée par une traduction des appels de méthodes en un ensemble de requêtes SQL. En ce qui concerne les technologies choisies, nous avons intégré par défaut la base de données HSQL qui permet d'avoir au choix une base de données relationnelle uniquement en mémoire (utilisation dans une applet Java), ou sur disque (cas d'une application). L'utilisation d'une autre base de données relationnelle est possible. En ce qui concerne l'index ation textuelle (en particulier pour permettre l'implantation de l'extension fulltext de XQuery) nous testons l'intégration d'Apache Lucene .

Pour l'implantation de XQuery (réduit à XPath 2.0 dans la première version), nous avons choisi de procéder par étapes. Dans un premier temps la requête est traduite vers le langage XQuery Core et c'est cette requête qui est utilisée pour construire l'arbre de requête. L'optimisation est réalisée par transformation de cet arbre. L'arbre de requête est interprété, les collections de fragments sont transmises en pipe-line à travers les opérateurs de l'arbre.

Il est à noter que certains opérateurs accèdent à l'index du modèle décrit dans la section précédente, pour produire de nouveaux fragments.

6. CONCLUSION

Dans cet article, nous avons abordé le problème de la multistructuration d'un document textuel. Après avoir présenté les difficultés liées à la représentation de ce type de documents, nous avons présenté un modèle et un schéma pour documents multistructurés. Notre proposition conserve l'aspect hiérarchique de chaque structure pour utiliser les outils XML existants. Nous étudions actuellement un algorithme efficace de validation de documents multistructurés. Ce travail constitue un premier pas vers la validation de hiérarchies multiples.

Parmi nos travaux en cours, nous avons déjà réalisé une extension de ce modèle pour prendre en compte les annotations d'un utilisateur : il s'agit de considérer une information sous la forme de valeurs éventuellement structurées que l'utilisateur associe à des fragments d'une structure (celle qui correspond à son analyse), la valeur textuelle et les autres structures ne sont pas affectées.

Nous travaillons également sur l'intégration des relations de Allen au langage XPath pour pouvoir sélectionner des fragments en utilisant leurs positions relatives par rapport à des fragments d'autres structures (voire leur modèle de contenu). Nous donnons quelques exemples de requêtes :

3 http://www.hsqldb.org/ 4 http://lucene.apache.org

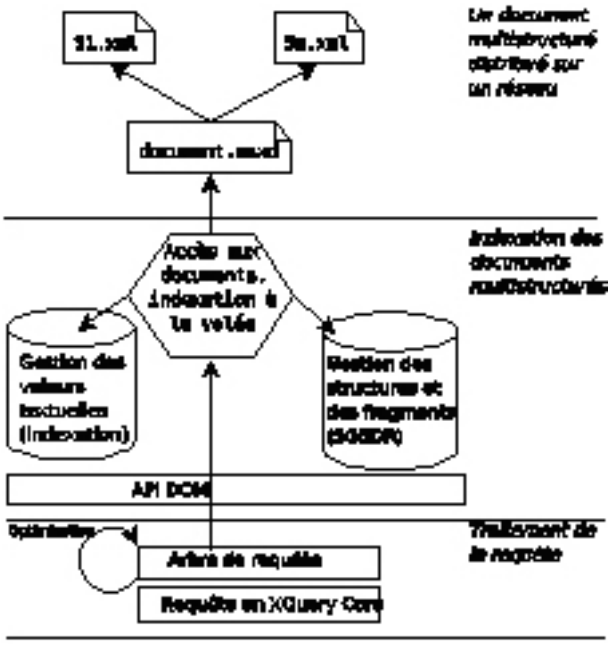

Figure 5 : Architecture du prototype

PROLOG

declare variable $doc := msxd:doc("example.msxd");

Q1 - Enfants de Manuscrit
$doc/manuscrit/*
retourne tous les enfants des fragments manuscrit dans toutes les structures concernées (i.e. page (dans S1), syntaxe (dans S2), et prescriptions (dans S3)).
Q2 - Lignes contenant des phrases
$doc//ligne[descendant::phrase]
Q3 - Première ligne suivant la première prescription
$doc/manuscrit//prescription[1]/following::ligne[1]
Q4 - Lignes qui sont exactement des phrases
for $v in $doc//ligne return
$doc//phrase[. is-equal $v]
Q5 - Phrases à cheval sur plusieurs lignes
for $v in $doc//ligne return
$doc//phrase[. is-overlapping $v]

Les auteurs remercient Madame M.-S. CORRADINI BOZZI, Philologue, Dipartimento di lingue e letterature romanze, Università di Pisa (Italia)Bozzi.

7. RÉFÉRENCES BIBLIOGRAPHIQUES :

[ABA03] R. Abascal et al., Modéliser la structuration multiple des documents. *In H2PTM Hypertexte et Hypermédia Créer du sens à l'ère du numérique. Hermès, septembre 2003.*

[ALL91] James Allen, Time and time again : The many ways to represent time. *International Journal of Intelligent Systems, 6(4):341-355, july 1991.*

[BIR01] P.-V. Biron and A. Malhotra, XML Schema Part 2: Datatypes. Rec., *W3C, 2001.*

[BOA03] S. Boag, XQuery 1.0 : An XML Query Language. Draft, W3C, 2003.

[COR90] M.S. Corradini, Etude des textes de matière medico-pharmacetique en langue d'oc. *Bulletin de l'Association Internationales d'Etudes Occitanes, VIII, pp 29-34, 1990.*

[BRA04] T. Bray., Namespaces in XML 1.1. Rec., W3C, 2004.

[BRA98] T. Bray, J. Paoli, and C.-M. Sperberg-McQueen, Extensible Markup Language (XML) 1.0. Rec., W3C, 1998.

[BRU06] E. Bruno, S. Calabretto, et E. Murisasco, *Documents textuels multistructurés : un état de l'art. In Rapport interne, LSIS (UMR 6168), Université du Sud Toulon-Var, 2006.*

[CLA99a] J. Clark, XSL Transformations (XSLT) V1.0. Rec., W3C, 1999.

[CLA99b] J. Clark and S. Derose, XML Path Language (XPath) V1.0. Rec., W3C, 1999.

[CLA01] J. Clark and M. Murata, RELAX NG Specification. Tech. *report, OASIS, 2001.*

[DEK05] A. Dekhtyar and I.-E. Iacob, A framework for management of concurrent XML markup. Data and Knowledge Engineering, 52(2):185-208, 2005.

[DER04] S. DeRose, Markup overlap : a review and a horse. *In Extreme markup language 2004 Conference Proceedings, 2004.*

[DUR01] P. Durusau, Implementing concurrent markup in XML. In Extreme Markup Languages Conference Proceedings, august 2001.

5 http://www.w3.org/TR/xquery-semantics/

[FER03] M. Fernandez et al., XQuery 1.0 and XPath 2.0 Data Model. Draft, W3C, 2003.

[GOL90] C.-F. Goldfarb and Y. Rubinsky. The SGML handbook. Clarendon Press, Oxford, 1990.

[GON87] G. Gonnet and F. -W. Tompa, Mind your grammar : a new approach to modelling text. In the 13th Conference on Very Large Data Bases (VLDB'87), pages 339-346, 1987.

[HOR04] A. Le Hors et al., Document object model (DOM) level 3 core spec. Rec., W3C, 2004.

[IAC04] I.-E. Iacob et al., Parsing concurrent XML. In Proceedings, 6th ACM International Workshop on Web Information and Data Management (WIDM 2004), pages 23-30, 2004.

[JAG04] H.-V. Jagadish et al., Colorful XML: One Hierarchy Isn't Enough. In SIGMOD Conference, pages 251-262, 2004.

[MUR00] M. Murata, Hedge automata: a formal model for XML schemata. 2000.

[SPE01] C.-M. Sperberg-McQueen and L. Burnard, Tei p4 guidelines for electronic text encoding and interchange. 2001.

[SPE00] C.-M. Sperberg-McQueen and C. Huitfeldt, Goddag : A data structure for overlapping hierarchies. In DDEP/PODDP, pages 139-160, 2000.

[TEN02] J. Tennison and W. Piez, Layered markup and annotation language (LMNL). *In The Late breaking paper presented at Extreme Markup, 2002.*

[VAN04] E. Van Der Vlist, Relax NG. O'reilly Et Associates, 2004.

[WIT02] A.Witt, Meaning and interpretation of concurrent markup. *In Joint Conference of the ALLC and ACH, 2002.*

[WIT04] A. Witt, Multiple hierarchies : news aspects of an old solution. *In Extreme markup language 2004 Conference Proceedings, 2004.*

Modèle d'accès multi-supports et multi-canaux aux "documents d'actualité" : transformations éditoriales et variété des modes de distribution et d'accès

Cécile PAYEUR
Manuel ZACKLAD

Université de Technologie de Troyes (UTT)
FRE CNRS 2848 - ICD - Equipe Tech-CICO
12, rue Marie Curie, BP 2060
F-10 010 TROYES CEDEX
cecile.payeur@utt.fr
manuel.zacklad@utt.fr

RÉSUMÉ

Au-delà des visions optimistes développées par certains analystes enthousiastes à l'idée de l'émergence d'une cyberculture, de nombreux acteurs, confrontés à la virtualisation des supports de diffusion documentaires, s'interrogent sur l'évolution des documents et sur les conditions de leur mise en circulation. Dans cet article, il s'agit pour nous de jeter les bases d'une explication générale, s'appuyant fortement sur l'exemple de la presse écrite. Après avoir proposé un modèle générique posant la question de l'accès aux " documents d'actualité ", nous verrons dans quelle mesure ce modèle peut s'adapter à des contextes particuliers.

MOTS-CLÉS :
document, diffusion, dématérialisation, presse écrite, éditorialisation.

ABSTRACT

Beyond the optimistic visions about the emergence of a cyberculture, both the evolution of the documents and the conditions of their circulation raise many questions. In this paper, we initiate a general explication based on the example of the press. In a first part, we propose a generic model representing the problem of access to up-to-date documents. In a second part, based on the case of news documents, we consider the question of adapting this model to a particular context.

INTRODUCTION

A l'heure de la dématérialisation, il existe toute une littérature qui insiste sur la virtualisation du monde. Dans cette littérature se dégagent des visions optimistes comme celle de Pierre Lévy. Ce dernier dresse un tableau raisonné, mais finalement optimiste, d'une cyberculture en émergence. Dans cette culture, ou ce nouvel espace de communication riche en potentialités, " la Poste, le Téléphone, la Presse, l'Édition, les Radios, les innombrables chaînes de Télévision forment désormais la frange imparfaite, les appendices partiels et tous différents d'un espace d'interconnexion ouvert, animé de communications transversales, chaotique, tourbillonnant, fractal, mû par des processus magmatiques d'intelligence collective " [LE98]. Cependant, cette vision tend à faire du réel et du virtuel deux univers dont l'un semble se détacher de l'autre de manière de plus en plus radicale. En même temps, on peut se demander si cette croissance exclut tout retour en arrière et condamne, à terme, les formes antérieures. L'évolution qui se dessine ne tendrait-t-elle pas plutôt vers une composition de plus en plus forte entre ces deux univers, qui s'imbriquent, se complètent, et se réinventent ? C'est pourquoi, il nous semble que le passage du réel au virtuel demande à être approfondi et à être mis à l'épreuve des faits. C'est ce besoin que traduit notamment le " Project for Excellence in Journalism" [JO06] dans le domaine des médias, et plus particulièrement de la presse. Ce projet, mené actuellement par des chercheurs à l'Université de Columbia, aux Etats-Unis, a pour objectif de clarifier l'illusion née de la multiplication des accès à l'information due au numérique. Cette multiplication, qui peut susciter l'enthousiasme sur le thème de l'entrée dans la cyberculture, reste complexe, ambiguë et parfois décevante.

C'est dans un tel contexte de tensions qu'il nous semble essentiel de nous interroger sur la question des modalités d'accès aux " documents d'actualité ", produits en particulier par des acteurs qui, à l'instar des organes de presse, peuvent hésiter entre le numérique et les supports matériels. Les articles de presse, offerts par des journaux existants en version papier et numérique, constituent un cas particulier, mais tout à fait significatif, de tous les documents mis en circulation, et qui rentrent dans des processus de réécriture, retranscription, transfert entre matérialité et dématérialisation. Pour aborder ce phénomène, nous nous intéresserons d'abord à la théorie des DopA comme cadre conceptuel permettant d'introduire la notion de " document d'actualité ", puis, en nous appuyant sur un double constat autour de l'accès à ce type de document, nous aborderons la question de la construction d'un modèle. Plus largement, cette recherche qui débute - et dont nous ne présentons ici que les premiers éléments disponibles - a pour finalité d'entrouvrir des perspectives d'innovations techniques dans le secteur de la presse écrite.

8. LES " DOCUMENTS D'ACTUALITÉ " PRIS DANS LE CADRE THÉORIQUE DES DOPA

Parmi les théories actuelles, le cadre conceptuel des Documents pour l'Action (DopA) [ZA04] propose une définition du concept de document et permet d'appréhender autrement des contenus multiformes, qui relèvent de moins en moins de la catégorie du texte classique pour inclure des images et des sons accessibles à partir de systèmes de navigation de natures de plus en plus variées. En insistant sur la dimension collective de l'activité rédactionnelle, il permet d'analyser les documents comme relevant de processus de communication pour partie différés, au sens des processus asynchrones décrits dans le champ du CSCW (Computer Supported Cooperative Work ou travail collaboratif assisté par ordinateur), entre des producteurs et des récepteurs liés par des intérêts communs. Cette vision communicationnelle du document s'inspire elle-même de l'approche en termes de transactions communicationnelles symboliques dans laquelle le document est analysé comme étant l'objet d'une transaction entre des acteurs impliqués dans un processus d'échange visant à la fois des engagements mettant en jeu leur " self " et des connaissances liées à la production d'une " œuvre " au moins pour partie commune [ZA04].

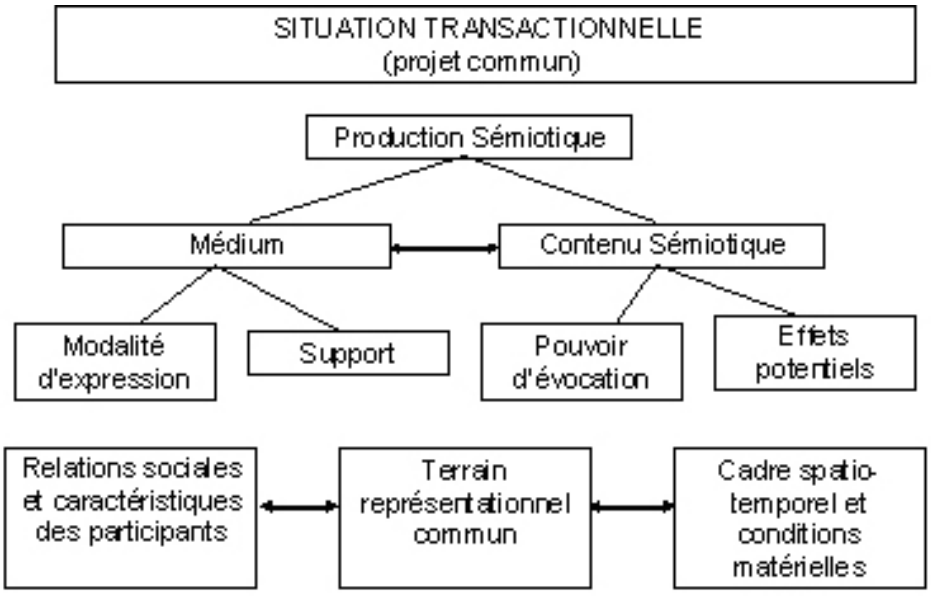

Figure 1 : Analyse de la production sémiotique [ZA04]

Dans la description d'un échange en termes de transactions communicationnelles, on considère qu'un self " réalisateur " et un self " bénéficiaire " ont une relation qui est médiatisée par des productions sémiotiques qu'ils coproduisent. La décomposition de la production sémiotique amène à établir une distinction entre, d'une part, la forme et le support de cette production ou médium de la production sémiotique et d'autre part, le contenu sémiotique " véhiculé " par ce médium. Une production sémiotique est donc l'association d'un contenu sémiotique - par exemple l'histoire de Blanche Neige - et d'un médium, qui se décompose lui-même en une modalité d'expression (récit oral, texte écrit, film) et un type de support (présence du narrateur ou enregistrement magnéto pour le récit oral, papier ou électronique pour le texte, DVD ou VHS pour le film, etc.). Bien sûr le choix du médium aura une influence sur le " pouvoir d'évocation " associé au contenu sémiotique.

Ce schéma inclut trois dimensions également proposées par l'auteur collectif Roger T. Pédauque [PE03] pour analyser un document: la " forme " qui correspond à la notion de support, et qui peut être matérielle ou immatérielle, le " signe " qui génère du sens en donnant lieu à une production sémiotique, et le " medium " qui le rend porteur de communication et lui confère une dimension sociale inscrite dans le cadre d'une situation transactionnelle.

Au sein du précédent schéma, deux éléments nous semblent particulièrement importants dans leur " instabilité " : il s'agit du contenu et du support. Pour nous, les " documents d'actualité " sont des documents qui intègrent l'actualisation de ces deux paramètres au sein d'un espace temporel réduit et d'un espace géographique distribué. Ces documents sont soumis à des actualisations successives qui sont de deux types : des réécritures touchant le contenu - effectuées notamment par l'ensemble de leurs auteurs, qui peuvent modifier le contenu, mais aussi l'annoter, le commenter - ou des changements de supports, ouvrant sur des espaces d'accès différents. Les documents culturels, mais aussi les Wiki , les blogs, les forums de discussion sur Internet, les documents de travail échangés au sein d'une entreprise, ou tout document comprenant ce double caractère sont des " documents d'actualité ". Cependant, le " document d'actualité " le plus représentatif reste le journal, au sens générique du terme, c'est-à-dire regroupant à la fois les notions de journal et de magazine. Il est donc significatif d'appliquer nos recherches au cas de la presse écrite. On le voit, une des contraintes qui pèse sur le " document d'actualité " est donc la diffusion en temps réel dans des lieux différents, sous des formes différentes et à destination de publics ciblés.

Un Wiki est un site Web dynamique permettant à tout individu d'en modifier les pages à volonté. Il permet non seulement de communiquer et de diffuser des informations rapidement, mais aussi de structurer cette information pour permettre d'y naviguer. L'encyclopédie Wikipédia se définit elle-même comme une " encyclopédie libre, gratuite, universelle, multilingue et écrite de manière collaborative sur Internet. Ce travail collaboratif est réalisé par des volontaires, sur un site Web utilisant la technologie Wiki, ce qui signifie que des articles peuvent y être ajoutés, complétés ou modifiés par pratiquement quiconque " [WK06].

9. L'ÉMERGENCE D'UN NOUVEAU PARADIGME

9.1. Un processus de transformations éditoriales

La théorie des DoPA exposée précédemment le montre: le document, à l'époque de la dématérialisation, n'est donc plus un document fixe dépendant d'un seul support, mais il peut désormais être défini comme un document vivant, soumis à de multiples transformations éditoriales au cours d'un cycle de vie de plus en plus rapide, et dans lequel il est sans cesse recomposé et manipulé. Un même document suit un " processus d'éditorialisation " [PR05] au cours duquel ses différentes formes sont manipulées par divers acteurs. Chaque médiation éditoriale inclut, à notre sens :

- une recomposition et une nouvelle hiérarchisation des éléments constitutifs du document,
- le choix d'un support et d'un mode de diffusion associé,
- une documentarisation, au sens de la théorie des DoPA [ZA04], qui consiste à équiper le support d'attributs spécifiques visant à faciliter les pratiques liées à son exploitation ultérieure dans le cadre de la préservation de transactions communicationnelles réparties. Ces attributs (nom de l'auteur, date, titre, sous-titres, version…) doivent permettre au document de circuler à travers l'espace, le temps, les communautés d'interprétation, pour tenter de prolonger les transactions communicationnelles initiées par ses réalisateurs.

Quand le document est écrit par un collectif réparti dans l'espace et le temps, il est constitué de multiples fragments eux- mêmes dotés d'attributs permettant leur articulation. Ces documents fragmentaires constituent les DopA.

Chaque nouvelle étape, qui associe une nouvelle composition de ces trois axes, est l'occasion d'une nouvelle réflexion éditoriale. Elle vient modifier, renforcer ou, à l'inverse, gommer les effets des médiations précédentes. Cette médiation passe soit par un intermédiaire humain, soit par un intermédiaire technique. Dominique Cotte décrit à ce titre les dangers de la tendance à l'automatisation des processus éditoriaux grâce à des logiciels ou à des commandes qui donnent l'illusion de pouvoir donner un sens en recomposant de manière mécanique les différentes " briques " de contenu, sans intervention humaine. Il met en garde contre " l'un des postulats couramment rencontrés (qui) consiste à affirmer que désormais la publication d'un texte (au sens large, c'est-à-dire aussi bien un rapport, qu'un site, ou qu'une plate-forme éditoriale…) peut se résumer à l'agencement de parties distinctes, pré-formées, et que le sens général se construit lors de cet assemblage." [CO04]. Ce type de processus est typiquement celui mis en œuvre dans les sites de syndication de contenu comme celui de Google-Actualités, sur lequel nous aurons l'occasion de revenir dans la suite de cet article.

9.2. Une multiplicité de modes d'accès

Nous l'avons vu, un même contenu d'actualité suit donc un processus de transformations éditoriales. De plus, il est également distribué sur plusieurs supports et par l'intermédiaire de plusieurs canaux. Les façons d'accéder au contenu sont multiples. Nous en relèverons trois principales : les annuaires, les moteurs de recherche et la syndication de contenu.

- Les annuaires :

Les annuaires, tout d'abord, s'appuient sur le principe du catalogue et permettent un accès par familles, par catégories ou - de plus en plus fréquemment dans le domaine de la consommation - par univers. Ce mode d'accès est présent aussi bien au sein des espaces physiques que des espaces virtuels. C'est le principe qui est adopté dans le merchandising des produits de la grande distribution ou sur les sites marchands de vente en ligne.

- Les moteurs de recherche :

Les moteurs de recherche, ensuite, supposent la notion additionnelle de référencement. Ils permettent d'accéder par mots-clés à l'ensemble d'un corpus et de faciliter la recherche au sein du contenu. Ils ouvrent sur des espaces virtuels ou hybrides.

Figure 2 : Accès couplé par familles et par moteur de recherche sur le site de la FNAC [FN06].

L'exemple des systèmes documentaires des bibliothèques est particulièrement significatif : accessibles aussi bien dans la bibliothèque physique que depuis un portail Internet, ils renvoient vers les livres en rayons - qui peuvent aussi, le cas échéant, être expédiés par la poste.

- La syndication de contenu :

La syndication de contenu, enfin, fait depuis peu irruption dans le monde de l'Internet. Le terme est issu du monde de la presse écrite américaine au début du vingtième siècle. A cette époque, les syndicats américains vendent leur production (bandes dessinées, avis, rubriques de jeux…) aux journaux écrits. Un bon exemple de syndicat est la société KFS (King Feature Syndicat) [KF06]. La notion de syndication est ensuite élargie et appliquée aux autres médias. Dans le domaine de la télévision, signer un contrat de syndication permet d'acheter le droit de diffuser un même contenu un certain nombre de fois sur une période donnée. Sur Internet, le terme syndication désigne un " procédé consistant à rendre disponible une partie du contenu d'un site web afin qu'elle soit utilisée par d'autres sites " [DI06].

Techniquement la syndication de contenu sur le web est basée sur l'utilisation d'un type particulier de fichiers XML: les fichiers RSS (Really Simple Syndication). Lorsqu'un site propose un flux RSS, cela rend alors possible l'extraction du contenu qu'il propose. Le contenu régulièrement mis à jour peut soit être récupéré par un autre site Internet, soit par un utilisateur utilisant un agrégateur de contenu.

En utilisant ce procédé, de nombreux sites donnent désormais accès à l'ensemble ou à une partie de leur contenu dans le but d'augmenter leur audience. C'est le cas de la plupart des sites de journaux en ligne. Par l'intermédiaire de la syndication de contenu, les éditeurs de presse offrent la possibilité de récupérer les actualités et les articles les plus récents et de les diffuser sur d'autres sites. Par exemple, le site du Monde Diplomatique ouvre un lien RSS vers l'ensemble des articles récemment publiés sur son site. A l'inverse, d'autres sites utilisent ce principe avec une fonction d'agrégation qui permet d'offrir un point d'entrée unique vers l'ensemble des articles traitant d'un thème bien précis. Google-Actualités propose ainsi un accès vers l'ensemble des articles d'actualité traitant d'un thème particulier et renvoie vers les sites Internet des éditeurs de presse. Ici, la notion de " péremption " du contenu d'actualité, et de temps réel, est particulièrement visible puisque Google indique à côté de chaque article le nombre d'heures depuis lesquelles il est publié, garantissant ainsi à ses lecteurs l'accès à une information de " première fraîcheur ".
Sur le même modèle, le nouveau service en ligne proposé par Microsoft, actuellement en version test, contient notamment un lecteur de fils RSS personnalisable. Il permet à l'internaute de construire facilement sa page d'accueil personnelle avec l'ensemble des contenus d'actualité qui l'intéressent.

Figure 3 : Site de syndication de contenu : Google-Actualités [GO06].

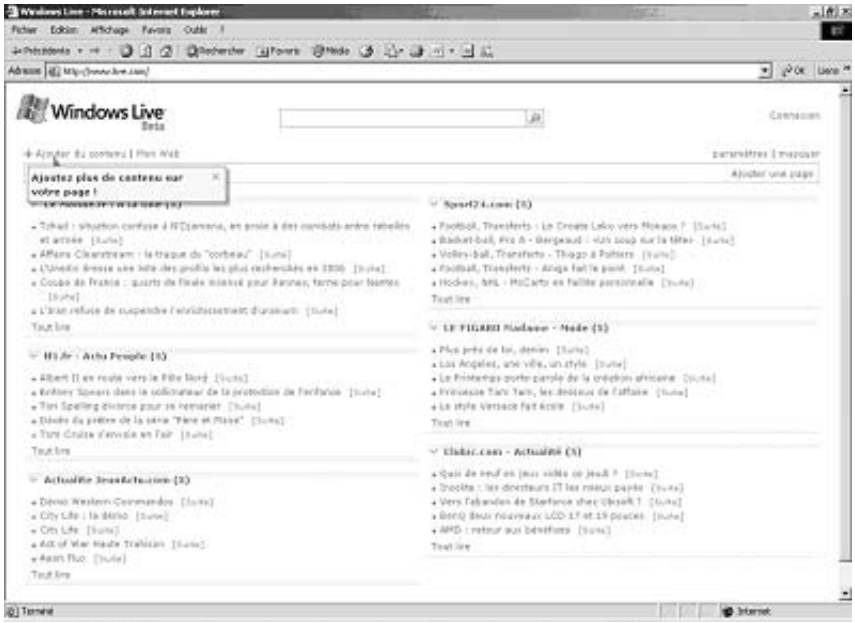

Figure 4 : Site de syndication de contenu : version bêta de Windows Live [WN06].

10. UN MODÈLE POUR PENSER CE PARADIGME

10.1. Un modèle générique

Pour nous aider à penser ce paradigme, nous proposons de construire un modèle d'accès aux " documents d'actualité " qui prenne en compte à la fois la problématique de la médiation éditoriale et celle des modalités d'accès. Voici la première version de notre travail :

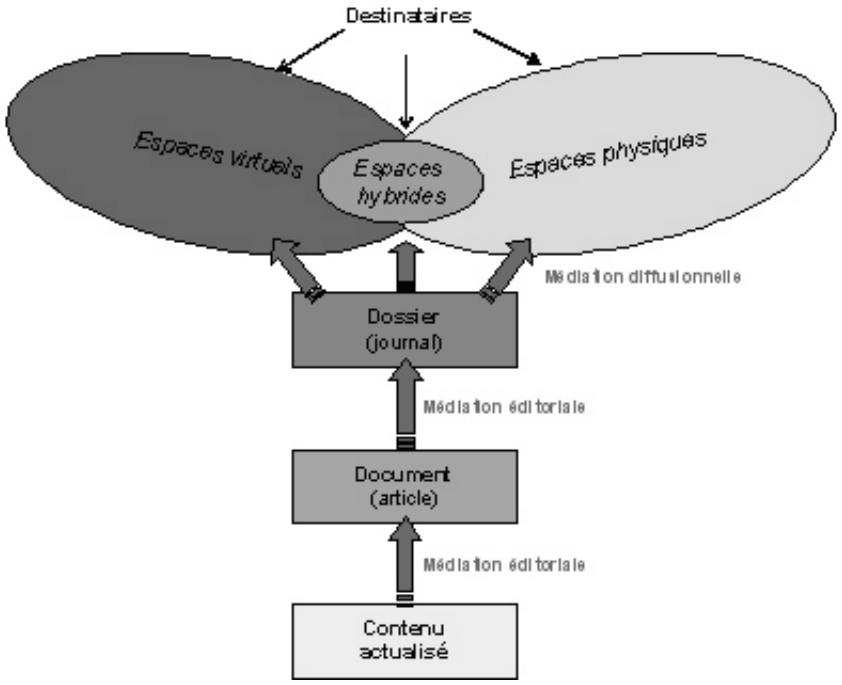

Figure 5 : Modèle générique d'accès aux " documents d'actualité ".

Le modèle fonctionne sur une lecture à deux niveaux :

- Le niveau du document,
- Le niveau du dossier, c'est-à-dire d'une collection documentaire cohérente, documentarisée, associée à un support unique ou à une structure de lien.

Correspondant à ces deux niveaux, un même contenu actualisé peut subir plusieurs stades de médiations successives :

- Une médiation éditoriale lors du passage du contenu au document,

- Une médiation éditoriale lors du passage du document au dossier,
- Une médiation de type diffusionnelle, qui a pour fonction d'ouvrir un espace d'accès jusqu'au destinataire du contenu.

Cette dernière médiation, même si elle transforme moins le document, est le point central du processus car c'est le niveau du lien avec le destinataire final. Ce lien s'effectue au sein d'espaces d'accès différents correspondant à deux types de canaux de distribution. Les canaux de distribution des produits physiques, tout d'abord, s'appuient sur des réseaux de transport et s'ouvrent sur des espaces physiques allant des espaces de vente au domicile du destinataire. Les canaux du virtuel, ensuite, s'appuient sur les réseaux du numérique et débouchent non seulement sur des espaces virtuels, mais potentiellement aussi sur des espaces que nous qualifierons d'espaces " hybrides " et qui émergent au cœur même des espaces physiques. C'est le cas par exemple des bornes d'achat de billets à l'entrée des cinémas, qui font le relais d'une réservation Internet, ou des dispositifs d'information sur les lieux de vente. Notre interrogation pourrait se résumer ainsi : quel canal emprunter et quel support choisir pour proposer au destinataire un contenu actualisé qui l'intéresse, sous la bonne forme, dans le bon espace et au bon moment?

10.2. Un modèle contingent appliqué au secteur de la presse écrite

Dans le contexte de la presse écrite, nous pouvons reformuler notre problématique de la manière suivante : quel mode de diffusion emprunter et quel support choisir pour proposer au lecteur-client un contenu journalistique qui l'intéresse, sous la bonne forme, dans le bon espace et au bon moment? Pour tenter d'apporter des éléments de réponse à cette question, nous travaillons sur un modèle qui permette d'expliciter les différents supports et les différents modes d'accès aux journaux de presse écrite, en faisant apparaître les processus de médiation éditoriale. Nous avons donc construit une première version de modèle contingent appliqué au domaine de la presse écrite.

On le voit, un même contenu journalistique est désormais accessible selon deux niveaux : celui du journal (ou dossier), qui est le mode traditionnel d'accès au contenu journalistique, mais aussi, et de plus en plus fréquemment avec les technologies du numérique, celui de l'article (ou document), en dehors du contexte éditorial du journal. Deux médiations éditoriales peuvent donc potentiellement intervenir: au niveau du passage du contenu à l'article (assemblage des différents types de contenus, mise en forme de l'article, hiérarchisation des thèmes, pérennisation du support…), et de l'article au journal (journal papier ou numérique, respect de la maquette, hiérarchisation des articles…). La médiation diffusionnelle finale intervient, quant à elle, soit au niveau de l'article, soit au niveau du journal. Elle ouvre, selon les combinaisons adoptées au cours de la ou des médiations précédentes, des espaces physiques, virtuels ou hybrides d'accès au contenu.

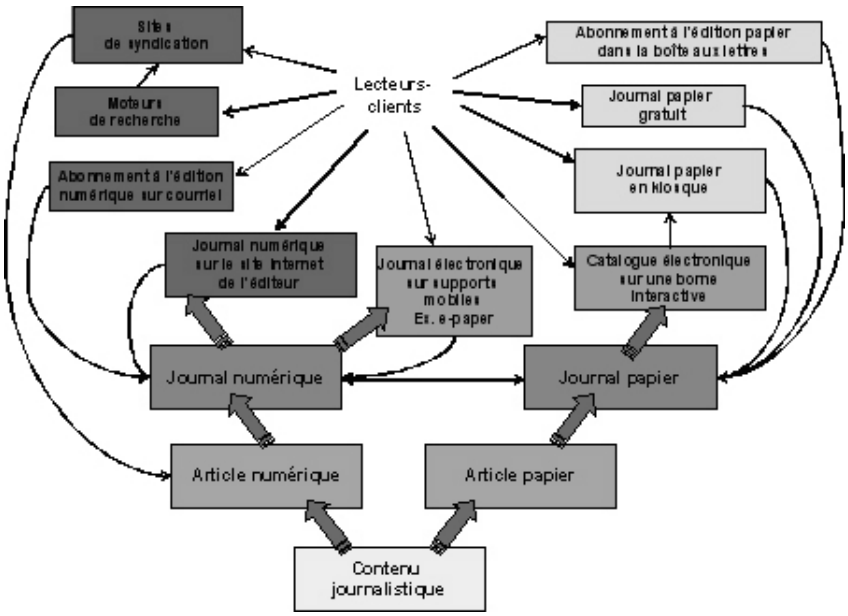

Figure 6 : Modèle contingent d'accès aux journaux dans le secteur de la presse écrite.

D'un côté, on retrouve le circuit distribution de la presse papier classique, objet physique tangible, et accessible depuis des espaces physiques tangibles. La vente au numéro des journaux papiers est basée sur un système conçu au sortir de la seconde guerre mondiale. Ce circuit spécifique, régi très précisément par un cadre légal, a pour objectif de garantir la liberté de la presse écrite. Il repose sur deux socles complémentaires, et sans lesquels elle n'existe pas : la liberté d'écriture et la liberté de diffusion. En résumé, tout éditeur est libre de diffuser lui-même son journal (c'est le choix que font les journaux gratuits et beaucoup de quotidiens régionaux) ou de le confier à un distributeur de presse, qui s'appuie sur un dense réseau de grossistes et de points de vente de proximité - les diffuseurs de presse - pour l'acheminer jusqu'au lecteur-client. Les journaux papiers vendus par abonnement sont expédiés par le système postal classique. Enfin, le portage permet d'acheminer directement les exemplaires chez les lecteurs-clients.

D'un autre côté, se trouvent les circuits des documents numériques en émergence, accessibles depuis les espaces virtuels que sont les réseaux. Ici, les grandes tendances qui se dessinent sont les suivantes : des journaux en ligne accessibles directement depuis le site des éditeurs de presse, un accès direct au contenu d'un article sorti du contexte du journal par l'intermédiaire

de sites de syndication ou l'abonnement à une édition numérique du journal, livrée directement sur le poste informatique du lecteur-client.

Au milieu de ces deux espaces d'accès naissent des espaces hybrides, dont l'accès repose sur des supports à inventer. Ainsi, le e-paper récemment annoncé devrait permettre à un lecteur-client mobile de recevoir en temps réel un contenu actualisé directement sur un journal électronique flexible au cours de ses déplacements, c'est-à-dire navigant potentiellement au sein des espaces physiques que sont les lieux de vente des produits physiques tangibles. Un autre exemple est celui d'une borne interactive qui pourrait proposer un système d'accès par catalogue renvoyant aux journaux papiers, ou encore l'achat, par la voie du téléchargement, de journaux dématérialisés directement sur le lieu de vente.

CONCLUSION ET PERSPECTIVES

La particularité du modèle que nous avons commencé de construire dans la première partie de cet article est d'ouvrir la voie sur des complémentarités à inventer entre le numérique et le papier, entre le virtuel et le réel. Il permet d'amorcer une réflexion prospective, voire stratégique, sur les modalités d'accès aux " documents d'actualité ". Au-delà encore : la question qui nous intéresse est celle que nous amorçons lors du passage au niveau du modèle contingent : c'est-à-dire celle de l'inscription d'un tel modèle dans la réalité au regard des possibilités techniques et des modes de valorisation économique, culturelle ou sociale. En effet, s'il est important, dans un premier temps, d'essayer de construire un modèle générique adapté à un grand nombre de problèmes du " document d'actualité ", passer d'une forme générique à une forme contingente oriente vers une nécessaire prise en compte de toutes les interactions observables entre les organisations et leurs environnements, dans lesquels les usages et les évolutions technologiques ont une large part. C'est l'analyse, mais aussi l'expérimentation en réel de ces différents paramètres qui nous permettra, dans un secteur donné, de concevoir et de mettre en oeuvre des dispositifs techniques innovants, créateurs de valeur, et acceptés par l'ensemble des acteurs.

BIBLIOGRAPHIE

[CO04] Cotte, D., Després-Lonnet, M., " Le document numérique comme " lego "® ou La dialectique peut-elle casser des briques? ", article de revue à comité de lecture, revue I3 4 1:159-172, 05/07/2004.

[DI06] Dictionnaire en ligne collaboratif Dico du Net, définition du terme syndication [en ligne], disponible à l'adresse suivante : http://www.dicodunet.com/definitions/creation-web/syndication.htm (consultée le 1er mars 2006).

[FN06] FNAC, site Internet de la FNAC [en ligne], disponible à l'adresse suivante : http://www.fnac.com (consultée le 1er mars 2006).

[GO06] Google Actualités, site Internet de Google Actualités [en ligne], disponible à l'adresse suivante: http://news.google.fr/ (consultée le 13 avril 2006).

[JO06] Journalism, " Project for Excellence in Journalism " [en ligne], disponible à l'adresse suivante : http://www.journalism.org/who/pej/about.asp (consultée le 04 avril 2006).

[KF06] King Features Syndicat (KFS), site Internet de la société KFS [en ligne], disponible à l'adresse suivante: http://www.kingfeatures.com/ (consultée le 1er mars 2006).

[LE98] Lévy, P., " L'universel sans totalité " [en ligne], disponible à l'adresse suivante: http://www.archipress.org/levy/cyberculture/universel.htm (consultée le 1er avril 2006), extraits de Cyberculture, rapport au Conseil de l'Europe, Paris, Odile Jacob, 1998.

[PE03] Pédauque, R.T. " Document : forme, signe et medium, le re-formulations du numérique ", 08 juillet 2003, working paper, archive SIC
[en ligne], disponible à l'adresse suivante:
http://archivesic.ccsd.cnrs.fr/sic_00000511.html
(consultée le 1er mars 2006).

[PR05] Peyrelong, M. F., Guyot, B., " Quelques résultats pour les sciences de l'information ", chapitre du rapport final Action Spécifique " document et organisation " RTP CNRS, 2005.

[WK06] Wikipédia, définition de l'encyclopédie libre Wikipédia [en ligne], disponible à l'adresse suivante: http://fr.wikipedia.org/wiki/Wikip%C3%A9dia (consultée le 1er mars 2006).

[WN06] Windows Live, site Internet de Windows Live [en ligne], disponible à l'adresse suivante: http://www.live.com (consultée le 5 avril 2006).

[ZA04] Zacklad, M. (2004), " Processus de documentation dans les Documents pour l'Action (DoPA) : statut des annotations et technologies de la coopération associées ", in actes du colloque " Le numérique : Impact sur le cycle de vie du document pour une analyse interdisciplinaire ", Montréal (Québec), 13-15 octobre 2004.

Construction d'une application vocale pour la sélection d'objets à l'aide d'un modèle basé sur les hypergraphes

Cyril BAZIN
Florent CHUFFART
Jacques MADELAINE

GREYC, Université de Caen, 14032 Caen Cedex, France
cyril.bazin@info.unicaen.fr
France Telecom R&D
42 rue des coutures 14000 Caen, France
florent.chuffart@francetelecom.com
GREYC, Université de Caen, 14032 Caen Cedex, France
jacques.madelaine@info.unicaen.fr

RÉSUMÉ

Cet article s'intéresse à l'élaboration de services vocaux. Il s'agit principalement du problème de l'accès multimodal à un document interactif. Le but du service est de permettre une sélection d'un élément d'un ensemble fini, mais de grande cardinalité, sans connaissance a priori fine de l'utilisateur de cet ensemble. L'apport principal du système est la production d'une interaction vocale efficace en vue de cette sélection. La solution s'appuie sur la construction d'un hypergraphe permettant d'affiner et de guider le choix de l'utilisateur. Cette solution est mise en œuvre dans une plate forme VoiceXML qui est également présentée.

Abstract

This paper presents the generation of vocal services. It adresses the problem of multimodal access to an interactive document. The goal of the service is to allow a selection into a large finite set of items, without any a priori knowledge of this set by the user. The main contribution of the system is the production of an efficient vocal interaction in order to drive that selection. The solution re on an hypergraph data struture that allows refinement of the user's choice. This solution is implemented in a VoiceXML frame work that is also presented in the paper.

MOTS-CLÉS :
hypergraphe, voiceXML, plate-forme vocale, application vocale

11. INTRODUCTION

Une expérimentation de service vocal pour le grand public a été menée à Caen dans le cadre du projet VOLCAN. Ce service permet d'avoir, par appel d'un simple numéro de téléphone, le temps d'attente avant le prochain passage d'un bus en donnant le numéro de l'arrêt. Ce service était destiné à l'origine pour être utilisé à l'arrêt du bus - lieu où est disponible l'arrêt. Une autre utilisation de ce type de service était prévue pour remplacer les panneaux lumineux disponible, par exemple au arrêt du tramway, pour les non voyants. Est apparu, à l'usage, encore une autre utilisation : téléphoner depuis son bureau ou son domicile afin d'éviter trop d'attente à l'arrêt.

Dans le souci d'élargir ce service, il a été étudié la possibilité de répondre à des requêtes du type demande d'itinéraires par des utilisateurs n'ayant pas de connaissances préalables sur le plan des bus. Le système et l'utilisateur doivent se mettre d'accord sur deux noms d'arrêts de bus. Le système se doit d'être suffisamment capable de guider l'utilisateur en lui posant des questions.

Si l'on reprend la classification de Allen [ALLEN01] donnée table 1, ce système est simplement un système de niveau 2 du genre instanciation de formulaire. Mais le problème adressé ici est celui où l'utilisateur n'a que peu de connaissances du domaine et le système ne doit pas simplement donner quelques clarifications ou éclairages, mais le guider pas-à-pas mais rapidement vers la solution.

	Technique Used	**Example Task**	**Dialogue Phenomena**
1	*Finite-state Script*	*Long-distance dialing*	*User answers questions*
2	*Frame-based*	*Getting train arrival and departure information*	*User asks questions, simple*
3	*Sets of Contexts*	*Travel booking agent*	*Shifts between predetermined topics*
4	*Plan-based Models*	*Kitchen design consultant*	*Dynamically generated topic structures, collaborative negotiation subdialogues*
5	*Agent-based Models*	*Disaster relief management*	*Different modalities (e.g. planned world and actual world)*

Table 1 : Dialogue and Task from Allen et all [ALLEN01].

11.1. 1.1 Exemple de dialogue

Le principe de notre approche a été de simuler au mieux un exemple de dialogue possible entre un opérateur téléphonique et un utilisateur du service.

L'exemple de la figure 1 permet de mieux illustrer le type de dialogue souhaité entre l'utilisateur et le système. À partir de cet exemple, il est possible de décomposer le dialogue en plusieurs parties. Tout d'abord l'introduction qui présente l'application et qui permet à l'utilisateur qui connaît le code hastus de l'arrêt de bus d'éviter le dialogue interactif. Ensuite, à la ligne 4 une question ouverte qui permet d'avoir une idée générale de l'arrêt de bus recherché par l'utilisateur. À partir de la ligne 6 et jusqu'à la ligne 9 se trouvent une suite de questions fermées qui servent à lever les ambiguïtés possibles. Enfin, une question finale, ligne 10, valide ou non la solution trouvée.

1 Opérateur :	Bonjour, bienvenue sur le serveur vocal Twisto...
2 Opérateur :	Connaissez-vous le code hastus de l'arrêt?
3 Utilisateur :	Non.
4 O :	D'où voulez-vous partir?
5 U :	Je souhaite partir de l'université.
6 O :	Voulez-vous partir du campus 1?
7 U :	Non.
8 O :	Voulez-vous partir du campus 2?
9 U :	Oui.
10 O :	Voulez vous partir de l'arrêt Maréchal Juin?
11 U :	Oui.

Figure 1 Exemple de dialogue entre l'utilisateur et le serveur vocal

11.2. 1.2 Conclusion partielle

Le problème que nous essayons de résoudre peut être découpé en plusieurs parties. D'une part, la problématique liée au dialogue téléphonique, d'autre part la problématique de choix des questions et au traitement des réponses. Notre approche a été de dissocier au maximum ces 2 parties. Pour ce faire, nous utilisons la plate-forme vocale mise au point par France Télécom.

Les problèmes de synthèse et de reconnaissance vocale est entièrement délégué à cette plate-forme. Nous utiliserons différents modules qu'elle met à notre disposition : reconnaissance à partir d'un très grand vocabulaire, reconnaissance des appuis de touches du téléphone, etc.

En se basant sur cet outil, nous avons développé un modèle de données et des algorithmes capable générer automatiquement des questions et de traiter les réponses.

La section 2 va présenter la plate forme vocale. La section 3 décrit la méthode de génération de dialogue utilisant un hypergraphe construit sur l'ensemble des objets à sélectionner. La section 4 aborde la façon de générer automatiquement l'hypergraphe.

12. DESCRIPTION DE LA PLATE-FORME VOCALE

Le groupe de travail MultiModal Interaction (MMI) du W3C propose la définition d'une architecture rendant possible l'interaction multimodale [W3C03] [W3C04a]. Elle décrit une architecture modulaire ou chacun des composants interagit avec les autres modules en s'appuyant sur les standards actuels du W3C. Cette architecture permet le traitement d'évènements multimodaux produits par l'utilisateur et assure un feedback multimodal de l'information traitée.

Un serveur vocal interactif est une instance particulière de cette architecture. Le cœur d'une plate-forme vocal tel que le conçoit le W3C est le navigateur VoiceXML [W3C04b]. Le VoiceXML est un dialecte de XML qui permet de décrire les scénarios des services vocaux.

Le langage VoiceXML permet de décrire des interactions vocales ou utilisant les touches du téléphone (DTMF). Il permet également le rendu audio de contenus numériques. Les modalités utilisées sont la lecture de fichiers audio et la synthèse de texte.

La synthèse vocale permet le rendu de contenus dynamiques. En effet, la ressource de synthèse (TTS) attend du navigateur VoiceXML du texte semi-structuré (texte ou SSML[W3C04c]). La ressource TTS synthétise le texte conformément aux instructions spécifiées en SSML (volume, vitesse, prosodie, hauteur, langue) [COTTO92] et retourne au navigateur VoiceXML le flux audio synthétisé.

Les ressources de reconnaissance automatique de la parole (ASR) permettent de transmettre des ordres, mais permettent aussi l'identification par biométrie vocale, l'enregistrement de messages vocaux ou la reconnaissance de la parole.

L'équipe du centre de recherche et développement de France Telecom basée à Caen a développé dans le cadre du projet PMX (Plate-forme Multimodale XML) un serveur vocal interactif conforme aux différents standards du marché (voir figure 2).
Notons que le module ASR de la plate forme PMX prend en charge, non seulement la reconnaissance de la parole, mais aussi celle des suites de touches du téléphone (DTMF), et plus généralement, se charge de transformer toutes les entrées en ordres pour le navigateur VoiceXML.

DTMF : Dual Tone Multi-Frequency
TTS : Text To Speech
ASR : Automatic Speech Recognition

Les grammaires SRGS [W3C04d] permettent de décrire les entrées DTMF ou paroles attendues par le système. Les grammaires les plus simples permettent de spécifier la reconnaissance d'un ordre donné par un mot ou par une touche DTMF. Des règles permettent de reconnaître des assemblages de ces éléments de base. Notons qu'il est possible de spécifier des menus à choix multiples déclenchables par mots clés ou par touche DTMF.

Dans le service décrit par le présent article, nous combinons les grammaires DTMF pour les réponses aux questions fermées et les grammaires speech pour le traitement de la question ouverte. La technologie de reconnaissance vocale permettant la sélection d'un item particulier parmi un grand choix d'items dans un délai acceptable correspondant aux contraintes temps réel du service est appelée TGV pour Très Grand Vocabulaire.

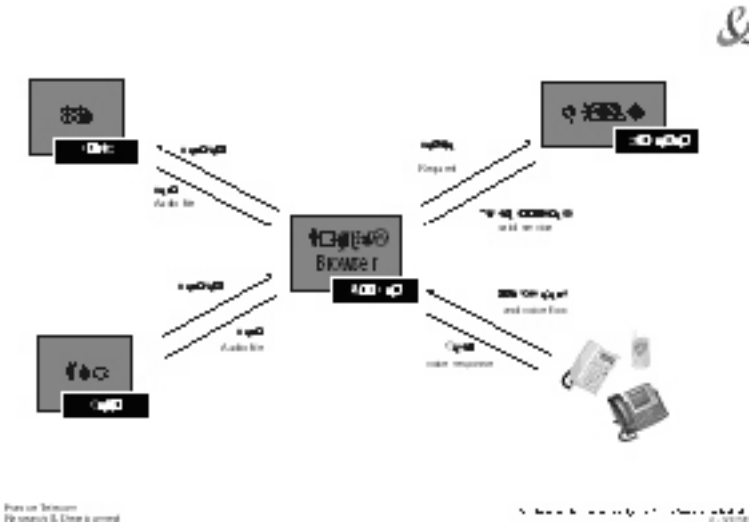

Figure 2 : Description modulaire d'un SVI.

Dans notre service l'utilisateur peut prononcer l'item (par exemple : piscine). Mais il peut également prononcer cet item dans une phrase construite (par exemple : je souhaite aller à la piscine). La technologie de reconnaissance vocale correspondant à ce second cas est appelée mot enrobé [HATON91].

Cette technique permet de reconnaître un mot ou une expression attendue enrobée dans une phrase comportant par exemple un verbe conjugué ou des mots de liaison. Notons que la structure de la phrase ou le temps du verbe n'impacte pas le comportement du service vocal. Seule la détection de l'item attendu est prise en compte par le service. Les technologies de reconnaissance

vocale prenant en compte la structure de la phrase, les temps des verbes, et plus généralement, le contexte global du dialogue sont dites "reconnaissance du dialogue naturelle". Cette technologie ne sera pas abordée dans notre étude.

La plate-forme PMX est également le cœur multimodal du Projet européen AURORA [CHUF05]. Ce projet a pour but, non seulement de promouvoir le protocole SIP et le concept de multimodalité distribuée mais également de démontrer la faisabilité du pilotage d'un service VoiceXML par le mode gestuel.

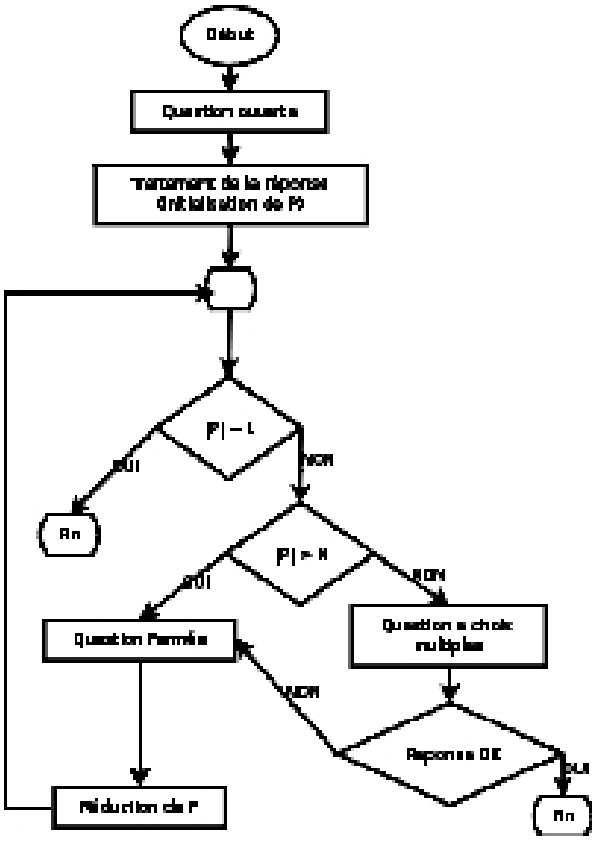

Figure 3 : Diagramme de l'application DialoBus

13. CHOIX DES QUESTIONS

Notre stratégie pour trouver un arrêt de bus est de réduire progressivement l'ensemble des solutions par une suite de questions. Pour cela, nous utilisons 3 types de questions : des questions ouvertes, des questions fermées et des questions à choix multiples.

La figure 3 illustre le comportement de notre algorithme. Nous commençons par poser une question ouverte pour obtenir un ensemble de solutions possibles de départ. Ensuite, grâce à une suite de question fermé, nous réduisons cet ensemble de solutions possibles. Lorsque l'ensemble de solutions est assez réduit, nous proposons l'ensemble des solutions à l'utilisateur.

13.1. 3.1 Rôle et agencement des composantes logiciels du système

La figure 2 montre les différentes ressources dont dispose la plate-forme vocale pour assurer l'extraction de données du système d'information. Dans cette partie nous détaillerons les échanges entre les différentes briques logicielles et l'utilisateur.

Dans un premier temps l'utilisateur initialise la session avec le serveur vocal au travers de l'interface téléphonique. Le navigateur VoiceXML charge la page de démarrage du service. Cela signifie qu'il exécute une requête HTTP à destination du serveur applicatif DialoBus. La composante DialoBus initialise une session et retourne une page VoiceXML contenant la question ouverte ainsi qu'un pointeur vers une liste de " Très Grand Vocabulaire ". Cette liste contient les mots ou expressions que le système DialoBus est capable d'interpréter.

Une fois la page VoiceXML chargée, le navigateur VoiceXML utilise sa ressource de synthèse (TTS) pour vocaliser la question ouverture envoyée par le système DialoBus. L'utilisateur peut alors répondre au système en lui précisant son lieu départ.

La réponse de l'utilisateur, traduite en signal audio est transmise à la plate-forme qui utilise sa ressource de reconnaissance de la parole (ASR) afin d'analyser le signal et d'extraire les mots pertinents pour la composante DialoBus. Cette analyse se fait par comparaison entre des motifs extraits du signal audio et le vocabulaire transmis par le système DialoBus. Les mots pertinents sont alors envoyés par le navigateur VoiceXML à la composante DialoBus en utilisant le protocole HTTP.

La composante DialoBus traite cette liste de mots afin de déterminer un ensemble d'arrêts de bus possibles. DialoBus détermine alors une nouvelle question fermée ou à choix multiples pour réduire cet ensemble des possibles.

Dans le cas d'un question à choix multiple, Le système énumère à l'utilisateur l'ensemble des choix possibles. Pour des raisons liés à la mémorisation

d'énumération dans le mode vocal, la taille des listes énumérées doit rester raisonnable[LIMAM02]. Une liste de vocabulaire, utilisée par la reconnaissance vocale, est associée à chacun des choix. Cette liste est considérablement réduite par rapport à la liste de vocabulaire qui initialise le dialogue.

Enfin dans le cas d'une question fermée, le système s'attend à une réponse affirmative ou infirmative de l'utilisateur.

Les résultats de la reconnaissance vocale sont transmis à la composante DialoBus via des requêtes HTTP. Tant que l'ensemble des possibles n'est pas réduit à un seul élément, DialoBus génère une nouvelle question selon le processus itératif décrit par la figure 3.

Le choix des questions à poser et le traitement des réponses de l'utilisateur seront obtenus grâce à une représentation des données sous forme d'hypergraphes [BERGE73].

13.2. 3.2 Modélisation par hypergraphes

Le modèle que nous proposons est basé sur des regroupements d'arrêts de bus ayant des caractères communs. La structure que nous avons choisie pour représenter ces données est l'hypergraphe (cf. figure 4).

La modélisation par hypergraphe est obtenue en définissant l'hypergraphe G=(V,E) tel que :

6. l'ensemble des sommets de l'hypergraphe V, représente les arrêts de bus;
7. chaque hyperarête de E représente un caractère commun entre les arrêts de bus qu'elle contient.

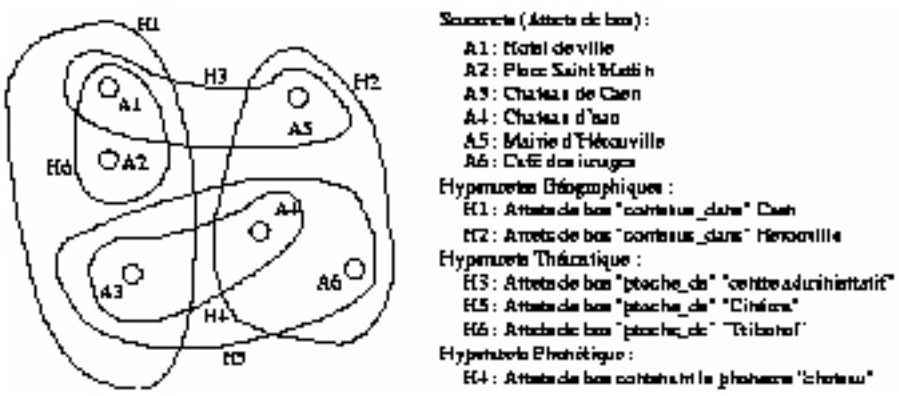

Figure 4 : Exemple d'hypergraphe à 6 sommets (A1 à A6) et à 6 hyperarêtes (H1 à H6).

Chaque hyperarête possède un type et des informations qui la caractérise. Par exemple, l'hyperarête qui regroupe les arrêts de bus proches du centre ville de Caen sera de type géographique avec les attributs suivant (dans_le_centre_de, " Caen "). L'hyperarête qui regroupe les arrêts proches des piscines sera de type thématique avec les attributs (proche_de, " piscine "). Il est possible de segmenter certaines hyperarêtes en plusieurs hyperarêtes, pour représenter les arrêts proches des piscines de Caen par exemple.

Outre les hyperarêtes de caractère géographique ou de caractère thématique, nous utiliserons des hyperarêtes phonétiques qui regroupent les arrêts de bus ayant des caractères phonétiques communs. Ces dernières hyperarêtes seront utilisées pour traiter la réponse à la première question ouverte.

13.3. 3.3 Traitement de la question ouverte

Les questions ouvertes sont utilisées pour réduire rapidement l'ensemble des solutions possibles. Elles utilisent le module de reconnaissance de très grand vocabulaire de la plate-forme vocale. Le très grand vocabulaire est généré à partir des informations portées par les hyperarêtes de type phonétique.

La plate-forme reconnaît un ensemble de mots clés dans la réponse. Les mots clés reconnus correspondent directement à un ensemble d'hyperarêtes phonétiques. Soit Ep l'ensemble des hyperarêtes phonétiques qui sont trouvées grâce à la plate-forme. Le calcul de l'ensemble des solutions potentielles initiales P0 est obtenu en effectuant l'intersection des hyperarêtes de Ep (équation 3.1).

(3.1) $$P_0 = \{a \mid a \in E_i, \forall E_i \in E_p\}$$

Le programme va ensuite réduire progressivement le nombre de solutions potentielles en posant une suite de questions fermées.

13.4. 3.4 Calcul des questions fermées

Les questions fermées que nous allons poser sont des questions qui attendent une réponse de type oui ou non. Le résultat de cette question doit permettre de réduire de façon dichotomique l'ensemble Pk. Nous choisissons la question à poser de façon à réduire l'ensemble des solutions de façon équitable indépendamment de la réponse donnée.

Le choix de la question est effectué de façon automatique en prenant une hyperarête de type géographique ou thématique. Une hyperarête H est choisie

de façon à respecter l'équation (3.2), c'est-à-dire qu'elle découpe le plus équitablement possible l'ensemble Pk en deux parties. Elle est telle qu'elle minimise la valeur absolue de la différence des cardinaux des deux sous-ensembles formés par les éléments de Pk qui sont contenus dans l'hyperarête ou non. Nous nous restreignons à l'ensemble E' des hyperarêtes qui possèdent au moins une solution potentielle. Cet ensemble E' peut être facilement calculé à partir du dual de l'hypergraphe.

(3.2)
$$H = \underset{E_i \in E'}{arg\,min}\left(\left|\left|\{a \mid a \in P_k, a \in E_i\}\right| - \left|\{b \mid b \in P_k, b \notin E_i\}\right|\right|\right)$$
$$avec\ E' = \{E_j \mid a \in E_j, \forall a \in P_k, \forall E_j \in E\}$$

Les attributs portés par chaque hyperarête permettent de générer automatiquement une question [LIMAM02][POL98]. Par exemple, une hyperarête ayant comme attributs (proche_de, " mairie ") générera le texte " Votre arrêt de bus est-il proche de la mairie? " .

Trois types de réponses sont possibles : oui, non, je ne sais pas. La réponse peut être donnée vocalement ou bien en utilisant les touches du téléphone. Si l'utilisateur répond qu'il ne sait pas, nous appliquons à nouveau l'algorithme de choix de question fermée en interdisant l'hyperarête précédente comme solution. Nous reprenons alors simplement la valeur de Pk pour pour Pk+1.

Si la réponse à la question générée à partir de l'hyperarête H est affirmative, nous déduisons un nouvelle ensemble de solution potentielles Pk+1 en retirant des solution potentielles toutes celles qui n'appartiennent pas à l'hyperarête (équation 3.3).

(3.3)
$$P_{k-1} = P_k \cap H$$

Par contre, si la réponse est négative, le nouvel ensemble des solutions potentielles Pk+1 est obtenus en enlevant de Pk les sommets de l'hyperarête H (équation 3.4).

(3.4)
$$P_{k-1} = P_k \setminus H$$

En cas de réponse positive ou négative, dans le cas où alors et . Ainsi, au fur et à mesure des réponses de l'utilisateur, le programme converge vers une solution.

13.5. 3.5 Choix final

Lorsqu'il ne reste que quelques solutions possibles, nous proposons à l'utilisateur le choix parmi la liste des solutions restantes . Chaque solution est associée à une touche du pavé numérique. La détection de la réponse de l'utilisateur peut donc être faite de façon non-ambiguë. Dans le cas où l'utilisateur ne pourrait pas répondre à la question, nous appliquons à nouveau l'algorithme en ne posant plus que des questions fermées.

14. GÉNÉRATION AUTOMATIQUE DES HYPERARÊTES

Les algorithmes de sélection des questions se basent sur un hypergraphe dont les sommets sont composés des différents arrêts de bus de la ville et dont chaque hyperarête regroupe des arrêts ayant des points communs.

La liste des arrêts de bus étant une donnée de départ, le principal problème inhérent à la génération de l'hypergraphe est la construction des hyperarêtes. Le nombre d'hyperarêtes construites étant très important (au moins 10000), il est impératif de mettre au point une méthode de construction automatique.

Le fait de générer des hyperarêtes automatiquement donne plusieurs avantages. Tout d'abord, si les arrêts de bus sont déplacés, ajoutés ou supprimés, il est possible de mettre à jour l'hypergraphe sans avoir à tout reprendre depuis le début. Par ailleurs, il devient possible d'adapter le programme rapidement à une autre agglomération.

La génération automatique des hyperarêtes ne peut pas gérer tous les cas possibles. Pour créer certaines hyperarêtes, il est possible de faire appel à un expert humain.

14.1. 4.1 Caractérisation des hyperarêtes

Les hyperarêtes sont classées selon trois types : géographiques, thématiques et phonétiques. En plus de ces classifications il est possible de catégoriser les hyperarêtes géographiques et thématiques. Par exemple, deux catégories d'hyperarêtes sont : les hyperarêtes de proximité, qui regroupent les hyperarêtes proches d'un endroit et les hyperarêtes d'inclusion qui regroupent les arrêts à l'intérieur d'une zone. La catégorie de chaque hyperarête est représentée par un attribut : proche_de, dans. Enfin, chaque attribut est paramétré par une valeur.

Pour générer une hyperarête automatiquement, il faut donc déterminer :

- les éléments qui la composent;
- les caractéristiques de cette hyperarête (attribut et valeur).

La figure 5 illustre le processus de création d'hyperarêtes géographiques et thématiques.

14.2. 4.2 Les données de départ

Pour générer les hyperarêtes, nous avons utilisé la base de donnée Géoroute de l'IGN pour le Calvados. Cette base de données géographique permet de connaître les contours des villes ainsi que les positions et les noms des routes. D'autre part, Twisto a mis à notre disposition les localisations géographiques des différents arrêts de bus de la ville.

Ces données permettent de créer automatiquement les hyperarêtes géographiques. Les hyperarêtes thématiques utiliseront en plus, pour leur génération, un annuaire électronique et une liste de lieux d'intérêts.

La construction des hyperarêtes phonétiques utilise, quant à elle, tous les noms et toponymes des données de base simplement phonétisées. Nous allons maintenant détailler chacune de ces constructions.

14.3.
14.4.
14.5.

14.6. 4.3 Génération des hyperarêtes géographiques

Les hyperarêtes géographiques sont obtenues en utilisant des opérations topologiques entre les positions des arrêts de bus et les géométries de certains objets obtenus via la base de données géographique. Ainsi, nous calculons les arrêts qui sont à l'intérieur de chaque ville en utilisant une opération d'inclusion entre la position de chaque arrêt de bus et le contour de la ville.

Nous avons aussi utilisé une approximation automatique du centre ville de chaque agglomération. Cette approximation consiste en une homothétie du contour original de chaque ville. Une opération d'inclusion permet de déterminer les arrêts qui se trouvent dans les centres villes de chaque ville.

De façon similaire, il est possible de déterminer les arrêts au nord, à l'est, à l'ouest et au sud de chaque ville [FRA91][FRA96].

Les hyperarêtes ainsi générées sont classées parmi les hyperarêtes de type géographiques et portent, comme attribut, l'opérateur topologique qui les a généré : dans, dans_le_centre_de, au_nord_de, etc. La valeur associée est le nom de la ville sur laquelle est effectuée l'opération : " Caen ", " Hérouville ", etc.

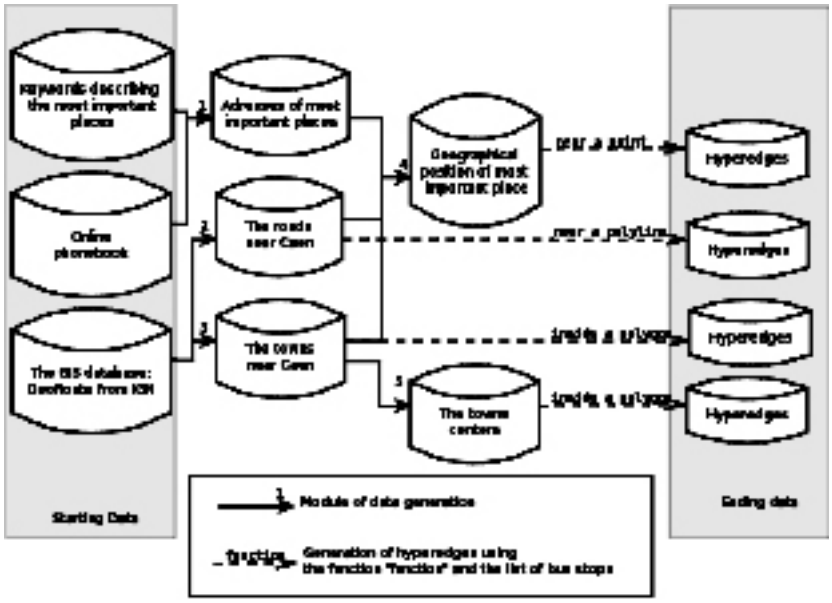

Figure 5 : processus de génération automatique de groupes d'hyperarêtes.

14.7. 4.4 Génération des hyperarêtes thématiques

En nous servant d'un robot utilisant un annuaire électronique sur le Web, nous avons pu obtenir les adresses de la plupart des lieux importants aux alentours de l'agglomération. Les lieux importants sont donnés à partir d'une liste de mots clés, par exemple : gare, mairie, piscine, église, restaurant, etc.

En utilisant la base de données " Géoroute ", nous avons conçu un logiciel de géocodage qui, à partir d'une adresse postale donnée retourne la position géographique correspondante. Un calcul de distance permet enfin d'obtenir les arrêts à proximité de cette position.

En fonction d'un thème donné, nous obtenons une liste des adresses des lieux correspondants à cette thématique. Les adresses sont géocodées et nous obtenons la liste des arrêts proches de chacun des lieux. Ces arrêts forment une hyperarête de type thématique appartenant à la catégorie des arêtes de proximité et ayant comme paramètre le nom du thème donné.

De façon similaire, nous utilisons une liste des noms des rues importantes de l'agglomération. Cette liste permet de générer des hyperarêtes correspondants aux arrêts de bus proche de chaque rue.

14.8.

14.9. 4.5 Génération des hyperarêtes phonétiques

Les hyperarêtes phonétiques sont générées en utilisant les hyperarêtes géographiques et thématiques. En se basant sur les valeurs de chaque hyperarête, nous construisons un dictionnaire qui associe à la phonétique de chaque mot, un groupe d'hyperarêtes qui comprennent cet élément phonétique. Pour chacun des groupe nous créons une hyperarête construite à partir de l'union des hyperarêtes contenues dans le groupe. L'hyperarête ainsi créée est paramétrée par la représentation phonétique du mot original.

Une table de hachage ayant pour clé le paramètre de chaque hyperarête. Permet de retrouver efficacement les hyperarêtes comprenant certains mots lors de l'analyse de la réponse à la question ouverte.

15. CONCLUSION

L'étude que nous avons menée a permis de mettre en avant une modélisation des données ainsi qu'une méthode pour obtenir ces données. Le modèle de choix des questions est assez général pour pouvoir être appliqué à différents problème de choix d'un élément au sein d'un ensemble fini par une interface vocale.

Par ailleurs, notre algorithme de génération des hyperarêtes rend possible le déploiement de notre application dans d'autres agglomérations en utilisant uniquement la position des arrêts de bus, une base de données géographique et un annuaire électronique. La génération des hyperarêtes permet aussi de mettre à jour l'hypergraphe de façon automatique lors des aménagements de lignes de bus ou lorsque l'on souhaite une catégorie de lieux importants.

Le système présenté repose sur la possibilité de construire un hypergraphe sur l'ensemble des objets à choisir. Il peut être ainsi appliqué à différents problèmes de choix d'un élément au sein d'un ensemble fini, par une interface vocale.

La possibilité de déploiement de l'application pour la recherche d'itinéraire de bus pour une autre agglomération demande comme ressource : la position géographique des arrêts de bus, une base de données géographique des rues, un annuaire électronique et une liste des lieux d'intérêts de l'agglomération.

La construction de l'hypergraphe se fait alors automatiquement. Notons qu'il est toujours possible d'adjoindre des hyperarêtes dépendant de lieux d'intérêts particuliers donnés par un expert.

Une amélioration du système pourrait reposer sur l'utilisation des données de fréquentation des différents lieux de manière à forcer le choix d'hyperarêtes les plus probables. Le système proposé peut aussi être adapté pour donner la possibilité à un lecteur de choisir interactivement parmi les points d'entrées d'un document multimodal et interactif.

16. Bibliographie

[ALLEN01] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, Towards Conversational Human-Computer Interaction, AI Magazine 22(4), pages 27-38, Winter, 2001.

[BERGE73] Claude Berge, " Introduction à la Théorie des Hypergraphes " Montréal, Presses de l'Université de Montréal, 1973.

[CHUF05] F. Chuffart. AURORA, a framework enabling multimodal interactions ICMI05 http://www.idiap.ch/ICMI05/pdf/MMMP_paper18.pdf, 2005.

[COTTO92] D. Cotto. Traitement automatique des textes en vues en vue de la synthèse vocale. Thèse, Université Paul Sabatier de Toulouse III, 1992.

[FRA91] A. Frank. Qualitative spatial reasoning about cardinal directions. In M. David and W. Dermis, editors, tenth international symposium on computer assisted Cartography, 1991.

[FRA96] A. Frank. Qualitative spatial reasoning : Cardinal directions as an example. International Journal of geographical Information Science, 10(3) pages 269-290, 1996.

[HATON91] J.-P. Haton et al. Reconnaissance Automatique de la parole Dunod édition, 1991.

[LIMAM02] M. Ould Ahmed Limam et M. Gaio, " Description textuelle de schèmas géographiques ", CIDE'02, 2002.

[POL98] A. Polguére, " La théorie sens-texte ", Dialangue - Université du Québec à Chicoutimi, 1998.

[W3C03]. W3C Multimodal Interaction Framework http://www.w3.org/TR/mmi-framework/, , 2003.

[W3C04a] W3C. EMMA: Extensible MultiModal Annotation markup language, http://www.w3.org/TR/emma/, 2004.

[W3C04b] W3C. Voice Extensible Markup Language (VoiceXML) Version 2.0, http://www.w3.org/TR/voicexml20/, 2004.

[W3C04c] W3C Speech Synthesis Markup Language (SSML) Version 1.0, http://www.w3.org/TR/speech-synthesis/ , 2004.

[W3C04d] W3C Speech Recognition Grammar Specification Version 1.0, http://www.w3.org/TR/speech-grammar/, 2004.

YML : une version épurée de XML pour faciliter une spécification rigoureuse des modèles et des transformations.

Catherine PUGIN
Rolf INGOLD

Département d'Informatique de l'Université de Fribourg
Boulevard de Pérolles 90
CH-1700 Fribourg, Suisse

{prenom.nom}@unifr.ch

ABSTRACT

This paper presents a critical point of view on XML and associated languages family. It focuses on the lack of integration of the three main languages : markup, schema and transformation language. According to the remarks made, it proposes a new markup language named YML. The focus is also put on the best-known schema languages: DTD and XML Schema. A new schema language dedicated to YML is introduced named DML. It is conceived as a YML application. YML and DML tend to simplify the existing concepts and their goal is to propose some clearer ones that will improve the development of common tools. These concepts will also provide a better integration of the new core languages.

KEYWORDS : Markup language, schema language, XML, YML, DTD, XML Schema, DML.

1. INTRODUCTION

Bâti sur des idées apparues dans le courant des années 1960-1970 et normalisé en 1986 (ISO 8879), SGML (Standard Generalized Markup Language) est le premier langage de balisage reconnu. Le but d'un tel langage est de permettre la création de documents structurés et modulaires. Le faible succès de SGML est une conséquence de sa trop grande complexité. Au début des années 1990, XML (eXtensible Markup Language) apparaît comme une tentative de simplification du langage SGML. La première recommandation est publiée en février 1998 par le W3C et le langage est rapidement adopté. La version 1.1 de XML paraît en février 2004 [XML04a].

L'objectif initial de XML est de proposer un langage simple et flexible pour publier des documents de grande taille. Il est ensuite également adopté comme langage d'échange de données sur le Web. Dès lors, c'est toute une

famille de langages qui se développe autour de XML. Le langage de modélisation XML Schema et le langage de transformation XSLT sont introduits.

XML Schema permet de définir la structure, le contenu et la sémantique d'une classe de documents XML. XSLT est une des deux composantes du langage XSL. Il est conçu initialement pour effectuer des opérations complexes sur des documents XML (génération de tables, par exemple) mais il devient petit à petit un langage généraliste pour le traitement du XML. XML Schema et XSLT forment avec XML le noyau de la famille des langages XML. Aujourd'hui, il n'est pas possible d'ignorer la position dominante de XML mais en travaillant avec ces langages, il apparaît qu'il subsiste des faiblesses.

Ainsi, l'intégration de ces trois langages est encore mauvaise. Ceci est dû à plusieurs facteurs. Tout d'abord, les spécifications des langages ne sont pas toujours rigoureuses et souvent incomplètes. Une marge d'interprétation est donc laissée au choix des développeurs. Ceci engendre des comportements non conformes suivant l'implémentation choisie et des résultats qui diffèrent selon le logiciel utilisé.

Pour résoudre les difficultés qui surviennent dans les applications XML telles que XML Schema et XSLT, nous pensons qu'il est primordial de repenser les concepts fondamentaux de XML. En simplifiant ces derniers et en proposant un nouveau langage appelé YML, nous voulons montrer que des concepts clairs dans le langage de balisage permettent de développer plus facilement les langages de modélisation et de transformation associés. De plus, une bonne intégration des trois langages peut être obtenue.

La section 2 détaille les faiblesses du langage XML et des langages de modélisation DTD et XML Schema puis la section 3 se concentre sur le langage YML qui répond à chaque reproche adressé à XML. La section 4 introduit le langage de modélisation DML dédié à YML. La section 5 présente les travaux en cours et les perspectives de ce projet. La section 6 conclut l'article.

2. CRITIQUE DES LANGAGES DE LA FAMILLE XML

Après avoir étudié et utilisé les technologies XML, il apparaît que certains points problématiques doivent être repensés. De manière générale, il semble que certains concepts aient été définis de manière pragmatique avant tout.

Dans le langage de balisage XML, les principales remarques concernent tout d'abord une mauvaise gestion des caractères blancs. L'impossible distinction entre les caractères destinés à la mise en page et ceux destinés au contenu du texte entraine des incohérences dans les traitements effectués sur les documents XML. Puis nous notons également que la définition des entités repose sur trois notions distinctes qui peuvent être simplifiées. L'intégration des espaces de nommage présente également des lacunes. En effet, une

définition d'espace de nommage ne fait aucune référence au modèle utilisé mais utilise une adresse URI abstraite. Finalement, certains concepts sont superflus et peuvent être redéfinis : les instructions de traitement, la déclaration XML, la présence d'un seul nœud racine.

Les langages de modélisation DTD et XML Schema présentent également un manque d'intégration important. Ainsi la syntaxe des DTD n'est pas XML et XML Schema possède une sémantique complexe. Leurs systèmes de typage respectifs sont soit insuffisants soit trop contraignants. De plus, le contenu mixte des éléments ne peut être clairement défini. Finalement, le manque d'intégration se retrouve dans le fait que ces deux langages ne définissent pas de modèles pour se valider eux-mêmes.

Le langage de transformation XSLT possède également une sémantique très compliquée ainsi que des lacunes dans sa spécification. Nous ne développons pas ci-dessous les remarques concernant ce langage mais nous nous concentrons essentiellement sur le langage XML puis évoquons les critiques contre les langages de modélisation DTD et XML Schema.

2.1. Le langage XML

2.1.1 La gestion des caractères blancs

La critique principale que nous formulons à l'encontre de XML concerne la définition et la gestion des caractères blancs. Sous " caractères blancs ", nous regroupons les quatre valeurs suivantes : l'espace simple (#x20, en valeur hexadécimale), la tabulation (#x9), le retour de chariot (#xD) et le saut de ligne (#xA).

XML est un langage de balisage pour l'annotation de textes et par conséquent les espaces blancs sont considérés comme du texte lorsqu'ils entourent les balises XML. Parallèlement, un modèle XML Schema ou DTD peut interdire la présence de nœuds textes à l'intérieur de certains éléments. Mais pour favoriser la lecture, si de tels éléments contiennent des sous-éléments, ceux-ci seront séparés par des espaces blancs considérés comme non pertinents (indentation). La spécification XML est incomplète sur ce point. Ceci permet aux développeurs d'implémenter leur propre sémantique et mène à des résultats incompatibles suivant les outils utilisés.

L'exemple suivant illustre ce problème de traitement des espaces blancs. Cet exemple considère une feuille de style XSLT qui transforme un document XML en un autre. La tâche est simple : la transformation doit rendre un document XML strictement identique à l'input. Le formatage doit donc être conservé. Etonnamment, ceci n'est pas aussi trivial qu'il n'y paraît. Pour simplifier le problème, nous décidons que le formatage doit être maintenu simplement pour le nœud racine et son contenu et non pour le prologue du document.

Nous observons tout d'abord que les différents processeurs testés rendent des résultats différents pour un même input et une même transformation. Pour l'exemple ci-dessous, nous utilisons le processeur Xalan-J (version 2.0.7, disponible à http://www.xalan.org) que nous considérons comme fiable. La figure 1 donne le code d'un document XML d'input et la figure 2 le code de la feuille de style qui, intuitivement, devrait permettre de recopier à l'identique un document XML.

```
01  <?xml version="1.0" encoding="UTF-8"?>
02  <document>
03  <element lang="fr">
04  salut
05  </element>
06  <element lang="en">
07   hello
08   </element>
09   <element lang="de">
10   hallo
11   </element>
12  </document>
```

FIGURE 1. Document XML d'input.

```
01 <?xml version="1.0" encoding="UTF-8"?>
02 <xsl:stylesheet
    xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
    version="1.0">
03  <xsl:output method="xml"/>
04
05  <xsl:template match="*|@*|text()">
06  <xsl:copy>
07  <xsl:apply-templates select="*|@*|text()"/>
08  </xsl:copy>
09  </xsl:template>
10
11 </xsl:stylesheet>
```

FIGURE 2. Feuille de style XSLT attendue.

La feuille de style XSLT de la figure 2 est indentée afin de garantir une bonne lisibilité pour l'utilisateur. Toutefois, cette mise en page ne permet pas de retrouver le formatage original du document XML. Le retour à la ligne entre le prologue et l'élément racine (figure 1, lignes 01-02) est ignoré et le résultat obtenu est illustré à la figure 3.

```
01  <?xml version="1.0" encoding="UTF-8"?><document>
02   <element lang="fr">
03     salut
04   </element>
05   <element lang="en">
06     hello
07   </element>
08   <element lang="de">
09     hallo
10   </element>
11  </document>
```

FIGURE 3. Résultat de la première transformation.

Pour éviter ce problème, nous ajoutons un template dédié au traitement du nœud racine dans la feuille de style XSLT. Pour obtenir une mise en page identique, le formatage de la feuille de style pour ce nouveau template doit être contrôlé et ne respecte plus l'indentation traditionnelle. L'attribut xml:space doit également être ajouté (Fig. 4). Le résultat de cette transformation modifiée donne exactement le document XML d'input.

```
01 <?xml version="1.0" encoding="UTF-8"?>
02 <xsl:stylesheet
     xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
     version="1.0">
03  <xsl:output method="xml"/>
04
05  <xsl:template match="/" xml:space="preserve">
06  <xsl:copy><xsl:apply-templates
     select="*|@*|text()"/></xsl:copy>
07   </xsl:template>
08
09   <xsl:template match="*|@*|text()">
10   <xsl:copy>
11   <xsl:apply-templates select="*|@*|text()"/>
12    </xsl:copy>
13   </xsl:template>
14
15  </xsl:stylesheet>
```

FIGURE 4. Feuille de style XSLT modifiée.

Le problème de la copie identique aurait également pu être résolu en ajoutant un attribut indent="yes" dans l'élément xsl:output. Cependant, c'est solution n'est pas satisfaisante puisque (1) il n'est s'agit pas d'une solution suffisante au traitement des espaces blancs, (2) la sémantique de cette construction n'est pas complètement spécifiée et (3) la recommandation XSLT elle-même ne la considère pas sûre pour les documents avec du contenu mixte.

2.1.2 Les différentes entités

La deuxième critique que nous formulons se rapporte à la notion d'entité définie en XML. Cette notion unique regroupe trois concepts différents : les entités générales, les entités caractères et les entités paramètres . L'accès à ces différentes entités se fait par référence. Une première différence majeure entre ces concepts vient du fait que les références vers des entités générales ou caractères se font à partir du document XML tandis que les références d'entités paramètres ne sont valides qu'à l'intérieur d'une DTD.
Si les entités générales et paramètres correspondent pleinement à la définition d'une entité telle que nous nous l'imaginons, les entités caractères ne sont qu'une représentation différente (décimale ou hexadécimale) d'un caractère.

Même si les références d'entités générales ne sont reconnues qu'à l'intérieur d'un document XML et les références d'entités paramètres que dans une DTD, le concept sous-jacent est le même. D'ailleurs, le traitement appliqué est sensiblement identique : le parseur substitue les références d'entités parsées par la valeur qui leur est attribuée dans la déclaration d'entité correspondante. Le processus est répété tant qu'il subsiste des références d'entité dans le texte.

Puisque le concept est le même, il serait judicieux de réunir ces deux types d'entités et d'utiliser une même notation et un même traitement dans les documents et leurs modèles. De plus, il faut un mécanisme simple qui permette de contrôler si une entité doit être substituée ou non. Les règles XML à ce sujet ne sont pas claires et certains comportements étranges peuvent être observés.

Prenons l'exemple suivant. Dans la figure 1, nous trouvons l'entité à. agrave est défini comme une entité générale dont la valeur est " à ". Le symbole & est une entité définie par défaut et sa représentation est &. Si nous remplaçons & par sa représentation dans l'entité à, nous obtenons à. En appliquant la même transformation que précédemment (Fig. 2), nous constatons alors que les règles de substitution des entités ne sont pas claires puisqu'une double substitution devrait être effectuée et rendre " à ". La dernière version de Xalan-J n'effectue aucune substitution. Avec les processeurs intégrés dans Internet Explorer 6 et Firefox 1.5, seul & est substitué ce qui rend le résultat à.

10 L'attention est portée ici exclusivement aux entités parsées.

2.1.3 Espaces de nommage

Une faiblesse du langage XML a trait à l'intégration des espaces de nommage (namespaces). Une première remarque touche la spécification des espaces de nommage [NS99] qui a été publiée après la première version de la recommandation XML. Dans cette dernière, le caractère ':' était autorisé pour les noms d'éléments. Avec la spécification des espaces de nommage, ce même caractère représente la séparation entre le préfixe et le nom d'élément. Afin d'assurer la compatibilité arrière, aucune restriction supplémentaire n'a été introduite dans la définition des noms d'éléments. Ainsi, un nom d'élément tel que un:element:simple doit être autorisé par un parseur conformant.

La seconde remarque est d'ordre syntaxique. En effet, la syntaxe est identique pour définir un espace de nommage puis pour l'utiliser. D'une part, XML offre la possibilité d'insérer dans n'importe quelle balise ouvrante une déclaration d'espace de nommage. Par exemple,
<test:element xmlns:test="uri"/>.

D'autre part, si un espace de nommage test est déclaré, il peut alors être utilisé pour préfixer les noms d'éléments et les attributs qui appartiennent à cet espace.
<test:element test:attribut="valeur"/>.

Ces deux exemples, bien qu'ils définissent des concepts différents - déclaration d'un espace de nommage et utilisation d'un espace de nommage - ont des syntaxes très proches qui empêchent de noter la spécificité de chaque élément. Dans le second, l'attribut attribut est véritablement issu de l'espace de nommage test, tandis que dans le premier, test n'est pas un attribut issu d'un espace de nommage hypothétique nommé xmlns, mais bien l'assignation du préfixe test à un espace représenté par la valeur d'attribut uri.
L'intégration des espaces de nommage dans les langages de modélisation est également problématique. DTD ne les considère simplement pas. XML Schema a été développé pour combler ce manque mais la solution proposée introduit de la redondance. En effet, dans un modèle, il est nécessaire de déclarer un attribut targetNamespace pour référencer l'espace de nommage associé. Puis, dans l'instance, il faut localiser le modèle et déclarer à nouveau l'espace de nommage associé.

2.1.5 Instructions de traitement

De notre point de vue, la syntaxe particulière des instructions de traitement (<? ... ?>) n'a pas de raison d'être spécifique. En effet, les instructions de traitement, telles que définies dans la syntaxe EBNF de XML [XML04] peuvent apparaître à la racine du document, comme contenu de n'importe quel élément ou encore après la balise fermante du nœud racine. Il serait alors tout à fait envisageable de simplifier cette syntaxe et de faire des

instructions de traitement des éléments simples dans un espace de nommage prédéfini.

En effet, un espace de nommage peut être facilement assigné à une application donnée. Dès lors, les éléments à l'intérieur de cet espace peuvent être référencés dans le document XML qui y fait appel. Le modèle du document en question pourra ainsi restreindre les portions du document dans lesquelles une instruction de traitement est autorisée. De plus, une confusion existe avec la déclaration XML qui possède une syntaxe semblable mais ne peut être considérée comme une véritable instruction de traitement.

2.1.6 Un seul nœud racine

XML n'autorise la présence que d'un seul nœud racine. Les fondements de cette décision sont pragmatiques et ne semblent reposer sur aucun principe clairement défini. Il serait plus judicieux d'autoriser la même structure pour le contenu d'un document que pour le contenu d'un élément, c'est-à-dire une séquence de nœuds.

2.2. Les langages de modélisation DTD et XML Schema

Il existe plusieurs langages de modélisation liés à XML dont deux font l'objet d'une normalisation par le W3C. Les DTD (Document Type Definition) [XML04] sont issus directement de SGML et ont toujours appartenu du langage XML. Ainsi, leur spécification a toujours fait partie intégrante des différentes recommandations du langage XML. Le second langage, XML Schema, est apparu en 1999 pour répondre à certaines faiblesses des DTD. Sa recommandation est publiée en 2001 par le W3C. La deuxième édition de celle-ci paraît en octobre 2004 [FAL04, THO04, BIR04].

2.2.1 Quelques critiques

La première critique s'adresse exclusivement aux DTD et a déjà été traitée lors de la conception de XML Schema. Il s'agit de la syntaxe non-XML utilisée par les DTD. Celle-ci empêche d'utiliser des outils communs pour les documents XML et leurs modèles.

Les systèmes de typage des deux langages présentent également des lacunes. D'un côté, les DTD ne permettent pas le typage ni des nœuds textes, ni des attributs. Un seul type est connu, le type #PCDATA, ce qui est largement insuffisant. A l'opposé, un nombre très important de types est prédéfini dans le langage XML Schema. Ceci génère un système de typage compliqué. De plus, type simple et type complexe ne définissent pas exactement la même notion. Un type simple se réfère au contenu d'un élément, tandis qu'un type complexe se réfère également au contenu d'un élément mais dans le sens de sa structure.

Une autre critique concerne la définition du contenu mixte. Les DTD ne permettent pas de définir un contenu mixte d'une meilleure façon qu'en ajoutant simplement le type #PCDATA parmi un choix d'élément (... | #PCDATA)+. La définition obtenue même si elle n'est pas fausse manque cruellement de précision. De plus, nous constatons que l'intégration est mauvaise puisqu'il n'est pas possible de définir un modèle DTD ou XML Schema pour valider chaque langage respectivement. Il est raisonnable d'attendre d'un tel langage qu'il soit capable de valider un modèle par un " méta-modèle ".

Finalement, la sémantique de ces langages, en particulier de XML Schema, est très complexe et comporte des lacunes. Ceci rend bien évidemment leur utilisation compliquée. Ainsi il est difficile de concevoir un modèle sans l'aide d'outils graphiques.

3. LE LANGAGE YML

3.1. Remarques préliminaires

Le langage YML pourrait aisément adopter une syntaxe XML. Cependant, pour bien distinguer les deux langages et pour des raisons pédagogiques, une nouvelle syntaxe a été adoptée. Le choix des nouvelles représentations de balises a été déterminé par la volonté d'utiliser des balises symétriques. Ainsi, une balise ouvrante se note <= ... > et la balise fermante correspondante < ... =>. Les commentaires également sont encadrés par des balises symétriques - <* et *> - de même que les entités - { et }. Les entités YML sont présentes dans les nœuds textes et les valeurs d'attribut. L'exemple ci-dessous représente un extrait de code YML (Fig. 5).

```
01 <* un commentaire *>
02 <=element attribut="{entité}">
03   <* un element vide *>
04   <=element_vide=>
05 <element=>
```

FIGURE 5. Illustration de la syntaxe YML.

3.2. Nœuds textes

Comme nous le soulignons dans la section 2.2, XSLT a introduit un artefact pour gérer les caractères blancs et ne considérer que les espaces significatifs. Prenons l'exemple suivant : dans une page HTML, nous souhaitons introduire des liens vers des ancres qui se présentent sous la forme #nom_de_l'ancre. Supposons que le document XML à transformer ne contient pas de caractères

blancs superflus. L'extrait de code ci-dessous devrait permettre d'effectuer cette opération (Fig. 6).

```
01 <a>
02  <xsl:attribute name="href">
03    #<xsl:value-of select="."/>
04  </xsl:attribute>
05 </a>
```

FIGURE 6. Créer une ancre HTML en XSLT.

Mais, interprété par un navigateur (Internet Explorer 6 ou Firefox 1.5), ce code rend une valeur d'attribut href qui débute par un retour de ligne. Le lien est donc inutilisable. Pour résoudre ce problème, il faut introduire une balise <xsl:text/>. Ainsi en remplaçant # par <xsl:text>#</xsl:text>, l'interpréteur ne va considérer que les caractères blancs entre les balises.

L'idée des nœuds textes YML est basée sur le même principe. Chaque nœud texte est délimité par une balise ouvrante [et une balise fermante]. Les caractères blancs (#x20 - espace simple - et #x9 - tabulation) entre ces deux balises sont significatifs et sont considérés comme appartenant formellement au contenu textuel de l'élément. Les caractères blancs à l'extérieur de ces balises sont simplement là pour la mise en page du texte : indentation des éléments et des nœuds texte, par exemple. Les retours à la ligne significatifs sont représentés par le symbole # situé juste après la balise fermante. L'exemple ci-après illustre ces nœuds textes (Fig. 7). En appliquant ces balises directement dans le fichier YML, document source d'une application, cette dernière ne doit plus avoir recours à des artefacts pour gérer les problèmes de caractères blancs.

Les nœuds textes YML présentent certaines similarités avec les sections CDATA de XML. Les deux constructions utilisent des délimiteurs clairs mais les nœuds textes YML ont la spécificité d'interpréter les références d'entités qui y apparaissent, au contraire des CDATA. A ce titre, les CDATA sont plus limités que les nœuds textes et par conséquent ne sont pas supportés en YML.

```
01 <=element>
02     [Cette phrase se trouve sur ]
03     [une seule ligne.]#
04     [Celle-ci sur une autre.]
05 <element=>
```

FIGURE 7. Illustration du nœud texte YML.

3.3. Entités

Les entités générales et les entités paramètres sont réunies sous un seul concept, celui d'entité. Une entité est définie dans un modèle, elle peut être ensuite référencée dans une valeur d'attribut ou dans un nœud texte indifféremment à partir d'un document YML ou à partir de son modèle. Les entités caractères de XML sont vues en YML comme une représentation différente d'un caractère (décimale ou hexadécimale). Seules les entités parsées sont traitées dans YML.

3.4. Intégration des espaces de nommage

Dans le langage YML, les espaces de nommage sont étroitement liés aux modèles. Un espace de nommage est défini avec son modèle au lieu d'une URI (Fig. 8). Comme en XML, un préfixe est associé à chacun. L'espace de nommage qui est déclaré sans préfixe devient l'espace de nommage par défaut.

L'intégration des espaces de nommage en YML répond à la problématique de différentiation entre la définition d'un espace de nommage et son utilisation. YML distingue ces deux mécanismes syntaxiquement. Toutes les déclarations d'espaces de nommage se font dans des éléments simples au début du document (Fig. 8). Les éléments et les attributs définis dans un espace de nommage sont utilisés avec des noms qualifiés comme dans le langage XML. Il n'existe ainsi plus de confusion entre la déclaration d'un espace et son utilisation.

```
01  <=yml:dml prefix="test" uri="./test.dml"=>
```

FIGURE 8. Déclaration d'un espace de nommage en YML.

Le préfixe " yml " est un préfixe réservé pour un espace de nommage qui définit certaines constructions basiques du langage YML : déclaration YML et déclaration d'espace de nommage.

YML : une version épurée de XML pour faciliter une spécification rigoureuse des modèles et des transformations.

3.5. Instructions de traitement et déclaration initiale

Dans la section 2.4.1, nous notons que la syntaxe particulière des instructions de traitement n'avait pas une utilité primordiale. Dans le langage YML, les instructions de traitement sont représentées sous la forme d'éléments simples. Un espace de nommage défini dans le document YML fera référence à l'application à utiliser.

Dans le langage XML, bien que la déclaration XML ne soit pas une instruction de traitement, elle possède une syntaxe identique. Là encore, il ne semble pas que cette syntaxe particulière apporte une information supplémentaire. En YML, la syntaxe de la déclaration YML est une syntaxe d'élément simple (Fig. 9). L'élément decl, comme l'élément dml qui permettait de définir un espace de nommage, appartient à l'espace de nommage réservé et prédéfini yml.

```
01  <=yml:decl version="1.0" encoding="UTF-8"=>
```

Figure 9. Déclaration YML.

3.6. Composition d'un document YML

Un document YML est donc composé de deux parties principales. La première partie contient les déclarations préliminaires comme exposées à la section précédente : déclaration YML et définitions d'espaces de nommage ; la second partie contient les éléments, les nœuds textes mais aussi les instructions de traitement qui composent le document YML. Les commentaires peuvent apparaître à n'importe quel niveau du document.

Une spécificité de YML, par rapport à XML, est la possibilité de définir plusieurs nœuds racines. Puisque les déclarations préliminaires sont considérées comme des éléments simples vides, elles figurent au même niveau que le ou les nœuds racines du document en tant que tel. Il n'existe ainsi plus de différence entre le contenu d'un élément et le contenu d'un document. La figure 10 illustre un document YML simple.

4. UN LANGAGE DE MODÉLISATION POUR YML

Les différentes critiques des langages de modélisation DTD et XML Schema (section 2.2) mènent à l'élaboration d'un langage de modélisation pour YML. Le langage DML en répondant aux faiblesses constatées est proche du langage RelaxNG [CLA01].

```
01 <=yml:decl version="1.0" encoding="UTF-8"=>
02 <=yml:dml uri="./modele-1.dml"=>
03 <=yml:dml prefix="mod" uri="./modele-2.dml"=>
04 <* début du document *>
05 <=document mod:titre="doc-1"=>
06   <=elt id="1"> [du texte !]# <elt=>
07   <=elt id="2">
08     [encore du ]
09     [texte !]#
10   <elt=>
11 <document=>
12 <=document mod:titre="doc-vide"=>
```

Figure 10. Document YML simple.

4.1. DML : DOCUMENT MODELING LANGUAGE

4.1.1 Une syntaxe YML et des déclarations simples

DML est une application YML dans le sens que sa syntaxe est YML. Un modèle DML peut être composée de quelques déclarations simples. Les séquences d'éléments sont exprimées par une déclaration de structures (<=struct=>) ; les déclarations de choix (<=choice=>) permettent de choisir un élément parmi leurs fils ; les déclarations d'éléments (<=elt=>) définissent la valeur des éléments et les déclarations d'attributs (<=attrs=>, <=attr=>) - qui sont des fils directs des déclarations d'éléments - définissent les attributs de ces éléments; finalement, une déclaration spécifique (<=text=>) permet de définir l'emplacement des nœuds textes dans l'instance. Ceci permet d'exprimer des contraintes sur les nœuds dont le contenu est mixte.

4.1.2 Le contenu mixte

Grâce à l'introduction d'une déclaration spécifique pour les nœuds textes (<=text=>), le contenu mixte est beaucoup mieux défini. C'est la seule solution qui permet de limiter le contenu mixte à une portion d'une séquence d'éléments, contrairement à la solution apportée par XML Schema qui définissait mieux le contenu mixte que les DTD mais ne permettaient pas de limiter ce contenu à une partie spécifique d'une structure.

4.1.3 Le système de typage

Le système de typage défini en DML est moins limité que celui des DTD mais également moins vague que celui de XML Schema. Un seul type de base est prédéfini, le type string. Ce type est défini dans l'espace de nommage yml

introduit plus haut. Les types simples de XML Schema sont remplacés par des définitions de types (<=type=>). Celles-ci sont composées d'un nom et d'un pattern. Le pattern est une expression régulière. Il n'y a pas de limites quant au nombre de types définis dans un modèle. Les définitions de types peuvent également être référencées dans d'autres modèles à partir du modèle courant comme expliqué plus bas. Quant aux types complexes de XML Schema, ils sont remplacés par les définitions de structures (cf. section 4.3.1).

4.1.4 La modularité

Chaque modèle est composé de trois parties distinctes. La première partie regroupe les déclarations préliminaires : comme dans les documents YML, celles-ci comportent la déclaration YML puis la définition d'un espace de nommage qui référence le méta-modèle (ymldml.dml). En plus, des importations permettent d'appeler d'autres modèles DML. La seconde partie, à l'intérieur de l'élément <=root=>, contient les déclarations qui forment le modèle en tant que tel. Finalement la troisième partie regroupe les définitions d'éléments, de structures ou d'attributs qui peuvent être référencées par les déclarations correspondantes.

Les déclarations d'éléments, de structures et d'attributs peuvent référencer une définition qui appartient au même modèle DML ou qui est importé d'un autre modèle, référencé par un préfixe. Ce mécanisme de référencement permet une grande modularité, puisque, un modèle DML peut être divisé dans plusieurs documents DML. Chaque document définit un groupe de constructions qui peuvent être réutilisées.

4.1.5 Un exemple simple

L'exemple simple ci-dessous (Fig. 11) illustre le modèle DML pour le document de la figure 10. Les lignes 01 à 04 regroupent les déclarations préliminaires. La racine du document (l. 5) contient deux références : l'une est fixe (l. 7), elle décrit la structure des déclarations préliminaires d'un document YML. L'autre (l. 8) fait appel à la définition d'élément de la ligne 12. A l'intérieur de cette définition, la ligne 13 fait appel à une définition d'attribut situé dans le modèle référencé par le préfixe mod. Puis la déclaration de structure (l. 14) décrit le contenu d'un élément document. Finalement la déclaration de type (l. 23) permet de décrire la valeur des attributs id (l. 17).

4.1.6 Le méta-modèle DML

Afin de répondre au critère d'intégration déjà formulé, un but lors du développement du langage DML était de pouvoir fournir un modèle au langage DML. Ce méta-modèle a pu être réalisé. Il est ainsi possible de

```
01 <*déclarations préliminaires*>
02 <=yml:decl version="1.0" encoding="UTF-8"=>
03 <=yml:dml uri="./ymldml.dml"=>
03 <=yml:import prefix="mod" uri="./modele-2.dml"=>
04 <* racine du document *>
05 <=root>
06   <=struct>
07     <=structref ref="yml:prolog"=>
08     <=eltref ref="document" occurs="many"=>
09   <struct=>
10 <root=>
11 <* definitions *>
12 <=elt name="document">
13   <=attrsref ref="mod:titre"=>
14   <=struct>
15     <=elt name="elt">
16       <=attrs>
17         <=attr name="id" type="number"=>
18       <attrs=>
19       <=text=>
20     <elt=>
21   <struct=>
22 <elt=>
23 <=type name="number" pattern="0 | [1-9][0-9]*"=>
```

FIGURE 11. Un modèle DML pour l'instance YML de la figure 10.

valider un document DML par rapport au méta-modèle avec le même outil qui permet de valider un document YML par rapport à son modèle. Le méta-modèle lui-même a pu être validé.

4.2. Comparaison avec RelaxNG

RelaxNG est le langage issu d'une réunion de TREX (Tree Regular Expressions for XML) et de RELAX (REgular LAnguage description for XML). Sa spécification est publiée en juin 2001 [CLA01]. RelaxNG propose une simplification du langage de modélisation XML Schema en se concentrant exclusivement sur l'aspect de validation. Les aspects d'interprétation (entités, valeurs d'attributs, etc.) n'existent pas en RelaxNG.

Un document RelaxNG est un document XML qui définit la structure et le contenu d'une classe de documents XML. Son rôle est de définir cette structure mais pas de fournir une interprétation des instances. Ainsi aucune entité n'est définie et les attributs n'ont pas de valeur par défaut. L'interprétation des caractères blancs n'est pas traitée non plus et finalement

aucun mécanisme ne permet de lier un modèle RelaxNG aux documents XML de la classe correspondante. Ceci ne favorise pas à notre sens l'intégration des langages. Le langage DML intègre les deux aspects de validation et d'interprétation comme le font les DTD et XML Schema.

De plus, le système de typage de RelaxNG reste compliqué puisqu'il est basé sur les types définis dans XML Schema [BIR04]. Enfin, RelaxNG est également capable de se définir lui-même puisqu'il existe un RelaxNG qui valide le langage RelaxNG, à l'image du méta-modèle DML.

5. RÉALISATIONS ET PERSPECTIVES

Un premier parseur qui teste la conformité des documents YML a été réalisé [SCH04] ainsi qu'un second parseur [PUG05] qui permet de vérifier la validité d'un document YML par rapport à son modèle DML. Il a également été possible de valider le méta-modèle avec ce second parseur. Ces deux outils ont été développés relativement aisément ce qui confirme la simplicité des concepts de YML.

Un pont vers le monde XML est également assuré grâce à un outil de conversion des documents YML vers des documents XML et inversement. Les deux langages ne sont pas entièrement compatibles (espace de nommage, nœuds textes, entre autres) et certaines adaptations doivent être faites. Malgré tout, ce pont est maintenu dans le but de pouvoir utiliser les nombreux outils qui existent dans le monde XML.

Les perspectives futures de ce projet sont de compléter les langages de balisage et de modélisation par un langage de transformation, déjà nommé DGL (Document Generation Language), qui reposerait également sur des concepts simplifiés. Un but poursuivi est que le trio formé de YML, de DML et de DGL puissent parfaitement s'intégrer l'un dans l'autre : ainsi le langage DGL possèdera également son propre modèle, à l'image du méta-modèle DML. De plus, un document DGL pourra être testé par rapport à un modèle DML pour vérifier en amont la validité du document YML qui sera produit. A terme, l'objectif est de définir de manière formelle la sémantique des langages DML et DGL.

6. CONCLUSION

La première partie de cet article a détaillé les différentes faiblesses que présente le langage XML à l'heure actuelle. La critique principale concerne une gestion inappropriée des caractères blancs qui entraîne des traitements non uniformes de la part des différentes applications XML, en particulier les processeurs XSLT. D'autres remarques ont aussi été faites à propos de la gestion des entités, du manque d'intégration des espaces de nommage, etc. Le langage YML qui a été introduit répond à ces critiques en proposant des concepts plus simples et plus clairs. En particulier, l'introduction de véritables

nœuds textes permet d'éviter des confusions entre les caractères blancs destinés à la mise en page du document et les caractères blancs qui font partie intégrante du contenu textuel d'un élément. Une nouvelle syntaxe est proposée pour bien différencier ce nouveau langage.

Quelques critiques ont également été faites contre les langages de modélisation les plus courants, DTD et XML Schema. Le langage de modélisation DML qui complète le langage YML répond à ces remarques et grâce, là encore, à l'introduction des nœuds textes, permet une meilleure gestion du contenu mixte qui posait problème avec les DTD et XML Schema. L'intégration du DML et du YML est améliorée grâce au méta-modèle DML et à une approche nouvelle de la définition des espaces de nommage.

Des outils pour traiter les documents YML et DML ont été implémentés. La facilité d'implémentation des parseurs conformant et validant montre que des concepts plus clairs et plus simples permettent effectivement de réduire les coûts de développement.

Les langages de la famille YML ont été conçus comme langages de recherche exclusivement. Le travail est maintenant axé sur le langage de transformation DGL qui suit les critères de simplicité et de clarté posés pour YML. Le but est de fournir des modèles DML pour les langages DML et DGL et de développer un typage efficace des transformations DGL. L'objectif principal qui est une intégration optimale des trois langages pourra ainsi être visé.

7. RÉFÉRENCES

[BIR04] Biron P.V., Malhotra A. (2004). XML Schema Part 2 : Datatypes. Disponible à http://www.w3.org/TR/xmlschema-2/

[CLA01] Clark J., Murata M. (2001). RelaxNG Specification.
Disponible à http://www.relaxng.org/spec-20011203.html

[FAL04] Fallside D.C., Walmsley P. (2004). XML Schema Part 0 : Primer. Disponible à http://www.w3.org/TR/xmlschema-0/

[HAR02] Harold E.R., Means W.S. (2002). XML in a Nutshell (Second Edition). O'Reilly & Associates.

[NS99] Bray T., Hollander D., Layman A. (1999). Namespaces in XML. Disponible à http://www.w3.org/TR/REC-xml-names/

[PUG05] Pugin C. (2005). YML+ : un parseur validant pour les langages YML et DML. Travail de Master en informatique, Département d'informatique, Université de Fribourg.

[SCH04] Schönbächler J. (2004). YML/DML : une révision de XML. Travail de Master en informatique, Département d'informatique, Université de Fribourg.

[THO04] Thompson H.S., Beech D., Maloney M. Mendelsohn N. (2004). XML Schema Part 1 : Structures. Disponible à http://www.w3.org/TR/xmlschema-1/

[XML04] Bray T., Paoli J. Sperberg-McQueen C.M., Maler E., Yergeau F. (2004). Extensible Markup Language (XML) 1.0 (Third Edition). Disponible à http://www.w3.org/TR/REC-xml

[XML04a] Bray T., Paoli J. Sperberg-McQueen C.M., Maler E., Yergeau F., Cowan J. (2004). Extensible Markup Language (XML) 1.1. Disponible à http://www.w3.org/TR/xml11

Session 02

Analyse et interprétation des documents

DOCUMENT INQUISITOR : un système de validation des structures et d'élicitation de modèles de documents

Florian EVÉQUOZ
Maurizio RIGAMONTI
Denis LALANNE
Rolf INGOLD

DIUF. Département d'Informatique de l'Université de Fribourg
Bd de Pérolles 90 CH-1700 Fribourg, Suisse
{prenom.nom}@unifr.ch

ABSTRACT

This paper introduces document inquisitor, a tool for the validation and correction of the results of document analysis systems. It presents the architecture of the system focusing on its internal data representation and exposes interactive features of the tool which are divided into two categories: (1) validation of document structures and (2) document model elicitation using explicit links and entities.

KEYWORDS : *document analysis, document structure, validation, ground truthing, document model.*

1. INTRODUCTION

Le groupe DIVA de l'Université de Fribourg travaille depuis plusieurs années sur le développement de techniques automatiques ou supervisées visant à reconstruire les structures physique et logique de documents. Des outils ont été développés dans le but de retrouver en plusieurs étapes, ces structures sous-jacentes et sont décrits notamment dans [Rig03], [Rig05b], [Had05] et [Blo06]. Néanmoins, le processus n'est pas trivial et les systèmes utilisés aux différents niveaux d'analyse ne parviennent pas toujours à des résultats optimaux. C'est pourquoi le besoin d'un outil léger permettant de valider et d'éditer facilement les résultats de différents niveaux d'analyse de documents s'est imposé et que l'application document inquisitor a été développée. En outre, certains systèmes d'analyse basés sur des règles, par exemple pour la détection de la structure logique d'un journal, requièrent l'appel à des modèles décrivant les caractéristiques fonctionnelles de la classe de document. Cependant, la définition de tels modèles n'est pas aisée et nécessite la prise en compte du jugement de l'utilisateur. Document inquisitor utilise un paradigme basé sur des liens explicites pour représenter les modèles de documents, permettant ainsi de les éliciter ou de les faire apparaître, par

la manipulation et l'expérimentation.

Un état de l'art récent des outils de validation de structure a été réalisé dans [Yac05]. Celui-ci donne un aperçu général de plusieurs systèmes, dont notamment xmillum [Rig03] qui a directement inspiré document inquisitor. PerfectDoc, système présenté dans l'article, est globalement le plus abouti de ces outils. Document inquisitor se distingue de ce dernier sur plusieurs points. Tout d'abord, l'utilisation d'un langage interne de représentation permet à document inquisitor de supporter différents formats de données issues de l'analyse de documents, moyennant un mécanisme de transformations bidirectionnelles. Ensuite, la manipulation de la structure physique au moyen de document inquisitor ne comporte pas d'outils de haut niveau dédiés à la correction d'OCR, car il a été utilisé exclusivement avec des systèmes d'analyse de documents électroniques, tels que XED, qui n'utilise pas d'OCR et assure des résultats de reconnaissances de mots supérieurs à 99% dans le meilleur des cas [Rig05b], et ce sur divers types de documents (journaux, articles scientifiques, présentations, etc.). Enfin, document inquisitor assure une plus grande souplesse de représentation de la structure logique que PerfectDoc qui ne semble proposer que deux niveaux de structure logique, représentés de manière ambiguë par des codes de couleur sur les blocs physiques ne donnant pas d'informations quant à leur organisation hiérarchique. Ces différences peuvent être attribuées à une orientation différente des systèmes : alors que PerfectDoc a été optimisé pour une tâche très spécifique, la correction de l'extraction d'archives du magazine " Times ", document inquisitor vise la généralité.

Cet article décrit l'architecture du système, particulièrement du point de vue du langage interne de représentation des données, et ses fonctionnalités de validation et de création de modèles. Après une présentation de l'héritage que document inquisitor doit à xmillum dans la section 2, la section 3 explore l'architecture du système, en mettant l'accent sur le langage interne de représentation des données de document inquisitor et ses avantages ainsi que sur le mécanisme de transformations bidirectionnelles. La section 4 décrit ensuite les fonctionnalités de l'application, tant au niveau de la correction et de la validation de la structure physique qu'à celui de la création de structures logiques et de l'élicitation de modèles de documents. La section 5 conclut l'article.

2. XMILLUM ET DOCUMENT INQUISITOR

Document inquisitor est le successeur de xmillum, dont il révise les concepts et l'architecture. Les deux systèmes proposent la même idée de base, c'est-à-dire de permettre aux chercheurs d'afficher, de valider et d'éditer n'importe quel type de résultats d'analyse de documents exprimés dans un format XML quelconque. Par contre, cette philosophie est adoptée d'une manière différente : xmillum privilégiait une architecture ouverte, permettant de configurer complètement l'affichage et le traitement des données à l'aide

du langage xmi (XMillum Internal), tandis que document inquisitor définit le format IML (Inquisitor Modeling Language, voir 3.2) pour une configuration limitée de l'éditeur. Ce choix a l'apparence d'une régression mais est largement justifié par la difficulté de configurer xmillum de la part des utilisateurs, qui étaient obligés d'écrire des transformations XSLT définissant en même temps la configuration du système et le traitement des données en entrée. La figure 1 compare l'architecture des deux systèmes et notamment : (1) la transformation des données en entrée dans le format interne (flèches épaisses) ; (2) leur manipulation par l'utilisateur à l'aide de l'éditeur (flèches fines à deux pointes) ; et (3) leur traitement après les tâches de validation et d'édition (flèches pointillées).

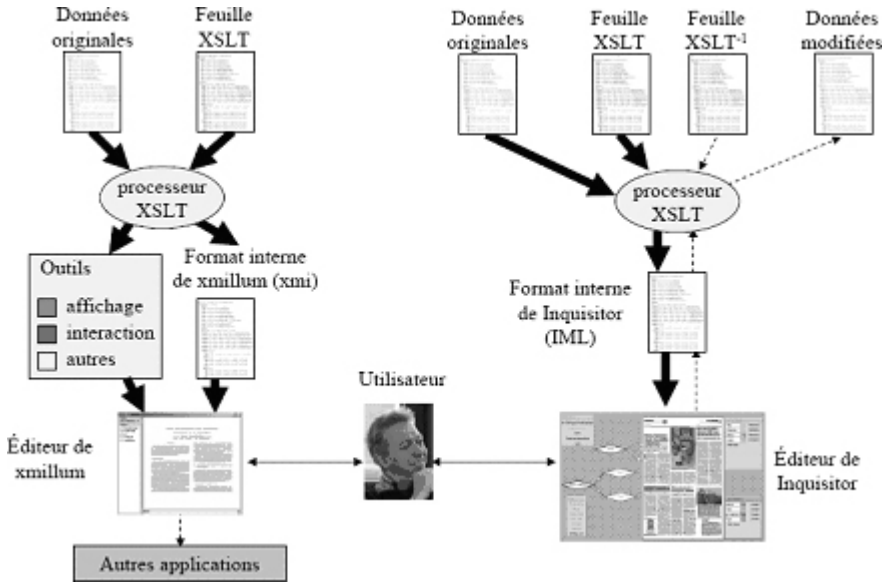

Figure 1 L'architecture de xmillum (gauche) et celle de document inquisitor (droite)

La figure 1 met en évidence un autre aspect qui différencie considérablement les deux systèmes : le cycle de vie des données après avoir été manipulées par l'utilisateur. Dans xmillum, les données n'étaient pas retransformées dans leur format d'origine et cette tâche était affectée soit à d'autres systèmes d'analyse de documents, soit à des extensions développées par les utilisateurs et qui étaient spécifiques au document XML en entrée. À l'opposé, document inquisitor propose un mécanisme de reconversion du document manipulé dans son format d'origine et résout un problème de transformation bidirectionnelle qui sera présenté à la section 3.4.

3. ARCHITECTURE

Cette section introduit dans un premier temps les notions de structures et de modèles de document. Elle présente ensuite le langage de représentation interne de document inquisitor, IML, et décrit la transformation appliquée pour le générer, puis détaille les mécanismes mis en œuvre pour garantir la bidirectionnalité des transformations.

3.1 Structures physique et logique, modèle de document

Selon notre acception, la structure physique d'un document décrit la mise en page. Elle consiste en une hiérarchisation de ses primitives en entités homogènes du point de vue morphologique, topologique, et typographique. Les aspects morphologiques permettent de différencier entre elles des entités de type image, tableau, graphique ou contenu textuel. Une distinction morphologique est donc littéralement une distinction de forme, ou de " type de donnée ". Les aspects topologiques sont quant à eux issus de la disposition des éléments sur la page. Les aspects typographiques ont trait aux polices de caractères et aux différents choix de mise en forme.

La structure physique se prête à une représentation unique et non ambiguë pour tout document contenant du texte, des images et des composants graphiques. Récemment [Blo06] a proposé un ébauche de format canonique pour représenter cette structure, définissant un langage XML propre nommé XCDF (eXhaustive Canonical Document Format), qui a été appliqué à la représentation de documents PDF analysés et restructurés. Plusieurs niveaux hiérarchiques y sont distingués, dont les principaux sont présentés sur la figure 2. La structure physique canonique divise ainsi le document en pages, lesquelles comprennent des images, des graphiques et des blocs de texte. Ces derniers contiennent eux-mêmes des lignes de texte, qui rassemblent à leur tour des entités de type token, représentant les unités lexicales. La tâche d'un outil d'extraction automatique de la structure physique est de produire un fichier conforme au format canonique. Pourtant un certain nombre d'erreurs peuvent apparaître au cours du processus, qu'un outil tel que document inquisitor permet de corriger.

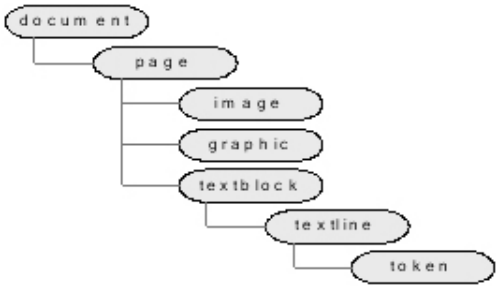

Figure 2 Schéma hiérarchique simplifié des éléments formant la structure physique canonique d'un document statique en XCDF

La structure logique d'un document peut être définie comme une hiérarchisation du contenu du document selon un modèle propre à la classe de documents à laquelle il appartient. Alors que la structure physique détermine des entités physiques cohérentes, la structure logique caractérise le rôle de celles-ci et décrit leurs relations, en particulier leur organisation hiérarchique [Sum95]. La reconstruction de la structure logique d'un document à partir de sa forme physique canonique comporte deux niveaux distincts : le premier consiste en un étiquetage des blocs physiques, le second en une projection des blocs étiquetés sur un modèle de document.

Le premier niveau, l'étiquetage, permet de caractériser le rôle logique de chaque entité physique. Chaque bloc de texte reçoit une étiquette, choisie parmi un ensemble déterminé, propre à une certaine classe de documents. Dans le cas d'un journal par exemple, les blocs de texte seront étiquetés comme titre, chapeau, corps d'article, (les caractères gras différencient dans la suite de l'article les entités physiques des entités logiques), etc. Le second niveau de reconstruction logique consiste à regrouper les entités liées en éléments logiques cohérents du point de vue de la classe de document considérée. Dans l'exemple du journal, le titre de l'article et le corps d'article associé constituent une nouvelle entité logique d'ordre supérieur : l'article. De même, le titre de rubrique, l'article qui vient d'être défini et les autres articles de la rubrique forment la rubrique.

Enfin, si la structure physique de tout document peut être représentée dans un format canonique unique, à l'inverse la structure logique reflète des informations propres à une catégorie de documents aux propriétés similaires, nommée classe de documents. A titre d'exemple, nous pouvons affirmer de manière quelque peu simplificatrice que les journaux forment une classe de documents. Cette classe propose une division en sujets, rubriques, articles, titres, paragraphes, etc., hiérarchisés selon des règles précises définies dans un modèle de journal. En toute généralité, la structure logique d'un document sera donc conforme à un modèle propre à la classe de documents à laquelle il appartient.

3.2 Représentation des données : le format IML

La structure de données interne de document inquisitor a été définie de manière à être exprimable dans un langage XML propre appelé IML (Inquisitor Modeling Language). La tâche première de document inquisitor étant de représenter des données visualisables issues de l'analyse de documents, la spécification du langage IML a été conçue de manière à faciliter l'affichage de celles-ci. Ses primitives les plus importantes sont donc les suivantes :

1) boîte (<box>)
2) couche (<layer>)
3) lien (<link>)
4) entités (<entity>)

La boîte est l'élément de base de représentation de la structure physique. Elle correspond à un simple rectangle positionné dans le document, doté de coordonnées (x,y) et de dimensions (w,h). Elle peut également posséder d'autres propriétés optionnelles, comme un contenu textuel par exemple. Les couches sont des " conteneurs " de boîtes, de liens, d'entités, ou d'autres couches. Elles peuvent être affichées ou masquées à la demande. Les liens permettent de joindre entre elles des boîtes et/ou des entités. Les entités, enfin, représentent des éléments porteurs de propriétés, mais pour lesquels il n'est pas pertinent de spécifier une position physique dans le document. Les éléments appartenant à la structure logique du document seront typiquement spécifiés au moyen d'entités.

Par exemple, pour décrire les structures physique et logique d'un journal dans le langage IML, les différentes primitives sont organisées comme suit. Les entités physiques (unités lexicales, lignes, blocs de texte, images dans la terminologie XCDF) sont représentées par des boîtes. Celles-ci sont regroupées en couches. Toutes les images partagent la même couche, tous les blocs de textes également, et ainsi de suite ; dans une structure hiérarchique analogue à celle du format canonique, la couche des mots est contenue dans celle des lignes, et cette dernière appartient à son tour à celle des blocs (cf. figure 3). Il serait possible de faire de même avec les résultats de la reconnaissance symbolique d'une carte : des boîtes engloberaient les segments, les symboles, les noms de lieux, et seraient regroupées dans différentes couches selon leur type ou leur niveau d'abstraction.

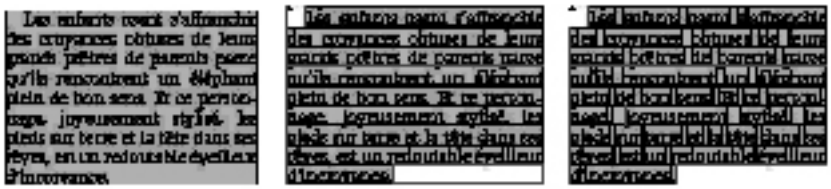

Figure 3 Affichage respectif des couches correspondant aux blocs de texte, aux lignes, et aux mots. La couche parente est visible en filigrane sur la deuxième et la troisième image.

La représentation de la structure logique utilise une couche supplémentaire réservée. Les entités logiques, qui peuvent être par exemple de type article, titre, date, etc. dans le cas d'un journal, possèdent des liens vers les blocs physiques correspondants, ou vers d'autres entités (cf. figure 4). Typiquement, une entité logique représentant un corps d'article aura des liens vers un certain nombre de blocs physiques englobant son contenu. Par contre, celle représentant un sujet pointera vers des articles thématiquement proches, qui sont eux-mêmes des entités logiques. Finalement, il faut encore signaler dans la structure IML interne l'utilité de deux listes globales : la liste des catégories d'étiquettes logiques et la liste des classes d'entités logiques disponibles.En effet, comme la structure logique d'un document s'exprime

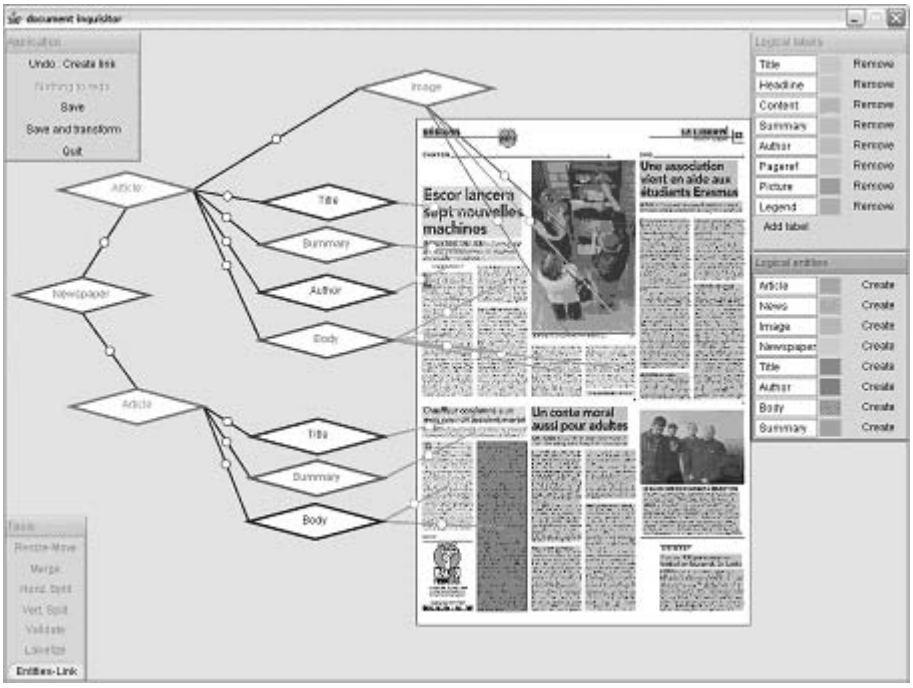

Figure 4 Aperçu de la structure logique partielle d'un document, avec entités et liens

en deux niveaux, i.e. l'étiquetage des blocs physiques et leur regroupement en entités logiques de plus haut niveau, il faut disposer d'étiquettes et de classes d'entités propres à la classe de document considérée, d'où l'utilité de ces deux listes. Par ailleurs, ces listes sont directement dépendantes du modèle de document courant. Elles apparaissent dans l'application sous la forme de deux panneaux distincts représentés sur la figure 5. La section 4.3 traite plus particulièrement de la création de modèles de document au moyen de document inquisitor.

3.3 Intérêt du format IML

Document inquisitor aspire à manipuler des résultats d'analyse quelconques. Son format interne doit donc être générique et aussi simple que possible pour faciliter la transformation depuis n'importe quel format de description de document. Pour permettre d'éditer facilement différents formats de données, il est en effet confortable de les transformer dans un langage générique. Xmillum a ouvert cette voie, mais souffre de la trop grande complexité du langage xmi, qui requiert l'écriture de feuilles de style absconses et qui n'est pas prévu pour subir des transformations inverses. IML résout en ce sens les défauts de xmi.

Figure 5 Les listes d'entités et d'étiquettes logiques du modèle de document courant.

Le premier avantage du langage IML est sa simplicité : il ne contient que les données à manipuler par l'application et non, comme c'est le cas de xmi, des constructions servant à représenter dans le langage-même les outils et le comportement de l'application. Deuxièmement, le langage IML est un langage dédié à la visualisation et à la manipulation de certaines structures : blocs, entités et liens. Il ne représente effectivement que les données utiles à cette fin et évite l'hétérogénéité de représentation pouvant apparaître dans les formats descriptifs originaux. Troisièmement, IML prévoit des constructions servant à isoler les données non éditées sans que le langage ait besoin d'être étendu. Ces constructions, présentées à la section 3.4, permettent de garantir la bidirectionnalité des transformations. Quatrièmement, le paradigme des entités et des liens, présenté en 4.3, permet de rendre explicites les structures logiques d'un document. Enfin, le langage IML assure à document inquisitor sa polyvalence : un format quelconque de représentation de données peut être utilisé en entrée et transformé en IML pour être directement interprété par l'application sans que l'utilisateur ait besoin de modifier l'implémentation du programme. L'écriture de l'adaptateur est déléguée à une feuille de style XSLT. Cette généricité du langage IML a été appréciée durant le développement de document inquisitor. Grâce à de simples adaptations au niveau des feuilles de style, il n'a pas été nécessaire de répercuter les changements du langage XCDF, développé en parallèle, ni dans les spécifications du langage IML ni dans l'implémentation de document inquisitor.

3.4 Transformations bidirectionnelles : contraintes et exemple

La généralité donnée par la transformation d'un langage XML dans un autre, comme IML pour document inquisitor, est contrebalancée par la difficulté de retrouver un document conforme au format original à partir du document

transformé. Cette réversibilité est pourtant nécessaire. En effet, si document inquisitor est utilisé pour valider ou corriger les résultats d'une extraction automatique de la structure physique, le fichier corrigé doit être disponible dans son format original pour un usage futur par exemple comme entrée d'un système de reconstruction de la structure logique ou en vue d'une intégration dans un système d'analyse tel que DocMining [Cla04]. Pour permettre une telle réversibilité, il faut néanmoins s'assurer que les processus de transformation dans un sens et dans l'autre n'induisent aucune perte d'informations. Deux contraintes d'ordre syntaxique sont à respecter pour garantir la bidirectionnalité des transformations :

1) Représentation, même implicite, de tous les éléments et attributs du langage original.
2) Représentation, même implicite, de la structure hiérarchique originale.

Pour résoudre la première contrainte, le langage IML met à disposition deux constructions ignore et original permettant respectivement d'ignorer des sous-arbres entiers du format original et de conserver le nom et les attributs d'éléments qui ne sont pas transformés dans un des quatre éléments principaux du langage IML décrits à la section 3.2. Pour que la deuxième contrainte de bidirectionnalité soit respectée, les éléments du format original doivent être pourvus d'un attribut id unique. Ainsi, lors de la transformation dans le langage IML, les éléments IML pointent vers leur parent dans la structure originale par le biais d'un attribut parent.

Pour illustrer le fonctionnement d'une transformation en IML et les exigences posées par les contraintes syntaxiques, nous allons prendre comme exemple la transformation d'un document XCDF en IML. Celle-ci procède selon les principes généraux suivants :

1) Chaque type d'élément XCDF amené à être représenté par des boîtes en IML (à savoir les éléments textblock, textline, token et image) se voit attribuer une couche propre. Les couches sont imbriquées de manière à représenter la structure hiérarchique.

2) Les éléments cités du format canonique sont transformés en boîtes. A chaque transformation, l'élément original est encapsulé au moyen de la construction original.

3) Les sous-arbres non traités sont isolés dans des nœuds ignore.

4) L'attribut parent garde la trace de la structure hiérarchique originale

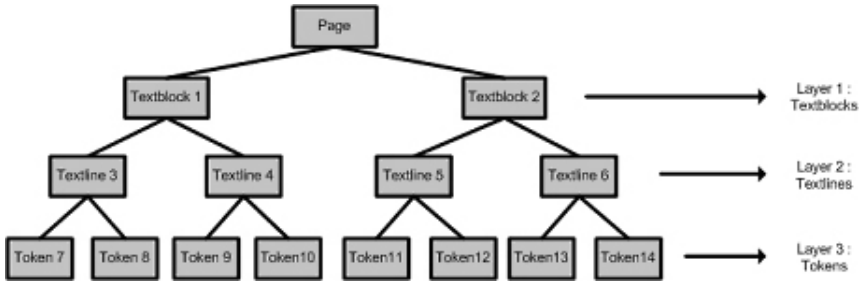

Figure 6 Un exemple de structure physique simple représentée dans le format canonique XCDF

La figure 6 montre un exemple de structure au format XCDF. Le résultat de la transformation de cette structure en IML est présenté sur la figure 7, qui indique par des traits pointillés la valeur de l'attribut parent. La structure canonique originale de la figure 6 apparaît visuellement par le biais des lignes pointillées.

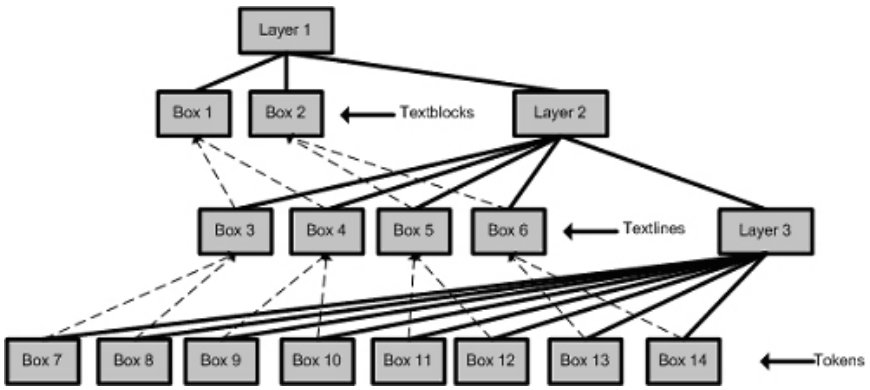

Figure 7 Structure interne correspondant à la structure au format canonique représentée à la figure 6. Les traits pointillés marquent la valeur de l'attribut parent du langage IML.

Nous avons vu que les unités lexicales sont représentées sous forme de boîtes dans le langage IML. Les seuls attributs requis pour une boîte définissent son placement physique (coordonnées et dimensions).

Nous avons également indiqué qu'une boîte pouvait posséder un contenu textuel optionnel. Les autres attributs d'une unité lexicale ainsi que le nom de l'élément sont, eux, encapsulés dans un nœud de type original. Ainsi, l'élément suivant :

<token x=… y=… w=… h=… content=… fontsize="…"/>

est représenté dans le langage IML de la sorte :

<box x=… y=… w=… h=… content=…>
<original>
<element name="token"/>
<attributes fontsize="…"/>
</original>
</box>

Cette construction isole dans l'élément boîte lui-même les attributs directement requis par l'interface. Le noeud original garde quant à lui toutes les informations relatives à l'élément original, qui a été transformé en cette boîte, pour permettre de le reconstituer lors de la transformation inverse.

Enfin, les documents au format canonique XCDF contiennent des éléments tels que la définition des polices de caractères qui ne sont pas visualisés par document inquisitor. Néanmoins, pour respecter la première contrainte de bidirectionnalité, il est nécessaire qu'ils apparaissent dans la structure interne IML. A cet effet, les sous-arbres XML concernés (les nœuds font et leurs enfants dans le format canonique) sont encapsulés dans des nœuds de type ignore et représentés à l'identique dans le langage IML, de la sorte :

<ignore>

<glyph …>
<name …/>
<code …/>
</glyph>

</ignore>

4. FONCTIONNALITÉS COUVERTES PAR DOCUMENT INQUISITOR

Document inquisitor est un outil de validation et de correction des résultats d'une extraction de structure, un outil dédié à la création de fonds de vérité et un outil permettant de découvrir et de créer des modèles de document. Cette section examine les mécanismes mis en œuvre pour remplir ces objectifs.

Elle présente tout d'abord les opérations applicables au niveau de la structure physique du document et les contraintes que celles-ci doivent respecter pour garantir le respect de l'intégrité du document. Enfin, elle décrit les opérations relevant de la structure logique du document et permettant la mise en lumière de modèles de document.

4.1 Opérations sur la structure physique

La première fonctionnalité couverte par document inquisitor est une tâche de validation et de correction des résultats d'analyse de documents. Différents types d'erreurs peuvent en effet survenir lors d'une extraction automatique de structures de documents électroniques, en particulier des erreurs de sur-segmentation et de sous-segmentation. Pour corriger manuellement ces erreurs, document inquisitor met à disposition un certain nombre d'opérations sur les boîtes, qui se divisent en opérations simples (ne touchant qu'une boîte par opération) ou complexes (touchant plus d'une boîte). Cette section décrit brièvement les différentes opérations possibles.

Les opérations simples sur la structure physique n'ont à faire qu'à un seul élément, et ne modifient pas la structure arborescente des données, mais seulement les propriétés ou attributs existants. Elles sont au nombre de quatre. (1) La première opération simple fournie par document inquisitor est le redimensionnement et le déplacement d'une boîte. Du point de vue de la structure physique du document, cela revient à changer les coordonnées et/ou les dimensions d'un bloc mal reconnu. (2) L'opération de validation permet d'attribuer une valeur de validité à un bloc, qui peut être utilisée pour entraîner un système dans le cas d'un apprentissage supervisé. L'opération de validation peut être considérée comme un étiquetage de boîte pour lequel les seules étiquettes possibles sont " valide ", " non valide " ou " indéterminé " tant que la validation n'a pas été effectuée. (3) L'édition manuelle des propriétés d'une boîte permet de modifier les propriétés optionnelles des boîtes, telles que le contenu textuel ou d'autres attributs directement hérités des éléments originaux dont elles sont issues. Cette opération permet d'éditer la valeur d'une propriété existante, mais pas d'ajouter de nouvelles propriétés ni d'en supprimer. (4) Finalement, la dernière opération simple est l'étiquetage des blocs physique. Il s'agit d'une opération d'annotation logique sur la structure physique d'un document. Elle sert à attribuer une étiquette à un bloc, spécifiant sa fonction logique. Les étiquettes peuvent être définies dans un modèle de document chargé au préalable (comme titre ou contenu peuvent être prédéfinies dans le modèle de document journal), ou peuvent avoir été définies manuellement par l'utilisateur.

La figure 8 montre une opération d'étiquetage, une opération de redimensionnement et une opération de validation. Le redimensionnement s'effectue par manipulation directe de l'enveloppe d'un bloc, alors que les opérations de validation et d'étiquetage sont effectuées par le biais de menus contextuels circulaires

Figure 8 Opérations d'étiquetage logique (gauche), de redimensionnement (centre) et de validation (droite)

Les opérations complexes manipulent plusieurs éléments et modifient l'arborescence IML interne. (1) La fusion consiste à regrouper deux blocs en un seul. Elle est nécessaire pour corriger les problèmes de sur-segmentation. (2) La division consiste à partager un bloc en deux, horizontalement ou verticalement. Elle sert à corriger les cas de sous-segmentation.

4.2 Contraintes sur les opérations

Les opérations décrites dans la sous-section précédente ne doivent pas engendrer d'états dans lesquels la structure physique du document ne serait plus cohérente. Pour garantir l'intégrité de la structure physique certaines contraintes particulières doivent être respectées dans le langage IML :

1) Contrainte d'inclusion : un bloc fils est géométriquement contenu dans son bloc parent.

2) Contrainte de respect de l'ordre de lecture : l'ordre des blocs fils d'un bloc donné respecte l'ordre de lecture.

La première contrainte est toujours respectée. Ainsi, l'interface ne permet pas de redimensionner ou déplacer un bloc de telle manière que cette contrainte soit bafouée. Similairement, la division n'est possible qu'aux endroits compatibles avec le respect de cette contrainte.

Le respect de la deuxième contrainte en toute généralité est soumis à une analyse dédiée à la découverte de l'ordre de lecture qui n'a pas sa place dans un outil de validation tel que document inquisitor. Aussi, pour résoudre les conflits apparaissant lors d'opérations complexes de fusion ou de division de blocs, document inquisitor demande à l'utilisateur de spécifier explicitement l'ordre de lecture des blocs physiques concernés. Il faut enfin noter que les contraintes citées ici ne concernent que la structure physique du document, et que document inquisitor n'impose aucune contrainte sur la structure logique dont le fonctionnement est présenté dans la section suivante.

4.3 Le paradigme du lien, les entités et les structures logiques

Une des particularités de document inquisitor est la possibilité de créer des structures logiques en utilisant des liens explicites et des entités logiques. Ces dernières groupent différentes composantes du document dans un même ensemble logique, défini par une étiquette. La relation d'appartenance à un même ensemble logique est mise en évidence par les liens explicites, à leur tour de deux types : (1) les liens entre entités physiques et entités logiques et (2) les liens entre différentes entités logiques. Par exemple, dans le cas d'un journal, des liens de la première sorte regroupent les régions textuelles d'une page de journal en articles et des liens de la deuxième sorte permettent par exemple de définir une table des matières pointant vers les titres des articles définis auparavant. Ces deux exemples sont explicités par la figure 9.

Figure 9 Exemple d'une entité article, avec des liens vers la structure physique du document (haut), et d'une entité table des matières, dont les liens pointent vers deux entités titre (bas).

Document inquisitor permet de créer de nouveaux liens et entités, ainsi que de définir une classe d'entités qui n'existe pas encore, pouvant être édités ou effacés dans un deuxième temps.

L'intérêt majeur de cette fonctionnalité réside dans les tâches qui peuvent être accomplies :

- validation de structures logiques : les résultats de systèmes d'analyses logiques supervisés ou automatiques peuvent être validés ou corrigés à l'aide de document inquisitor, en éditant les étiquettes des entités et des blocs logiques et en modifiant leurs liens. Par exemple, l'utilisateur pourrait soit changer l'étiquette d'un titre en chapeau, soit affecter un bloc de texte à un article en déplaçant le lien qui pointait précédemment vers un autre article ;

- création de données de fonds de vérités : la rapidité de création de liens et de nouvelles instances d'entités permet à l'utilisateur de créer manuellement des archives de données hiérarchisées et étiquetées à utiliser comme fonds de vérités pour des systèmes d'apprentissage. Par exemple, l'ensemble d'entités attribuées par un système d'analyse à un type de document particulier pourrait être réutilisé en chaîne sur d'autres documents similaires afin d'incrémenter la consistance de la base de donnée ;

- création de nouveaux modèles de document : la possibilité de définir ses propres entités, de les intégrer dans une structure d'arbre et enfin de les lier à des blocs physiques implique que l'utilisateur puisse créer ses propres modèles de documents, à utiliser ensuite avec des systèmes de reconnaissance de structures logiques. Un utilisateur peut donc utiliser document inquisitor comme un langage graphique pour décrire la structure d'une page de journal en définissant une hiérarchie d'articles, composée de titres, auteurs, etc. Les processus de création et de regroupement d'entités permettent ainsi, à partir de la structure physique d'un document, de découvrir progressivement, ou d'éliciter, un modèle de structure logique pouvant par la suite s'appliquer à tous les documents de sa classe et pouvant également être partagé avec la communauté de recherche ou avec des utilisateurs du même métier.

5. CONCLUSION

Document inquisitor est un système permettant de valider et d'éditer les résultats d'analyses de documents. Il a été employé avec succès dans la correction d'une cinquantaine de documents au format XCDF extraits de journaux et de présentations PowerPoint, destinés à être utilisés dans des navigateurs multimédia [Rig05]. Cet article a présenté le système et ses spécificités, dont en particulier la capacité d'accepter en entrée divers formats XML transformés en un format interne nommé IML, de permettre à l'utilisateur d'éditer le document dans ce format et de restituer en sortie un document modifié conforme au format original. L'éditeur de document inquisitor permet de manipuler des résultats d'analyse et de créer de nouvelles données de fonds de vérité, un système de contraintes garantissant l'intégrité de la structure interne. Document inquisitor utilise par ailleurs des

primitives simples pour la visualisation des données, en superposant des couches et des boîtes à l'image du document. En outre, la création de liens et d'entités au moyen de l'interface facilite l'élicitation et la définition de nouveaux modèles de documents.

Actuellement, document inquisitor est composé d'un processeur XSL pour la transformation dans le format IML et d'un éditeur spécialisé pour des documents contenant une majorité de régions textuelles. Les prochaines extensions prévues touchent à l'intégration d'outils d'édition dédiés aux graphiques et aux images, à l'élaboration des contraintes d'intégrité adaptées, et à la mise à jour du format IML pour le rendre plus flexible quant à la gestion des nœuds ignore et pour lui associer un espace de nommage adapté. Enfin, l'utilisation actuelle du système dans le cadre de la préparation de données destinées à des navigateurs multimédia suggère une évaluation de l'ergonomie de l'éditeur, de l'efficacité des contraintes dans la détection d'éventuelles d'erreurs et de la validité des transformations bidirectionnelles appliquées à de nouveaux formats de représentation.

6. BIBLIOGRAPHIE

[Blo06] J.-L. Bloechle, M. Rigamonti, K. Hadjar, D. Lalanne, R. Ingold, XCDF: A Canonical and Structured Document Format. In Horst Bunke, A. Lawrence Spitz (eds.), LNCS: "7th International Workshop, DAS 2006, Nelson, New Zealand, February 13-15, 2006, Proceedings", Springer-Verlag, vol. 3872, pp. 141-152.

[Cla04] E. Clavier, G. Masini, M. Delalandre, M. Rigamonti, K. Tombre, J. Gardes, DocMining: A Cooperative Platform for Heterogeneous Document Interpretation According to User-Defined Scenarios. In Josep Lladós, Young-Bin Kwon (eds.), LNCS: "Graphics Recognition, Recent Advances and Perspectives, 5th International Workshop, GREC 2003, Barcelona, Spain, July 2003, Revised Selected Papers", Springer-Verlag Berlin Heidelberg, vol. 3088, ISBN:3-540-22478-5, Barcelona (Spain), 2004 , pp. 13-24.

[Had05] K. Hadjar, R. Ingold, Logical Labeling of Arabic Newspapers using Artificial Neural Nets. In proc. of 8th International Conference on Document Analysis and Recognition (ICDAR'05), Seoul (Korea), August 09 - September 01 2005 , pp. 426-430.

[Rig03] M. Rigamonti, O. Hitz, R. Ingold, A Framework for Cooperative and Interactive Analysis of Technical Documents. In proc. of 5th IAPR International Workshop on Graphics Recognition (GREC'03), Barcelona (Spain), July 30 - 31 2003 , pp. 407-414.

[Rig05] M. Rigamonti, D. Lalanne, F. Evéquoz, R. Ingold, Browsing Multimedia Archives Through Intra- and Multimodal Cross-Documents Links.

In Steve Renals, Samy Bengio (eds.), LNCS: "Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers", Springer-Verlag, vol. 3869, pp. 114-125.

[Rig05b] M. Rigamonti, J.-L. Bloechle, K. Hadjar, D. Lalanne, R. Ingold, Towards a Canonical and Structured Representation of PDF Documents through Reverse Engineering. In proc. of 8th International Conference on Document Analysis and Recognition (ICDAR'05), Seoul (Korea), August 29 - September 01 2005 , pp. 1050-1054.

[Sum95] K. Summers, Towards a Taxonomy of Logical Document Structures, In Electronic Publishing and the Information Superhighway: proc. of the Dartmouth Institute for Advanced Graduate Studies (DAGS), 1995, pp 124-133.

[Yac05] S. Yacoub, V. Saxena, S. Nusrulla Sami, PerfectDoc: A Ground Truthing Environment for Complex Documents. In proc. of 8th International Conference on Document Analysis and Recognition (ICDAR'05), Seoul (Korea), August 29 - September 01 2005, pp. 452-457.

Découvrir les thèmes d'un document pour en améliorer la segmentation thématique

Olivier FERRET

CEA-LIST/LIC2M
18, route du Panorama - BP6, 92265 Fontenay-aux-Roses Cedex
ferreto@zoe.cea.fr

RÉSUMÉ

La segmentation thématique et l'identification des thèmes d'un document sont souvent traitées comme des problèmes séparés, même si elles relèvent toutes deux de l'analyse thématique. Dans cet article, nous proposons d'examiner comment l'identification thématique peut contribuer à améliorer la segmentation de documents lorsque celle-ci ne s'appuie que sur la récurrence lexicale. Nous présentons d'abord une méthode non supervisée de découverte des thèmes d'un document ; puis nous détaillons comment ces thèmes sont utilisés dans la segmentation pour aider à reconnaître les similarités thématiques entre des segments de documents. Nous montrons enfin, au travers d'une évaluation faite à la fois pour le français et pour l'anglais, l'intérêt effectif de la méthode proposée.

MOTS-CLES : *analyse thématique de documents, segmentation thématique.*

17. INTRODUCTION

Le problème auquel nous nous intéressons dans cet article est celui de la segmentation thématique, problème consistant à découper linéairement un document en une suite de segments thématiquement homogènes. Cette partie de l'Analyse du Discours a fait l'objet de nombreux travaux ces dernières années à la suite notamment de celui de Hearst [HEA94]. Sur le plan applicatif, elle intervient dans différentes tâches d'accès à l'information thématique dont la plus notable est le résumé automatique, comme l'illustrent des travaux tels que [BAR97], [BOG00] ou [CHA04]. L'intérêt porté à la segmentation thématique s'est manifesté en particulier par sa présence dans le cadre des évaluations Topic Detection and Tracking (TDT), consacrées plus généralement à différentes dimensions de l'analyse thématique à la fois de documents écrits et de transcriptions de parole.

Un des critères permettant d'appréhender les travaux dans le domaine de la segmentation thématique est le type de connaissances sur lesquelles ils reposent. La plupart d'entre eux s'appuient sur les seules caractéristiques intrinsèques des documents : la récurrence lexicale dans le cas de [HEA94],

[CHO00], [UTI01] ou [GAL03] ; la réitération des multi-termes et des entités nommées dans le cas de [KAN98] ; la présence de marques discursives dans [PAS97] ou [GAL03]. Ces méthodes ne faisant pas appel à des connaissances externes, elles ne sont pas limitées à un champ thématique particulier. En revanche, leur application est restreinte à un certain type de documents : la récurrence lexicale n'est en effet un indice thématique fiable que si les concepts du document considéré ne sont pas exprimés sous des formes trop diverses (synonymes, hyperonymes, etc.) ; pour leur part, les marques discursives sont souvent rares et spécifiques d'un corpus.

Une des pistes suivies par certains systèmes pour surmonter ces limitations est d'exploiter des connaissances sur les relations de cohésion lexicale, connaissances qui présentent elles aussi l'avantage de ne pas dépendre d'un domaine particulier. Elles prennent la forme d'un réseau lexical construit à partir d'un dictionnaire dans [KOZ93], d'un thésaurus dans [MOR91] ou encore d'un large ensemble de cooccurrences lexicales dans [FER98], [KAU99] ou [CHO01]. D'une certaine façon, ces connaissances permettent aux systèmes de segmentation thématique de détecter les récurrences à un niveau plus conceptuel. Cependant, leur nature lexicale et leur absence de structuration thématique explicite font que les systèmes les utilisant sont parfois mis en échec par l'ambiguïté sémantique des mots ou par l'impossibilité d'identifier des relations comme spécifiquement thématiques.

La solution la plus simple concernant ce dernier point est la possibilité d'exploiter des connaissances sur les thèmes susceptibles d'être rencontrés dans les documents analysés. C'est en particulier l'approche retenue par les participants aux évaluations TDT qui construisent une représentation des thèmes rencontrés à partir des documents exemples fournis. Cette approche est typiquement représentée par le travail décrit dans [YAM98] et se retrouve dans une partie de [BEE99] ou de [TÜR01]. Le travail de Bigi [BIG98] s'inscrit dans la même perspective mais en se focalisant sur des thèmes plus larges que ceux considérés dans TDT tandis que [FER00] opère avec des représentations des thèmes construites de façon non supervisée. Plus généralement, les connaissances thématiques apprises par ces systèmes leur permettent une plus grande précision mais restreignent parallèlement leur champ d'action à des documents portant sur une certaine thématique.

Enfin, des systèmes hybrides combinant différentes approches parmi celles présentées ci-dessus ont également été développés et ont prouvé leur intérêt : [JOB98] associe ainsi la récurrence lexicale, l'utilisation de conccurrences et celle d'un thésaurus ; [BEE99] et [TÜR01] s'appuient à la fois sur une modélisation statistiques des thèmes et sur l'utilisation de marques discursives ; [GAL03] exploite conjointement la récurrence lexicale et des marques discursives.

Le travail que nous présentons dans cet article s'inscrit dans la première catégorie de systèmes que nous avons distinguée : il ne s'appuie pas sur des

connaissances a priori et se fonde sur l'usage des mots plutôt que sur le repérage de marques discursives. Plus précisément, nous ne proposons pas une méthode de segmentation radicalement nouvelle mais nous montrons comment une méthode fondée sur la récurrence lexicale peut être améliorée en identifiant les thèmes d'un document sans faire appel à des connaissances de référence.

18. PRINCIPES

Les algorithmes de segmentation s'inscrivant dans la filiation plus ou moins directe de l'algorithme TextTiling de Hearst prennent comme point de départ une représentation des documents sous la forme d'une séquence d'unités de discours. Dans le cas de documents écrits, il s'agit généralement de phrases, approche que nous avons également adoptée dans notre travail. Chaque unité est transformée en un vecteur de mots suivant les principes du modèle Vector Space [SAL83]. La similarité entre deux unités peut ainsi être évaluée en faisant appel à une mesure de similarité vectorielle. Dans ce contexte, une telle mesure est considérée comme représentative de la proximité thématique des unités impliquées. Compte tenu des caractéristiques du modèle Vector Space, ce principe s'étend à des regroupements d'unités, comme des segments de document. La détection des changements de thème s'identifie alors à la détection des zones dans lesquelles la similarité entre unités ou entre regroupements d'unités est faible.

Cette vue d'ensemble souligne le rôle central de l'évaluation de la similarité entre les unités de discours dans ce type de méthode. Lorsque aucune connaissance externe n'est utilisée, cette similarité ne repose que sur la réitération lexicale. Mais il est possible d'y intégrer la prise en compte de relations sémantiques entre les mots. C'est ce qui est implicitement réalisé dans le système CWM [CHO 01], une variante de l'algorithme C99 dans laquelle chaque mot est remplacé par sa représentation dans un espace issu de l'Analyse Sémantique Latente. Le même type de démarche se retrouve dans [PON 97] et [CAI 04] où un mot est représenté par sa proximité par rapport à un ensemble de concepts de référence construits automatiquement à partir d'un corpus, en utilisant dans le premier cas la méthode Local Context Analysis et dans le second cas, l'algorithme de classification X means.

Dans cet article, nous proposons d'améliorer la détection de la similarité thématique entre des segments de texte sans exploiter de connaissances externes. Dans un premier temps, nous identifions les thèmes de chaque document à segmenter en procédant à une classification non-supervisée de son vocabulaire. Cette classification s'appuie sur les cooccurrences enregistrées au sein du document entre les mots de ce vocabulaire. À la suite de cette phase, chaque thème est représenté par un sous-ensemble du vocabulaire du document. Lors de la segmentation, l'évaluation de la similarité entre deux segments repose d'abord sur la proportion des mots communs à ces deux segments mais elle prend aussi en compte la proportion de leurs

mots appartenant à un même thème parmi ceux identifiés précédemment dans le document. Ainsi, deux segments peuvent être jugés proches parce qu'évoquant un même thème sans pour autant partager un nombre important de mots. Plus globalement, cette prise en compte des thèmes des documents vise à faire diminuer le taux de détection de " faux " changements de thème.

19. DÉCOUVRIR LES THÈMES D'UN DOCUMENT

Pour découvrir les thèmes d'un document sans utiliser de connaissances a priori, nous faisons l'hypothèse que les mots les plus représentatifs de chaque thème apparaissent dans des contextes similaires. Étant donné cette hypothèse, nous collectons les cooccurrents de chaque mot du document possédant une fréquence d'occurrence minimale, nous évaluons la similarité deux à deux de ces mots en nous appuyant sur leurs cooccurrents pour finalement construire une représentation des thèmes du document en appliquant à ces mots une méthode de classification non supervisée.

19.1. Évaluer la proximité thématique des mots d'un texte

La découverte des thèmes d'un document commence par le prétraitement linguistique de ce dernier. Ce prétraitement découpe le document en phrases et représente chacune d'elles comme la séquence de ses mots pleins normalisés, c'est-à-dire ses noms (noms communs et noms propres), ses verbes et ses adjectifs. Il est réalisé par l'outil TreeTagger [SCH94], utilisé à la fois pour les documents en français et en anglais. La normalisation des mots est la première façon, réalisée au niveau morphologique, de favoriser la détection de la similarité entre segments de document. Après le filtrage des mots de plus faible fréquence, les cooccurrents des mots restant du document sont collectés suivant les principes décrits dans [CHU90], en enregistrant les cooccurrences au sein d'une fenêtre de taille fixe déplacée sur tout le document prétraité. À la suite de cette étape, chaque mot sélectionné est représenté par le vecteur de ses fréquences de cooccurrence avec les autres mots du document. La similarité deux à deux de tous les mots sélectionnés est alors évaluée pour constituer leur matrice de similarité. Classiquement, nous appliquons la mesure Cosinus entre les vecteurs représentant ces mots pour réaliser cette évaluation.

19.2. De la proximité des mots aux thèmes d'un document

L'étape finale de la découverte des thèmes d'un document est la classification non supervisée de ses mots à partir de la matrice de similarité construite précédemment. Nous nous appuyons pour cette tâche sur une adaptation de l'algorithme Shared Nearest Neighbors (SNN), décrit dans [ERT01]. Cet algorithme s'accorde particulièrement avec nos besoins dans la mesure il détermine automatiquement le nombre de classes, dans notre cas le

nombre de thèmes d'un document, et qu'il détecte les éléments ne s'accordant pas avec les classes qu'il constitue. Ce dernier point est particulièrement important étant donné que tous les mots pleins d'un document ne sont pas spécifiques de ses thèmes.

L'algorithme SNN s'inscrit dans la mouvance des algorithmes ramenant le problème de la classification à celui de la détection de composantes de forte densité dans un graphe de similarité. Dans un tel graphe, chaque nœud représente un élément à classer et une arête relie deux nœuds lorsque la similarité entre les éléments qu'ils représentent est non nulle. Dans notre cas, le graphe de similarité est construit directement à partir de la matrice de similarité des mots du document. Dans son principe général, l'algorithme SNN comporte deux grandes étapes : la première vise à mettre en évidence les éléments les plus représentatifs de leur voisinage en masquant les relations les moins importantes du graphe de similarité. Ces éléments constituent les embryons des futures classes, formées dans un second temps en agrégeant les autres éléments à ceux sélectionnés lors de la première phase. L'algorithme SNN tel que nous l'avons adapté à la découverte des thèmes d'un document se décompose plus précisément comme suit :

1. Éclaircissement du graphe de similarité. Pour chaque mot sélectionné du document, seules les arêtes en direction des k (k = 15 en l'occurrence) plus proches mots sont conservées.

2. Construction du graphe des plus proches voisins partagés. Cette étape consiste à remplacer dans le graphe éclairci la valeur portée par chaque arête par le nombre de voisins directs que les deux mots reliés par l'arête ont en commun.

3. Calcul de la distribution en liens forts des cooccurrents. L'objectif de cette étape est, comme lors de l'étape 1, de procéder à une sorte d'éclaircissement. Il s'agit de repérer les mots du document autour desquels s'organisent un ensemble d'autres mots, i.e. des germes de thème, mais aussi de repérer ceux qui sont visiblement sans connexion véritable avec les autres. Pour ce faire, un seuil minimum est fixé concernant le nombre de voisins partagés par deux mots du document, seuil au-dessus duquel on considère les deux mots comme fortement liés. On caractérise ensuite chaque mot du document par le nombre de liens forts qu'il possède.

4. Détermination des germes de thème et élimination du bruit. Les germes de thème et les mots du document laissés de côté sont déterminés par simple comparaison de leur nombre de liens forts par rapport à un seuil.

5. Construction des thèmes. Cette étape consiste principalement à associer aux germes de thème trouvés à l'étape précédente les mots du document non déjà sélectionnés comme germe de thème ou bruit pour former des classes représentant les thèmes du document. Pour associer un mot à un germe de

thème, la force du lien qui les unit doit être supérieure à un seuil. Si plusieurs germes de rattachement sont possibles, le choix se porte sur celui avec lequel la force du lien est la plus grande. Par ailleurs, cette étape est aussi l'occasion de rassembler plusieurs germes de thème considérés comme trop proches pour former des thèmes distincts : le rattachement des mots fait donc également intervenir les germes de thème.

6. Élimination des thèmes non représentatifs. Lorsque le nombre de mots rassemblés par un thème est trop petit, ce thème est considéré comme non représentatif et il est supprimé. Ses mots sont alors rattachés à l'ensemble des mots non affectés à l'issue de l'étape précédente.

7. Élargissement des thèmes. À l'issue des étapes précédentes, un nombre plus ou moins important de mots du document n'ayant pas été considérés comme du bruit se retrouvent néanmoins sans affectation à un thème. Ce nombre dépend bien entendu de la sévérité du seuil de rattachement à un germe de thème mais l'objectif étant de former des classes homogènes, celle-ci doit être nécessairement assez forte. Néanmoins, il est également intéressant que les thèmes puissent être décrits de la façon la plus complète et la plus précise possible. Les thèmes à ce stade étant caractérisés de façon plus sûre qu'à l'issue de l'étape 4, il est possible de leur rattacher des mots du document dont la force de lien avec leurs constituants est plus faible de façon plus sûre :

Par rapport à l'algorithme SNN décrit dans [ERT01], nous avons ajouté l'étape 6 car nous avons observé que malgré la possibilité de fusionner des classes lors de l'étape 5, certains thèmes restent divisés en plusieurs classes. Dans un nombre significatif de cas, le thème " divisé " se répartit entre une ou plusieurs classes ne regroupant que 3 à 4 mots et une classe de plus large ampleur. Pour regrouper ces classes " minoritaires " avec la classe la plus importante, nous avons choisi de laisser l'algorithme SNN le faire en détruisant ces classes et en remettant leurs éléments dans l'ensemble des cooccurrents non rattachés. La dernière étape de l'algorithme permet alors de rattacher ces cooccurrents à la classe " majoritaire " dans la plupart des cas. De plus, ce mécanisme permet d'obtenir une plus grande stabilité des thèmes formés lorsque les paramètres de l'algorithme sont modifiés.

Concernant les différents seuils de l'algorithme, nous avons opté pour un mode unique de fixation s'adaptant à la distribution des valeurs considérées : chaque seuil est exprimé comme un quantile de ces valeurs. Pour le seuil de détermination des germes de thème et de celui de définition du bruit (cf. étape 4), il s'agit d'un quantile du nombre de liens forts des mots du document. Pour le seuil définissant la notion de lien fort (cf. étape 3), celui de rattachement des mots aux germes (cf. étape 5) et celui de rattachement des mots aux thèmes (cf. étape 7), le quantile est appliqué à la force des liens entre les mots dans le graphe des plus proches voisins partagés.

À l'issue de l'application de l'algorithme SNN, un ensemble de thèmes, éventuellement vide si le niveau global de récurrence lexicale est trop faible, est donc associé au document à segmenter, chacun d'entre eux étant défini par un sous-ensemble du vocabulaire de ce document.

20. UTILISER LES THÈMES DÉCOUVERTS POUR LA SEGMENTATION THÉMATIQUE

Dans cette section, nous commencerons par présenter une méthode de segmentation thématique s'appuyant sur la récurrence lexicale et se situant dans la filiation directe de celle de Hearst. Nous indiquerons ensuite comment utiliser les thèmes d'un document tels que découverts au moyen de la méthode décrite ci-dessus pour améliorer cette méthode de segmentation.

20.1. Principes de la méthode de segmentation thématique

La méthode de segmentation thématique proposée par Hearst dans [HEA94], TextTiling, se décompose en trois grandes parties :

- Le prétraitement linguistique des documents.
- L'évaluation de la cohésion lexicale au sein du document.
- L'identification des changements de thème.

La méthode que nous proposons ici reprend ces trois grandes étapes mais avec des modalités un peu différentes de celles adoptées par Hearst. Le prétraitement linguistique des documents employé pour la segmentation est le même que celui décrit pour la découverte de thèmes. L'évaluation de la cohésion lexicale s'appuie comme dans TextTiling sur l'utilisation d'une fenêtre glissante de taille fixe. Cette fenêtre se déplace sur le texte de phrase en phrase. À chaque station de cette fenêtre, la cohésion lexicale est évaluée en son sein et affectée à la fin de phrase sur laquelle elle est centrée. Cette évaluation est réalisée suivant le principe proposé dans [JOB98] : la cohésion est mesurée par l'application du coefficient de Dice entre les vecteurs représentant les deux moitiés de la fenêtre glissante. Plus précisément, si Fg désigne le vocabulaire de la moitié gauche de la fenêtre et Fd, celui de la moitié droite de cette même fenêtre, la cohésion au sein de celle-ci est donnée par :

(4.1)
$$cohésion = \frac{2 \cdot card(F_g \cap F_d)}{card(F_g) + card(F_d)}$$

Cette mesure a été utilisée plutôt que la mesure Cosinus retenue pour TextTiling car sa définition ensembliste rend son extension plus facile pour intégrer d'autres relations que la simple récurrence lexicale, à l'instar de [JOB98]. La cohésion est ainsi évaluée pour chaque frontière inter-phrastique du document considéré et le résultat global est une courbe de cohésion couvrant l'ensemble du document.

La troisième partie de l'algorithme s'inspire quant à elle de son homologue dans le système LCseg [GAL03]. Elle comprend elle-même trois étapes :

- Le calcul d'un score évaluant la probabilité pour chaque minimum de la courbe de cohésion de correspondre à un changement de thème.

- La suppression des candidats segments de trop petite taille.

- La sélection des bornes de segments thématiques.
Le calcul du score initial d'un minimum commence par la recherche de la paire de maxima g et d qui l'entourent. En notant, CL(x) la valeur de la cohésion lexicale à la position x,
le score d'un minimum m est donné par :

(4.2) $$score(m) = \frac{CL(g) + CL(d) - 2 \cdot CL(m)}{2}$$

Ce score, compris entre 0 et 1, est d'autant plus élevé que la différence entre le minimum considéré et les maxima qui l'entourent est plus importante. Il favorise ainsi en tant possible lieu de changement de thème les minima caractérisés par une chute très nette de la cohésion lexicale.

La suppression des candidats segments trop petits s'effectue quant à elle par une simple comparaison par rapport à un seuil de référence : les minima se trouvant à P phrases au plus du minimum qui les précèdent (P étant égal à 2 dans le cas des expérimentations de la Section 5) sont éliminés en tant que possibles changements de thèmes. Finalement, la sélection des bornes de segments thématiques est réalisée par l'utilisation d'un seuil s'adaptant à la distribution des scores des minima. Un minimum m est ainsi retenu comme borne de segment si :

(4.3) $$score(m) > \mu - \alpha \cdot \sigma$$

où *u* correspond à la moyenne des scores de minima, , à l'écart-type de ces scores et , à un coefficient de modulation.

20.2. Prise en compte des thèmes d'un document

Le cœur de l'algorithme présenté ci-dessus, que nous appellerons F06 dans ce qui suit, est l'évaluation de la cohésion à l'intérieur de la fenêtre glissante, incarnée par l'équation (4.1). Cette évaluation est également son point faible car le terme ne s'appuie que sur la notion de récurrence lexicale. Ainsi, deux mots différents appartenant respectivement à Fg et Fd mais faisant partie du même thème ne peuvent en aucune façon contribuer à identifier une éventuelle similarité thématique entre les deux volets de la fenêtre.

L'algorithme F06T reprend les principes de F06 mais en étendant l'évaluation de la cohésion au sein de la fenêtre glissante pour y ajouter la prise en compte des proximités thématiques entre les mots. Les thèmes de référence sont bien entendu les thèmes du document découverts par la méthode exposée à la Section 3. Dans cette version étendue, l'évaluation de la cohésion au sein de la fenêtre glissante s'articule en trois étapes :

- Le calcul de la cohésion s'appuyant sur la seule récurrence lexicale.
- La détermination du ou des thèmes de la fenêtre.
- Le calcul de la cohésion s'appuyant sur les thèmes de la fenêtre et sa combinaison avec la cohésion fondée sur la récurrence lexicale.

La première étape est strictement identique au calcul de cohésion réalisé dans F06. La deuxième a pour objectif de restreindre les thèmes utilisés lors de la dernière étape aux thèmes véritablement représentatifs du contenu de la fenêtre glissante, c'est-à-dire représentatifs du contexte courant du discours. Ce problème de représentativité est particulièrement sensible dans les zones de changement de thème. Pour détecter l'instabilité thématique qui les caractérisent, il ne faut pas en effet amplifier les manifestations des thèmes environnants. Pour ce faire, un thème est considéré comme représentatif du contenu de la fenêtre seulement s'il s'apparie avec chacun des deux volets de cette fenêtre. En pratique, cet appariement est évalué en appliquant la mesure Cosinus entre le vecteur représentant une moitié de la fenêtre et le vecteur représentant le thème . Le thème est jugé représentatif de la fenêtre lorsque la valeur de cette mesure est supérieure à un seuil fixé a priori (égal à 0,1 dans les expérimentations de la Section 5). Comme la découverte des thèmes se fait de façon non supervisée et sans connaissances externes, il arrive qu'une thématique du document se retrouve répartie sur plusieurs représentations de thème. L'évaluation de la similarité avec la fenêtre est donc réalisée avec tous les thèmes du document à chaque nouvelle station de la fenêtre et tous ceux remplissant la condition mentionnée ci-dessus sont retenus pour l'étape suivante.

11 Dans le vecteur représentant un thème, tous les mots du thème se voient attribuer un poids égal à 1. Dans celui représentant une moitié de la fenêtre, ce poids est égal au nombre d'occurrences du mot dans la fenêtre.

Cette dernière étape consiste en premier lieu à déterminer pour chaque volet de la fenêtre glissante le nombre de ses mots présents dans l'un des thèmes représentatifs du contenu de cette fenêtre. La cohésion de la fenêtre évaluée sur la base des thèmes du document est ensuite obtenue grâce à l'équation (4.4), qui rapporte le niveau de représentation des thèmes de la fenêtre dans ses deux parties au nombre total de mots de la fenêtre.

(4.4) $$cohésion = \frac{card(F_g \cap T_f) + card(F_d \cap T_f)}{card(F_g) + card(F_d)}$$

où Tf désigne l'union des représentations des thèmes associés à la fenêtre glissante. En final, la cohésion globale au sein de cette fenêtre est donnée par l'addition de la cohésion évaluée sur la base de la récurrence lexicale (cf. équation (4.1)) et de la cohésion évaluée à partir des thèmes du document (cf. équation (4.4)). On notera que du fait de cette addition, cette cohésion globale accorde une importance particulière aux mots récurrents faisant partie des thèmes associés à la fenêtre.

21. RÉSULTATS ET ÉVALUATION

21.1. Corpus de travail et d'évaluation

L'objectif principal de l'évaluation que nous avons menée était de vérifier que la prise en compte des thèmes d'un document découverts sans recours à des connaissances externes à ce document peut améliorer de façon effective un algorithme de segmentation thématique initialement fondé sur la seule récurrence lexicale. Comme nous l'avons vu à la Section 3, la découverte des thèmes s'appuie sur la distribution des mots dans le document. Par conséquent, le cadre d'évaluation proposé par Choi [CHO00], qui est maintenant classiquement utilisé pour l'évaluation des segmenteurs thématiques, n'est pas directement applicable ici. Ce cadre propose en effet de construire artificiellement les documents de référence destinés à évaluer les résultats d'un segmenteur thématique en assemblant des extraits de " vrais " documents. Dans le cas précis du travail rapporté dans [CHO00], chaque document d'évaluation est ainsi constitué de 10 extraits de documents issus du corpus Brown. Chaque extrait, dont la taille est comprise entre 3 et 11 phrases, provient d'un document différent. Cette procédure présente un double avantage : elle ne nécessite pas l'intervention d'un jugement humain et elle permet de contrôler très précisément les caractéristiques du corpus d'évaluation (taille des segments, des documents, etc.). Son inconvénient principal est évidemment que les documents ainsi constitués sont artificiels et que la tâche évaluée s'apparente plus à une segmentation en documents qu'à une segmentation des documents. Dans notre cas, ce cadre présente un inconvénient supplémentaire : notre découverte des

thèmes d'un document s'appuie sur le fait que les mots d'un thème ont tendance à cooccurrer à l'échelle du document. Cette hypothèse n'a plus vraiment de sens pour des documents construits suivant les principes proposés par Choi. C'est la raison pour laquelle nous avons adapté ces principes pour nous rapprocher d'une forme des documents plus réaliste tout en conservant les avantages de ce cadre d'évaluation.

	nombre de documents sources	**nombre de topics sources**	**nombre de segments / document**	**nombre moyen de phrases / doc.**	**nombre moyen de mots pleins / doc.**
français	128	11	10 (84%) 8 (16 %)	65	797
anglais	87	3	10 (97%) 8 (3%)	68	604

Tableau 1. Caractéristiques des corpus d'évaluation

Cette adaptation concerne la façon dont les extraits de documents agencés sont sélectionnés. Au lieu de tirer chaque extrait d'un document différent, nous n'utilisons que deux documents. Chacun d'entre eux est divisé comme dans le cas de Choi en segments de 3 à 11 phrases et le document d'évaluation est constitué en prenant, à partir du début des documents, alternativement un segment dans un des deux documents et le suivant dans l'autre et ce, jusqu'à obtenir 10 segments ou jusqu'à ce que le processus de construction atteigne la fin d'un des deux documents. Pour nous assurer que deux segments consécutifs font référence à deux thèmes différents et que le changement de thème entre les deux est donc effectif, les deux documents sont sélectionnés de manière à appartenir à des thématiques différentes.

Pour ce faire, nous nous sommes appuyés sur les données constituées pour l'évaluation CLEF, dédiée à la recherche d'information multilingue. Les documents source utilisés pour réaliser notre corpus étaient ainsi des documents issus du corpus CLEF pour lesquels nous disposions d'un jugement de pertinence par rapport à un des topics d'interrogation définis pour l'évaluation, que nous assimilons ici à des thèmes. Chaque document de notre corpus a ainsi été construit à partir de deux documents du corpus CLEF jugés pertinents pour deux topics différents. Outre l'existence de cette forme d'annotation thématique, le corpus CLEF présente l'avantage de comporter des documents comparables pour différentes langues, propriété

que nous avons exploitée pour constituer des corpus d'évaluation en français et en anglais assez proches. Plus précisément, ces documents sont des articles de journaux des années 1994 et 1995 - Le Monde pour le français, Los Angeles Times and Glasgow Herald pour l'anglais - et des dépêches de l'agence de presse SDA couvrant la même période dans les deux langues. Le Tableau 1 résume les caractéristiques de ces deux corpus. Chacun d'entre eux est formé de 100 documents. On voit donc qu'un document source de CLEF contribue généralement à la construction de plusieurs documents d'évaluation.

21.2. Découverte des thèmes d'un document

La méthode que nous proposons exploitant les thèmes des documents, il est intéressant d'avoir une évaluation de la méthode permettant de les découvrir afin de mettre en relation ses résultats avec ceux de la segmentation thématique. Nous utilisons pour ce faire les corpus que nous avons présentés à la Section 5.1. Dans ce contexte, la représentation de référence d'un thème d'un document est constituée par le vocabulaire contenu dans les segments relevant de ce thème et n'apparaissant pas dans les segments des autres thèmes. Le Tableau 2 donne l'exemple des thèmes découverts pour un document de test construit à partir d'un document à propos des risques induits par le problème la vache folle en Suisse et d'un document évoquant les difficultés des fabricants de ski suisses. On constate que les deux thématiques sont effectivement bien séparées mais que certains mots des thèmes découverts (en italique) sont absents des thèmes de référence qui leur correspondent. Un mot comme Suisse peut être considéré comme pertinent mais sa présence dans les deux thèmes explique son absence des thèmes de référence. En revanche, des mots tels que devenir, année ou dernier n'ont pas de spécificité par rapport aux thèmes considérés et ne sont présents dans les thèmes découverts que du fait de leur forte présence dans un des deux documents d'origine.

Thème 1	folle, fédéral, cas, devenir, vache, bovin, infecter, maladie, ESB, humain, déclarer
Thème 2	fabricant, Streule, marché, paire, production, ski, Stöckli, Suisse, indiquer, directeur, année, entreprise, dernier

Tableau 2. Exemple de thèmes découverts pour un document du corpus d'évaluation

Pour juger plus formellement de la pertinence des thèmes découverts par rapport aux thèmes de référence, nous faisons appel à trois mesures complémentaires. La principale est la mesure de pureté, classiquement utilisée pour évaluer les résultats d'une classification non supervisée. La pureté d'un thème découvert est donnée par la proportion de son vocabulaire correspondant au vocabulaire du thème de référence auquel il est associé. Un thème

découvert est associé au thème de référence avec lequel il partage le plus de mots. La pureté globale des thèmes découverts est donnée quant à elle par :

$$Pureté = \sum_{i=1}^{k} \frac{v_i}{V} P(Td_i) \quad (5.1)$$

où P(Tdi) est la pureté du thème découvert Tdi, V est l'ensemble du vocabulaire des thèmes découverts et vi est le vocabulaire de Tdi. La deuxième mesure que nous utilisons évalue dans quelle mesure les thèmes de référence sont effectivement représentés parmi les thèmes découverts, chacun d'entre eux étant associé comme précédemment au thème de référence partageant avec lui le plus large vocabulaire. Cette mesure est donnée par le rapport entre le nombre de thèmes découverts associés à un thème de référence et le nombre de thèmes de référence. Finalement, la dernière mesure permet de savoir si le vocabulaire des thèmes découverts couvre de façon significative ou pas le vocabulaire des thèmes de référence en calculant le rapport entre la taille du vocabulaire des thèmes découverts faisant partie des thèmes de référence et celle du vocabulaire des thèmes de référence.

	pureté	**% de thèmes représentés**	**couverture du vocabulaire des thèmes**
français	0,771 - 0,117	0,895 - 0,239	0,299 - 0,078
anglais	0,766 - 0,082	0,99 - 0,1	0,316 - 0,053

Tableau 3. Résultats de la découverte des thèmes des documents

Le Tableau 3 donne pour chaque mesure sa moyenne, suivie de son écart-type. Les résultats sont globalement comparables pour le français et l'anglais même si l'on peut observer une pureté un peu meilleure en français et une représentation des thèmes et de leur vocabulaire un peu meilleure en anglais. À un niveau plus général, on constate que la méthode de découverte des thèmes d'un document produit des représentations de thèmes assez peu bruitées, i.e. assez pures, que chaque thème de référence y est habituellement représenté mais que cette représentation ne couvre qu'une petite partie de sa forme de référence. Les thèmes découverts semblent donc pertinents mais leur description est sans doute trop lacunaire.

21.3. Segmentation thématique d'un document

Concernant la segmentation thématique, la procédure d'évaluation consiste à appliquer l'algorithme de segmentation considéré sur les documents construits dont on a supprimé les marques de séparation des segments et à comparer les changements de thème détecté avec ces marques de référence. Cette comparaison s'effectue en utilisant principalement la mesure d'erreur Pk [BEE99], conformément aux évaluations récentes faites dans ce domaine. WindowDiff [PEV02], dont nous donnons aussi les résultats, est une variante de Pk corrigeant certaines de ses insuffisances. Pk évalue la probabilité que deux mots choisis aléatoirement dans un document et séparés par k mots soient jugés comme appartenant au même segment alors qu'ils sont dans des segments différents (faux négatif) ou qu'ils soient jugés comme appartenant à des segments différents alors qu'ils sont dans le même (fausse alarme). k est égal à la moitié de la taille moyenne des segments au niveau du corpus de référence. L'objectif est bien entendu de minimiser Pk.

systèmes	P_k			WindowDiff		
	erreur	p (F06)	p (F06T)	erreur	p (F06)	p (F06T)
U00	25,91	**0,003**	**1,3 e-07**	27,42	0,799	**0.032**
C99	27,57	**4,2 e-05**	**3,6 e-10**	35,42	**8,6 e-07**	**6,5 e-13**
TextTiling*	21,08	0,699	**0,037**	27,43	0,803	**0,032**
LCseg	20,55	0,439	0.111	28,31	0,767	**0.007**
F06	21,58		**0,013**	27,83		**0,016**
F06T	18,46	**0,013**		24,05	**0,016**	

Tableau 4. Résultats de la segmentation thématique pour le corpus en français

Le Tableau 4 et le Tableau 5 donnent non seulement les résultats obtenus par les segmenteur F06 et F06T sur les corpus français et anglais décrits à la Section 5.1 mais également les résultats sur ces mêmes corpus de certaines méthodes de référence : U00 est ainsi la méthode décrite dans [UTI01], C99, celle proposée dans [CHO00] et LCseg est présentée dans [GAL03]. TextTiling* est une variante de TextTiling dans laquelle la troisième étape d'identification des changements de thème est reprise de [GAL03]. Il est à noter que toutes ces méthodes sont utilisées comme F06 et F06T, sans fixer le nombre de changements de thème à trouver et que leurs paramètres ont été adaptés au corpus d'évaluation pour en tirer le meilleur parti. Pour chaque résultat de ces méthodes, nous donnons en outre le niveau p de signification de sa différence avec F06 et F06T, niveau évalué grâce à un test de Student unilatéral. Ce niveau correspond plus précisément à la probabilité d'erreur quant au fait de rejeter l'hypothèse stipulant que la différence entre les résultats testés n'est pas significative. La probabilité maximale en dessous de laquelle deux résultats sont considérés comme significatifs est

classiquement fixée à 0,05. Les différences significatives sont marquées en gras dans les deux tableaux.

systèmes	P_k			WindowDiff		
	erreur	p (F06)	p (F06 T)	erreur	p (F06)	p (F06 T)
U00	19,42	**0,048**	**4,3 e-05**	21,22	0,826	**0,039**
C99	21,63	**1,2 e-04**	**1,8 e-09**	30,64	**1,4 e-12**	**0**
TextTiling*	15,81	0,308	0,111	19,80	0,355	0,253
LCseg	14,78	**0,043**	0,496	19,73	0,325	0,271
F06	16,90		**0,010**	20,93		**0,046**
F06T	14,06	**0,010**		18,31	**0,046**	

Tableau 5. Résultats de la segmentation thématique pour le corpus en anglais

Le premier enseignement à tirer des deux tableaux ci-dessus est que l'hypothèse que nous avions formulée sur l'intérêt de prendre en compte pour la segmentation les thèmes des documents découverts automatiquement est confirmée. Aussi bien pour le français que pour l'anglais, les résultats de F06T sont significativement et même très significativement supérieurs à ceux de F06. Le deuxième enseignement notable est le fait que ces résultats sont assez stables, même si les deux corpus considérés sont assez proches : la mise au point de F06 et de F06T a été faite sur le corpus français et les résultats obtenus pour l'anglais sont comparables, aussi bien au niveau de la comparaison entre F06 et F06T qu'au niveau de la comparaison de leurs résultats avec les méthodes de référence. Concernant cette comparaison, on retrouve au niveau des résultats les parentés entre méthodes : TextTiling*, LCseg, F06 et F06T partagent un nombre important de principes, ce qui se caractérise ici par des résultats significativement plus élevés que ceux de U00 ou C99. Cette tendance vient d'ailleurs partiellement à rebours des résultats obtenus sur le corpus de Choi, ce qui laisse à penser que les méthodes performantes sur ce corpus ne le seront pas nécessairement sur de " vrais " documents. Enfin, on pourra noter sans véritablement l'expliquer que les performances de toutes ces méthodes sont plus élevées en anglais qu'en français. Compte tenu de la similarité des deux corpus, la différence semble en première analyse imputable à la langue, peut-être du fait de la moindre réticence stylistique à la répétition en anglais.

22. Conclusion et perspectives

Dans cet article, nous avons présenté une méthode non supervisée de découverte des thèmes d'un document ne s'appuyant sur aucune connaissance externe. Nous avons également montré comment les thèmes ainsi identifiés peuvent servir à améliorer une méthode de segmentation thématique fondée sur la récurrence lexicale. Nous avons en outre proposé une adaptation du cadre d'évaluation proposé par Choi pour la segmentation

thématique afin de mettre en évidence cette amélioration et nous avons pu la démontrer sur un corpus en français et un corpus en anglais.

Bien que la capacité à ne s'appuyer sur aucune connaissance externe soit intéressante, elle est également porteuse de limites sérieuses, même si nous avons montré comment les repousser partiellement. Dans le prolongement de travaux antérieurs sur l'utilisation des réseaux de cooccurrences lexicales [FER98], qui constituent une forme de connaissances facile à construire, nous envisageons d'étendre la méthode de découverte de thèmes en combinant le graphe de cooccurrence des mots d'un document avec le graphe formé par un réseau de cooccurrences lexicales constitué à partir d'un vaste corpus. Ce réseau pourra par ailleurs être utilisé directement pour la segmentation, à l'instar de [JOB98].

23. Références bibliographiques

[BAR97] Barzilay R., Elhadad M., " Using Lexical Chains For Text Summarization ", ACL 97 Workshop on Intelligent Scalable Text Summarization, p. 10-17, 1997.

[BEE99] Beeferman D., Berger A., Lafferty J., " Statistical Models for Text Segmentation ", Machine Learning, Vol. 34, n°1, p. 177-210, 1999.

[BIG98] Bigi B., Mori R. d., El-Bèze M., Spriet T., " Detecting topic shifts using a cache memory ", 5th International Conference on Spoken Language Processing, p. 2331-2334, 1998.

[BOG00] Boguraev B., Neff M. S., " Discourse Segmentation in Aid of Document Summarization ", HICSS, 2000.

[CAI 04] Caillet M., Pessiot J.-F., Amini M., Gallinari P., " Unsupervised Learning with Term Clustering for Thematic Segmentation of Texts ", 7th Conference on Rercherche d'Information Assistée par Ordinateur (RIAO'04), p. 1-11, 2004.

[CHA04] Châar S. L., Ferret O., Fluhr C., " Filtrage pour la construction de résumés multi-documents guidée par un profil ", Traitement Automatique des Langues, Vol. 45, n°1, p. 65-93, 2004.

[CHO00] Choi F. Y. Y., " Advances in domain independent linear text segmentation ", NAACL'00, p. 26-33, 2000.

[CHO01] Choi F. Y. Y., Wiemer-Hastings P., Moore J., " Latent Semantic Analysis for Text Segmentation ", EMNLP'01, p. 109-117, 2001.

[CHU90] Church K. W., Hanks P., " Word Association Norms, Mutual Information, And Lexicography ", Computational Linguistics, Vol. 16, n°1, p. 22-29, 1990.

[ERT01] Ertöz L., Steinbach M., Kumar V., " Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach ", Text Mine'01, Workshop of the 1st SIAM International Conference on Data Mining, 2001.

[FER00] Ferret O., Grau B., " A Topic Segmentation of Texts based on Semantic Domains ", ECAI 2000, p. 426-430, 2000.

[FER98] Ferret O., " How to thematically segment texts by using lexical cohesion? ", ACL-COLING'98, p. 1481-1483, 1998.

[FIS99] Fiscus J., Doddington G., Garofolo J., Martin A., " NIST's 1998 Topic Detection and Tracking ", DARPA Broadcast News Workshop, 1999.

[GAL03] Galley M., McKeown K., Fosler-Lussier E., Jing H., " Discourse Segmentation of Multi-party Conversation ", 41st Annual Meeting of the Association for Computational Linguistics (ACL'03), p. 562-569, 2003.

[HEA 94] Hearst M. A., " Multi-paragraph segmentation of expository tex "t, 32th Annual Meeting of the Association for Computational Linguistics, p. 9-16, 1994.

[JOB98] Jobbins A. C., Evett L. J., " Text Segmentation Using Reiteration and Collocation ", ACL-COLING'98, p. 614-618, 1998.

[KAN98] Kan M.-Y., Klavans J. L., McKeown K. R., " Linear Segmentation and Segment Relevence ", 6th International Workshop of Very Large Corpora (WVLC-6), p. 197-205, 1998.

[KAU99] Kaufmann S., " Cohesion and Collocation: Using Context Vectors in Text Segmentation ", 37th Annual Meeting of the Association for Computational Linguistics (Student Session), p. 591-595, 1999.

[KOZ93] Kozima H., Text Segmentation Based on Similarity between Words, 31th Annual Meeting of the Association for Computational Linguistics (Student Session), p. 286-288, 1993.

[MOR91] Morris J., Hirst G., " Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text ", Computational Linguistics, Vol. 17, n°1, p. 21-48, 1991.

[PAS97] Passonneau R. J., Litman D. J., " Discourse Segmentation by Human and Automated Means ", Computational Linguistics, Vol. 23, n°1, p. 103-139, 1997.

[PEV02] Pevzner L., Hearst M. A., " A critique and improvement of an evaluation metric for text segmentation ", Computational Linguistics, Vol. 28, n°1, p. 19-36, 2002.

[PON97] Ponte J. M., Croft B. W., " Text segmentation by topic ", First European Conference on research and advanced technology for digital libraries, 1997.

[REY94] Reynar J. C., " An Automatic Method of Finding Topic Boundaries ", 32th Annual Meeting of the Association for Computational Linguistics (Student Session), 1994.

[SAL83] Salton G., " Introduction to Modern Information Retrieval ", McGraw-Hill, 1983.

[SCH94] Schmid H., " Probabilistic Part-of-Speech Tagging Using Decision Trees ", International Conference on New Methods in Language Processing, 1994.

[TÜR01] Tür G., Tür D. H., Stolcke A., Shriberg E., " Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation ", Computational Linguistics, Vol. 27, n°1, p. 31-57, 2001.

[UTI01] Utiyama M., Isahara H., " A Statistical Model for Domain-Independent Text Segmentation ", ACL 2001, p. 491-498, 2001.

[YAM98] Yamron J., Carp I., Gillick L., Lowe S., van Mulbregt P., " A hidden Markov model approach to text segmentation and event tracking ", IEEE Conference on Acoustics, Speech and Signal Processing, p. 333-336, 1998.

Sémantique des liens hypertextes

Moustafa AL-HAJJ
Gilles VERLEY
Hubert CARDOT

Université François-Rabelais de Tours
Laboratoire d'Informatique (EA 2101), 64, Avenue Jean Portalis,
37200 TOURS - France
http://www.li.univ-tours.fr
*prenom.nom@univ-tours.fr

RÉSUMÉ

Les auteurs qui publient sur le Web des connaissances sous la forme de documents électroniques lisibles sur un écran utilisent de plus en plus la technologie des liens hypertextes pour améliorer l'ergonomie de leur sites et pour les enrichir par des informations provenant d'autres sites Web. Nous nous intéressons à la sémantique des liens hypertextes, en termes d'extraction et d'exploitation, dans le but de faciliter la recherche d'information sur le Web. Dans cet article, nous proposons une méthodologie originale d'extraction de la sémantique des liens hypertextes par des moyens manuels et semi-automatiques. Dans une première partie, nous montrons comment nous avons constitué un corpus de documents sur le Web, qui sera par la suite notre base de test. Cette constitution consiste à extraire un sous-ensemble du Web, regroupant des pages ayant des critères utiles à l'étude de la sémantique des liens hypertextes. Ensuite nous proposons une méthode d'analyse de la sémantique des liens hypertextes. Celle-ci consiste à faire l'analyse sémantique du contexte appelant du lien et du contexte appelé par le lien, et à expliciter de manière formelle la relation sémantique entre le contexte appelant et le contexte appelé. La dernière partie est consacrée à l'élaboration d'outils d'aide à l'analyse, nous proposons une automatisation de la reconnaissance des formes littéraires des contextes appelant des liens et des contextes appelés par des liens avec les treillis de Galois.

MOTS CLÉS : Constitution de corpus ; analyse sémantique de liens hypertextes ; ontologie ; RDF(S) ; treillis de Galois ; K-means.

24. INTRODUCTION

Les auteurs qui publient sur le Web des connaissances sous la forme de documents électroniques lisibles sur un écran utilisent la technologie des liens hypertextes pour améliorer l'ergonomie de leurs sites et pour les enrichir par des informations provenant d'autres sites Web.

La différence entre les approches d'écriture et de lecture hypertextuel et de celles pour les papiers réside dans la structure non linéaire de l'hypertexte.

Le fait qu'à partir d'un nœud le lecteur peut se retrouver sur d'autres nœuds grâce à l'activation des liens hypertextes supposerait que l'hypertexte contienne une multiplicité de parcours de lecture. Cette multiplicité des parcours risque d'égarer le lecteur habitué à une approche traditionnelle de l'information écrite. Il est donc intéressant d'offrir aux utilisateurs un moyen de naviguer dans les réseaux hypertextes, qui limite les risques de désorientation et de surcharge cognitive. Il s'agit d'offrir aux lecteurs des outils lui permettant d'accéder à la sémantique de l'information.

La mise en évidence de la sémantique des nœuds d'un hypertexte et de relations sémantiques entre ces nœuds aide le lecteur à s'orienter dans sa recherche d'information. La sémantique des nœuds et la relation sémantique entre ces nœuds permettent au lecteur d'avoir un contexte de navigation, elles sont très utiles pour cibler l'information pertinente plus rapidement. En associant à chaque nœud un sous-ensemble de termes pertinents et à chaque lien la nature de la relation entre les nœuds qu'il lie, on peut offrir au lecteur un moyen de se déplacer sémantiquement dans le graphe hypertextuel.

Cela nous a motivé à nous intéresser à la sémantique des liens hypertextes. Dans cet article, nous proposons une méthodologie originale d'extraction de la sémantique des liens hypertextes par des moyens manuels et semi-automatiques. Pour vérifier la validité de la méthode, nous l'avons testée sur les liens hypertextes d'un corpus spécifique sélectionné sur le Web.

Dans une première partie, nous montrons comment nous avons constitué le corpus. Ensuite nous proposons une méthode d'analyse de la sémantique des liens hypertextes. Celle-ci consiste à faire l'analyse sémantique du contexte appelant du lien et du contexte appelé par le lien, et à expliciter de manière formelle la relation sémantique entre le contexte appelant et le contexte appelé. La dernière partie est consacrée à l'élaboration d'outils d'aide à l'analyse, nous proposons une automatisation de la reconnaissance des formes littéraires des contextes appelants des liens et des contextes appelés par des liens avec les treillis de Galois.

25. CONSTITUTION DU CORPUS ET ONTOLOGIE DU DOMAINE

Le corpus va constituer notre base de test, on l'obtient en extrayant un sous-ensemble du Web, regroupant des pages ayant des critères utiles pour l'étude de la sémantique des liens hypertextes. Le thème retenu est la biographie de personnages célèbres.

Nous présentons d'abord les critères de sélection des documents et ensuite une partie de l'ontologie du domaine représentée en RDF(S) (Resource Description Framework Scheme).

25.1. CRITÈRES DE SÉLECTION

Les documents du corpus ont été sélectionnés par rapport à plusieurs critères :

25.1.1. Sujet et langue des documents

Notre corpus est constitué de documents traitant des biographies de personnages célèbres, ce choix est dû à la richesse de ces documents en liens hypertextes, les auteurs les utilisent en effet pour améliorer l'ergonomie de leur sites Web et pour les enrichir par des informations provenant d'autres sites sur le Web, etc. [VAN00].

Pour faciliter l'annotation sémantique des liens hypertextes, le corpus a été limité aux pages écrites en français.

25.1.2. Variétés des auteurs et de serveurs de documents et type de mise en page

Les documents du corpus sont issus de serveurs différents, les auteurs des biographies varient d'un serveur à un autre. Ceux-ci ont beaucoup de raisons différentes de poser des liens hypertextes dans leur propos, de plus les documents du corpus ont des formes littéraires très diverses, ceci constitue une grande richesse pour le corpus.

L'interrogation des outils de recherche francophones (annuaires et moteurs de recherche) comme Google (www.google.fr) et Altavista (www.altavista.com) et Yahoo (www.yahoo.fr), avec les requête " biographie " et/ou " nom d'un personnage célèbre " réduite aux pages satisfaisant les critères de sélection ainsi définis, a permis d'obtenir un ensemble de 140 biographies provenant d'au moins 19 serveurs différents, l'ensemble de biographies contient plusieurs milliers de liens hypertextes.

Les documents du corpus sont des documents HTML, ce choix a été fait car la majorité des documents sur le Web sont encore en format HTML, et permet la standardisation de certains traitements afin d'étudier la sémantique des liens hypertextes.

25.1.3. Liens natifs et répondant à des besoins variés

On entend par lien natif, un lien voulu par l'auteur, contrairement à un lien calculé par des automates.
Les auteurs posent des liens hypertextes dans leurs propos pour satisfaire différents besoins, ces besoins sont tels que :

- La généralisation du propos de l'auteur, comme un lien vers un sommaire.
-La spécialisation du propos de l'auteur, comme un lien situé dans un sommaire.

- L'illustration du propos de l'auteur, tel qu'un lien vers une photo.
- Etc.

Toute relation sémantique d'un propos de l'auteur avec un autre propos, peut être à l'origine de la pose d'un lien hypertexte dans le premier propos vers le second.
Nous avons sélectionné les sites constitués essentiellement de liens natifs et dont la sémantique était variée. Voici une typologie des relations sémantiques portées par les liens de notre corpus et que nous utiliserons dans la seconde partie.
- Une personne citée dans le propos de l'auteur s'est opposé à, a connu, a été maître de, a fréquenté, est l'élève de, soutient, a été soutenu par, est la fille de, est la grand-mère de, est le fils de, est le grand-père de, est le mari de, est parent d' une personne dont la biographie est traitée dans la cible du lien.
- Une personne citée dans le propos de l'auteur a étudié, a écrit, a joué, a adapté, a fondé, a lu, a réalisé, a découvert, a traduit, a utilisé, a commenté, a détaillé, a retranscrit, s'est inspiré de, est à l'origine d' une chose citée dans le propos cible du lien.
- Une chose citée dans le propos de l'auteur a été étudiée par, a été écrite par, a été jouée par, a été adaptée par, a été fondée par, a été lu par, a été réalisée par, a été découverte par, a été traduit par, a été utilisée par, a été commentée par, a été détaillée par, a été retranscrite par une personne dont la biographie est traitée dans la cible du lien.
- Une personne citée dans le propos de l'auteur a participé à un évènement cité dans la page ciblée par le lien.
- Une personne citée dans le propos de l'auteur a travaillé, a vécu, a voyagé, est mort, est né dans un lieu cité dans la cible du lien.
- Une chose citée dans le propos de l'auteur a été jouée en un lieu cité dans la cible du lien.
- Une chose citée dans le propos de l'auteur donne accès à, est apparenté à, est comparé à, est détaillé par, est illustré par, a influencé, explique, fait partie de, illustre, a été commentée par, parle de, est représenté par, représente, est généralisée par, contient une chose citée dans le propos cible du lien.
- Un évènement cité dans le propos de l'auteur a eu lieu dans un lieu cité dans la cible du lien.
- Un évènement cité dans le propos de l'auteur est illustré par une chose citée dans le propos cible du lien.
- Un lieu cité dans le propos de l'auteur est l'endroit d'un évènement cité dans la cible du lien.

EXEMPLE D'UN LIEN NATIF

Considérons une page de la biographie de François Mitterrand et une page qui a pour sujet la convention de Lomé IV. La dernière est, soit faite par le même auteur, soit par un autre.

Dans la première page, l'auteur cite les oeuvres économiques étrangères de François Mitterrand, et parmi elles, la convention de Lomé IV. Dans la partie qui cite la convention de Lomé IV, l'auteur pose un lien hypertexte vers la page qui a pour sujet la convention de Lomé IV. Le contexte appelant du lien est supposé être la partie de la page du lien qui cite les œuvres économiques étrangères, et le contexte appelé par le lien est supposé être la cible du lien qui traite la convention de Lomé IV.

La relation sémantique entre le contexte appelant du lien et du contexte appelé par le lien est la suivante : " Les oeuvres économiques étrangères de François Mitterrand " contient " La convention de Lomé IV ".

Si l'auteur de la page biographique n'est pas le même que celle de la convention de Lomé IV, la découverte par l'auteur de l'existence de la page sur la convention de Lomé IV et de son adresse, l'aura motivé à poser un lien vers celle-ci dans sa page, plus précisément dans la partie qui traite des oeuvres économiques étrangères, dans la sous-partie qui cite la convention de Lomé IV.
Et si l'auteur de la page de la convention de Lomé IV est aussi celui de la page biographique de François Mitterrand, ce sont les avantages d'utilisation des liens hypertextes, pour améliorer l'ergonomie des sites Web, qui l'auront motivé à créer la page sur la convention de Lomé IV à part, et à poser un lien vers cette page dans le texte principal.

25.1.4. Importance de la relation sémantique portée par les liens natifs

Nous nous intéressons dans cet article à expliciter la relation sémantique des liens natifs car nous faisons l'hypothèse que cette relation sémantique est suffisamment importante pour avoir motivé la pose d'un lien par l'auteur lui-même et qu'ainsi elle est hautement pertinente, nous pensons également qu'il est plus simple d'expliciter formellement les relations sémantiques portées par les liens natifs que celles entre les autres phrases du texte.

25.1.5. Variété des formes littéraires des contextes des liens et des contextes appelés par les liens

Pour la suite, on nomme " contexte appelant d'un lien " l'ensemble minimal de textes, caractères et objets, autour du lien et qui constituent une seule idée, concept ou sujet [VIG01].

De même, on nomme " contexte appelé par un lien " l'ensemble minimal de textes, caractères et objets de la page ciblée par le lien et qui constituent un sujet en rapport avec le " contexte appelant du lien " [VIG01].
Les contextes, qu'ils soient contextes appelants ou appelés, peuvent être de diverses formes littéraires. On les a regroupés de la manière suivante :
Forme Sommaire ; Forme Illustration Graphique ; Forme Définition ; Forme Citation ; Forme Liste ; Forme Référentielle ; Forme Récit ; Forme Description ; Forme Détail ; Forme Résumé. Cet aspect est traité en §4.2.2.

25.2. Ontologie des liens hypertextes du corpus

représentation par RDFS
Une fois notre corpus constitué, nous avons défini une ontologie des liens hypertextes du domaine considéré.
Cette partie a pour objectif de présenter une partie de l'ontologie, et de montrer qu'elle est représentable par les technologies RDF(S) (Resource Description Framework Scheme) [RDFS] [RDF] [CHAR03]. Les concepts de l'ontologie sont : Personne, Chose, Lieu, Evènement, Date, FormeContexte. Les relations sont celles du §2.1.3. La représentation de l'ontologie par RDFS est la suivante :

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:owl="http://www.w3.org/2002/07/owl#"
        xmlns:dc=http://purl.org/dc/elements/1.1/ >
<owl:Ontology
        rdf:about="URIOntologieLiensHypertextes"
        dc:title="l'ontologie de liens hypertextes"/>
<rdfs:Class rdf:about="URIOntologieLiensHypertextes#Personne">
                <rdfs:label>Personne</rdfs:label>
<rdfs:comment>
        Une personne décrite par son nom et son prénom
</rdfs:comment>
</rdfs:Class>
…
<!-- Les autres concepts : Chose,  Lieu, Evènement, Date sont représentés de la même manière -->
<rdfs:Class rdf:about="URIOntologieLiensHypertextes#FormeContexte">
<rdfs:label>FormeContexte</rdfs:label>
        <rdfs:comment>
La forme littéraire du contexte appelant du lien ou du contexte appelé par le lien
</rdfs:comment>
</rdfs:Class>
<rdf:Property rdf:ID="a travaillé">
         <rdfs:domain rdf:resource="#Personne" />
         <rdfs:range rdf:resource="#Lieu" />
</rdf:Property>
…
<!--Toutes les relations sémantiques du §2.1.4 sont représentées de la même manière -->
</rdf:RDF>
```

26. Méthode proposée pour effectuer l'analyse manuelle sémantique d'un lien

Pour effectuer l'analyse sémantique manuelle d'un lien hypertexte, la méthode proposée consiste à faire l'analyse sémantique des deux contextes, contexte appelant du lien et contexte appelé par le lien, et à trouver la relation sémantique entre le contexte appelant et le contexte appelé.
On fait l'hypothèse que la raison pour laquelle l'auteur a posé ce lien est contenue dans ces trois analyses sémantiques.

26.1. ANALYSE SÉMANTIQUE DES CONTEXTES APPELANTS ET APPELÉS

'analyse sémantique des deux contextes, contexte appelant du lien et contexte appelé par le lien, consiste à les décrire dans une phrase composée de trois parties :

- La première pour dire qu'il s'agit d'un contexte du lien ou d'un contexte appelé par le lien.
- La deuxième pour décrire la forme littéraire du contexte - appelant ou appelé - qu'on analyse.
- La troisième pour décrire, par quelques mots clés reliés dans une phrase dans langage naturel, le contexte appelant (resp. appelé) en cours d'analyse.

La forme littéraire peut être choisie dans la liste de formes parmi celles citées en §2.1.5. Les mots clés les plus représentatifs du contexte - appelant ou appelé - peuvent être dérivés de l'ontologie du domaine, c'est-à-dire, ils peuvent être des instances des concepts de l'ontologie que nous avons vu en §2.2. Sinon, c'est-à-dire au cas où un mot clé ne peut être dérivé d'aucun concept de l'ontologie, il faudra ajouter de(s) nouveau(x) terme(s) à cette dernière pour que le mot clé puisse être dérivé de celle-ci.

26.2. Relation sémantique entre les deux contextes

Une fois l'analyse sémantique des deux contextes, appelant et appelé, terminée, on cherche à trouver une relation entre ces deux contextes. On l'explicite dans une phrase selon le modèle présenté en § 2.1.3.

Pour vérifier la validité de notre méthode d'analyse sémantique des liens h
L'expérience montre que la méthode est opérationnelle, voici un exemple d'analyse d'un lien hypertexte selon la méthode, il sera suivi par une représentation de la sémantique du lien par RDF (Resource description Framework).

26.3. Exemple

Soit le lien hypertexte situé dans la page biographique de George Bush dans la partie qui raconte son repêchage dans l'océan pacifique, la cible du lien est une illustration par une photo du repêchage (voir figure 1), la sémantique est la suivante :

a) Sémantique du contexte appelant du lien :
" La source est le récit du repêchage de George Bush dans l'océan pacifique le 2 septembre 1944 ".

- " Source " : il s'agit de l'analyse du contexte appelant.
- " Récit " est la forme littéraire du contexte source de lien.
- " Repêchage ; George Bush ; océan pacifique ; 2 septembre 1944 " sont les mots clés décrivant le contexte appelant du lien.

b) Sémantique du contexte appelé par le lien :
" La cible est une illustration par une photo du repêchage de George Bush dans l'océan pacifique le 2 septembre 1944 ".
-" Cible " : il s'agit de l'analyse du contexte appelé.
-" Illustration graphique " est la forme littéraire du contexte cible de lien.
-" Photo ; Repêchage ; George Bush ; océan pacifique ; 2 septembre 1944 " sont des mots clés décrivant le contexte appelé par le lien.

c) Sémantique de la relation entre le contexte appelant et le contexte appelé :
" Le repêchage est illustré par la photo ".
-" Repêchage " est un mot clé du contexte appelant.
-" photo " est un mot clé du contexte appelé.
-" est illustré par " est une relation sémantique entre les deux contextes.

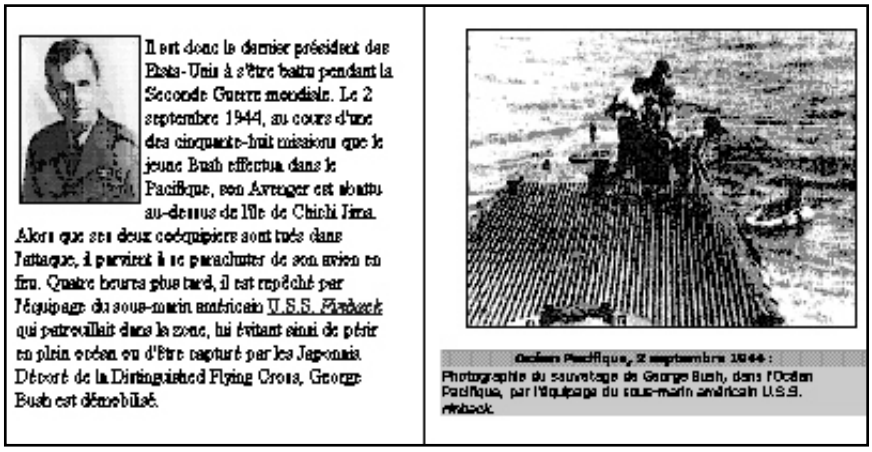

Figure 1 - à gauche : contexte appelant du lien " U.S.S. Finback " ;
à droite : contexte appelé par le lien

Représentation de la sémantique du lien par la technologie RDF [RDF] :

```
<rdf:RDF
                    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
                    xmlns:ol="URIOntologieLiensHypertextes"
  >
<rdf:Description about="URIduContexteAppelantDulien">
                    <ol:FormeContexte>Récit</ol:FormeContexte>
                    <ol:Evenement>Repêchage</ol:Evenement>
                    <ol:Personne>George Bush</ol:Personne>
                    <ol:Lieu>Océan Pacifique</ol:Lieu>
                    <ol:Date>2 septembre 1944</ol:Date>
<rdf:Description/>
<rdf:Description about="URIduContexteAppeléParLelien">
<ol:FormeContexte>
          Illustration Graphique
</ol:FormeContexte>
                    <ol:Chose>Photo</ol:Chose>
                    <ol:Evenement>Repêchage</ol:Evenement>
                    <ol:Personne>George Bush</ol:Personne>
                    <ol:Lieu>Océan Pacifique</ol:Lieu>
                    <ol:Date>2 septembre 1944</ol:Date>
<rdf:Description/>
<ol:est illustré par>
                    <ol:Evenement>Repêchage</ol:Evenement>
                    <ol:Chose>Photo</ol:Chose>
<ol:est illustré par/>
</rdf:RDF>
```

27. AIDE À L'ANALYSE DE LA SÉMANTIQUE D'UN LIEN HYPERTEXTE

Dans une perspective d'automatisation de l'extraction, selon notre méthode, de la sémantique d'un lien hypertexte, nous proposons une automatisation de la reconnaissance des formes littéraires des contextes appelants des liens et des contextes appelés par des liens. Tout d'abord, nous présentons une réflexion sur la délimitation formelle des contextes appelants et appelés, puis nous choisissons une méthode permettant de faire une telle délimitation, ensuite nous présentons des travaux sur la classification des pages du Web par leurs profils, nous définissons nos profils de contextes appelants et appelés avec leurs paramètres, une expérience de classification des contextes sera ensuite menée avec les treillis de Galois.

27.1. Délimitation des contextes appelants et appelés : réflexion et choix de méthode

A partir des définitions données aux contextes au § 2.1.5, nous avons eu plusieurs idées de délimitation formelle des contextes appelants et appelés, nous en citons quelques unes :

1) La première est de considérer le contenu de la cible du lien qui correspond à l'écran, comme support du contexte appelé par le lien, cette idée est due au point de vue suivant :

> Les auteurs posent des liens dans leurs documents pour améliorer la lecture à l'écran. L'écran cible serait révélateur de ce que l'auteur a voulu faire.

2) La deuxième est de considérer le paragraphe contenant le lien comme support du contexte appelant du lien, car un paragraphe représente une idée ou un sujet.

3) La troisième est de considérer le contenu entre les deux balises " a name " précédant et succédant immédiatement le lien comme support du contexte appelant du lien, et le contenu de la cible du lien entre le début de la cible et la première balise " a name " comme support du contexte appelé par le lien, cette idée est le fruit du raisonnement suivant :

> La présence d'un " a name " dans une page Web, signifie l'existen ce d'un lien hypertexte, interne ou externe à la page, pointant vers la partie de la page qui commence par le " a name ". Cela signifie l'existence d'un sujet pour celle-ci, en rapport avec le sujet du contexte appelant du lien qui pointe vers elle. De ce fait, l'existence de deux " a name " successifs dans une page, signifie l'existence de deux sujets pour ces deux parties. Ces deux sujets, l'un par rapport à l'autre, sont :
> 8. Indépendants.
> 9. Dépendants et dissociables, comme un sujet et un sous-sujet.
> 10. Dépendants et indissociables, dans ce cas l'idée du sujet de la première partie n'est pas complète en l'absence de la deuxième. Ce cas est rarement rencontré.

Dans les deux premiers cas, la partie comprise entre les deux " a name " constitue un sujet, et donc un contexte, appelant ou appelé, selon que cette partie contient le lien ou sa cible.
Nous avons opté pour la troisième idée, étant donné qu'on ne dispose pas de moyens pertinents pour pouvoir définir le rendu-écran (qui dépend des tailles des écrans, des navigateurs, etc.), et qu'il est facile d'extraire automatiquement les parties entre les balises " a name ". La deuxième idée de délimitation, à savoir, considérer le paragraphe comme étant le support du contexte appelant du lien, est en cours d'application.

4.1.1. DISCUSSION :

Notre méthode de délimitation formelle des contextes appelants et appelés donne des résultats satisfaisants dans le cas ou les parties de la page à découper sont des cibles d'autres liens. Cependant, dans plusieurs pages, il n'existe aucun " a name ", donc le seul contexte que nous pouvons repérer par la méthode, est le sujet de toute la page, bien que ce contexte puisse être découpé en plusieurs contextes.

27.2. Classification des pages web par leurs profils

4.2.1. GÉNÉRALITÉS

Pour indexer les documents web, trois types d'information peuvent être utilisées :

- Le contenu lui-même des pages web : c'est-à-dire l'ensemble du code source de la page, le texte, les balises, les liens hypertextes, les liens vers les images ou d'autres ressources multimédias, la taille des fichiers, etc.
- Le graphe créé par les liens hypertextes reliant les pages les unes aux autres.
- Les données provenant de l'usage comme les fichiers de log, les "cookies", etc.

Cette classification est proposée par la communauté du " web mining " [KOSA00].

Il existe plusieurs approches pour aider l'utilisateur à naviguer sur le Web mais aucune ne prend en considération la notion de profil syntaxique des documents. Pourtant ces profils permettent d'identifier les types de données qu'ils contiennent. Les balisages utilisés dans les documents écrits par exemple en HTML, fournissent ces types de données.

HTML définit un ensemble de balises de base. On cite les balises de structure, puis celles qui permettent d'agencer et de composer du texte. L'autre catégorie de balises est celle qui permet de mettre en place des hyperliens. Une page Web peut être définie par un ensemble de caractéristiques (domaine du site, structure (frames, etc.), liens internes, liens externes, quantité et poids des images intégrées, rapport balise/contenu, …)

On part de l'idée qu'une page HTML peut être intéressante par sa forme descriptive et par son aspect. Celle-ci est intéressante si elle contient des liens vers le site lui même, des liens externes vers d'autres serveurs. Une page Web peut contenir des formulaires ce qui permet de comprendre qu'il s'agit d'une interface de saisie.

Il est aussi important de signaler que le poids d'une page est un élément très

significatif car il peut permettre de déduire l'importance du contenu de la page quantitativement. La présence d'images dans une page est un élément qui permet aussi de dégager une idée sur la dimension esthétique de la page.

Les documents sur le Web sont hétérogènes (sites commerciaux, pages personnelles, livres, articles, annuaires), ne possèdent aucune véritable structure.
Les contenus des sites peuvent varier d'un site à un autre par rapport aux objectifs de chaque site.

Papy F. et Bounai N. [PAPY03] proposent une approche fondée sur la classification de pages. Ils prennent en considération les balisages utilisés dans les pages Web pour élaborer des profils des pages Web. Cette approche est fondée sur les caractéristiques de pages HTML. Cette catégorisation permet alors :

- d'améliorer les navigations en réduisant l'espace de recherche en montrant seulement les pages pertinentes par rapport aux souhaits de l'usager.

- d'éviter la situation de surcharge cognitive à laquelle l'usager est souvent confronté au fil de ses lectures.

- de signaler à l'usager les types de pages auxquels aboutit sa requête.

- de donner des possibilités à l'usager de filtrer et de choisir les types de pages qu'il désire consulter.

Ils distinguent trois catégories de sites Web par rapport à leurs contenus :

- Les sites textuels privilégient les contenus textuels avec plusieurs liens internes et des liens externes car leur objectif est de diffuser les informations auprès des utilisateurs (les sites institutionnels, bibliothèques, universitaires, entreprises). Dans ceux-ci, les images ou les illustrations offrent des informations complémentaires et n'interviennent le plus souvent qu'à un deuxième niveau de recherche.

- Les sites visuels : privilégient les contenus visuels (images, graphiques d'illustration, etc.). Ainsi, ils intègrent souvent des formulaires (champs de saisies), par exemple les sites commerciaux, publicitaires, commerces électroniques, musées. L'image joue un rôle important, elle participe à l'attractivité du site et pour les commerciaux, elle est une valeur ajoutée indispensable. Pour les sites "plus techniques", l'image a une fonction différente. Elle permet à l'utilisateur de mettre rapidement ses attentes en correspondance avec l'information présentée. Dans ces sites, les textes offrent des informations complémentaires et n'interviennent qu'à un deuxième niveau de recherche.

- Les sites portails (annuaires) : privilégient plutôt les liens externes.

Pour établir une catégorisation de classification automatique des pages, ils se sont appuyés sur les travaux d'Alain Lelu ([LELU99], [BALPE96]) en utilisant l'algorithme de K-means axiales.

Une fois la méthode de K-means appliquée sur leur corpus, cinq types de pages ont été distingués automatiquement, et leur degré de typicité visualisé par une échelle à trois degrés (*, **, ***). En effet, ces cinq catégories constituent des pôles flous, plus que des classes bien distinctes :

- Page informative textuelle : Le contenu de la page est un texte.

- Page informative avec texte illustré : Le contenu de la page est une illustration visuelle, ce peut être des images, des figures, des boutons, etc.

- Page carrefour interne au site : le contenu de la page est un ensemble de liens internes au site.

- Page carrefour externe au site : le contenu de la page est un ensemble des liens externes au site.

- Page interface à la saisie : le contenu de la page est un ensemble de champs de saisie.

4.2.2. NOTRE CONTRIBUTION

Nous nous sommes inspirés de ces travaux pour construire nos classes de formes littéraires de contextes appelants et appelés, nous en avons retenu certaines et en avons rajouté d'autres spécifiques au domaine des biographies de personnages célèbres. Nous avons opté pour les classes suivantes :

- Classe sommaire : Le contenu du contexte est un résumé qui comporte les titres des parties des sites, c'est la même chose que la page carrefour interne. On les reconnaîtra principalement grâce à l'adjacence des liens.

- Classe illustration graphique : Le contenu du contexte est une illustration graphique par une image. On les reconnaîtra principalement grâce à la présence d'images de taille importante dans le contexte.

- Classe récit : Le contenu du contextes est en majorité du texte, on les reconnaîtra principalement grâce à la présence de texte en grande quantité dans le contexte.

- Classe citation : Le contenu du contexte est un texte qui fait référence directe à une oeuvre dans sa totalité ou en partie. On les reconnaîtra principalement grâce à la présence de texte en quantité moyenne et sans liens hypertextes.

- Classe liste : Le contenu du contexte est une suite d'articles inscrits les uns

à la suite des autres. On les reconnaîtra principalement grâce à la présence des puces ou numéros aux débuts des articles.

27.3. Paramètres

En partant des caractéristiques citées auparavant et en observant une page Web sous ces deux angles, il est possible d'établir le profil d'un contexte appelant ou appelé en constituant un vecteur d'informations.
Le profil est construit par une analyse et un traitement statistique de balises HTML. Les données les plus significatives obtenues à partir de notre échantillon de contextes sont :

nbHref : nombre de liens, nbImg : nombre d'images, TGimg : taille de la plus grande image, SMoyImg : surface moyenne des images, nbMot : nombre de mots hors balise, nbLEH : nombre de lignes entre balises " a href ", nbLigne : nombre de lignes hors balise, nbBListe : nombre de balises qui définissent des listes et/ou listes avec puces et/ou les énumérations, nbBPg : nombre des balises qui définissent les paragraphes, nbBSLigne : nombre de balises de saut de lignes, cit : prend 1 si des mots tels que " citation " figurent en balise méta name et 0 sinon, def : prend 1 si des mots tels que " définition " figurent en balise méta name et 0 sinon, desc : prend 1 si des mots tels que " description " figurent en balise meta-name " et 0 sinon, sommaire : prend 1 si des mots tels que " sommaire, résumé " figurent en balise meta-name et 0 sinon.

L'agent Web recueille les indicateurs quantitatifs, et les stocke sous forme d'une matrice (cf. tableau 1), chaque ligne correspond à un contexte, appelant ou appelé, et chaque colonne correspond à l'un des paramètres cités précédemment.

nbHref	nbImg	TGimg	SMoyImg	nbMot	nbLEH	nbLigne	nbBListe	nbBPg	nbBSLigne	cit	Def	Desc	Sommaire
10	1	4628	4628	2770	23	239	40	47	0	0	0	0	0
9	2	0	0	308	0	40	0	0	0	0	0	0	0

Tableau 1. Deux lignes de la matrice documents / paramètres

27.4. Découpage de la base de données

Pour la phase d'expérimentation, nous avons choisi 1029 contextes parmi les contextes appelants de liens hypertextes et des contextes appelés par les liens hypertextes de notre corpus. Ensuite nous avons annoté ces contextes manuellement par leurs formes littéraires.

A partir de cet ensemble de contextes, nous avons tiré au hasard 852 contextes pour la base d'apprentissage et ce qui reste (177 contextes) sera pour la

base de test.
Le tableau 2 est un récapitulatif des effectifs des formes littéraires dans les deux bases.

	Citation	Illustration	Liste	Sommaire	Récit
Base d'apprentissage	376	13	59	130	274
Base de test	80	3	14	18	62
% de classes dans les 2 bases	44,3	1,6	7,1	14,4	32,6

Tableau 2 - Effectifs de forme littéraire dans les bases

La classe citation est fortement représentée du fait du domaine d'application de biographies de personnages célèbres.
Ensuite nous avons mené une expérience de classification supervisée avec les treillis de Galois.

27.5. Classification supervisée avec les treillis de Galois

L'analyse formelle des concepts (AFC) [GAN99] offre un cadre théorique aux applications nombreuse et reconnues. Elle permet de représenter des données définies par une relation binaire entre deux ensembles, représentation encore appelée treillis de Galois [MEP02].

4.5.1. DISCRÉTISATION ET RÉSULTATS

Notons que cette phase est un point très important pour l'efficacité du treillis de Galois. Il existe plusieurs méthodes permettant de faire une telle discrétisation. Nous avons d'abord appliqué la méthode de discrétisation que nous décrivons dans [HAJ03] et nous avons obtenu un ensemble d'environ mille cent intervalles, ce qui fait un très grand nombre d'attributs. En conséquence la complexité en temps de la construction du treillis de Galois est très grande car elle est exponentielle en fonction du nombre d'attributs.

Nous avons alors choisi une autre méthode de discrétisation. Nous avons défini pour chaque paramètre quantitatif cité auparavant quatre intervalles, le premier correspond à des valeurs très petites du paramètre, le deuxième à des valeurs petites, le troisième à des valeurs grandes, le quatrième à des valeurs très grandes. Cela nous donne une quarantaine d'intervalles, ce qui fait que la complexité en temps de la construction du treillis de Galois est assez raisonnable.

Pour passer de la matrice précédente " contextes de la base d'apprentissage "/ " attributs quantitatifs " (tableau 1) au tableau binaire " contextes de la base d'apprentissage "/ " attributs qualitatifs ", qui sera utilisé comme entrée pour

l'algorithme de construction du treillis de Galois, nous avons procédé de la façon suivante :

Dans ce tableau, les premiers attributs qualitatifs de chaque contexte sont obtenus par échantillonnage de chaque valeur des paramètres du contexte (§4.3) dans les quatre intervalles qui lui sont définis. A ces attributs s'ajoutent cinq attributs binaires dont chacun correspond à une de nos classes et prend la valeur 1 si le contexte est de la classe qui correspond à l'attribut et 0 sinon. Les contextes de la base de test sont représentés de la même manière mise à part les cinq derniers attributs qui sont tous des zéro. Le problème de classification d'un contexte de la base de test revient alors à lui donner un attribut de classe.

Nous avons utilisé les deux techniques de reconnaissance se basant sur le treillis de Galois que nous avons déjà utilisées dans [HAJ03] : "Global Validation" et "Local Validation".

Par application de la méthode "Global Validation" sur les 177 contextes de la base de test, nous avons pu classer 108 contextes et ils sont tous correctement classés. Par application de la méthode "Local Validation" sur l'ensemble de test, nous avons pu classer 154 contextes dont 139 sont correctement classés. Le tableau 3 récapitule les résultats obtenus avec les treillis de Galois.

		Total	Citation	Illustration	Liste	Sommaire	Récit
	Effectifs	177	80	3	14	18	62
VG	Classés	108	57	1	3	12	35
	Correctement classés	108	57	1	3	12	35
VL	Classés	154	70	2	8	18	56
	Correctement Classés	139	67	2	6	13	51

Tableau 3 - Résultats obtenus avec les treillis de Galois

La même expérimentation a été réalisée avec d'autres outils de classification supervisée (k-ppv, réseaux de neurones, arbres de décisions) avec des résultats moins bons [HAJ06].

28. Conclusion et perspectives

Cette étude visait à l'extraction de la sémantique de liens hypertextes natifs pour aider à la navigation et à la recherche d'informations sur le Web. Nous avons proposé une méthode pour effectuer l'analyse sémantique des liens hypertextes, et nous avons montré la compatibilité de notre travail avec les

formalismes RDF(S). L'analyse dans un premier temps se fait manuellement. Dans un souci d'automatisation, nous proposons une aide à cette analyse qui consiste en l'extraction par des outils de reconnaissance de formes, ici nous avons utilisé les treillis de Galois, des formes littéraires des contextes appelants et appelés. Nous avons discuté des délimitations formelles des contextes appelants et appelés, puis nous avons choisi une méthode permettant de les délimiter. Les résultats de classification obtenus avec les treillis de Galois montrent leur efficacité pour la classification de contextes appelants et appelés selon leurs formes littéraires. Parmi nos perspectives, nous envisageons une autre manière de délimiter les contextes en les limitant aux paragraphes contenant les liens. Concernant les mots clés décrivant les contextes, nous sommes en train de tester des approches d'extraction de mots clés sur les contextes comme le TFIDF [SALT88] et des approches exploitant les informations autour des liens entrant aux contextes.

29. Références bibliographiques

[BALPE96] Balpe J.P., Lelu A., Saleh I., Papy F., " Techniques avancées pour l'hypertexte ", Editions Hermès, 1996.

[CHAR03] Charlet Jean, Bachimont Bruno et Troncy Raphaël (2003), Prié, Y., " Les ontologies pour le Web sémantique ", Web sémantique, Rapport final - Action spécifique 32 CNRS/STIC.

[GAN99] Ganter B. & Wille R. (1999). "Formal concept analysis, Mathematical foundations". Springer Verlag, Berlin.

[HAJ03] Al-hajj M., Bertet K., Gay J., and Ogier J.-M.. "Using the Concept Lattice for Graphic Understanding". In the proceedings of the Fifth IAPR International Workshop on Graphics Recognition (GREC'2003) , pages 329-340, Barcelona, Spain July 2003.

[HAJ06] AL-HAJJ M., VERLEY G., CARDOT H., " Une approche de caractérisation des contextes appelants et appelés des liens hypertextes ". XIIIème Rencontres de la Société Francophone de Classification SFC'06.

[KOSA00] Kosala R., Blockeel H., "Web Mining Research: A Survey", SIGKDD Explorations, vol. 2 (1) 2000, p. 1-15.

[LELU99] Lelu A., Hallab M., Rhissassi H., Papy F., Bouyahi S., Bouhaï N., He H., Qi C., Saleh I., " Projet NeuroWeb : un moteur de recherche multilingue et cartographique ", 5e conf. Int. H2PTM'99, Paris, France, septembre 1999.

[MEP02] Mephu Nguifo et Njiwoua, 2002, " Treillis de concepts et classification supervisée : un état de l'art ". CRIL rapport de recherche.

[PAPY03] Papy F., Bounai N., " Navigation et recherche par catégorisation floue des pages HTML ", Actes des JET'2003, 2003.

[RDF] http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[RDFS] http://www.w3.org/TR/rdf-schema/

[SALT88] Salton G. and Buckley C., "Term weighting approaches in automatic text retrieval". Inf. Process. Manage. 24(5): 513-523, 1988.

[VAN00] Vandendorpe C., " Du papyrus à l'hypertexte : essai sur les mutations du texte et de la lecture ", Ed. la découverte, Paris, 2000, p. 113-138.

[VIG01] VIGNAUX G., " L'hypertexte. Qu'est-ce que l'hypertexte. Origines et histoire ", Laboratoire Communication et Politique, CNRS-UPR 36 (juin 2001).

LA COMMANDE SÉMANTIQUE :
une navigation conceptuelle pour le cartable électronique.

Stephan RENAUD
Georges VIGNAUX
Charles TIJUS

Laboratoire Cognition &Usages, Université
stephan.renaud@cognition-usages.org
tijus@univ-paris8.fr
Maison des Sciences de l'Homme de Paris Nord
gvignaux@mshparisnord.org

RÉSUMÉ

Dans le cadre du développement du Cartable Electronique ", nous avons conçu et testé un nouveau type de navigation pour les manuels scolaires électroniques : la commande sémantique. L'hypermédia est conçu de telle manière que la navigation s'opère dans un espace de connaissances organisées à partir d'une ontologie formalisée par un treillis de Galois et structurée selon 3 axes :
"est une sorte de vs. a comme cas particulier ", "est la cause de vs. est la conséquence de ", " est composant de vs. est composé de ". Dans cet hypermédia, les documents possèdent la même structure que celle des connaissances et sont représentatifs d'un concept. Les relations conceptuelles existant entre les documents sont ainsi rendues explicites par le choix même des actions via le module de navigation qui représente ces relations sous forme schématique. L'utilisateur apprenti est alors libre de construire son discours dans cet hyper texte signifiant où chaque parcours est pédagogique. Les résultats des observations expérimentales montrent que la commande sémantique facilite l'apprentissage et rend plus efficace la recherche d'information.

MOTS-CLES : Hypermédia ontologique, commande sémantique, relations conceptuelles, parcours pédagogiques, manuels scolaires,

ABSTRACT

We present here the conception and the evaluation of a new kind of navigation in electronic schoolbook documents: the semantic control command. This research takes place in the context of the electronic schoolbag project. The hypermedia is built such as the navigation operates in an organized knowledge-space. In this hypermedia, an hypertext is an ontology (Galois lattice) shaped in three axes: " is a kind of vs. has instances ', " is the cause

of vs. is a consequence of", " is component of vs. is composed of". In this hypermedia the documents keep the structure of the organization of knowledge pieces and are representative of a given concept. Conceptual relations between documents appear explicitly with the choice of actions via the navigation module that represents those relations within a schematic shape. The electronic schoolbook user is then free to build his discourse in this significant hypertext where each way is a teaching one. The semantic control facilitates the learning process and improves the information search.

KEY-WORDS: Ontological hypermedia, semantic control command, conceptual relations, schoolbooks

INTRODUCTION

Cette recherche s'inscrit dans une continuité de travaux sur le Cartable Electronique menés entre autres dans le cadre du projet ACEDU (Adaptation du Cartable Electronique à ses Divers Utilisateurs) [BOU 04] et du projet LUTIN [Laboratoire des Usages en Technologies d'information Numériques] financé par le Réseau National de Recherche en Télécommunications (RNRT). La notion de " cartable électronique " recouvre l'idée de numérisation des ressources, outils et dispositifs scolaires et leur accessibilité à partir de terminaux via le réseau Intranet de l'établissement scolaire, ou via Internet. La recherche concerne l'implémentation d'un mode de navigation conceptuelle profitable aux apprentissages scolaires. Ce nouveau type de navigation à partir de commandes sémantiques est basé sur une ontologie dont la construction s'inspire des travaux sur la logique et la sémantique naturelles de Grize [GRI 02], Piaget [PIA 87] et Vignaux [VIG 05], c'est-à-dire au plus près de la manière dont les connaissances se construisent chez l'apprenti.

Au regard des connaissances à acquérir, l'apprentissage peut être peut être plus ou moins favorisé par le discours de l'auteur, c'est-à-dire la manière dont les connaissances sont exprimées dans les unités d'affichage que sont les pages et les fenêtres (lisibilité et clarté du texte, choix des illustrations, etc.), par la structure des pages et par le parcours de l'apprenti dans l'ensemble des pages (la navigation) [BOU 04]. Dans un document, le discours de l'auteur est linéaire. La structure de l'ensemble des pages dépend du support. Dans un livre, la structure est linéaire avec la pagination. Dans un manuel électronique classique, la structure peut être linéaire, mais aussi arborescente lorsqu'il n'y a pas de précédence entre chapitres, par exemple. La structure des documents contraint évidemment le parcours de l'apprenti qui peut avancer, reculer, sauter, lorsque la structure est linéaire et, en plus, monter et descendre dans une structure arborescente. Avec un hypertexte qui permet de naviguer d'une page à l'autre, la navigation s'affranchit de la structure de l'ensemble des pages.

Les navigations hypertextuelles, qui s'affranchissent alors de la structure, ne

sont dès lors plus guidées ni par la structuration du document, ni par les commandes de l'interface, ni par la sémantique du contenu. D'un côté, un tel hypertexte permet à l'apprenti de s'adapter au discours de l'auteur en révisant par exemple ce l'auteur suppose être connu par l'apprenti, voire en allant voir des pages qui sont à étudier plus tard ; par exemple en allant visiter la page sur " la seconde guerre mondiale " à partir d'un texte qui se termine par " Les survivants de 1914-1918 ont alors juré que la Grande Guerre serait "la der des ders". Vingt ans plus tard, Hitler, déclenchait la seconde guerre mondiale ". D'un autre côté, il va de soi que l'apprenti peut être rapidement perdu.

Dans cet article, nous rapportons nos travaux sur la conception et le test d 'un nouveau type de module de navigation pour les manuels scolaires électroniques en intervenant sur les 3 niveaux que sont discours, structure et navigation. L'objectif est de fournir un mode de navigation propre à faciliter l'apprentissage et de rendre chaque navigation signifiante à tous les niveaux [REN 06].

Considérant qu'une part des navigations hypertextuelles est d'ordre sémantique, c'est-à-dire qu'elles relèvent de navigations conceptuelles, de concept en concept, l'idée est d'offrir à l'apprenti une navigation sémantique formalisée, c'est-à-dire dans un ensemble de pages structurées du point de vue de leurs relations sémantiques ; ici dans une ontologie. Notons que dans le cadre du document numérique, la navigation sémantique que nous proposons s'ajoute aux autres modes de navigation linéaire, arborescente et hypertextuelle lorsque c'est possible. Enfin, il s'agit d'une structure virtuelle, c'est-à-dire offrant la possibilité de parcours sémantiques indépendante de la structure visible du document à partir de l'interface.

L'HYPERMÉDIA ONTOLOGIQUE

L'ontologie comme organisation des connaissances

Pour donner du sens à la navigation supervisée dans un ensemble de connaissances, il faut organiser ces connaissances et pour organiser ces connaissances, il faut des critères d'organisation. Les travaux en psychologie développementale et cognitive [KEI 79 et 89], [TIJ 03], montrent que la classification ontologique est une méthode d'agencement des connaissances que l'on retrouve dans la cognition naturelle.

En effet, si l'ontologie, en philosophie, est l'étude de l'être en tant qu'être, les enfants sans vouloir faire de métaphysique demandent souvent " qu'est ce que c'est ? " ou " pourquoi ? ", ils s'informent sur leur environnement en questionnant sur les raisons d'être des choses. Dans son sens large,

l'ontologie est un type d'organisation qui possède de nombreux avantages vis-à-vis des autres classifications et autres taxonomies [DES 02]. La relation conceptuelle entre deux connaissances se fait par les articulations langagières de relation d'inclusion (est une sorte de) ou d'appartenance (est un cas particulier de). Elle permet d'organiser les connaissances sur un axe bijectif du général vers le particulier. Elle fonctionne par inclusion hiérarchique de catégories de nature sémantique (les oiseaux sont des animaux), ou empirique (les oiseaux sont des animaux qui font des nids) ; ce qui facilite l'intégration des connaissances et leur manipulation. Elle peut intégrer comme descripteurs des catégories les prédicats et arguments qui sont présents dans le discours. Enfin, cette organisation est évolutive et la compatibilité des ontologies entre elles peut être évaluée [POI 05].

La typologie des relations conceptuelles de Georges Vignaux

Les travaux de Georges Vignaux [VIG 03] concernent les relations entre les notions et leur genèse dans les articles scientifiques. Il a ainsi défini une typologie des relations conceptuelles qui permet d'organiser les concepts selon différents critères ontologiques qui contextualisent les connaissances.

Cette typologie est basée sur la distinction de six types de relations conceptuelles :

- Relation ONTOLOGIQUE stricte (vue avant)

- Relation de DÉFINITION ou de REDÉFINITION (réitération)

- Relation de COMPOSITION

- Relation d'ASSOCIATION

- Relation de DÉVELOPPEMENT/CONSÉQUENCE

- Relation d'OPPOSITION

De la typologie de Georges Vignaux, nous avons retenu les 3 types de relations généralisables aux notions des corpus scolaires et, de ce fait, nous obtenons un principe de navigation selon 3 axes d'organisation des connaissances :

- Le premier axe est dit " ontologique strict ". Il organise les connaissances en catégories, du général vers le particulier et inversement, les articulations langagières utilisées sont " est une sorte de " dans le sens ascendant, et " a comme cas particulier " dans le sens descendant.

- Le deuxième axe est dit " temporel ". Il organise les connaissances en cause et en conséquence. Les articulations langagières sont " est la cause de " et " est la conséquence de ".

- Le troisième axe montre la granularité. Il est dit axe de composition. Il organise les connaissances en fonction de leurs composants, les déplacements sur cet axe sont semblables à un changement de zoom, du plus grand au plus petit et vice-versa. Les articulations langagières sont " est composant de " et " est composé de ".

Chacun de ces axes est au moins bijectif, c'est-à-dire que ce qui est vrai dans un sens est vrai dans l'autre sens, ce qui donne à la fois cohérence et orientation à la navigation. Ainsi si X est une sorte de Y, alors Y a comme cas particulier X. Si X est la cause de Y, alors Y a pour conséquence X. Si X est composant de Y, alors Y est composé de X. Ces trois dimensions des relations entre connaissances permettent de structurer l'ensemble des notions, c'est-à-dire de relier une notion à une autre, à travers tout un texte ou toute une collection de textes.

La conception de la structure sémantique

La méthode a été appliquée pour mettre en relation les connaissances délivrées dans deux chapitres de Sciences de la Vie et de la Terre de niveau 4ème des éditions Bordas. La structure obtenue en utilisant la typologie des relations de G.Vignaux est le résultat de l'analyse de l'information sémantique contenue dans les pages. Cette structure est une représentation formelle des informations fournies par le discours de l'auteur. Elle représente également les notions et l'articulation des notions que doit acquérir un élève.

Considérons par exemple les énoncés :

- Le granit est une roche composée de cristaux qui se transforment en arène granitique

- L'argile est une roche composée de feldspath et de mica altéré

- Le granit est composé de feldspath et de mica

- Le feldspath et le mica sont des cristaux qui peuvent exister sous la forme cristalline et sous la forme altérée

Ces énoncés sont représentés sous forme d'ontologies locales et réduites, qui synthétisent les informations relatives aux concepts et surtout les relations interconceptuelles présentes dans chaque phrase (figure 1-gauche). Les ontologies locales sont alors intégrées jusqu'à obtenir une ontologie unique (figure 1-droite) qui contient toutes les notions du corpus à représenter. Le formalisme utilisé est celui du treillis de Galois [POI 05]. Le treillis complet correspondant aux deux chapitres du manuel scolaire de SVT 4ème est donné dans la figure 2.

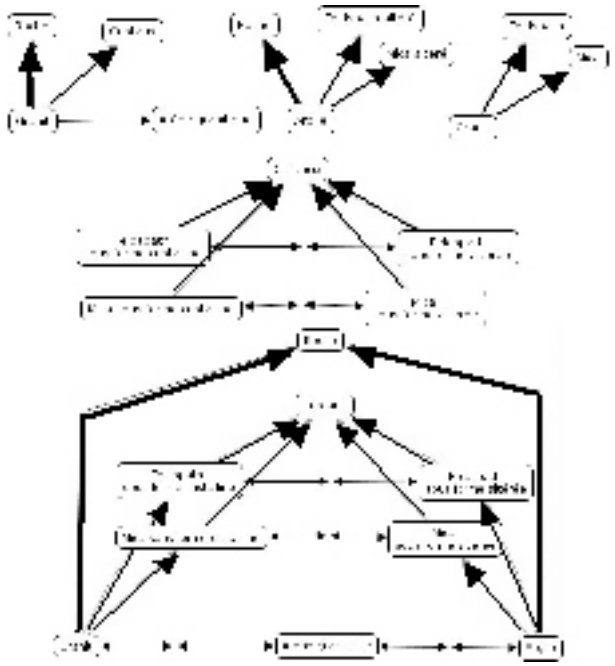

Fig. 1. Par convention, en trait fin sont indiqués les concepts reliés par une relation ontologique, en trait épais, ceux qui sont reliés par une relation de composition et, en trait très épais, ceux qui sont reliés par une relation de causalité et de conséquence. Le graph de droite correspond aux associations conceptuelles de chacun des 4 énoncés. Le graphe de gauche correspond au contenu des 4 énoncés.

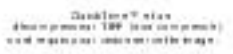

Fig. 2. L'ontologie des connaissances de deux chapitres d'un manuel de SVT 4ème.

La structure de la figure 2 a été utilisée comme " fond d'hypertexte ", c'est-à-dire que nous avons calqué l'organisation des unités d'affichage que sont les pages sur celle des connaissances. De la sorte, les liens et déplacements entre les différentes pages correspondent à des déplacements dans le treillis de la figure 2. Pour cela, nous avons sélectionné la page la plus représentative pour chacune des notions, représentée par un nœud du treillis, et nous lui avons attribué la place et les relations (liens hypertextes) correspondantes dans le treillis qui figure l'ontologie des connaissances des deux chapitres. De ce fait, la structure de la figure 2 représente à la fois l'organisation des connaissances et la structure des pages. Notre hypothèse est alors que les liens, et les déplacements que ceux-ci engendrent, sont explicites et porteurs de sens.

LA COMMANDE SÉMANTIQUE

La navigation conceptuelle

La navigation dans cet hypermédia " conceptuellement augmenté " s'opère de page-notion en page-notion via " la commande sémantique ". Les déplacements se font dans l'environnement conceptuel immédiat de la page-notion sur les 3 axes : " ontologique strict ", " temporel " et " granulaire ". Chaque déplacement a lieu sur une dimension et est orienté (avant-après, zoom avant-zoom arrière, monter-descendre), cette représentation s'inscrit dans une métaphore spatiale du fait de la linéarité de l'organisation des concepts sur les 3 axes (figure 3). Chaque page source est ainsi potentiellement liée de manière explicite avec 6 autres documents contenus dans les 6 pages et qui " représentent " son origine, son développement, sa catégorie super-ordonnée, ses cas particuliers, le concept qu'il compose et ses composants. De plus, chaque type de déplacement est caractéristique d'un type unique de relation logique et langagière, ce qui fait que chaque action correspond à une signification, d'où l'appellation de commande sémantique pour ce module de navigation, car l'action est en elle-même signifiante.

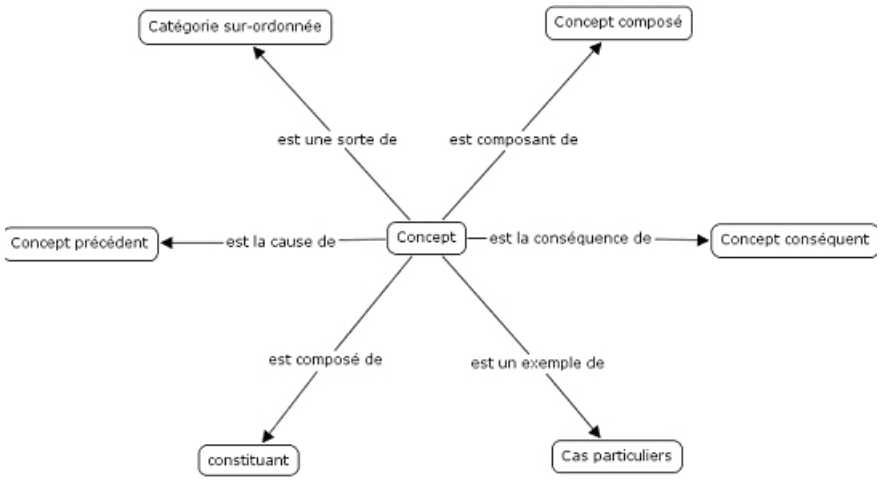

Fig. 3 : Les 6 déplacements possibles avec la commande sémantique

Ainsi, " la commande sémantique " utilise des articulations langagières orientées spatialement, ce qui rend chaque clic sur un lien, significatif sur 3 niveaux :

" Sémantique : les relations ontologiques sont sémantiques de par leur contenu et celui des documents.

" Logique : les relations ontologiques sont proprement logiques du fait de la catégorisation qu'elles proposent.

" Action : le sens de déplacement est doublement porteur de sens, il signifie à la fois un mouvement dans les connaissances et dans l'hypermédia.

Enfin, il y a une adéquation structurelle entre :
- " l'organisation des notions,
- " les documents à parcourir,
- " et la représentation de l'action,

Il s'ensuit qu'une seule représentation du fonctionnement de l'hypermédia est nécessaire ; ce qui facilite fortement l'apprentissage du fonctionnement du dispositif puisque les 3 niveaux de signification et les 3 types de représentation sont rendus explicites sous le même format qu'est " la commande sémantique ". Cette navigation est ainsi cohérente du point de vue ergonomique et significative du point de vue cognitif.

A chacun son hyper-parcours pédagogique

Les hypermédias offrent par nature plusieurs parcours possibles, donnant ainsi la liberté à l'utilisateur de choisir son document cible. Toutefois, d 'un point de vue pédagogique, l'utilisateur n'est guidé que dans un seul parcours, celui de l'auteur. Cette limitation est due à deux contraintes. La première contrainte est de nature structurelle. Le système d'organisation par arborescence empêchant les mises en relation des documents entre plusieurs branches, la navigation aux documents est montante et descendante. La seconde contrainte est que la raison d'être des manuels scolaires étant de fournir un support pour le professeur, l'objectif n'est pas de favoriser plusieurs parcours. Ainsi, le choix des documents, les intitulés des chapitres et les contenus sémantiques sont orientés par le discours de l'auteur et correspondent à un parcours qui n'est qu'un parcours parmi d'autres. En dehors de ce parcours, l'élève n'est plus guidé et ses parcours deviennent nettement moins pédagogiques.

Avec la commande sémantique et la représentation des connaissances sous forme d'ontologies, la navigation se fait dans un treillis de Galois qui permet de rendre compte de tous les types d'implications que possède un concept. L'accès à un document peut se faire de 6 manières avec la commande sémantique. Le document cible est nécessairement un document appartenant à l'environnement immédiat du concept source et possède une relation particulière avec lui. Les parcours sont ainsi constitués de successions de mises en relations entre un concept source et un concept cible, décidées par l'utilisateur. La navigation que propose la commande sémantique est alors véritablement hypertextuelle puisque l'élève, en utilisant la commande sémantique, est libre du choix de ses documents cibles comme dans tout

hypertexte. Il est également l'auteur de son propre parcours pédagogique dans un cadre qui fait que tous les parcours réalisés avec la commande sémantique sont pédagogiques.

EXPÉRIMENTATION

Le principe de navigation à partir de la commande sémantique, appliqué à deux chapitres du manuel scolaire de SVT 4ème a été testé auprès d'élèves de 5ème, nous assurant ainsi de la nouveauté du contenu pour les élèves.

PARTICIPANTS

Les participants sont 21 élèves de 5ème, recrutés dans un collège de l'ouest parisien avec l'accord parental. Les élèves ont tous une pratique de l'ordinateur.

MATÉRIEL

Les documents utilisés sont extraits des chapitres 1 et 2 d'un manuel hypermédia Sciences de la Vie et de la Terre 4ème. Les thèmes de ces 2 chapitres sont la variété des paysages et leurs évolutions, la notion cible sur laquelle porte le test est que l'argile est le résultat de transformations successives à partir du granit. A cette notion correspond la plus longue lignée conceptuelle de l'ontologie présentée dans la figure 2. Nous avons déterminé 3 conditions expérimentales d'observation :

- C1 : condition faite sur le manuel existant (condition contrôle)
- C2 : Condition faite sur le manuel avec la commande sémantique, version liens textuels
- C3 : condition faite sur le manuel avec la commande sémantique, version liens fléchés.

Les 3 conditions utilisent une même version allégée du manuel actuellement commercialisé. Sa structure est arborescente et nous avons supprimé les animations pour ne pas ralentir les recherches d'informations.

LA CONDITION CONTRÔLE

La condition contrôle a comme mode de navigation le clic sur les titres de la structure arborescente (figure 4). La navigation se fait ainsi de documents en documents selon l'organisation définie par l'auteur.

Fig. 4. Les différentes composantes de la navigation dans le manuel existant utilisé comme condition contrôle

Les deux manuels avec la commande sémantique

Les manuels sont ceux de la condition contrôle auxquels a été ajoutée la navigation par commande sémantique qui explicite la nature des relations entre les notions. Les liens sont de deux sortes : liens textuels pour la condition 2 (figure 5-haut) et liens fléchés pour la condition 3 (figure 5-bas).

Fig. 5: Exemples d'écrans du manuel électronique SVT 4ème pour la commande sémantique : liens textuels (figure du haut) et liens fléchés (figure du bas)

Procédure

Le niveau de connaissance sur le domaine des élèves a été évalué préalablement par un questionnaire. Les participants disposaient alors de 5 minutes pour se familiariser avec la navigation dans le manuel scolaire. L'expérimentateur vérifiait que les participants manipulaient correctement les fonctions de bases du navigateur. Cette phase de familiarisation était strictement la même pour tous les participants. Les élèves qui passaient les conditions avec l'une des deux commandes sémantiques étaient prévenus juste avant la passation de son existence. Il n'y avait pas de familiarisation pour la navigation par commande sémantique.

La tâche demandée à tous les participants consistait dans un premier temps à expliquer à l'expérimentateur comment se réalisait " la formation de l'argile " en faisant des recherches d'informations dans le manuel électronique. Les réponses, qui consistaient à fournir l'explication demandée, étaient données oralement et se précisaient au cours de la recherche. L'expérimentateur, en fonction des explications fournies, provoquait un conflit cognitif (mise en évidence de phénomènes inexpliqués, d'incohérence de l'explication). Cette méthode d'entretien critique permettait de relancer la recherche sans donner d'indication. Les élèves disposaient de 30 minutes pour faire la recherche et la passation s'arrêtait quand le participant estimait qu'il avait trouvé l'explication. Nous avons considéré comme correctes les réponses à cette question qui satisfont le contrôle scolaire. Dans un troisième temps, il était demandé aux participants s'ils avaient une idée du rapport entre l'argile et le granit.

Enfin, toutes les passations sont enregistrées sur vidéo pour le dépouillement des données et l'analyse (système Noldus installé au Laboratoire LUTIN).

Prédictions expérimentales

L'hypothèse générale est que les hypermédias ontologiques avec la commande sémantique facilitent la navigation, rendent plus efficace la recherche et améliorent l'apprentissage.

En raison de la forte cohérence et explicitation des liens entre les documents-notions :

" Le nombre de pages visitées non pertinentes devrait être réduit,
" le temps de recherche devrait être plus court,
" Le nombre de recherches réussies devrait être augmentées,
" L'apprentissage et la compréhension des interactions entre les concepts devraient être facilitées,
" La présentation par liens fléchés devrait avoir une influence bénéfique sur la compréhension.

Résultats

Le niveau de connaissance sur le domaine des élèves a été évalué préalablement par un questionnaire qui montre que les élèves n'ont pas de difficulté avec la notion de paysage et de ses composants. La majorité admet volontiers que les paysages subissent des évolutions. Par contre, ce n'est que très rarement qu'ils donnent une origine géologique aux évolutions. Les exemples de modification de paysage donnés par les élèves sont généralement des transformations faites par l'Homme. Les roches sont peu connues des élèves de 5ème. Ainsi, ils ne les distinguent pas en granit, en calcaire, ou en une autre roche mais plutôt en caillou, rocher ou pierre....

TAUX D'EXPLICATION DE LA FORMATION DE L'ARGILE

Tous les participants ayant utilisé la commande sémantique ont donné la réponse correcte, alors qu'avec le manuel existant le taux de réponse correcte n'est que de 50%. Dans le cadre de nos conditions expérimentales, en rendant plus certaine la recherche, la commande sémantique permet manifestement de trouver la réponse.

TAUX D'EXPLICATION DE L'ORIGINE GRANITIQUE DE L'ARGILE

Tous les participants ayant réussi à expliquer le rapport de l'argile et du granit sont des participants qui ont fait la recherche avec la commande sémantique. Le manuel existant à structure arborescente ne semble pas permettre de faire aisément ce rapprochement, tandis que la condition C2 le rend possible dans 25% des cas et la condition C3 permet dans 50% des passations la compréhension du lien entre les deux roches. La condition C2 permet des recherches d'informations plus rapides que la condition C3, mais cette dernière semble rendre plus explicites les relations entre les connaissances

NOMBRE DE PAGES VISITÉES

Le taux de pages revisitées est de 47 % avec le manuel arborescent tandis qu'il est de 31 % avec les manuels dont la navigation se fait via la commande sémantique. La navigation optimale pour expliquer la formation de l'argile nécessite un parcours sur 4 pages et la consultation d'un seul document. Le nombre de pages visitées avec le manuel arborescent est de 51 pages en moyenne, alors qu'avec la commande sémantique ce chiffre descend à 23 pages pour les deux conditions C2 et C3 confondus. La navigation dans les hypermédias ontologiques apparaît nettement plus efficace : elle réduit de plus de moitié le nombre de pages consultées.

TEMPS DE RÉSOLUTION

Les temps moyens de recherche sont de 7'21 pour C1 (condition contrôle), de 5'12 pour C2 (condition liens textuels) et de 6'18 pour C3 (condition liens fléchés). La commande sémantique rend plus fiable mais aussi plus rapide la recherche. Alors que les participants n'étaient pas familiarisés avec cette nouvelle navigation, puisqu'ils l'ont découverte au début de la passation, ils parviennent à une utilisation satisfaisante en 3 à 4 minutes. Le temps de consultation pour chaque page est de 5,6 secondes avec le manuel arborescent, ce temps de lecture augmente nettement avec l'hypermédia ontologique : 9,3 secondes. Ce résultat peut s'expliquer par la différence des taux de pages revisitées. La consultation d'une page varie de 10 à 30 secondes selon sa nouveauté et sa pertinence. Certaines pages ne sont pas ou plus lues : c'est le cas des pages revisitées et non pertinentes. Ce sont des pages de transit ou de passage : dès que le participant a reconnu qu'il s'agit de pages déjà vues et qui ne sont pas pertinentes, il continue sa navigation.

EFFETS DES LIENS ARBORESCENTS (C1)

Une lecture en survol est observable en condition C1. C'est, selon nous, une conséquence des organisations arborescentes. On peut voir ces comportements sur la figure 6 en J. Ce mode de lecture consiste à consulter les contenus des chapitres sans ouvrir les documents. Les participants consultent les vignettes et leurs intitulés sans aller plus en profondeur (la figure 4 en donne un exemple). Cette navigation est justifiée par une démarche rationnelle et systématique de la recherche. Elle donne au participant une fausse impression d'exhaustivité de la consultation du manuel alors même qu'elle reste superficielle du point de vue des notions et des documents. Les élèves qui sont incités à utiliser cette stratégie reproduisent à l'identique des cycles (de pages) de consultations ; ce qui montre une difficulté à se repérer et à avoir des repères dans l'hypermédia. Il s'agit d'une désorientation puisqu'il s'agit plus de difficultés dues à la navigation pour trouver ou retrouver les pages pertinentes que de cycles de vérifications nécessaires à la recherche d'informations complexes.

EFFETS DE LA COMMANDE SÉMANTIQUE À LIENS TEXTUELS (C2) À LIENS FLÉCHÉS (C3)

Nous avons vu que la commande sémantique avec liens textuels (C2) et liens fléchés (C3) permettent des recherches plus rapides et plus efficaces que la commande arborescente (C1). Il existe néanmoins des différences entre C2 et C3 : C2 offre une plus grande rapidité alors que C3 facilite davantage la compréhension. La condition C2 ne pose pas de problème d'utilisation car les participants sont habitués au système de liens textuels généralisé dans les hypermédias. On remarque toutefois, après analyse des enregistrements vidéos, que l'information sémantique fournie par C2 dans les liens textuels est un facteur important dans la restitution des explications. Ces liens textuels contiennent intrinsèquement l'information sous forme langagière. De ce fait, les élèves ayant à leur disposition le label des liens, ils peuvent plus facilement verbaliser les informations sur la formation de l'argile, sans que les relations entre les différentes notions soient nécessairement bien comprises. Ils peuvent ainsi facilement expliquer le rapport entre deux notions parce qu'ils ont " les mots pour le dire ". Ils sont toutefois plus en difficulté que les participants de C3 pour expliquer le phénomène de la formation de l'argile dans sa globalité. S'ils voient bien les liens entre les concepts établis par les relations, ils ne voient pas les liens entre couples de concepts. Ces derniers, sous forme de liens contextuels, sont présentés en liste dans la condition 2 (voir figure 5). Cette présentation ne permet pas de rendre compte visuellement de l'organisation sur 3 axes qui existe entre les " couples conceptuels " ; ce qui a pour conséquence de limiter la taille des regroupements conceptuels en général à deux associations et de provoquer des erreurs de navigation dues à la non-orientation des liens textuels.

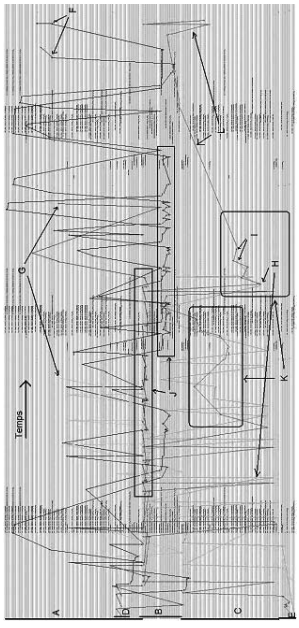

Fig. 8. Exemple de traces des parcours analysées avec le logiciel d'analyse des traces LogViz [DAM 06] qui permet de comparer plusieurs traces de conditions différentes (ici 7 traces des conditions C1 et C2) et d'observer les similitudes et différences dans les navigations des participants. LogViz a été utilisé pour analyser l'influence et la facilitation qu'apporte la commande sémantique sur la recherche d'information et la compréhension. Le temps est représenté horizontalement et chaque ligne horizontale représente une page du manuel. A : Pages de bas niveau correspondant à des documents (condition C1). B : Pages de haut niveau dans l'arborescence (têtes de chapitre et de sous-chapitre équivalent à la figure 4). C : Pages de bas niveau correspondant à des documents (condition C2). D : Page de démarrage pour C1. E : Page e démarrage pour C2. F : Page menant à la réussite C1. G : Page menant à une erreur (impasse). H : Page menant à la réussite C2 ET C3. I : Manipulations pour expliquer le rapport avec le granit. J : Lecture en survol. K : Episode de l'utilisation de la commande sémantique. L : Résidu logiciel de trace non coupé en fin de passation.

Discussion-conclusion

Les expérimentations sur un corpus de deux chapitres d'un manuel SVT de 4ème ont montré la pertinence et la puissance de la commande sémantique en tant qu'outil pédagogique et cognitif :

" Elle propose et permet un formalisme efficace pour délimiter les concepts en jeu (calibration conceptuelle) ;

" En permettant de visualiser, avec l'orientation des axes, l'environnement conceptuel, elle favorise la catégorisation des connaissances et par là, contribue à les structurer et à les comprendre. Le découpage standardisé qu'elle offre, facilite en effet, les manipulations langagières et le raisonnement centré sur les concepts.

Deux conséquences majeures en découlent :

" La conceptualisation se traduit en termes de spatialisation permettant à la fois une représentation en 3D (selon les axes) et une vue d'ensemble des connaissances organisées, mettant en évidence les lacunes à combler et les gradations à respecter (cartographie conceptuelle) ;

" Les parcours personnels sont d'autant favorisés que ce guidage demeure à tout moment disponible, appuyant la navigation dans les connaissances. Il s'agit, en quelque sorte, d'une navigation réfléchie :

on peut conclure qu'avec un tel dispositif, l'élève est acteur du manuel.

Nos résultats montrent que la commande sémantique à partir de l'hypermédia ontologique permet des performances en recherche et en compréhension qui sont nettement supérieures aux performances observées avec la commande hiérarchique. Néanmoins, au regard des résultats et des verbalisations obtenus au cours des passations, il apparaît que le module de navigation peut être optimisé en termes de présentation.

De ce fait, une autre version de la commande sémantique est en cours de réalisation. Cette version présente les liens à la fois de manière fléchée et textuelle. De la sorte, elle bénéficiera de la cumulation des effets observés en C2, avec la commande sémantique textuelle, qui offre une l'information langagière, et des effets observés en C3, avec la commande sémantique fléchée, qui offre une présentation schématique qui rend explicite les relations logiques, les déplacements et réduit la désorientation.

Références bibliographiques

[BOU 04] Bouchon-Meunier B., Tijus C., Demarcy C., Leproux C., Poitrenaud S., Renaud S., Giraudon, V. & De Vulpillière T. " Conception guidée utilisateur et aides aux apprentissages ", Actes des journées scientifiques du réseau des sciences cognitives d'Ile de France, 2004.

[DAM 06] Damez M., " Méthode de récolte de traces de navigation sur interface graphique et visualisation de parcours ". Article pour démonstration de logiciel, EGC 2006

[DES 02] Desmoulins C. & Grandbastien M., " Des ontologies pour la conception de manuels de formation à partir de documents techniques ", Sciences et techniques éducatives, n°3/4, Vol 9, 2002, pp 291-340.

[GRI 02] Grize J.B., Logique et langage, Paris, Ophrys, 2002.

[KEI 89] Keil F.C., Concepts, kinds and cognitive development, Cambridge, ma, the MIT press, 1989.

[KEI 79] Keil F.C. Semantic and conceptal development : an ontological perpective, Cambridge, Cognitive sciences series, 1979.

[PIA 87] Piaget, J., & Garcia, R. Vers une logique des significations. Murionde, Genève, 1987.

[POI 05] Poitrenaud, S., Richard, J.F., & Tijus, C.A., " Properties, Categories and Categorization". Thinking and Reasoning, 11, 2005, p151-208.

[REN 06] Renaud S., Conception d'hypermédia ontologique pour une navigation conceptuelle dans le cartable électronique, 2006.
http://plate-forme-ast.mshparisnord.org/Conception-d-hypermédia.

[TIJ 03] Tijus, C. & Cordier, F., " Psychologie de la connaissance des objets : catégories et propriétés, tâches et domaines d'investigation ".
L'année Psychologique, 103, 2, 2003, p 87-120.

[VIG 05] Vignaux G., Construire le sens : catégories, frontières, ajustements. Québec, Presses de l'Université Laval, 2005.

[VIG 03] Vignaux G., Du signe au virtuel : les nouveaux chemins de nos intelligences, Paris, Seuil, 2003.

Session 03

Production collaborative de documents et partage de connaissances

Maintien de la cohérence des intentions de communication dans la rédaction coopérative

Saïd TAZI
Khalil DRIRA
Kamal ESSAJIDI

LAAS-CNRS, 7, Avenue Colonel Roche, 31077 Toulouse Cedex 4
{tazi,drira,kamal}@laas.fr
Université des Sciences Sociales Toulouse I

RÉSUMÉ

Dans un contexte de production de documentation technique ou pédagogique, l'utilisation de systèmes de rédaction coopérative est confrontée au défi de produire des documents cohérents ; c'est-à-dire, dans lesquels il n'y a ni omission, ni redondance ni contradiction. Dans cet article nous présentons un modèle de l'intention qui semble être approprié pour rendre l'édition coopérative plus efficace dans ce contexte. Nous démontrons comment l'intention, exprimée sous forme de metadonnées, permet de partager des informations utiles pour aider à maintenir cette cohérence. Dans cet article nous présentons également l'architecture et les fonctionnalités de XSEdit, un système de rédaction coopérative, basé sur ce modèle, et développé à cet effet. Cette approche permet aux auteurs de maintenir la conformité des leurs actions par rapport à des intentions collectives prédéfinies en utilisant un ensemble de règles.

MOTS-CLES : Intention, Métadonnées, annotation, rédaction coopérative, systèmes répartis.

1. INTRODUCTION

L'édition coopérative est le processus de développement de documents par un groupe des personnes de façon simultanée ou de manière asynchrone. L'utilisation de systèmes synchrones d'édition coopérative permet aux auteurs de rédiger des documents sans être nécessairement réunis géographiquement. Cependant, il arrive que dans le document produit final, on trouve des informations manquantes, redondantes ou même contradictoires. Ces systèmes ne permettent pas aux auteurs d'éviter ces incohérences sans avoir à multiplier les corrections et les révisions. Les recherches présentées ici sont appliquées sur des documents procéduraux comme les supports de cours ou la documentation technique. Si dans les documents développés par les Wikis, on peut permettre d'avoir des idées redondantes, manquantes ou même contradictoires, on ne peut pas permettre ces incohérences dans des documents techniques ou pédagogiques. En effet, selon [BURROW04], un

Wiki fournit un modèle pour la collaboration, parce qu'il enlève tous les obstacles au partage des idées. Cependant, ces systèmes favorisent la communication, la dialectique, le partage et la diffusion plutôt que la rédaction ou la production de documents. Les incohérences, telles que les omissions, les redondances ou les contradictions dans les documents de type " Web ", sont parfois admises parce qu'elles peuvent être source de créativité par la dialectique. En revanche dans un contexte professionnel où les facteurs de productivité, et parfois de sécurité, sont importants, ces incohérences sont indésirables et peuvent causer des situations catastrophiques. Par exemple, si le pilote d'un avion, lors d'une situation urgente trouve des informations manquantes dans un document embarqué, les conséquences peuvent être non souhaitables.

L'idée principale de nos recherches est de permettre aux auteurs d'exprimer les intentions de communication relatives à différentes parties du document, appelées dans la suite fragments, d'une manière telle qu'elles permettent d'améliorer la recherche et la réutilisation de ces fragments par les intentions. Le travail présenté dans cet article vise particulièrement à aider les co-auteurs à maintenir la cohérence de leur document par rapport aux intentions de communication de ces fragments.

Par intention, nous entendons ici ce que souhaitent obtenir les auteurs comme effet sur leurs lecteurs. Dans un contexte de documents procéduraux, pédagogiques ou techniques, où la lecture doit être faite de manière perspicace et déterministe, les lecteurs peuvent se poser des questions concernant l'intention des auteurs, par exemple, un lecteur peut se poser les questions suivantes : " Que veut l'auteur que je fasse, que je comprenne ou que j'apprenne par ce passage ? " ou " Où sont les parties des documents où les auteurs argumentent, expliquent ou démontrent une idée ? ". Or les informations permettant de répondre à ces questions peuvent parfois être implicites ou inexistantes dans le document. Par exemple, l'intention d'expliquer une idée n'est pas toujours explicitement mise par écrit ; seulement l'interprétation du texte permet la compréhension qu'il s'agit d'une explication. Une description explicite des intentions permet leur découverte sans effort d'interprétation ; en outre, elle permet de faire des traitements informatisés de ces intentions. Dans l'état actuel de notre recherche, la description des intentions est ajoutée manuellement par les auteurs comme annotations en tant que métadonnées appropriées. Un modèle de l'intention individuelle et collective a été défini, il est utilisé comme un cadre conceptuel pour le développement de XSEdit, un éditeur coopératif permettant aux auteurs de rédiger, de modifier et d'annoter simultanément leurs documents. Chaque auteur annote sa partie, constitué d'un ou plusieurs fragments en ajoutant des métadonnées pour décrire ses intentions concernant chaque fragment. Un des buts de XSEdit est de permettre aux auteurs de maintenir la cohérence des intentions des auteurs en les guidant pour éviter, les répétitions et les manques d'information par rapport aux intentions. Cette vérification se fait grâce à un ensemble de règles prédéfinies et décrites pour l'ensemble du

document. La vérification se fait sur les annotations des intentions et non sur le contenu.

Dans la section suivante, le problème de l'édition coopérative sera exposé en comparant notre approche aux travaux existants. La section 3 est consacrée au modèle d'intention. La section 4 présente l'outil XSEdit, son architecture et ses fonctionnalités, elle est suivie d'une conclusion.

2. L'ÉDITION COOPÉRATIVE

Les avantages de l'édition coopérative sont multiples. Le plus visible est probablement la possibilité à un groupe de travailler sur un document partagé avec des machines distribuées géographiquement. En utilisant des systèmes d'édition coopérative, les auteurs peuvent rédiger des documents sans être nécessairement réunis géographiquement. Toutefois, la rédaction coopérative est une tâche difficile. Il n'est pas suffisant d'apprendre comment écrire ensemble [MITCHELL95], mais il est nécessaire au départ de partager la motivation et avoir les instruments qui soutiennent la communication, la coordination et la coopération. Dans [CERRATO99], l'auteur a montré que les logiciels d'écriture collaborative peuvent jouer un rôle de diffraction de prisme au lieu de soutenir la convergence de la coopération. Cela veut dire que les systèmes peuvent jouer le rôle inverse de celui pour lequel ils sont, ces systèmes permettent aux auteurs de produire simultanément des documents dont l'objectif peut être divergeant. La divergence des idées ou l'inconsistance des concepts peuvent être tolérés pour permettre une dialectique favorisant la créativité, c'est-à-dire l'émergence de nouvelles idées dans des documents généraux. En revanche, dans les documents professionnels à vocation technique ou pédagogique les intentions doivent être convergentes de manière à rendre la lecture efficace pour des raisons économiques ou de sécurité. Pour cette raison il est essentiel d'avoir un protocole de rédaction coopérative, d'envisager une gestion de la coopération qui peut consister en la définition d'un auteur privilégié qui joue le rôle d'administrateur du document et d'un protocole. Ce protocole permet de définir des règles de coordination des structures souhaitées qui rendent la vérification de la cohérence du document final possible. Les travaux de recherche sur les systèmes distribués et les collecticiels sont également concernés par l'objectif des travaux présentés ici [SUN98]. Toutefois, la plupart des recherches existantes s'occupent de la maintenance de la cohérence syntaxique, c'est-à-dire que seules la synchronisation et la sérialisation des opérations parallèles sont étudiées sans tenir compte de la cohérence sémantique, [HONG05], [DOURISH96], [ELLIS90], [GREENBERG94]. Par exemple, l'effet syntaxique d'une opération d'édition de texte est simplement insérer/effacer des caractères aux positions spécifiques d'un document de texte considéré comme une séquence de caractères. Dans [SUN02] les auteurs font la distinction entre la cohérence syntaxique et la cohérence sémantique. Cette dernière pourrait être " insérer/effacer " une phrase en français dans un document en gardant la cohérence du texte.

La notion de cohérence dont nous nous occupons se distingue de la cohérence syntaxique et sémantique. Nous voulons garantir le suivi du document rédigé par plusieurs auteurs, c'est-à-dire éviter les omissions les redondances. Le contrôle de cohérence porte sur les annotations décrivant les intentions. Les auteurs annotent leurs documents en ajoutant des métadonnées qui décrivent l'intention de ce qu'ils écrivent. Le système vérifie si toutes les intentions conjointes sont concordantes avec l'intention du groupe (appelé intention collective). Le système ne vérifie pas la sémantique de ce qui est écrit, mais il vérifie la cohérence des annotations. Le système XSEdit compare des annotations par rapport à un ensemble de règles qui décrivent les intentions conjointes du groupe.

La notion d'intention est utilisée dans les champs multiples comme le modèle BDI (Believe, Desire Intention) dans les systèmes Multi-agents [FERGUSON95]. Elle est aussi utilisée par la recherche sur la maintenance de la cohérence syntaxique dans les systèmes de rédaction coopérative [SUN02], [SUN98], [XUE03].

3. LES INTENTIONS DE COMMUNICATION ÉCRITE

3.1 L'intention individuelle

La notion d'intention qui nous intéresse est celle que l'auteur a en tête quand il planifie, quand il fait la première ébauche, quand il rédige ou quand il révise son document. Nous supposons que tous les auteurs se demandent au moins les questions suivantes, que nous appellerons les questions de l'écriture : " A qui est destinée cette partie du document ? ", " Qu'est ce qu'on veut dire/accomplir en écrivant ? ", " Pourquoi on veut écrire cela ? ", " Comment formuler une proposition ? ", " Pourquoi écrire de cette manière plutôt que d'une autre ? ". Nous estimons que l'efficacité de communication écrite dépend des réponses que les auteurs donnent à leurs questions d'écriture. Le modèle d'intention que nous avons développé est basé sur une notation de prédicats de la logique du premier ordre [TAZI01].

Selon la théorie des actes de discours, le concept d'acte de discours peut être appliqué non seulement à la communication parlée, mais également à la communication écrite, voir [AUSTIN62] et [SEARL69]. Un auteur comme un orateur, a un ensemble d'effets qu'il veut produire chez le destinataire. Quand un auteur écrit, il réalise (ou essaie de réaliser) un ensemble d'actes qui expriment simultanément :

- l'acte d'écriture, par exemple l'utilisation du clavier, ou un stylo ;
- les buts de cette action, par exemple " publier des résultats scientifiques dans un article de journal ", ou " argumenter pour défendre une idée ", etc. ;

- l'acte d'utiliser un moyen qui réalise l'action d'écriture, par exemple le choix

de la langue, le médium, ou le style linguistique ; nous appelons cet acte " Moyen " et ;
- un acte qui peut exprimer la raison de l'action et argumenter sur l'acte d'écriture.

L'intention dépend aussi du destinataire de la communication par exemple à qui l'écriture est destinée. La forme générique suivante a été adoptée pour représenter une intention :

Intention (Agent1, Agent2, Action, But, Moyen, Raison)

Où :

Agent1 : est l'auteur de l'action ;
Agent2 : est l'agent à qui l'action est destinée ; généralement c'est le lecteur ;

Action : est un acte qui exprime ce que l'auteur de l'intention veut réaliser ;

But : est un acte qui exprime ce que l'auteur veut faire/atteindre en exécutant l'action ;

Moyen : est l'expression des moyens utilisés pour exécuter l'action. Il est représenté comme un acte qui exprime le type des moyens utilisés pour accomplir l'acte d'écriture ;

Raison : est un acte qui exprime pourquoi l'auteur réalise l'action afin d'argumenter et justifier l'acte d'écriture.

3.2 Intention de groupe

La notion d'intention de groupe peut être étudiée selon l'un des points de vue social ou philosophique. Raimo Tuomela est un des philosophes qui ont étudié l'intention de groupe dans la perspective d'offrir un cadre conceptuel pour comprendre cette notion cognitive [TUOMELA06]. Il distingue entre l'Intention Conjointe et l'intention collective qu'il appelle " We-intention ". La We-intention est l'intention qu'un groupe de personnes peut formuler comme intention propre au groupe qui se réalise par l'exécution d'un ensemble d'actions. Les Intentions Conjointes sont des intentions individuelles réalisées par des actions complémentaires qui, si elles sont réunies, exécutent l'action collective voulue. Les intentions conjointes peuvent être illustrées par "Deux auteurs A et B souhaitent conjointement réaliser les actions X par A et Y par B, dans le but de réaliser l'action collective Z". Quand les gens écrivent ensemble le même document, la collaboration peut être l'une ou la combinaison des trois situations suivantes : (1) Les auteurs écrivent sans aucune coordination ; chaque auteur écrit sa propre partie du document partagé comme s'il écrivait seul. Dans cette première situation si aucune coordination n'est faite, le résultat peut être un document où les idées sont divergentes, et pire, elles peuvent être contradictoires ou incongrues.

(2) Les gens essaient de coordonner en communiquant leurs intentions et des plans. Le problème avec cette deuxième situation est que les auteurs peuvent passer beaucoup de leur temps dans la communication essayant de définir ou coordonner leurs actions d'écriture au lieu de les réaliser et d'augmenter la production (3) Les gens essaient de collaborer conformément aux recommandations et conseils d'un auteur privilégié considéré comme un coordinateur de la rédaction. Pourtant, même s'il y a un coordinateur de document, chaque auteur peut continuer à écrire comme s'il est seul, parce que les habitudes sont souvent difficiles à changer. Le résultat peut être comme la concaténation des différents fragments. Dans ce cas-là nous estimons que le coordinateur du document peut prendre beaucoup de temps dans la reconsidération et les nouvelles recommandations des changements aux auteurs. Les auteurs ont besoin des moyens robustes de communication et des mécanismes de contrôle basés sur des règles pour aider dans le processus de coordination de l'écriture coopérative [CERRATO99].

Quand des auteurs décident d'écrire ensemble un même document, ils doivent être coordonnés par une personne considérée comme l'administrateur de document. Nous supposons alors qu'il y a un auteur privilégié qui joue ce rôle. Il doit définir le but global du document et discute avec les autres auteurs leurs rôles respectifs et définit à priori leurs intentions conjointes (c'est-à-dire l'ensemble des intentions individuelles qui permettent de réaliser l'intention collective). Il définit aussi un ensemble de règles qui délimite la frontière des rôles des auteurs. Cela signifie par exemple que l'auteur A1 n'a pas à définir un concept dans l'introduction, puisque cette définition doit être faite par l'auteur A2 dans une autre section, et dans ce cas A2 ne doit pas oublier de faire cette définition. Ces contraintes sont formulées grâce à une interface spécifique et sont traduites comme des règles logiques que le système utilise pour vérifier la cohérence du document.

Le modèle d'intention de groupe que nous proposons ici est fondé sur le modèle d'intention individuelle défini ci-dessus et une adaptation du modèle de Tuomela [TUOMELA06]. Pour un document écrit par N auteurs A1, A2, … AN ; les intentions conjointes conformes aux différentes actions faites par les différents auteurs sont :

Intention (A1, Lecteur, Action1, But1, Moyen1, Raison1)
Intention (A2, Lecteur, Action2, But2, Moyen 2, Raison 2)
Intention(AN, Lecteur, ActionN, ButN, Moyen N, Raison N)

Ces intentions doivent vérifier un ensemble de règles qui expriment la consistance du document entier. Une règle est un prédicat du premier ordre qui exprime les contraintes pour que les intentions conjointes ne soient pas en conflit avec l'intention collective.

Les règles sont définies par rapport à la structure logique du document. Chaque nœud de l'arborescence du document est attribué à un auteur ; un ensemble de règles sont alors attribuées au nœud pour que l'auteur respecte

les intentions collectives. Supposons que la section S est attribué à l'auteur Ai pour sa rédaction et son édition, une règle par exemple peut stipuler que l'auteur n'est pas autorisé à développer les concepts définis dans cette section, la règle stipulera alors que l'intention de " développer " n'est pas autorisée. L'ensemble des actions intervenants dans l'expression des intentions ont été définies comme une ontologie des intentions, de manière que tous les auteurs parlent le même langage, et de manière à pouvoir trouver des équivalences synonymiques, par exemple.

3.3 Les règles

Chaque règle a la structure suivante :

Règle : ÉLÉMENT, AUTEUR
ACTIONS PERMISES
ACTIONS INTERDITES

L'élément spécifie un chemin décrivant dans la structure logique l'élément dont l'auteur a la responsabilité de rédaction. Les intentions seront explicitées par l'auteur relativement à différents fragments de cet élément ; Les actions sont des listes de termes que l'auteur a le droit d'utiliser ou ne pas utiliser pour exprimer l'intention des fragments. Des opérateurs logiques sont utilisés pour pouvoir combiner des termes. On utilise les opérateurs de conjonction noté ET, et de disjonction noté OR ainsi que la négation noté ^. Le mécanisme d'édition des règles a été implémenté dans XSEdit selon l'interface de la figure 5. Les règles sont implémentées en XML, Voir la figure 6, pour un exemple de représentation de ces règles avec xml. L'intention collective est exprimée en termes de ces intentions conjointes et des règles qui les soutiennent.

4. XSEDIT

4.1 Les fonctionnalités de XSEdit

XSEdit (Xml Shared Editor) est un système de rédaction coopérative qui a été implémenté pour répondre aux besoins suivants :

- L'éditeur doit implémenter les fonctionnalités de traitement de texte élémentaires sur des documents xml, selon le principe du WYSIWG ;

- Il doit permettre d'insérer des métadonnées à n'importe quel endroit du document pour permettre aux co-auteurs d'exprimer leurs intentions ;

- Il doit permettre de contrôler les métadonnées par des règles prédéfinies ;

- Il doit implémenter le modèle d'édition coopérative, qui consiste à gérer les

versions des documents, la synchronisation du contenu entre les différents collaborateurs, ainsi que la gestion des modifications concurrentielles ;

- Il doit permettre à un auteur privilégié d'éditer des règles de contrôle de cohérence des annotations ;

- Il doit stocker les documents dans une base de donnés qui offre un moyen d'interrogation similaire au langage SQL.

4.2 L'architecture de XSEdit

XSEdit permet à un ou plusieurs utilisateurs d'éditer et annoter un document et de partager son contenu, ainsi que les métadonnées entre des collaborateurs. XSEdit est implémenté en Java suivant l'architecture hybride Client/Server et Producteur/Consommateur pour la diffusion. Il se fonde sur le service d'événement JSDT de SUN.

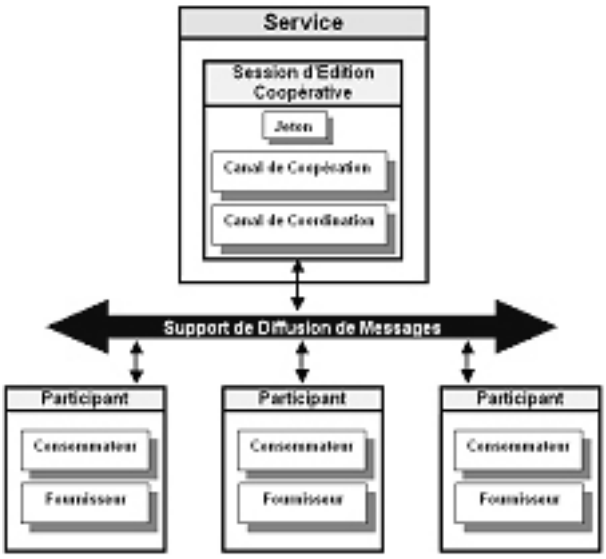

Figure 1 : Architecture de XSEdit.

Il est possible aux utilisateurs éloignés de définir et joindre des sessions coopératives. À chaque session coopérative on définit des canaux de communication selon l'architecture de la figure 1.
XSEdit traite des documents structurés dans le format xml. Chaque document est analysé avec l'API SAX [SAX04] sur la machine cliente. SAX

offre un ensemble d'opérations d'édition donnant la possibilité d'accès à chaque élément du document et ses métadonnées, pour les éditer ou en créer de nouveaux. Dès que les opérations d'édition sont réalisées, le document peut être conservé localement sur la machine de l'utilisateur ou sur une base de données native xml d'un serveur central. Le serveur garantit la connexion entre les clients et le partage du contenu et des métadonnées, en notifiant des mises à jour de modification aux coauteurs. Le modèle de coordination implémenté dans XSEdit est le contrôle par passage de privilège sur le document entier alloué seulement à un seul utilisateur à la fois. La possibilité de verrouiller un document (ou une partie du document) repose sur un mécanisme que nous avons adopté pour sérialiser les opérations d'accès à ce document. En utilisant des mécanismes de gestion de droit de parole ("floor control"), XSEdit garantit la sécurité aux coauteurs. A tout moment, si un auteur écrit dans un document alors on peut être sûr que le reste du groupe lit seulement cette partie du document et ne peut pas la modifier. Un mécanisme de mise en conscience des coauteurs a été implémenté en XSEdit par l'envoi de messages textes et des boîtes de dialogue. Ceci permet d'avertir un auteur des modifications faites par un autre auteur sur un élément partagé.

4.3 Fonctionnalités de XSEdit

L'API Swing de Java a permis de construire des fonctionnalités fondamentales de l'éditeur, qui sont : le chargement, la visualisation, la révision, l'annotation, le contrôle du partage et la sauvegarde d'un document ou d'une partie d'un document. Comme représenté dans la figure 2, le client permet la création d'un document localement ou sur une base de données xml sur le serveur. Nous avons utilisé Exist comme système de gestion de base de données native xml [EXIST06].

Chaque document est accompagné de sa feuille de style écrite en XSLT, cette feuille détermine comment chaque nœud de l'arborescent doit traité par les auteurs.

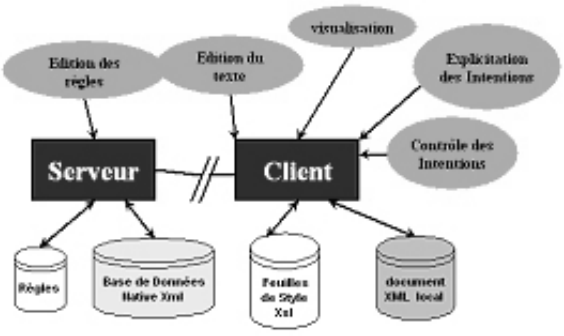

Figure 2 : Fonctionnalités de XSEdit.

L'originalité de XSEdit est due à son module d'annotation des intentions (représenté dans le schéma de la figure 2 par " explicitation des intentions ". L'annotation se fait interactivement, en sélectionnant un fragment du document, l'auteur choisit un ensemble de mots provenant d'une ontologie pour expliciter son intention, si par exemple un auteur souhaite expliciter qu'un tel fragment a pour objectif d'expliquer un concept, qui aurait été défini dans le document, le système ajoute des métadonnées exprimant cette intention. Si un des moyens de cette explication est un exemple que l'auteur a choisi, il l'explicite de cette manière, et si la raison de l'explication est de rendre le concept plus clair, il peut le mentionner à travers ce module d'annotation. Les modules d'édition et de visualisation s'exécutent de manière classique en mode WYSIWYG. Les documents peuvent être restitués en PDF, HTML ou XML. La figure 3 représente l'interface principale de XSEdit, où on peut voir la région d'édition et la région où l'annotation peut être éditée.

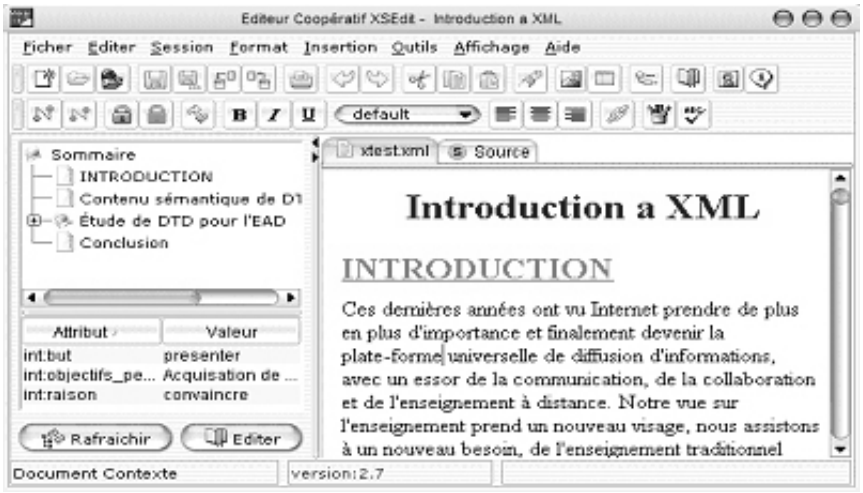

Figure 3 : Interface principale de XSEdit

La seconde originalité de XSEdit est due à son module d'édition des règles pour vérifier la cohérence du document. Quand un ensemble de règles est défini et associé à un document, une feuille de style XSLT est produite pour accompagner le document chez tous les clients. Si un auteur annote le document (sur la machine cliente) le contrôle de cohérence est exécuté selon cette feuille de style sur la machine cliente.

5. MÉTADONNÉES POUR LES INTENTIONS

Nous appelons " Explicitation des Intentions " (EI), le processus par lequel un auteur associe à ses fragments de texte, des descripteurs sous forme d'an-

notations. L'EI est une activité supplémentaire de l'auteur, qui peut augmenter le temps normalement investi pour l'écrit.

Le processus d'Explicitation des Intentions (EI) est une fonctionnalité principale de XSEdit. Il peut être accompli sur les documents existants ou sur les documents en création. Dans les deux cas, la description est faite par l'insertion de balises xml et d'attributs/valeurs dans le document. Dans le cas de description de documents existants, l'auteur est invité d'abord à choisir une portion de texte (le fragment) et choisir l'action et ainsi que les autres éléments constituant l'intention comme le but, la moyen et la raison afin de décrire l'intention associée au fragment. La description d'une intention peut être créée, insérée, modifiée, et/ou effacée. La figure 4 montre comment est représentée l'intention " Informer " dont le but est " Définir le concept du paragraphe ". On remarque l'utilisation d'un espace de nom " xmlns:int " et l'utilisation de métadonnées du standard Dublin Core (DC) Les métadonnées définissant les intentions sont stockées sous forme d'ontologie des intentions. Cette ontologie a été modélisée par rapport au domaine d'application. Nous avons défini une ontologie des intentions pédagogiques à utiliser dans les supports de cours [ALTAWKI02]. Une étude précédente nous a permis de stipuler les rudiments pour la construction d'une ontologie dans le domaine de l'aviation civile, concernant les FCOM (Flight Crew Operating Manuals) qui sont des documents embarqués dans les avions de type A320 [NOVICK98].

```
<?xml version="1.0" encoding="UTF-8" ?>
<Document xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:int="http://www.laas.fr/xsedit/int/elements/1.0/">
<docinfo>
 <dc:title> My first document </dc:title>
  <int:version>1.0</int:version>
  <int:dateversion>2004-04-23</int:dateversion>
  <int:datecreation>2004-04-23</int:datecreation>
 </docinfo>
<chapter>
  <dc:creator>Kamal Essajidi</dc:creator>
  <dc:title>INTRODUCTION</dc:title>
 <int:action>informer</int:action>
  <int:but>definir</int:but>
<para>Le concept ici . . . </para>
</chapter>
 </Document>
```

Figure 4 : Représentation du document avec les intentions en xml

5.1 Expression des règles de cohérence

Nous supposons que, pendant la création d'un document, un auteur privilégié définit une structure de base, qui sera le point de départ pour la création du document. Cette structure définit, d'une part la structure logique du document et, d'autre part, un ensemble de règles par auteur et par nœud associé.

Cette structure de base contiendra principalement les environnements que l'auteur considérera comme significatif pour le contrôle. Pour chaque nœud nous définissons un ensemble de règles pour contrôler les annotations. La figure 6 présente une règle simplifiée qui est appliquée à la première section sous le premier chapitre. Par exemple, chaque fois que l'auteur annote un segment comme étant une introduction (l'action de l'intention est " introduire "), alors il n'a pas le droit de le développer néanmoins dans cette section (l'action de l'intention doit être différente de " développer ").

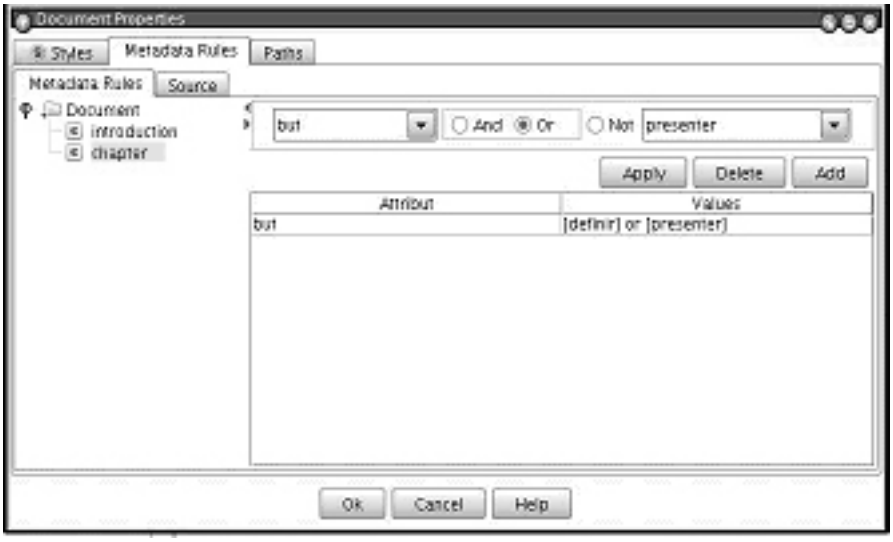

Figure 5 : Edition d'une règle

Pour chaque règle on définit un ensemble d'éléments :

id : l'identification de la règle ;
elementpath : le chemin (en syntaxe XPATH) de l'élément du document qui sera contrôlé par cette règle.
metadata : contient toutes les annotations permises pour l'élément défini dans elementpath ainsi que les règles pour les contrôler. Chaque règle se compose d'un ensemble des valeurs entre [ ] et contient des opérateurs logiques, AND, OR et l'opérateur de négation dénoté ^.

```
<?xml version="1.0" encoding="UTF-8"?>
<rules xmlns:int="http://www.laas.fr/namespaces/int/">
<rule>
<id>1</id>
<elementpath>/Document/chapter[1]/section[1] </elementpath>
<metadata>
  <int:action> [introduire]and ^[ développer] </int:action>
  <!-- ...  -->
</metadata>
</rule>
   <!-- etc...  -->
</rules>
```

Figure 6 : Une règle en xml.

Par exemple la règle " [introduire] and ^[développer] " implique que les annotations peuvent se faire avec l'action " introduire " et que n'importe quelle autre intention doit être différente de " développer " dans cette section. Par ailleurs, nous supposons que " l'Explicitation des Intentions ", c.-à-d., l'expression de l'intention par des annotations et le contenu du texte sont concordants. Nous ne vérifions pas si l'annotation correspond (ou ne correspond pas) au texte annoté, notre intérêt a été porté seulement sur la vérification de la cohérence des annotations des intentions avec celles qui sont indiqués par les autres auteurs du document. La vérification si l'annotation correspond au contenu n'est pas traitée dans l'état actuel de cette recherche.

5.2. Transformation de règles en feuille de style XSLT

L'implantation du module de contrôle peut être faite selon deux techniques possibles :

- La première consiste à implémenter l'algorithme en JAVA. Il s'agit d'un algorithme qui analyse le document pour chercher les contradictions basées sur le fichier des règles.
- La deuxième technique consiste à bénéficier de la puissance du langage XSL basé sur XPATH pour vérifier la validité du document en effectuant une transformation XSLT.

Notre choix était en faveur de la deuxième solution, parce qu'il est plus facile de modifier une feuille XSL qu'un programme de JAVA.
Le processus de la commande des annotations mises en application ici consiste à effectuer deux transformations XSLT comme montré dans la figure 7. La première transformation permet de générer une feuille XSL à partir du fichier contenant les règles, la feuille générée sera utilisée dans la deuxième transformation qui permet d'analyser les annotations et reconnaitre celles qui ne sont pas conformes aux règles initiales.

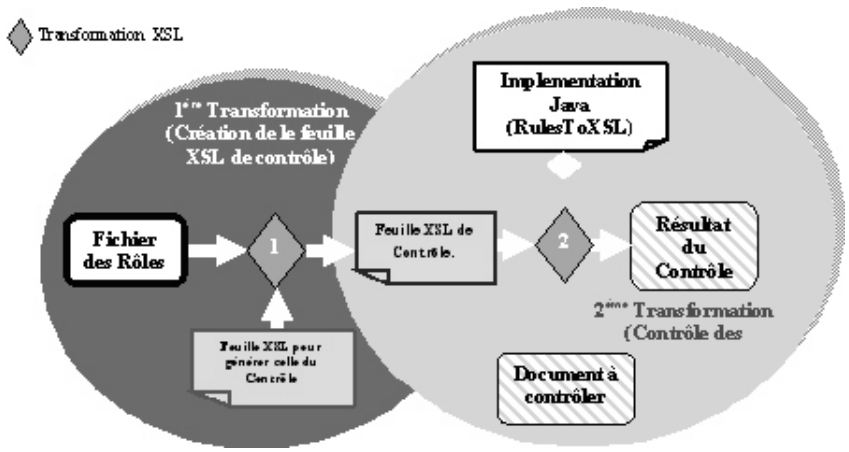

Figure 7 : Transformation des règles en feuilles de style.

Le Code de la figure 8 montre la feuille de contrôle généré à partir du fichier des règles utilisé dans la figure 6.

Dans la figure 8 on note que chaque règle se compose de deux tests :

- Le premier test permet de vérifier que l'annotation dont il est question existe dans le document.
- Le deuxième test vérifie la validité de l'annotation comparée à la règle qui contrôle le nœud, et puis il fait appel à la fonction isValide () développée dans une classe JAVA externe.

Le résultat de ce processus de contrôle d'annotations est un fichier XML contenant les erreurs trouvées. Si aucune erreur n'a été détectée on obtient un fichier vide. La figure 9 montre un exemple du résultat obtenu, c'est-à-dire le fichier xml qui est également transformé par une feuille XSLT pour un affichage adéquat.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
 <xsl:stylesheet exclude-result-prefixes="java_call" version="1.0"
xmlns:int="http://www.laas.fr/namespaces/int/" xmlns:java_call="xsedit.ser-
ver.rules.RulesToXSL"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
 <xsl:output encoding="ISO-8859-1" indent="yes" method="xml" />
  <xsl:template match="text()">
    <xsl:apply-templates />
 </xsl:template>
  <xsl:template match="/Document/chapter[1]/section[1]">
       <!-- on teste l'existence de la métadonnée -->
     <xsl:if test="count(child::int:action) = 0">
       <Erreur>
         <xsl:text>Erreur pas de metatdonnée:/Document/chapter[1]/sec-
tion[1] /int:action
          </xsl:text>
       </Erreur>
     </xsl:if>
      <!-- on crée une instance de la classe externe RulesToXSL -->
     <xsl:variable name="test2" select="java_call:new('[introduire] AND
^[presenter]', int:action)"/>
       <!-- on teste la validité de la métadonnée -->
    <xsl:if test="not(count(child::int:action) = 0) and not
(java_call:isValide($test2))">
          <Erreur>
            <xsl:text>Erreur de
contenu:/Document/chapter[1]/section[1]/int:action</xsl:text>
            <xsl:value-of select="java_call:getValue($test2)" />
          </Erreur>
    </xsl:if>
  </xsl:template>
</xsl:stylesheet>
```

Figure 8 : Génération de feuille de contrôle.

7. CONCLUSION

La recherche menée ici a permis de montrer que la notion d'intention est difficile à cerner si on ne limite pas le domaine d'application. Parmi les difficultés on peut citer que dans le cas des intentions individuelles les auteurs ne sont pas tous favorables à l'expliciter, et ce pour plusieurs raisons : d'abord il est souvent difficile qu'ils soient conscients de toutes les actions qu'ils font ; et même si s'ils l'étaient ils ne seraient pas forcément d'accord pour les expliciter. Toutefois dans un domaine où les documents sont procéduraux, les difficultés annoncées semblent s'atténuer. Pour cela, il faut

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<Erreurs>
<Erreur>Erreur pas de métadonnées : /Document/chapter[1]/sec-
tion[1]/int:but </Erreur>
<Erreur>Erreur de contenu :
/Document/chapter[1]/section[2]/int:but</Erreur></Erreurs>
```

Figure 9 : Un fichier d'erreur généré à partir d'une violation d'intention.

que le domaine soit délimité, analysé et modélisé de façon appropriée. Pour permettre aux auteurs d'avoir un langage commun des intentions, et aussi pour les aider à les trouver, l'emploi d'une ontologie des intentions semble être indispensable. Cette ontologie doit être dynamique, c'est-à-dire il faut qu'elle puisse être enrichie par les nouveaux termes exprimant de nouvelles actions apportées par les auteurs. Si par exemple un auteur souhaite exprimer une intention qui ne figure pas dans l'ontologie, il peut enrichir celle-ci par la nouvelle intention. La vérification de la cohérence des intentions peut être alors possible, à condition de spécifier les règles. L'affectation de droit d'édition des règles a différents auteurs doit permettre de rendre le système plus souple, mais le risque de ne plus maîtriser la vérification de la cohérence devient tangible. Si la notion de DTD ou de Schéma XML permet de contrôler la syntaxe des éléments d'un document, elle ne peut pas vérifier que les annotations exprimant l'intention des auteurs ne présentent pas des omissions ou des redondances des actions souhaitées par les auteurs. Cette recherche constitue une première étape vers la compréhension et l'évaluation des difficultés rencontrées pendant la conception et l'implémentation de XSEdit. Une validation auprès d'utilisateurs potentiels est en cours.

8. RÉFÉRENCES BIBLIOGRAPHIQUES

[ALTAWKI02]Al-Tawki Y. et Tazi S., "Sabre, an authoring system based on reuse of documents". Formal Ontology, Knowledge Representation and Intelligent Systems for the World Wide Web, SemWeb Workshop, 19-20 Avril 2002, Toulouse SEMWEB@KR2002.

[AUSTIN62] Austin, J. How to do things with words. Cambridge, MA: Harvard University Press, 1962.

[BURROW04] Burrow, A. L. 2004. Negotiating access within Wiki: a system to construct and maintain a taxonomy of access rules. In Proceedings of the Fifteenth ACM Conference on Hypertext and Hypermedia (Santa Cruz, CA, USA, August 09 - 13, 2004), 77-86.

[CERRATO99] Cerrato, T., I., 1999, Phd Thesis of Paris VIII University (1999), France. http://www.student.nada.kth.se/~tessy/ visited in December 2005.

[DOURISH96] Dourish, P. 1996. Consistency guarantees: Exploiting application semantics for consistency management in a collaboration toolkit. In Proceedings of the ACM Conference on Computer- Supported Cooperative Work (Nov. 1996). 268-277.

[ELLIS90] Ellis C., Gibbs, S. and Rein, G., Design and Use of a Group Editor Cockton (ed.), Engineering for human-computer Interaction, North-Holland, 1990.

[EXIST06] eXist on the site http://exist.sourceforge.net/, visited on February 2006.

[FERGUSON95] Ferguson, I.A. 1995. On the role of BDI modeling for integrated control and coordinated behaviour in autonomous agents. Applied Artificial Intelligence, Vol. 4, No. 9, 421-448.

[FORSIC04] FORSIC, 2004 http://www.urfist.cict.fr/forsic.shtml , visited in December 2005.

[GREENBERG94] Greenberg, S. and Marwood, D. 1994. Real time groupware as a distributed system: Concurrency control and its effect on the interface. In Proceedings of the ACM Conference on Computer Supported Cooperative Work, ACM, New York, 207-217.

[HONG05] Hong S., Tolone W., Ahn, G. and Pai T. Access Control in Collaborative Systems ACM Computing surveys Vol 37, N°1, March 2005, 29-41

[MITCHELL95] Mitchell, A., Posner, I., and Baecker, R. 1995. Learning to write together using groupware. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Denver, Colorado, United States, May 07 - 11, 1995). I. R. Katz, R. Mack, L. Marks, M. B. Rosson, and J. Nielsen, Eds. Conference on Human Factors in Computing Systems. ACM Press/Addison-Wesley Publishing Co., New York, NY, 288-295.

[NOVICK98] Novick, D. et Tazi S. (1998). "Flight crew operating manuals as dialogue: The act-function-phase model", Proceedings of HCI-Aero'98, Montreal, Mai 1998, pp 179-184.

[SAX04] SAX, 2004 http://www.saxproject.org/, Retrieved in December 2004.

[SEARL69] Searle, J. Speech acts. Cambridge: Cambridge University Press, 1969.

[SUN02] Sun, C. and Chen, D. 2002. Consistency maintenance in real-time collaborative graphics editing systems. ACM Transaction on Computer-Human. Interaction. 9, 1 March. 2002, 1-41.

[SUN98] Sun, C, Jia X., Zhang, Y., Yang, Y. and Chen, D., 1998. Achieving convergence, causality preservation, and intention preservation in real-time cooperative editing systems ACM Transactions on Computer-Human

Interaction, Vol. 5, No. 1, (Mar 1998), 63-108.

[TAZI01] Tazi, S. and Evrard, F. 2001. Intentional structures of documents. In Proceedings of the Twelfth ACM Conference on Hypertext and Hypermedia (Århus, Denmark, August 14 - 18, 2001). HYPERTEXT '01. ACM Press, New York, NY, 39-40.

[TUOMELA06] Tuomela, R. 2006 Joint Intention, We-Mode and I_Mode In the site Visited in February 06 http://www.valt.helsinki.fi/staff/tuomela/papers/

[XUE03] Xue, L., Orgun, M., and Zhang, K. 2003. A multi-versioning algorithm for intention preservation in distributed real-time group editors. In Proceedings of the Twenty-Sixth Australasian Computer Science Conference on Conference in Research and Practice in information Technology - Volume 16 (Adelaide, Australia). M. J. Oudshoorn, Ed. ACM International Conference Proceeding Series, vol. 35. Australian Computer Society, Darlinghurst, Australia, 19-28.

Analyse de forums dans la formation à distance

Nadine LUCAS
Mohamed SIDIR
Emmanuel GIGUET

Laboratoire Paragraphe (Paris8), université de Picardie Jules Verne Amiens
GREYC, CNRS UMR 6072, Université de Caen, 14032 Caen Cedex
Nadine.Lucas@info.unicaen.fr
Sidir@u-picardie.fr
Emmanuel.Giguet@info.unicaen.fr

RÉSUMÉ

Deux forums d'étudiants en situation d'enseignement à distance sont iés dans le cadre de la rédaction collective de devoirs par trinômes. Les discussions accompagnant la réalisation de la tâche ont été analysées manuellement, du point de vue pédagogique pour évaluer le déroulement de la collaboration et du point de vue linguistique pour évaluer le discours collectif. Les forums ont également été analysés automatiquement par le logiciel ThemAgora. Ces trois éclairages convergent et se complètent dans la mise en lumière de phases de développement des forums collaboratifs. L'analyse automatique souligne les moments conflictuels des forums avant l'établissement d'un consensus.

MOTS-CLÉS : forums, documents collaboratifs, pédagogie, EAD, analyse de discours, mise en forme matérielle, TAL robuste, thématique, communication.

Abstract

E-learning students' fora were studied both manually and by computer. Physically distant students were asked to collaborate to write a program in small groups. Tutors evaluated the records of their discussions to assess the collaborative process and its phases. Discussions were also studied as a collective discourse directed by a goal.
Computer parsing based on a discourse model was achieved. Results back some of the human findings but also contribute to show how collaborative discussion evolves through disagreement before reaching consensus.

KEY-WORDS : discussion groups, e-learning, collaborative writing, computer discourse parsing, text structure, robust parsing.

1. INTRODUCTION

La diffusion croissante des environnements virtuels partagés, de type synchrone et asynchrone, a renouvelé les problématiques de la communication dans les communautés. Le recours à la collaboration à distance revêt un grand intérêt tant dans le monde professionnel que dans le cadre de l'éducation et de la formation. Ceci pose une série de questions de recherche insistantes du point de vue des changements pouvant survenir dans la communication et leurs impacts sur les processus de collaboration et d'acquisition des connaissances dans des situations éducatives formelles ou informelles. L'analyse automatique des forums est menée dans le cadre d'une équipe de recherche en technologie de l'éducation, Calico1. Elle a pour objectif de faciliter le suivi d'un nombre important de formations sur une plate-forme d'enseignement entièrement à distance ou dans un dispositif hybride (enseignement en présence et à distance). Au stade actuel, c'est une aide attendue pour faciliter la réflexion a posteriori des tuteurs sur une formation en ligne.

Si les nouvelles formes de communication électronique commencent à faire l'objet de recherche [Anis 1998, Véronis & Guimier 2004, Karlgren 2006], les problèmes de collaboration sont principalement abordés dans une visée pédagogique [Lewkowicz & Marcoccia 2004, Bruillard & Baron 2006]. Très peu de travaux portent sur l'analyse automatique de forums de discussion, qui posent de nombreux problèmes aux analyseurs classiques du fait de leur fantaisie tant orthographique que de ponctuation [Torzec 2004]. Les forums sont ainsi traités essentiellement de manière statistique [Farrel 2001, Newman 2002, Bigi & Smaïli 2002]. La recherche d'information dans les forums à fin d'indexation ou de résumé suit [Vert 2001, Klaas 2005]. D'autres analyses à partir des échanges de contributions mettent en valeur les réseaux sociaux [Reffay 2005]. À notre connaissance, il n'existe pas de traitement automatique portant sur le discours des forums et a fortiori intégrant l'aspect collectif du discours à plusieurs voix.

Nous avons abordé cet aspect et mené des expériences d'analyse automatique sur des forums d'apprenants. Nous nous intéressons ici à l'analyse qualitative de deux mini-forums orientés par une tâche d'écriture d'un cahier des charges informatique. Ces forums ont la particularité de concerner trois participants seulement et d'être limités dans le temps (un mois). Le nombre de contributions échangées par trinôme dépasse cependant largement la centaine (157 et 129 respectivement), ce sont donc des forums très actifs. Un outil d'analyse automatisée, ThemAgora, a été mis en oeuvre. Les résultats de l'analyse automatique sont confrontés à l'analyse de discours manuelle et à l'analyse de la collaboration menée par les formateurs.

Nous présentons le contexte des forums et le cadre de l'expérimentation dans un dispositif de Formation Ouverte et à Distance (FOAD) de niveau universitaire, puis les présupposés de l'analyse de discours appliquée aux

forums et la mise en oeuvre du logiciel ThemAgora. Les résultats sont discutés et comparés avec ceux de l'analyse manuelle. En conclusion, nous envisageons différentes pistes pour améliorer le suivi des activités collaboratives à travers les forums à visée pédagogique.

1 COMMUNAUTÉS D'APPRENTISSAGE EN LIGNE, INSTRUMENTATION, COLLABORATION. CETTE ERTÉ RÉUNIT CINQ LABORATOIRES DE RECHERCHE ET SIX IUFM.

2. DE LA COMMUNICATION AUX PROCESSUS COLLABORATIFS

2.1. Le forum de discussion :

Un média, une forme écrite et une pratique de communication L'écriture collective en réseau est le résultat d'une tension entre un espace privé d'écriture individuel et un espace public de diffusion et de communication. Le premier espace est le système d'écriture à travers un écran d'ordinateur, basé essentiellement sur un logiciel de traitement du texte et le deuxième espace est présenté ici par le forum de discussion ; un lieu permettant une écriture commune, construite au travers d'une collaboration qui aurait pour propriété de profiter des possibilités techniques du réseau, tout en fonctionnant dans une logique de groupe. Ainsi, le forum de discussion se présente comme un outil simple d'utilisation et favorisant la collaboration. Il renvoie à des situations de communication interpersonnelles n'obligeant pas le scripteur à manipuler des logiciels d'écriture complexe et de mise en formes spécifique. Le forum de discussion met cependant en scène une situation de communication très particulière qui repose à la fois sur une communication écrite asynchrone sous forme de messages qui tend à rapprocher ce type de communication des situations classiques de production écrite comme le courriel et d'une forme de conversation interpersonnelle qu'on retrouve dans le chat.
Il présente une forme d'écriture informatisée qui se conjugue intimement à des processus de communication se trouvant inscrits sur l'écran. Ainsi, les écrits dans les forums de discussion présentent à la fois des marques d'oralité communicationnelle (style informel) et des marques d'écrit soutenu (bibliographie). Marcoccia [1998] explique la relation entre les écrits médiatisés par ordinateur et la communication écrite par un cadrage cognitif permettant " de faire du face à face avec l'écrit ".

Pourtant les marques d'une partie du matériau sémiotique disponible dans la conversation en face à face disparaissent naturellement avec la conversation médiatisée par ordinateur pour laisser la place à un certain nombre de procédés permettant de représenter le non-verbal et le para-verbal (l'intonalité et la mimogestualisation) ayant des fonctions comparables aux données en face à face : les smileys ou émoticons, la ponctuation expressive …

Une autre caractéristique du forum de discussion réside dans son asynchronicité qui relève des dimensions de cotemporalité et de simultanéité. Utiliser dans un cadre éducatif cette asynchronicité du forum couplé à une permanence des messages permet de parler à la fois d'extériorisation et de partage de la cognition qui semblent favoriser l'apprentissage [Sharples et Pemberton 1990, Legros et Crinon 2002, Mangenot 2002]. Mais la propriété particulière du forum de discussion réside dans son caractère public. En effet, l'écrit asynchrone public fait du forum l'équivalent d'un texte en perpétuelle évolution et enrichissement où les caractéristiques stylistiques de la langue écrite semblent être conservées en majorité, sans doute à cause de ce caractère permanent, public des textes produits [Peraya, 2005], ce qui modifié profondément le cadre communicationnel comme le signale Marcoccia [1998, p. 17] : " dans un forum de discussion, il est impossible de sélectionner un destinataire. Toute intervention est " publique ", lisible par tous les participants au forum, même si elle se présente comme la réaction à une intervention initiative particulière. L'aparté est impossible : le polylogue est la forme habituelle du forum et le multi-adressage en est la norme ".

Reprenant à son compte les positions de Marcoccia, Mangenot [Mangenot 2002] considère que le polylogue constitue une des caractéristiques majeures de ce type de dispositif de communication. En revanche, Peraya faisant référence à [Ducrot 1980] et à ses analyses des interactions verbales dans une classe virtuelle, nuance cette affirmation dans la mesure où une approche polyphonique de la communication permet de faire la distinction entre les allocutaires qui sont ceux à qui les paroles sont dites et les destinataires auxquels s'adressent réellement les actes de langages. Cette description montre l'intérêt des forums de discussion comme outil de communication et de collaboration. Il nous semble important d'observer ces dispositifs d'écriture et de lecture dans un contexte de communication éducative particulière que nous décrivons dans la section suivante.

2.2. Contexte de l'étude

Le dispositif de formation ouverte et à distance (FOAD) exploité à l'université de Picardie Jules Verne a été mis en place depuis 1995. Il est composé de contenus en ligne et de services pédagogiques dont le tutorat constitue la majeure partie. Ce dispositif compte dix formations diplômantes pour plus de 1000 apprenants dispersés géographiquement en France et dans quelques pays francophones. Le service pédagogique est basé sur un tutorat individuel où le tuteur est considéré comme un accompagnateur, suffisamment disponible pour répondre aux questions de l'apprenant lorsque celui-ci rencontre des difficultés dans sa formation. Or, comme le remarquent [Bruillard, D'Halluin et Weidenfeld 2003], dans un dispositif de formation à distance, la mise en ligne des ressources pédagogiques, assortie d'un simple échange de questions-réponses entre apprenant et tuteur n'est qu'une forme très réductrice du processus d'apprentissage. C'est la raison pour laquelle des expérimentations ont été mises en place pour mettre à l'épreuve des modèles pédagogiques

notamment l'Apprentissage Collectif Assisté par Ordinateur (CSCL) [Sidir 2004, 2005, 2006].

La démarche pédagogique consiste à faire travailler les apprenants par groupes de trois sur des projets planifiés sur des périodes de un à six mois, basés exclusivement sur le travail collaboratif à distance. Pour soutenir leurs travaux, nous avons mis en place un environnement technique attaché à la plate-forme INES [Sidir, 2003] permettant aux membres du même groupe de communiquer, d'organiser et de partager des idées, des contributions et des informations. Cet environnement se présente sous la forme d'un ou plusieurs forums de discussion, intégrant à la fois un service d'édition et de publication de documents et un système de vote pour une prise de décision " démocratique " et collective. Le corpus de notre étude est extrait de ce type de forums inséré dans un scénario d'écriture collective d'un document numérique sous forme d'un cahier de charges. La suite de cet article s'intéresse aux formes caractéristiques du discours médiatisé s'inscrivant dans une perspective d'analyse pourtant sur les processus collaboratifs. On considère alors l'ensemble des interactions dans le forum comme un texte, un corpus unique prêt à être analysé. La section suivante décrit la méthodologie relative à cette analyse.

3. ANALYSE DU DISCOURS COLLECTIF

3.1. Analyse linguistique du discours collectif

Les forums sont envisagés dans notre approche non comme une somme de contributions individuelles, mais comme un discours collectif [Peraya 1999, 2005].
Les cadres théoriques que nous avons utilisés sont basés sur les propriétés du discours écrit, qu'il soit individuel ou collectif. Les forums sont ainsi analysés de la même manière que des articles ou ouvrages qui peuvent avoir plusieurs auteurs. Pour mettre en valeur la construction d'un développement collectif à partir d'un problème soumis par les tuteurs, nous nous sommes basés sur la théorie de l'exposition : le modèle de référence est celui de Yamada (1873-1958), un linguiste japonais qui définit des opérations de mise en discours inspirées par la philosophie allemande. Il présente des opérations cognitives, qui peuvent être vues comme une série de contraintes qui limitent de facto un exposé qui serait sinon un développement infini [Yamada 1936]. On va donc délimiter des passages de texte, comme ayant une cohésion interne, mais il faudra aussi hiérarchiser ces passages, puisqu'il est possible qu'un exposé contienne des développements subordonnés ou incis, qui interrompent momentanément la progression du discours. L'intérêt principal d'une telle théorie est qu'elle fait appel aussi à la perception (de la mise en forme du texte et du style). Elle a une portée générale puisque le discours est un phénomène universel et s'applique aisément à des textes en langues occidentales. Pour mettre en valeur les aspects dialectiques, la construction d'un récit collectif d'expérience, nous nous sommes appuyés sur la théorie

de l'énonciation et sur la poétique de Jakobson [Jakobson 1971, 1973]. Ce qui nous intéresse est davantage le rapport des apprenants à l'objet traité que les relations des apprenants entre eux. Dans cette théorie, le problème (l'énoncé) est considéré en tant que protagoniste du discours, quoiqu'il ne s'agisse pas d'une personne mais d'un objet de transaction. Cependant il est moteur de la discussion. Autrement dit, le problème posé est à résoudre et " actif " pendant une certaine période, en particulier, il peut diviser les apprenants.

À la fin des forums, on observe le changement de statut du problème, en tant que protagoniste du discours, autrement dit le problème posé devient résolu donc " passif ". Ainsi les divergences et le consensus sont envisagés comme des moments signalant l'émergence de difficultés et de solutions.

3.2. Algorithme et implémentation

Les forums sont formatés par une DTD partagée minimale, permettant l'échange de forums, XMLForum2. Elle décrit le forum comme document, constitué de contributions, elles-mêmes segmentées en paragraphes. La ponctuation n'est pas très stable, aussi nous ne donnons pas une définition normative mais contextuelle de ces segments. Le logiciel ThemAgora traite des forums sous ce format-pivot. Il est intégré à la plate-forme " Wims " [Giguet 2005] et il fonctionne en ligne. ThemAgora est un logiciel d'étude dérivé à l'origine du logiciel UniThem, basé sur la théorie de l'exposition de Yamada et utilisé pour les articles de presse [Lucas & Giguet 2005]. Il a été adapté à la segmentation particulière des forums et à leur ponctuation. L'analyseur est descendant du discours jusqu'aux composants des contributions, le grain le plus fin équivalant à des phrases.

L'algorithme fonctionne sur le principe de la division, en respectant le modèle cognitif théorique qui suppose une structuration ordonnée ayant un rapport de forme " raisonné " : le thème se trouve avant le rhème et il est proportionnellement plus court que le rhème ; le rhème peut faire l'objet de développements subordonnés, et il peut à ce titre faire l'objet d'une nouvelle division. Cependant, le nombre de sousdéveloppements doit également rester cohérent avec l'ensemble. Le programme établit un segment thématique au niveau global et son développement ou rhème (niveau 1). Le rhème est divisé en sous-thèmes et sous-rhèmes de niveau 2, euxmêmes divisés le cas échéant en thème et rhème de niveau 3. L'affichage des sorties sous forme graphique indentée et coloriée utilise le module d'UniThem. La frontière entre les segments thématiques et rhématiques est concrètement déterminée par une fonction d'identification de marques contrastives (en présence/absence). Ces marques sont de nature typo-dispositionnelle (ponctuations, mises en forme matérielle,…) et linguistique (connecteurs de discours). ThemAgora a été conçu pour traiter des forums. Pour tenir compte de l'absence d'orthographe, les contraintes morphologiques sur les connecteurs ont été relâchées et les règles exploitant les ponctuations particulières, comme les " émoticons " ou les ponctuations multiples, fréquentes et répétées dans les forums, ont été renforcées.

Cette adaptation est possible car l'algorithme n'est pas dépendant d'un lexique fourni en entrée. Les seuils de fréquence sont gérés automatiquement. Les règles reflètent les attendus des modèles théoriques utilisés, mettant l'accent sur des propriétés du discours interprétables à la fois dans le cadre de l'exposition et dans le cadre de la communication à travers un problème. Dans la conduite de l'expérience, le document analysé est variable, pour tenir compte des attendus de l'analyse de discours et permettre un aller-retour entre interprétation et modèle d'analyse [Lahlou 1995]. Ainsi, les forums ont été analysés séparément par tâche ou ensemble pour un trinôme, et avec ou sans l'énoncé de départ.

2 LES SPÉCIFICATIONS ADOPTÉES DANS LE CADRE DE L'ERTÉ CALICO SERONT PROCHAINEMENT PUBLIÉES.

4. RÉSULTATS ET DISCUSSION

4.1. Analyse automatique

Les résultats de l'analyse corroborent en grande partie les analyses manuelles. Lorsque l'énoncé du problème est intégré au document, il est en fonction thématique, mais lorsque les forums sont analysés indépendamment comme tels ou par étapes, la structure de niveau 1 remonte au niveau global. Faute de place, étant donné la taille des forums, nous nous concentrons sur l'analyse des sous-ensembles établis par étapes et en l'absence de l'énoncé de départ. Les résultats sont ici présentés pour les étapes 1 correspondant à l'écriture collaborative d'un cahier des charges pour les deux trinômes. Nous envisageons ici uniquement les niveaux 0 (global) 1 et 2. Les unités de niveau 3 sont plus fines et leur discussion est omise.

Résultats de l'analyse du trinôme 1
Les résultats de l'analyse automatique sur l'étape 1 du trinôme 1 travaillant sur l'" école " (85 contributions pour l'étape 1 sur 157 au total) montrent que le segment thématique détecté est très long. Il correspond à la phase d'exploration ou d'initialisation de la discussion collective, à l'appropriation du problème posé. L'analyse automatique distingue 4 unités de niveau 2 ou sous-thèmes à l'intérieur de ce rhème, correspondant à des moments dans le déroulement chronologique du forum. Les deux premières unités de niveau 2 correspondent à un moment de discussion ou à l'élaboration d'une solution. Les unités 1 et 2 Ce moment se termine par un accord. La troisième unité correspond au remaniement de cette proposition, elle marque un désaccord, c'est le moment de " dramatisation ".
Dans cette partie, les participants se disputent. La quatrième et dernière unité marque le consensus final. La comparaison est peu développée et n'apparaît pas en tant qu'unité autonomisée par l'analyse automatique. L'analyseur souligne une répétition de déclaration Je valide à priori et par avance vos modifications du cahier des charges.

Résultats de l'analyse du trinôme 2
Pour le trinôme travaillant sur la gestion documentaire de bibliothèque (81 contributions pour l'étape 1 sur 129 au total), le schéma est similaire. Le segment thématique détecté en l'absence de l'énoncé de départ est également très long (29 contributions sur 81). Le développement ou rhème commence lorsque la discussion reprend des questions déjà abordées, ce qui est conforme à la sémantique du modèle linguistique de l'exposition et souligne l'intérêt du modèle stylistique sous-jacent.
On y trouve des reprises (Dans ton document…) interprétables à la lecture, mais rappelons que le programme ne calcule pas d'anaphores. On remarque aussi que le rhème correspond bien au changement de statut cognitif du problème, qui n'est plus " sujet " mais " objet " pour les apprenants.
Le rhème (52 contributions) est subdivisé en 5 unités très inégales en longueur recouvrant, à quelques contributions près par rapport à l'analyse manuelle, des moments principaux : la discussion (G1 et G2), la dramatisation (G3), la comparaison (G4) et la clôture (G5). On note en particulier que le premier thème de niveau 2 (le début du rhème) coïncide en fait avec les premières confrontations sur la solution, par exemple je suis absolument contre … t-ai-je convainqu ??. Le dernier moment correspond au contraire à l'accord collectif sur la dernière version du cahier des charges.
G0 Thème niveau global
bonjour ! quand est-ce qu'on se voit pour en parler ?
le plus tôt possible !!!! […]
je n'ai qu'un type de compte.
il n'y aura qu'un compte ADMIN (id=1 ou 0) par exmple, le reste sera des membres.
Dans ton document, il semble que les emprunteurs soit également des utilisateurs indentifié du système ? [...] Si la bibliothéquaire doit modifier les informations de ses membres (une adresse par exemple) , faut pas qu'elle court appeler l'administrateur., idem pour ajouter un auteur...
t-ai-je convainqu ??
G1
[#]
Si c'est le cas cela v nous compliquer l'affaire pour la gestion des disponibilités des ouvrages ... [...] et si oui combien.
-> Je ne pense pas que ca soit si simple que cà: si on veut gerer des dates d'emprunts et de retour pour chaque exemplaires d'une même ouvrages, [...]
Impressionnant :) Très joli travail
G2
[#]
j'ai modifie le CDC et j'ai mis mes remarques dedans en rouge et jaune.
tu peux
dire si t'es d'accord ?
Ben apparement déja on est pas d'accord sur ce qu'on doit faire ... [...]
qu'un login pour accéder à la partie admin
G3
[#]

c'est un petit projet. je ne pense pas qu'il faut faire une chose comme ça. [...] à nous de le réaliser au mieux... comme des pros ^^
Je pense que nous pourrions combiner la localisation des ouvrages dans le centre (qui pourrais paraitre superflu) et le numéro unique apposé sur les ouvrages (pas très parlant pour le public), [...] roman pour un roman qui parlerai de la vie quotidienne des paysans au 14ème siècle par exemple...
G4
[#]
pas plus que le placement dans le centre ... [...] les bibliothèquaires
serai libre de mettre ce qu'il veulent suivant la classification qu'il veulent ...
(c juste qu'on leur donne la possibilité ...)
Voici la nouvelle version de mon cahier des charges integrant quelques unes des suggestions exprimés par Xing. [...] Oui.
G5
[#]
maintenant c'est ok pour moi [...] mais pas 2002 ^^
Figure 1. Résultat compacté de l'analyse automatique étape 1 trinôme 2
Dans l'extrait de la figure 1 compressé faute de place, les segments sont évidés. Le symbole [#] indique des coupures de plusieurs contributions entre la fin d'un thème et la fin du rhème correspondant, le symbole […] indique des coupures de une ou plusieurs contributions à l'intérieur d'une unité.

Ces résultats ou plutôt leur intérêt sémantique peuvent paraître étonnants, ils sont cependant corroborés par d'autres approches stylistiques [Karlgren 2000, 2006].

4.2. Discussion

L'analyse automatique de forums par le logiciel d'étude ThemAgora a permis de dégager utilement des épisodes structurés dans des forums clos, malgré les difficultés posées par ce type d'écrit. Notre étude est centrée sur des forums guidés par la tâche et limités dans le temps. Le résultat d'analyse des forums de trinômes est satisfaisant et conforme aux résultats précédents obtenus sur une étude de cas, avec davantage de participants (une trentaine). Les analyses automatiques sur cinq forums de ce type ont été favorablement évaluées par les formateurs dans le cadre de Calico, en confrontant leurs analyses manuelles et les résultats de ThemAgora. Cela montre l'adéquation du modèle à ce type de forum au cours desquels le discours progresse. Un plus grand nombre de cas devrait être analysé, dans les mois à venir, grâce à la mise à disposition du logiciel sur la plate-forme Calico. En revanche, pour des forums libres, ressemblant à une liste de questions/réponses, sans progression dans le temps, l'analyse avec cette grille est moins pertinente, ce qui semble normal puisque le présupposé de l'analyse n'est pas respecté [Sidir et Lucas, en préparation]. La discussion à bâtons rompus ne correspond pas au modèle théorique de l'exposition. Nous avons donc entrepris d'implémenter une version différente du logiciel pour ces forums ouverts.

5. CONCLUSION

L'analyse automatique de forums est tout à fait possible, sans passer par une normalisation de l'écrit, lorsque l'algorithme est fondé sur un modèle théorique d'analyse de discours. Elle est satisfaisante lorsque le modèle est adéquat au type de forum analysé. L'analyse automatique par le logiciel d'étude ThemAgora est basée sur les principes de l'exposition, elle segmente les forums guidés par la tâche en unités inégales en longueur mais très informatives sur le déroulement de la discussion collaborative. Elle met en valeur les moments de dissension et de consensus. Lue dans la grille de Jakobson, l'analyse est également informative sur le rapport des participants à l'objet de la discussion, elle montre les étapes d'appropriation puis de résolution du problème posé. Nous ne prétendons pas avoir résolu tous les problèmes d'analyse automatique de forums. Outre l'implémentation d'une version pour forums libres, beaucoup de travail reste à faire pour adapter cet outil déjà utile aux besoins des formateurs. Ceux-ci se réfèrent souvent à des grilles d'analyse empiriques issues de leur pratique professionnelle ou à des grilles inspirées par des travaux de didacticiens ou encore par des modèles variés d'analyse de discours. Nous avons entrepris l'étude des grilles d'interprétation familières pour les utilisateurs, dans le cadre de Calico. Une autre direction de recherche plus générale est le calcul dynamique du grain d'analyse pour traiter des forums, en fonction de leur taille et des objectifs desutilisateurs.

Les perspectives de cette recherche sont à court terme de fournir des outils objectifs et partageables permettant d'asseoir et mutualiser les pratiques pédagogiques en ligne. À long terme, le but pratique serait de suivre l'évolution d'un groupe en contexte d'apprentissage pour gérer les formations en cours et non plus seulement d'étudier les traces de la discussion a posteriori.

6. RÉFÉRENCES BIBLIOGRAPHIQUES

[Anis 1998] Texte et ordinateur, l'écriture réinventée J. Anis. Bruxelles, De Boeck.

[Bigi & Smaïli 2002] "Identification thématique hiérarchique : application aux forums de discussion". TALN 2002, Nancy. pp. 115-124.

[Bruillard & Baron 2005] "Study and design of new modalities of learning and work for preservice teachers assisted by ICT: collaborative learning and case studies" WCCE 2005, IFIP, juillet 2005.

[Bruillard & Baron 2006] Technologies de communication et formation d'enseignants : vers de nouvelles modalités de professionnalisation G. L. Baron et E. Bruillard (coord.) Paris, INRP.

[Bruillard, D'Halluin et Weidenfeld 2003] "Comment appliquer l'apprentissage collaboratif assisté par ordinateur à la formation à distance ? L'exemple du campus numérique Ape-Lac" Actes Campus numériques et Universités numériques en région Montpellier, France.

[Ducrot 1980] Les mots du discours. O. Ducrot. Paris, Minuit.

[Farrel 2001] "Text Summarization and Question Answering: Summarization of discussion groups" R. Farrell, P. G. Fairweather, K. Snyder Proceedings of the tenth international conference on Information and knowledge management October 2001 ACM.

[Giguet 2005] "Modélisation de l'activité expérimentale du chercheur en traitement des langues sur corpus multilingues" E. Giguet Journée de l'ATALA Articuler les traitements sur corpus, 12 février 2005.

[Jakobson 1971] Word and Language R. Jakobson. The Hague, Paris, Mouton.

[Jakobson 1973] Questions de poétique R. Jakobson. Paris, Seuil.

[Karlgren 2000] Stylistic Experiments for Information Retrieval, J. Karlgren, PhD thesis, Stockholm, Université de Stockholm, 2000.

[Karlgren 2006] "Workshop on New Text" April 3, 2006 Trento, Italy J. Karlgren (ed) http://www.sics.se/jussi/newtext

[Klaas 2005] Toward indicative discussion fora summarization M Klaas UBC CS Technical Report TR-2005-042005 [Lahlou 1995] "Vers une théorie de l'interprétation en analyse des données textuelles." S. Lahlou JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data.

S. Bolasco, L. Lebart and A. Salem Eds.. CISU, Roma, 1995. 221-28. Vol. I.

[Legros & Crinon 2002] Psychologie des apprentissages et multimédia. D. Legros et J. Crinon (eds) Paris, Colin université.

[Lewkowicz & Marcoccia 2004] "The Participative Framework as a design model for newsgroups" M. Lewkowicz et M. Marcoccia. PartRoOM. COOP 2004 pp. 243-256

[Lucas 2005] "Les procédés d'exposition et de développement collectif dans un forum pédagogique : le cas Maxime" N. Lucas Symposium Symfonic, 20-22 janvier 2005 Amiens.

http://www.dep.u-picardie.fr/sidir/articles/index.php

[Lucas & Giguet 2005] "UniTHEM, un exemple de traitement linguistique à couverture multilingue" N. Lucas et E. Giguet Cide 8 Conférence internationale sur le document électronique Beyrouth 25-28 mai 2005. K. Zreik (ed). Paris, Europia, pp. 115-132.

[Mangenot, 2002] "Forums et formation à distance : une étude de cas ", in Education permanente 152, pp. 109-119.

[Marcoccia 1998] "La normalisation des comportements communicatifs sur Internet : étude sociopragmatique de la netiquette" M. Marcoccia In Communication, société et Internet, N. Guéguen & L. Toblin (éds.) pp. 15-22. Paris, L'Harmattan.

[Newman 2002] Exploring discussion lists: steps and directions. P. S· Newman. Proceedings of the secondACM/IEEE-CS joint conference on Digital libraries, 2002, pp. 126-134.

[Peraya 1999] "Vers les campus virtuels. Principes et fondements

techno-sémio-pragmatiques des dispositifs de formation virtuels" D. Peraya In Le Dispositif. Entre Usage et concept. G. Jacquinot et L Montoyer (eds) Paris, CNRS Editions Hermès, pp. 153-168.

[Peraya 2005] "Axes de recherches sur les analyses de communication dans les forums" Daniel Peraya Symfonic, 20-22 janvier 2005 Amiens. http://www.dep.upicardie. fr/sidir/articles/index.php

[Reffay, 2005] "Réseaux sociaux et analyse de traces des forums d'une communauté d'apprentissage - Calculer la cohésion. Symposium Symfonic, 20-22 janvier 2005 Amiens.
http://www.dep.u-picardie.fr/sidir/articles/index.php

[Sharples & Pemberton 1990] "Starting from the Writer : Guidelines for the Design of Usercentred Document Processors" In Computer Assisted Language Learning Vol. 2 / 1990,. Oxford, Intellect. pp.37-57.

[Sidir, 2003] "e-formation : quel choix technique ?" In l'enseignement à distance, théories et pratiques, Acte 8 (ed,) Paris, France. pp. 47-58

[Sidir, 2004] "Modes de collaborations au sein de groupes d'apprentissage dans une formation à distance universitaire" M. Sidir TICE 2004, Compiègne.

[Sidir, 2005] "La médiation par les TIC de la communication éducative" M. Sidir H2PTM'05, Hermès-Lavoisier, pp. 395-408.

[Sidir, 2006] "Interaction communicationnelle à travers les forums de discussion dans un dispositif de e-formation" M. Sidir In G-L., Baron et E., Bruillard (coord.). Technologies de communication et formation d'enseignants, INRP, pp. 235-249.

[Torzek 2004] "Contribution à l'étude des messages électroniques francophones Quelques résultats et leurs conséquences pour le TALN". N. Torzec Journée de l'ATALA du 4 juin 2004. (http voir Veronis)

[Vert 2001] Text Categorization Using Adaptive Context Trees J-P. Vert Proceedings Computational Linguistics and Intelligent Text Processing CICLing 2001, Mexico-City, Mexico, February 18-24 2001,A. Gelbukh (Ed.) Berlin / Heidelberg Springer.

[Veronis & Guimier de Neef 2004] "Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.)" J.

Veronis et E. Guimier de Neef (eds) Journée de l'ATALA du 4 juin 2004 http://www.up.univ-mrs.fr/veronis/jenfce/index.html
[Yamada 1936] Nihon bunpôgaku gairon [Somme sur la grammaire japonaise] Y. Yamada, Tôkyô, Hôbunkan, 1936 (ré-imp. 1989).

Modèle de représentation sémantique des documents électroniques pour leur réutilisabilité dans l'apprentissage en ligne

Nathalie HERNANDEZ
Josiane MOTHE
Bachelin RALALASON
Patricia STOLF

IRIT, 118 route de Narbonne, 31062 Toulouse Cedex 09,
{hernandez,bachelin,mothe,stolf}@irit.fr
Institut Universitaire de Formation des Maitres, Av. de l'URSS, 31078, Toulouse

RÉSUMÉ

Cet article propose un modèle de représentation sémantique pour les documents électroniques. Nous situons notre étude dans le cadre de l'apprentissage en ligne ; les documents électroniques auxquels nous nous intéressons sont des objets pédagogiques. Le modèle proposé vise à améliorer la réutilisabilité d'un objet en considérant ses différents aspects: sa description par des méta-données, son usage dans les scenarii d'apprentissage, son découpage et sa représentation sémantique. Le modèle est en conformité avec les normes ce qui assure entre autres interopérabilité et pérennité. Le modèle est illustré par un exemple de représentation d'objets pédagogiques.

MOTS-CLÉS : Document pédagogique électronique, re-utilisabilité des objets pédagogiques, représentation des ressources pédagogiques, ontologies, normes d'apprentissage en ligne.

ABSTRACT

This article proposes a model of semantic representation for electronic documents. We place our study within the framework of the on-line learning; the electronic documents in which we are interested are educational objects. The proposed model aims at improving the re-use of an object by considering its various aspects : its description by meta-data, its usage in the scenario of learning, its division and its semantic representation. The model is in keeping with the standards what insures interoperability and perpetuity. The model is illustrated by an example of representation of educational objects.

Key words : electronic educational document, re-use of educational objects, representation of educational resources, ontologies, standards of e-learning.

1. INTRODUCTION :

Les objets pédagogiques sont des documents électroniques créés dans l'objectif d'être intégrés dans un environnement technologique dédié à l'apprentissage en ligne. Ce type de documents reflète donc les enjeux majeurs de tout document électronique : utilisation dans différents scénarii d'usage (ici pédagogiques), ré-utilisabilité (utilisation de tout ou partie d'une ressource pour en construire une autre, pour d'autres objectifs ou d'autres utilisateurs), confrontation aux normes en cours d'élaboration et aux environnements technologiques qui les manipulent (ici les plates formes d'apprentissage en ligne ou même Internet). Dans cet article, nous nous intéressons donc à ces documents électroniques spécifiques que sont les objets pédagogiques utilisés dans les environnements d'apprentissage en ligne.

De nombreuses ressources pédagogiques sont accessibles via des moteurs de recherche sur Internet mais celles-ci correspondent souvent à de simples présentations en ligne de documents qui n'ont pas été créés spécifiquement pour leur exploitation dans des environnements d'apprentissage. Ce même problème se retrouve lorsqu'il s'agit de plate forme d'apprentissage qui deviennent des espaces organisés de ressources mais auxquelles ne sont pas rattachés de véritables situations d'utilisation. [PSYCHE05] constate l'insuffisance ou l'absence de l'application d'une approche pédagogique, que ce soit au niveau de la présentation des ressources pédagogiques ou du séquencement des activités d'apprentissage dans les outils actuels. Pourtant, parmi les solutions proposées par les organismes de normalisation, IMS-LD [IMSLD03] qui est en charge de la pédagogie d'apprentissage et de son déroulement intègre la notion de scénario pédagogique. D'autres normes telles que SCORM [SCORM04] et LOM [LOM02] aident à l'homogénéisation des représentations de ce type de documents et facilitent l'interopérabilité. LOM rassemble les différentes méta-données nécessaires pour la description des ressources pédagogiques mais n'inclue pas la représentation sémantique des contenus ; SCORM permet la structuration des contenus d'objets pédagogiques et leurs relations avec l'environnement d'utilisation. Ces représentations ne sont pas suffisantes pour permettre et assurer la ré-utilisabilité de ressources ou de parties de ressources. La ré-utilisation de parties de ressources nécessite d'une part que la structure du document initial soit suffisamment marquée pour être exploitée, d'autre part que le contenu sémantique ainsi que la portée d'usage soient suffisamment explicites pour chacune des parties.

Dans cet article, nous proposons un modèle de représentation des documents électroniques de type " objet pédagogique " [WILEY01] en portant plus particulièrement notre attention sur l'aspect ré-utilisabilité de ces documents. Nous nous appuyons pour cela sur les normes actuelles de l'apprentissage en ligne. Notre apport concerne en particulier l'enrichissement de ces normes par leur représentation sous forme d'ontologies qui permet

d'introduire un raisonnement sémantique. Une autre originalité concerne la représentation sémantique des contenus à travers une ontologie de thème. Ainsi, nous représentons différentes formes de connaissances qui nous permettent de faciliter la communication entre la machine et l'utilisateur. Plus spécifiquement, les domaines des formations en ligne visées sont représentés sous forme d'ontologies et sont utilisées pour indexer sémantiquement les granules issus des documents pédagogiques et rendant possible une recherche efficace ultérieure. De la même façon, les théories d'apprentissage sont représentées sous forme d'ontologies, permettant ainsi des mécanismes d'inférence qui facilitent le travail de conception des scénarii pédagogiques à partir des granules d'information. Enfin, les méta-données associées aux documents permettent une représentation générique de ceux-ci, en accord avec les normes actuelles du domaine.

Cet article est structuré de la façon suivante : dans la section 2 nous présentons les principes qui régissent les systèmes d'apprentissage en ligne, l'objectif des différentes normes associées aux documents qui les composent ainsi qu'une revue des travaux reliés. Dans la section 3 nous présentons les différents aspects de la représentation des documents pédagogiques et le modèle sous-jacent. Le modèle permet une représentation sémantique des contenus et des usages grâce à des ontologies tout en respectant les normes de l'apprentissage en ligne. La section 4 illustre ces représentations par un exemple d'application. Enfin, nous concluons notre article en indiquant les perspectives à ce travail.

2. REPRÉSENTATION DES RESSOURCES PÉDAGOGIQUES D'UN SYSTÈME D'APPRENTISSAGE EN LIGNE

Dans cette partie, nous détaillons les caractéristiques des systèmes d'apprentissage en ligne. Après une présentation de différentes normes du domaine, nous verrons l'intérêt de l'utilisation des ontologies sur lesquelles se sont appuyés les travaux du domaine que nous présentons.

2.1. Systèmes d'apprentissage en ligne

L'apprentissage en ligne est une activité pédagogique qui vise à acquérir ou à approfondir des connaissances tout en repoussant les contraintes de temps et d'espace entre l'apprenant et l'enseignant, par l'utilisation des nouvelles technologies de l'information et de la communication [E-TUD] [BOUTEMEDJET04]. Il s'agit donc d'une méthode d'apprentissage reposant sur la mise à disposition de contenus pédagogiques à travers des scénarii pédagogiques dans un environnement numérique. Un système d'apprentissage en ligne doit permettre :

- l'accès aux ressources pédagogiques pertinentes grâce à une bonne indexation des ressources. [GASEVIC05], [PSYCHE05], [LENNE05], [ABEL03],

- une interaction et navigation suivant une pédagogie d'apprentissage adéquate mise en place. [PSYCHE05]

- la réutilisabilité des objets et des scenarii pédagogiques. [KNIGHT05]

- la conception et la mise à jour du contenu des cours par les enseignants. [LENNE05], [ABEL03]

- le suivi individualisé des apprenants. [IMSLD03]

Chacun de ces aspects fait référence à des problématiques spécifiques. Dans cet article, nous nous intéressons plus particulièrement à la modélisation des documents électroniques créés pour être intégrés dans de tels systèmes. Plus spécifiquement, nous nous intéressons donc aux trois premiers points.

2.2 Les normes associées aux systèmes d'apprentissage en ligne

L'application des normes du domaine de la formation en ligne garantit non seulement l'interopérabilité mais également la qualité du système. Parmi les normes de la formation en ligne, on peut citer SCORM, LOM et IMS-LD. LOM s'intéresse à la description des ressources pédagogiques, SCORM à la structure du contenu des objets, et IMS-LD au scénario d'apprentissage.

LOM

LOM [LOM02] ou Learning Object Meta data est une norme utilisée pour l'annotation des objets pédagogiques par les méta-données comme : le type de contenu, son auteur, l'utilisation préconisée, etc... Elle est particulièrement utile pour assurer l'accessibilité des ressources pédagogiques. L'ensemble des méta-données de LOM a été regroupé en neuf catégories :

1. Général : ensemble des caractéristiques générales,
2. Cycle de vie : informations relatives à l'historique et à l'état courant,
3. Méta-méta-données : informations sur les méta-données elles-mêmes,
4. Technique : exigences et caractéristiques techniques requises ,
5. Pédagogique : caractéristiques pédagogiques,
6. Droits : caractéristiques exprimant les droits sur la propriété intellectuelle et les conditions légales d'utilisation de la ressource,
7. Relation : caractéristiques exprimant les liens avec d'autres ressources,
8. Commentaire : commentaires libres sur l'utilisation de la ressource,
9. Classification : description de la ressource à partir de classes.

Concernant le contenu sémantique des objets pédagogiques, la norme laisse un degré important pour leur description. Dans notre approche, nous préconisons une représentation des contenus qui est clairement établie en s'appuyant sur une ontologie de domaine.

SCORM

SCORM (Sharable Content Object Reference Model) [SCORM04] de Advanced Distributed Learning (ADL) est un modèle de référence pour le partage de contenus et d'objets. SCORM est un modèle pour l'assemblage des contenus web et un environnement d'apprentissage pour les objets pédagogiques. Il a pour vocation la mise en place de la bonne structuration du contenu du cours et de ses interactions avec son environnement.
SCORM traite les éléments suivants :

- Packaging : il a pour objectif la transmission d'un contenu d'une plate-forme vers une autre, l'importation ou l'exportation de contenus d'objets pédagogiques pour les mettre à disposition d'autres. Il s'intéresse également à la structuration des objets pédagogiques.

- Méta-données : elles sont issues de LOM et ont pour objectif de partager les informations standards qui décrivent la nature et l'objectif du contenu.

- Communication ou environnement d'exécution : détermine la communication avec un environnement web. La notion d'environnement est également présente dans IMS-LD.

- Séquencement et navigation : définit une méthode de représentation de la navigation entre objets d'apprentissages.
La structuration du contenu des modules d'enseignement suivant le modèle SCORM permet de les réutiliser dans d'autres modules pour différentes formations ou systèmes. De plus, elle améliore le dialogue entre les objets pédagogiques et le système d'une part, et entre les acteurs et le système d'autre part.Dans notre modèle, nous utilisons SCORM pour représenter les structures des modules d'enseignement d'une formation et ainsi garantir leur interopérabilité.

IMS-LD

En complément de SCORM, IMS-LD [IMSLD03] ou Instructional Management System Learning Design est une norme qui vise à apporter des éléments de pédagogie dans un système d'apprentissage en ligne. Il s'agit d'un langage de modélisation des processus d'apprentissages. Il a été conçu pour la définition de scénarii d'apprentissages et d'interaction pour les créateurs de contenu ou de cours. Il aide les concepteurs à modéliser : qui fait quoi, quand et avec quelles ressources et quels services pour réaliser des objectifs d'apprentissages. En effet, il définit la structure d'une unité d'apprentissages comme " pièce " : un ensemble " d'actes " composés de " partitions " associant des " activités " à des " rôles " (enseignant, apprenant,…).
Dans notre modèle, nous nous appuyons sur IMS-LD pour définir le déroulement des interactions Homme-Machine pendant la phase d'exécution et d'utilisation des objets pédagogiques.

Les normes définies dans le contexte de l'apprentissage en ligne permettent de s'assurer d'une certaine inter-opérabilité et utilisabilité (au travers de scénarii). Cependant, l'utilisation des méta-données telle qu'elle est préconisée ne suffit pas et ne résout pas les problématiques des systèmes d'apprentissage en ligne : réutilisabilité et accessibilité. Un problème complémentaire relève du fait que le système et les acteurs doivent partager le même sens accordé aux valeurs des méta-données. D'autre part, les liens et relations comme la composition, ordre d'apprentissage, et dépendances de pré requis entre chaque objet pédagogique doivent être mentionnés pour permettre non seulement de réaliser des traitements ou tâches automatiques sur ces objets. Différents systèmes d'apprentissages en ligne s'appuient sur les ontologies pour tenter de résoudre ces problèmes.

2.3. Ontologies dans les systèmes d'apprentissages en ligne

Une ontologie regroupe les concepts qui représentent l'ensemble des connaissances d'un domaine en une spécification explicite et formelle [STUDER98]. Elle montre les relations ainsi que les règles d'associations qui existent entre ces concepts en vue de permettre d'une part à l'ordinateur la production de nouvelles connaissances par le biais d'une inférence, et d'autre part à l'homme et à l'ordinateur d'accorder des sens communs aux termes utilisés dans un domaine d'activité afin de lever toute ambiguïté pendant les traitements. Différents travaux de la littérature s'appuient sur des ontologies pour indexer et accéder aux ressources pédagogiques.

Memorae (pour MEMoire Organisationnelle Appliquée à l'apprentissage en ligne) [LENNE05] [ABEL03] est un outil d'apprentissage en ligne et d'indexation de ressources. Cet outil met à disposition des ressources pédagogiques aux apprenants, soit au sein d'une banque de ressources locales, soit dans un emplacement distant sur le Web, référencé par son URI. Memorae a pour objectif de faciliter l'autorégulation de l'apprentissage en explicitant les connaissances à appréhender ainsi que les relations qui existent entre elles et en leur associant des ressources appropriées. Par rapport à Memorae qui présente des cours structurés suivant les relations d'inclusion, d'utilisation , de référence et de pré requis entre les notions à appréhender, [GASEVIC05] propose, en plus, pour la recherche de ressources pédagogiques un outil permettant aux utilisateurs de formuler des requêtes libres, dans un champ prévu à cet effet. Cela permet ainsi aux utilisateurs de rechercher de l'information dans une banque de ressource distante. Par ailleurs, il respecte la norme LOM quant à l'annotation et l'indexation des ressources pédagogiques. Ces deux études représentent la connaissance du système à l'aide d'ontologies. Une ontologie de domaine de la formation décrivant les concepts tels que les personnes (étudiants, tuteurs, secrétaires,…), les documents (livres, supports de présentation, pages web…), appelé ontologie du domaine de la formation pour [LENNE05] et ontologie cible pour [GASEVIC05], et une autre ontologie pour les notions à appréhender (ontologie d'application pour

[LENNE05] et ontologie source pour [GASEVIC05]). [HERNANDEZ05] met l'accent sur la séparation des aspects de tâche et de thème tout en les mettant en relation. Chaque aspect est modélisé par une ontologie de domaine. Tandis que l'ontologie de thème spécifie les notions qui doivent être assimilées par des étudiants pour une formation donnée, l'ontologie de tâche a pour but de préciser les contextes de l'apprentissage en spécifiant les ressources disponibles (ouvrage, logiciel,…), les modules qui composent ces ressources, leur type (cours, exercices, évaluation) ainsi que l'ordre dans lequel ils doivent être étudiés. Cette formalisation permet d'établir les connaissances associées à ces deux aspects à travers des relations sémantiquement riches. Le système d'apprentissage présente cette formalisation à l'utilisateur par un mécanisme d'exploration du corpus pédagogique reposant sur les deux ontologies. L'utilisateur appréhende ainsi le contexte associé aux objets pédagogiques.

[GASEVIC05], pour la recherche de ressources pédagogiques dans une banque de ressource distante, considère de plus une ontologie de correspondance, qui sert à décrire la correspondance ou la similarité entre les concepts de l'ontologie source qui prend en compte le contexte du cours présenté avec ceux de l'ontologie cible qui décrit la banque de ressource distante. Toutefois, l'inconvénient de ce système est que son efficacité dépend fortement de l'ontologie de mise en correspondance car une connaissance préalable de la structure de l'ontologie cible est nécessaire pour la mise en place de cette ontologie de correspondance. Par conséquent, ce système n'est pas facilement réutilisable du fait que l'ontologie de correspondance doit être mise à jour à chaque modification de l'ontologie cible ou de l'ontologie source.

[LENNE05] comme [GASEVIC05] orientent leurs études autour de l'apprenant alors que [KNIGHT05] et [PSYCHE05] considèrent aussi une assistance à l'auteur de la ressource pédagogique. La prise en compte de l'apprenant et de l'enseignant parait essentielle pour une utilisation optimale par ces deux types d'acteurs. [GASEVIC05], [PSYCHE05] et [KNIGHT05] intègrent, contrairement à [LENNE05], la notion de scénario pédagogique grâce à une ontologie basée sur la norme IMS-LD. Seul, [PSYCHE05] prend en compte les théories éducatives afin de pallier au manque de relation entre le scénario d'apprentissage (Learning Design ou LD) et les théories des pédagogies d'apprentissage. Les différents types de théories de l'éducation sont représentés grâce à une troisième ontologie " ontologie de théorie éducationnelle ".

Dans toutes ces études ([GASEVIC05], [PSYCHE05], [LENNE05]) le contexte d'utilisation des objets pédagogiques n'est pas pris en compte. Afin d'augmenter la réutilisabilité des scénarii et des objets pédagogiques, seul [KNIGHT05] introduit une " ontologie de contexte ". Il associe aux objets pédagogiques (appelés LO pour Learning Object) un à plusieurs objets de contexte (appelés LOC pour Learning Object Context). Une séquence

d'activités de l'apprenant est ainsi composée d'activités. Une activité étant

.lités demandées à un outil d'apprentissage en ligne comme la réutilisabilité, l'accessibilité, l'interopérabilité et la durabilité [SCORM04][FAGE05], nous souhaitons prendre en compte les théories de l'éducation ainsi que les contextes d'apprentissage et d'utilisations des ressources pédagogiques par l'utilisation des ontologies, tout en respectant les normes en apprentissage en ligne en vigueur. Notre approche se veut d'une part être plus complète dans la mesure où elle couvre tout le cycle de vie d'un outil d'apprentissage (réutilisation/conception de ressources par les enseignants, recherche d'information et utilisation par les apprenants) et d'autre part elle est applicable à n'importe quel domaine de formation. Dans cet article nous nous focalisons sur les aspects correspondants à la représentation des documents et leurs usages.

3. Représentation multi-facette des documents pédagogiques et usage

Afin d'une part de disposer d'un système d'apprentissage qui utilise des approches pédagogiques adéquates pour mieux apprendre les notions et connaissances relatives à un domaine d'études particulier, et d'autre part de permettre la réutilisabilité des objets pédagogiques et des Learning Design (LD), nous proposons de modéliser les différents aspects permettant de décrire les objets pédagogiques. Nous distinguons la description propre à l'objet pédagogique et celle liée à ses usages.

Les documents pédagogiques sont composés d'objets pédagogiques et de composants élémentaires, ils abordent des notions d'un domaine donné, et sont inclus dans des scenarii pédagogiques. Pour représenter un document pédagogique nous devons considérer différentes connaissances (figure 1) :

- Connaissance sur la structuration du document (norme SCORM)
- Connaissance sur la ressource elle-même (norme LOM)
- Connaissance sur le thème abordé par le document
- Connaissance sur l'ensemble des théories éducatives existantes
- Connaissance sur le scénario pédagogique (norme IMS-LD)

Afin de représenter ces connaissances, nous proposons d'utiliser des ontologies. L'intérêt d'utiliser des ontologies dans ce contexte réside d'abord dans une représentation non ambiguë de la connaissance (en particulier levée des ambiguïtés terminologiques) [MIZOGUCHI04]. Ensuite, en associant les concepts des ontologies aux documents pédagogiques ou aux usages de ces documents (scénarii d'apprentissage), il est possible d'induire un raisonnement grâce aux axiomes associés à celles-ci. Le schéma figure 1 représente le modèle utilisé pour représenter les différents aspects d'un document et ses usages. Ces aspects sont décrits dans les sections suivantes.

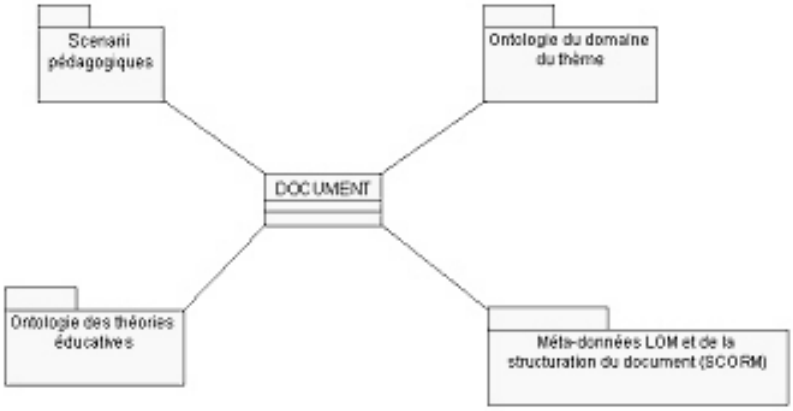

Figure 1 : Connaissances utiles pour représenter un document et son usage

3.1. Description SCORM et LOM

Un Objet pédagogique est une unité sémantique de ressource d'apprentissage. Il peut être un exercice, un sujet d'examen, une définition, des exemples, ou bien une leçon, etc… Chaque objet pédagogique peut rassembler des composants élémentaires comme une image nommés Composant (appelé "asset" dans la norme SCORM) qui peuvent être de format numérique (.DOC, .PDF, .JPG etc) ou physique différents. Un objet pédagogique pouvant par ailleurs être composé d'autres objets pédagogiques.

La description des méta-données associées à un objet pédagogique correspond à celle qui est prévue par LOM. Comme dans [DUVAL02], nous proposons l'utilisation d'un Profil d'application. Le profil d'application permet de définir pour une application donnée ou une formation donnée quelles sont les méta-données (issues d'un ou plusieurs schémas) qui sont d'intérêt pour cette application. Dans notre proposition, une description LOM est rattachée à chaque objet pédagogique. Les méta-données utiles pour l'application donnée peuvent donc ensuite être filtrées via le profil d'application, en fonction de celle-ci. De la même façon, lorsqu'un objet est utilisé dans une formation donnée, certaines valeurs des méta-données associées à la formation elle-même sont automatiquement renseignées pour les objets pédagogiques associés. De notre point de vue, ces informations ne correspondent pas à une connaissance nécessitant leur modélisation à travers une ontologie. Aussi, ces informations sont-elles simplement rattachées à la formation d'une part, aux objets pédagogiques d'autre part.

LOM est de mise pour l'annotation et l'indexation des ressources pédagogiques. Les méta-données associées permettent de renseigner, d'une manière bien classifiée, les différentes informations nécessaires sur chaque objet

d'apprentissage, de façon à ce que les recherches ultérieures soient rendues plus efficaces. Cependant, comme nous l'indiquions plus haut la représentation sémantique des contenus n'est pas suffisante pour permettre leur ré-utilisation, complète ou partielle dans d'autres applications ou d'autres systèmes.

3.2. Description thématique

Les objets pédagogiques sont également représentés par rapport aux thématiques ou notions qu'ils abordent. Dans le but de pouvoir réutiliser des documents abordant une notion traitée dans le cadre de plusieurs formations ou de plusieurs modules, les objets sont indexés à partir des concepts d'une ontologie de domaine du thème décrivant les thématiques abordées dans la matière considérée (comme par exemple l'informatique). Cette ontologie décrit l'ensemble des notions en lien avec cette matière et les représentent à partir de leur lien sémantique (par exemple, la notion de " base de données relationnelles " se " conceptualise " à partir d'un "modèle Entité-Association"). L'apprentissage d'une notion donnée pouvant demander un certain nombre de pré requis de connaissances, les notions servant à introduire une notion particulière sont également représentées dans l'ontologie (par exemple la notion de cardinalité d'une relation doit être assimilée pour appréhender le modèle Entité-Association "). Les ontologies du thème sont construites à partir de méthodes manuelles ou semi-automatiques présentées dans [MAEDGE04] et [HERNANDEZ05].

La représentation sémantique du contenu des objets pédagogiques à partir d'une ontologie du thème présente différents avantages. Pour un module donné (par exemple un module de bases de données), les notions à assimiler sont précisées dans l'ontologie du thème considérée. Lorsque l'enseignant souhaite concevoir sa leçon, il peut ainsi avoir accès à l'ensemble des objets pédagogiques qui ont été indexés à partir des notions spécifiées pour le module. Il peut alors réutiliser les objets ou décider d'en concevoir de nouveaux s'ils ne lui conviennent pas, éventuellement à partir des composants présents. Du point de vue de l'apprenant, l'accès aux différentes notions en lien avec la formation et le module suivis lui permet de situer ses connaissances (acquises ou à acquérir) dans le contexte d'apprentissage.

3.3 Description des théories pédagogiques

Différents types d'approches pédagogiques existent :

- Empiriste : Pour l'empiriste, comprendre une réalité donnée, c'est avant tout savoir de quoi elle est faite, quels sont les faits qui la constituent.

- Rationaliste : Pour le rationaliste, comprendre une réalité donnée, c'est de saisir la loi d'organisation de cette réalité, sa structure, abstraction faite du contenu particulier des faits.

- Interactionniste : L'apprentissage est fondamentalement abordé comme le processus par lequel le savoir circule, se construit et se transforme au sein d'une communauté, d'un groupe social. Dans cette perspective, apprendre, pour l'individu, c'est participer à ce processus collectif de co-construction du savoir.

Selon les théories d'apprentissages suivies, chaque pédagogie appartient à un type d'approche pédagogique (Empiriste, Rationaliste, Interactionniste) [LEBRUN02] et est normalement constituée de plusieurs étapes à suivre. Une pédagogie choisie pourra donner lieu à plusieurs scenarii pédagogiques (Méthodes).

Une Etape désigne le découpage théorique d'une approche pédagogique donnée. Une étape peut être une phase d'information, de motivation, d'interaction, de production ou d'analyse, etc …Une étape est associée à plusieurs actes dans le scénario pédagogique. En ce qui concerne l'ontologie des théories pédagogiques, le concept Pédagogie décrit l'ensemble des théories d'apprentissages qui peuvent être utilisées pour bien mener des formations.

Les connaissances associées aux théories pédagogiques sont représentées sous forme d'ontologies. Cette représentation se justifie par le fait que nous souhaitons pouvoir associer des aspects raisonnements. Plus spécifiquement, l'aide à la construction d'un scénario à partir d'objet pédagogique sera guidé par la connaissance préalable de la théorie d'apprentissage sous jacente. Bien que les objets pédagogiques ne soient pas directement représentés à partir de cette ontologie, elle influence leur intégration dans le scénario pédagogique.

3.4. Le scénario pédagogique

IMS LD propose de modéliser la séquence des activités d'apprentissage attribuées à chaque rôle pour que l'objectif visé par l'apprentissage soit réalisé, tout en suivant une pédagogie bien déterminée. Les connaissances nécessaires pour prendre en compte les scenarii d'apprentissages sont les suivantes :

- Connaissance sur l'ensemble de tous acteurs qui participent à l'aboutissement d'une formation. Il est représenté par le Rôle dans notre modèle :" Enseignant ", " Apprenant "," Tuteur ", ou un " Administratif ". A chaque rôle est associé un ensemble d'activités à réaliser.

- Connaissance sur le déroulement de l'apprentissage d'un cours dans lequel le document est utilisé (scénario). IMS-LD l'appelle Méthode, il peut contenir une ou plusieurs pièces. Une Pièce est composée d'Actes qui sont exécutés séquentiellement. Les actes sont composés de Partitions qui associent un rôle à une activité effectuée dans un Environnement composé d'objets pédagogiques et de services (chat, forum, supports de cours …).

- Connaissance sur les activités dans lesquelles le document est utilisé. Dans notre modèle, l'Activité décrit les tâches interactives qui se déroulent

entre les différents acteurs à travers le système pour l'apprentissage d'une notion donnée (lecture d'une ressource pédagogique, test, simulation, une auto-évaluation, un exercice, un dialogue ou interaction directe entre apprenant et tuteur, etc...) Elle traite un ensemble de notions et de compétences.

- Connaissance sur le Contexte d'utilisation de l'objet pédagogique :
la réalisation d'une activité peut utiliser ou manipuler des Objets Pédagogiques comme support ou référentiel dans un contexte d'utilisation donné. Ainsi, un même objet pédagogique peut être considéré ou valorisé différemment d'une activité (d'une formation) à l'autre. Le Contexte nous permet de décrire l'usage de l'objet pédagogique dans l'activité.

L'ensemble de ces connaissances est représenté grâce à une ontologie. Des relations entre concepts sont introduites. Par exemple, le concept Pédagogie de l'ontologie des théories éducatives est relié au concept Méthode de l'ontologie du scénario pédagogique, cela nous permet suivant la pédagogie choisie par l'auteur du cours, de le guider dans la conception du document. De même le concept Notion de l'ontologie du domaine est relié au concept Activité car l'apprentissage d'une notion peut se réaliser dans une ou plusieurs activités et cela nous permet de prévoir la réutilisabilité de la ressource. Un même objet peut être utilisé pour plusieurs notions et dans plusieurs activités. On exprime ainsi la réutilisabilité des objets pédagogiques. C'est l'agencement séquentiel des objets pédagogiques dans différentes activités qui assure la conformité avec le scénario pédagogique choisi.

3.5. Modèle global

Ainsi notre modèle se base sur une représentation multi-facettes des documents.

- Nous utilisons une ontologie de thème [HERNANDEZ05] de la formation qui rassemble les thèmes, notions et connaissances à appréhender.

- Une ontologie des tâches décrit les différentes activités d'apprentissages et d'enseignements, les organisations mises en place ainsi que les objets pédagogiques utilisés. Elle respecte la norme IMS-LD.

- Inspirée d'EML-OUNL [KOPER01], l'ontologie des théories pédagogiques décrit l'ensemble des différentes approches pédagogiques existantes.

- Enfin, la description LOM permet de créer des profils d'applications pour la description des méta données utilisées aussi bien pour l'annotation des objets pédagogiques que pour la recherche de ces dernières, en réponse à des requêtes utilisateurs et/ou celles du système.

La construction des ontologies des tâches et des théories pédagogiques ne pose pas de problème particulier et son résultat peut être ré-exploité pour différentes applications. En revanche, une nouvelle ontologie de thème doit

être construite pour chaque nouveau domaine de formation abordé. Notons toutefois que des méthodologies existent afin d'aider cette élaboration [HERNANDEZ05].

En utilisant ces quatre aspects tout en respectant les normes relatives à l'apprentissage en ligne que nous avons citées plus haut, nous arrivons au modèle de l'application suivant (Figure 2).

Le modèle que nous proposons traite tous les aspects du contexte de l'objet pédagogique :

- le niveau dans lequel l'objet est abordé est mentionné grâce aux méta-données de LOM,

- son utilisation dans différents domaines spécifiques se traduit grâce aux différentes formations,

- les différents pré-requis sont considérés grâce aux notions pré-requises et aux compétences pré-requises,

- son usage dans différentes activités est précisé dans la classe contexte.

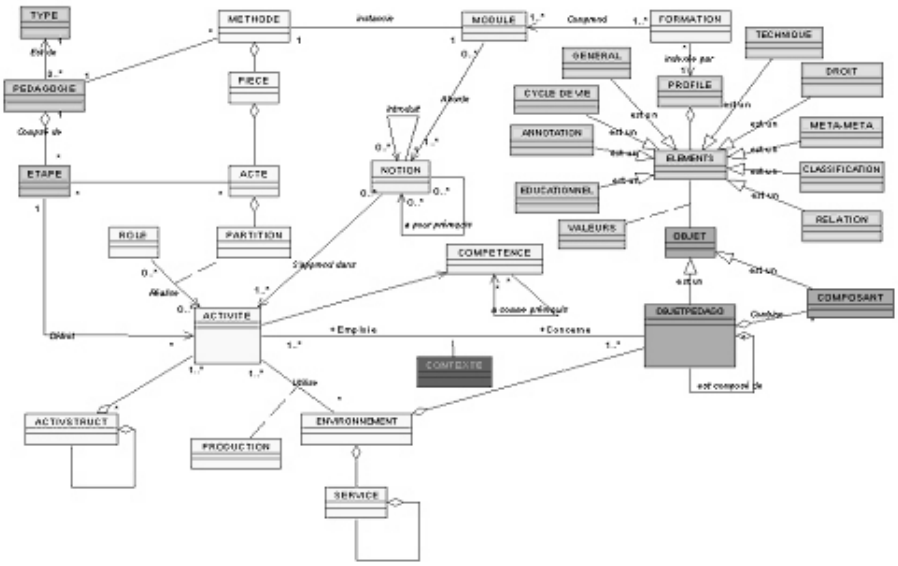

Figure 2 : Modèle complet intégrant les différents aspects
de représentation d'un document dans son contexte d'utilisation

4. Illustration dans le cadre d'une formation

Le modèle proposé a été instancié dans le cadre d'une formation en informatique. Nous présentons dans cette section, un exemple de représentation d'un objet pédagogique préparé pour le module Base de Données Relationnelles des L3. L'objet pédagogique représenté est un exercice portant sur l'indexation de fichiers de données. Sa représentation à partir de notre modèle à facettes est illustrée afin de montrer l'intérêt d'une telle représentation lors de son intégration dans le système d'apprentissage (que ce soit pour une utilisation par l'enseignant ou par l'apprenant).

Par rapport à sa structure, cet objet est constitué de trois objets élémentaires : deux images de b-arbres en jpg et un énoncé. La décomposition de l'objet à partir de sa structure permet d'envisager la réutilisabilité de chacun des éléments.

Les éléments élémentaires ainsi que l'objet pédagogique qu'ils constituent sont indexés par rapport aux méta-données LOM. Par exemple, la méta-donnée "droit" des deux images est définie avec la valeur " public ". Ceci permet à n'importe quel enseignant ou apprenant d'y accéder et de l'utiliser. Par contre, pour l'énoncé cette méta-donnée est définie comme propre aux personnes de la formation. Ceci implique que l'énoncé ainsi que l'objet pédagogique exercice ne pourra être réutilisé que par des enseignants de la formation et consulté par des apprenants inscrits. La méta-donnée pédagogie-niveau relative à l'exercice est fixée à initiation. Ceci indique que l'exercice s'adresse à des étudiants n'ayant jamais étudié l'indexation de fichiers et qu'il pourra être réutilisé dans le cadre d'autres modules s'adressant à un public ayant le même niveau.

L'objet pédagogique est également représenté à partir des notions qu'il aborde. Un extrait de l'ontologie du thème de l'informatique est présenté dans la figure 3. Les concepts sont représentés par des rectangles contenant les différents labels ou termes permettant de définir les notions, les flèches légendées représentent les relations sémantiques entre concepts. Dans l'extrait proposé, les concepts doublement encadrés représentent les notions à aborder dans le cadre du module Base de Données Relationnelles. Une des images représente un b-arbre + , et l'autre un b-arbre *. Les deux images sont donc dans notre modèle représentées à partir de ces deux concepts de l'ontologie de l'informatique. L'exercice quant à lui aborde les notions d'organisation indexée des données à partir de b-arbres tout en insistant sur le temps d'accès aux données. Ces trois concepts de l'ontologie sont donc utilisés pour l'indexer. L'avantage de spécifier ces concepts est qu'un enseignant pourra ré-utiliser cet exercice lorsqu'il voudra travailler sur les notions précédemment citées.
Dans le cadre de son utilisation dans un scénario pédagogique, l'exercice proposé pourra être utilisé à différents niveaux. Si l'on considère la pédagogie de

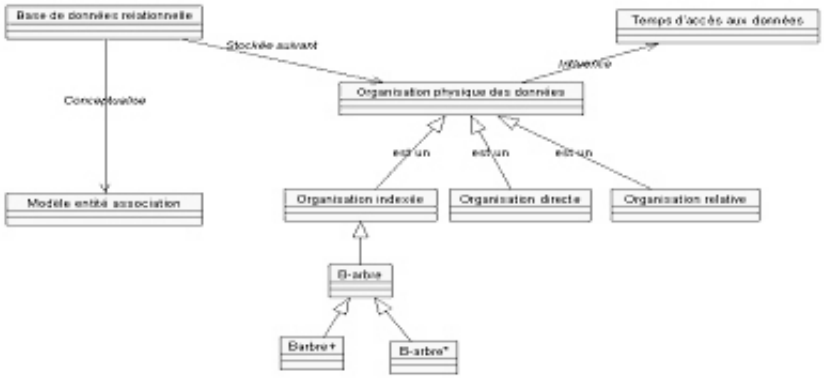

Figure 3 : Extrait de l'ontologie du thème de l'informatique

Gagné, il pourra être utilisé pour provoquer la performance, fournir la rétroaction et évaluer la performance. Son utilisation est indiquée dans le diagramme d'activités présenté figure 4 au niveau des actes en gras (4 et 5). L'enseignant présente les notions à appréhender aux étudiants. Ces derniers lisent les chapitres correspondants et demandent des explications à l'enseignant. L'enseignant donne des explications à l'aide des exemples et attendent une nouvelle réaction des étudiants, et donne éventuellement des suppléments d'explications au cas où les étudiants en demandent. Puis, le délai écoulé, l'enseignant donne l'exercice que les étudiants doivent traiter. Ensuite, l'enseignant corrige des exercices tout en fournissant des explications précises sur chaque point non maîtrisé par l'étudiant, par courrier électronique, dialogue en direct ou sur forum. Afin d'assurer la compréhension et la rétention des notions à appréhender, des exercices d'évaluation et d'auto évaluation sont fournis aux étudiants. Finalement, l'enseignant donne un résumé des points à retenir sur la notion étudiée.

5. CONCLUSIONS ET PERSPECTIVES

Le modèle que nous avons proposé peut être appliqué à tout document électronique à objectif pédagogique. Il s'intègre dans tout modèle d'apprentissage (formation à distance synchrone et asynchrone, présentiel ou non,…). Il présente aussi la possibilité d'utilisation et d'adaptation sur tout type de formation car il n'a pas été conçu dans le cadre d'une formation spécifique. L'originalité du modèle est la représentation multi-facettes des documents en utilisant trois ontologies : ontologie de thème, ontologie des tâches, ontologie des théories pédagogiques et une description LOM/SCORM. Les ontologies que nous proposons seront représentées en OWL (langage Web Ontology Language) [W3C04] basé sur la syntaxe RDF/XML. Il permet de représenter explicitement la signification des termes de vocabulaires et les relations entre les concepts associés.

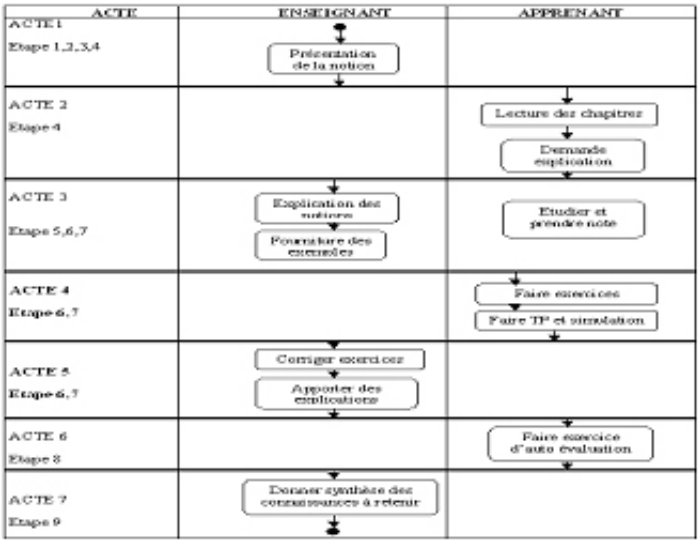

Figure 4 : Diagramme d'activité du scénario pédagogique intégrant l'exercice

Un des points forts du modèle est que les ontologies permettent de couvrir tout le cycle de vie d'un document : depuis sa conception par l'enseignant jusqu'à la recherche d'information de l'apprenant sur une notion donnée. Le modèle permet de développer un système permettant d'assister les utilisateurs dès la conception même des ressources. Les différents acteurs du système (enseignants, apprenants) et leurs différentes tâches sont considérées. Il fournit la souplesse d'une représentation sémantique et de la recherche d'objets pédagogiques pertinents par l'utilisation des profils d'application du LOM. Le respect des normes et standards d'apprentissage en ligne garantit l'interopérabilité avec d'autres systèmes.

Pour poursuivre notre étude, nous envisageons de travailler sur l'indexation des documents à partir de l'annotation sémantique que nous proposons. Nous travaillons sur une implantation dans le cadre de documents pédagogiques créés dans le cadre d'un environnement pour la formation à distance. Nous poursuivrons alors notre étude en mettant en place un système qui permet le suivi personnalisé des activités menées par chaque apprenant et qui permet l'utilisation d'interfaces variables et dynamiques suivant le modèle pédagogique choisi.

6. RÉFÉRENCES :

[PSYCHE05] Psyché V., Bourdeau J., Nkambou R. et Mizoguchi R. (2005). Making Learning Design Standards Work with an Ontology of Educational Theories. AIED 2005.

[IMSLD03] http://www.imsglobal.org/learningdesign/

[SCORM04] Le modèle SCORM (2004). http://www.adlnet.org

[LOM02] LOM (2002). LOM standard, document IEEE 1484.12.1-2002.
[WILEY01] Wiley D.A. (2001). Connecting Learning Objects to Instructional Design Theory: a Definition, a Metaphor, and a Taxonomy. Wiley, David A. (Ed.) The Instructional Use of Learning Objects, p 1-35.

[MIZOGUCHI04] Mizoguchi R. (2004). Le rôle de l'ingénierie ontologique dans le domaine des EIAH. Sticef Volume 11 (Numéro spécial : Ontologies pour les EIAH).

[E-TUD] http://www.e-tud.com/encyclo_e-learning.htm

[BOUTEMEDJET04] Boutemedjet S. (2004). Web Sémantique et e-Learning. Cours FT6261.

[GASEVIC05] Gasevic D. et Hatala M. (2005). Searching context relevant learning resource using ontology mappings. International Workshop on Applications of Semantic Web Technologies for E-learning (SW-EL), Winston-Salem State University, p 45-52.

[LENNE05] Lenne D., Abel M.-H., Moulin C. et Benayache A. (2005). Mémoire de formation et apprentissage. EIAH 2005, Montpellier, p 105-116.
[ABEL03] Abel M.-H., Lenne D., Moulin C. et Benayache A. (2003). Gestion des ressources pédagogiques d'une e-formation. Document Numérique 7(1-2), p 111-128.

[KNIGHT05] Knight C., Gasevic D. et Richards G. (2005). Ontologies to integrate learning design and learning content. Journal of Interactive Media in Education (07).

[STUDER98] Studer R., Benjamins VR. et Fensel D. (1998). Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering (DKE), 25(1-2), p 161-197.

[HERNANDEZ05] Hernandez N. (2005). Ontologies de domaine pour la modélisation du contexte en Recherche d'Information. Thèse de l'Université Paul Sabatier.

[FAGE05] Fage C. (2005). Vous avez dit SCORM. eLearning Agency, p 1-14.

[DUVAL02] Duval E., Sutton S. et Weibel SL. (2002). Metadata Principles and Practicalities. D-Lib Magazine 8(4), p 1-16.

[MAEDCHE04] Maedche A. et Staab S. (2004). Ontology Learning. Handbook on Ontologies, S Staab, R. Stubers (Eds.), p 173-190.

[LEBRUN02] Lebrun M. (2002). Des technologies pour enseigner et apprendre, De Boeck (2ème édition).

[KOPER01] Koper R. (2001). EML-OUNL (Open University of the Netherlands' Educational Modeling Language), Modeling Units of Study from a Pedagogical Perspective.

[W3C04] W3C Consortium (2004), OWL Specification Development, http://www.w3.org/2004/OWL/#specs Feb 2004.

Document pour l'Action comme media pour la Gestion de Connaissances

Samuel PARFOURU
Alain GRASSAUD
Sylvain MAHE
Manuel ZACKLAD

ISTIT - Laboratoire Tech-CICO,
12 Rue Marie Curie 10010 TROYES, France
{samuel.parfouru,manuel.zacklad}@utt.fr
EDF Recherche & Développement,
6 quai Watier 78400 CHATOU, France
{alain.grassaud,sylvain.mahe}@edf.fr

RÉSUMÉ

Dans cet article, nous proposons une approche originale d'analyse des Systèmes de Gestion de Connaissances (SGC) basée sur un rapprochement avec la notion de document numérique. Dans un premier temps, nous abordons les théories en rapport avec le document numérique sur lesquelles nous nous appuyons. Ainsi, nous introduisons la notion de document et ses reformulations par son passage au numérique puis la notion de Document pour l'Action. Nous proposons ensuite une analyse du SGC selon trois axes en le considérant successivement comme un medium, comme un signe et comme une forme. Cette étude est étayée par des éléments concrets relatifs à un projet de gestion de connaissances réel dans lequel nous avons mis en œuvre et implémenté une approche basée sur l'écriture et la manipulation de documents numériques. L'ensemble de l'analyse permet ainsi de dégager une parenté entre le développement de SGC et la notion de document. Ainsi, les différents acteurs gravitant autour du SGC, sur l'ensemble de son cycle de vie, prennent successivement un statut d'auteur, d'éditeur et de lecteur qui nous place dans un contexte " hyper rédactionnel ". La notion de Document pour l'Action (DopA) prend alors toute sa place : le DopA devient un media pour la gestion de connaissances.

http://www.utt.fr/labos/TECH-CICO

MOTS-CLES : Gestion de Connaissance, Coopération Médiatisée (CSCW), Document Numérique, Document pour l'Action (DopA), Multi domaines, Multi points de vue.

32. INTRODUCTION :

Les Systèmes de Gestion de Connaissances (SGC) visent généralement deux objectifs complémentaires [DIEN01]. D'une part la " capitalisation " des connaissances expertes détenues par un ou plusieurs spécialistes d'un domaine et d'autre part le traitement et la restitution de ces connaissances au moyen d'un " dialogue homme machine " (DHM) permettant de guider l'utilisateur dans le contexte d'une " situation problème " particulière.

Dans le domaine de l'intelligence artificielle, la vue qui a le plus longtemps prévalue était celle de la conception d'un programme qui " simulerait des raisonnements experts ". Ce point de vue privilégiait la recherche d'un niveau d'expression adéquat pour la connaissance. Une autre vue alternative, défendue notamment dans le cadre des conférences Cooperative System Design (COOP) était celle d'un programme permettant une forme de "coopération" entre le système et l'utilisateur. Ce point de vue privilégiait la problématique de la restitution et de l'appropriation. L'approche que nous présentons ici, plus inspirée du Computer Supported Cooperative Work (CSCW) [SCHI96] et des Sciences de l'Info@rmation, est de considérer que les Systèmes de Gestion de Connaissances sont des systèmes qui médiatisent l'activité collective d'un réseau d'acteurs coopérant de manière asynchrone et délocalisée sur la durée grâce à la médiation du système.

Dans ce papier, nous proposons une analyse du SGC en argumentant qu'il peut être appréhendé comme un document numérique. Dans un premier temps, nous abordons les théories relatives au document numérique sur lesquelles nous nous appuyons. Nous analysons ensuite le SGC selon qu'il peut être vu comme un medium, comme un signe ou comme une forme. Chaque partie de notre étude est étayée par des éléments concrets d'analyse et d'implémentation s'inscrivant dans un projet réel de gestion de connaissances. Ce projet renvoie à une problématique de capitalisation des connaissances théoriques et des pratiques métiers de diagnostic liées à de grands ouvrages hydrauliques. Ce système, que nous nommerons guide par la suite, est construit de façon à représenter un support à la conservation et la diffusion d'une représentation des connaissances et des pratiques liées au diagnostic. Enfin, nous terminons par tracer un bilan de l'analyse décrite dans ce papier, en appuyant l'intérêt d'une approche documentaire et le rapprochement possible entre SGC et Document pour l'Action. Finalement, nous dressons quelques perspectives à ce travail.

33. THÉORIES ACTUELLES DU DOCUMENT NUMÉRIQUE

Notre intention ici n'est pas de proposer une revue complète des théories actuelles sur le document numérique. Nous introduisons seulement un aperçu du texte fédérateur traitant de la notion de document et de sa reformulation au travers du passage au numérique [PEDA03]. Nous abordons ensuite la

notion de document pour l'action (DopA) [ZACK04], approche sociale de la notion de document avec un point de vue CSCW.

33.1. La notion de document dans son passage au numérique

Nous proposons ici de rappeler les 3 entrées sous lesquelles peut être étudiée la notion de document ainsi que l'impact de son passage au numérique [PEDA03]. Ainsi, le document peut être étudié selon qu'il est analysé comme forme, comme signe ou comme medium. Il ne s'agit pas là d'une " partition " et ces 3 aspects sont évidemment interdépendants.

Le document comme forme est appréhendé comme un objet de communication où les règles de mise en forme définissent un contrat de lecture. On peut définir l'équation : Document traditionnel = support + inscription. Cette équation peut se traduire dans le monde numérique par : Document numérique = structure + données. La mise en forme de la structure n'est alors pas partie prenante dans l'équation, et se trouve reléguée à offrir une lisibilité partagée entre concepteur et lecteur. Les évolutions dans le domaine du document numérique, et notamment toute la philosophie accompagnant les technologies XML, redéfinissent cette dernière équation pour aboutir à : Document XML = données structurées + mise en forme. La mise en forme (le " style ") est définie indépendamment des données structurées, elle devient modulable et prend donc une place importante dans le sens du document. Le " style " traduit alors un mode de représentation, une objectivation, des données structurées.

Le document comme signe traite le document comme un objet signifiant. Une inscription est associée à un sens : Document = inscription + sens. Dans le monde numérique, on peut modifier l'équation et aboutir à : Document numérique = texte informé + connaissances. La substitution " d'inscription " par " texte informé " tend à signifier que le texte a été soumis ou pourrait être soumis à un traitement permettant d'en repérer les unités d'information. Le remplacement de " sens " par " connaissances " voudrait introduire la notion de personnalisation pour un lecteur ou un usager donné. L'important travail sur le Web Sémantique tendrait à aboutir à l'équation Document Web Sémantique = texte informé + ontologies.

Le document comme medium peut se traduire initialement par l'équation : Document = inscription + légitimité qui semble représenter le processus social de mise en document. Le statut de document s'acquérrait sous deux conditions : l'inscription doit dépasser la communication intime (entre quelques personnes privées) pour devenir légitime et la légitimité doit s'affranchir de l'éphémère (dépasser le moment de son énonciation) et donc être enregistrée, inscrite. L'équation précédente avec le passage au numérique devient Document numérique = texte + procédure. Le document numérique correspond alors à une trace d'une relation sociale ou d'une

pratique calculée au travers d'un processus informatique. Le document WEB traduit sous l'équation Document Web = publication + accès repéré introduit la notion de repérage dans l'accès à l'information pouvant traduire une certaine légitimité.

33.2. Le document pour l'action (DopA)

Le document pour l'action (DopA) [ZACK04] permet de redéfinir le concept de document et d'appréhender autrement leurs contenus qui relèvent de moins en moins de la catégorie du texte classique. En insistant sur la dimension collective de l'activité rédactionnelle il permet d'analyser les documents comme relevant de processus de communication pour partie différés, au sens des processus asynchrones décrits dans le champ du CSCW, entre des producteurs et des récepteurs liés par des intérêts communs. Alors que la conceptualisation en termes d'hypermédia visait surtout à rendre compte des nouvelles pratiques de lecture associées aux hypertextes, le DopA vise à rendre compte des processus " d'hyper-rédaction " associés aux documents numérisés. Le DopA se définit alors comme un ensemble de fragments portés par des auteurs divers dont le contenu final reste largement indéterminé alors même que sa circulation rapide lui fait déjà jouer un rôle majeur d'information, d'aide à la décision et de preuve [ZACK04].

Notons bien que cette définition vient pour partie en contradiction avec le document vu comme medium évoqué précédemment. En effet, le DopA peut d'une part renvoyer à une communication " intime " : dans une situation de travail, la feuille de brouillon exploitée par une personne dans une situation de résolution de problème peut prendre le statut de DopA. D'autre part, il n'exclut pas le caractère " éphémère ", ou tout au moins temporaire, ce qui vient en contradiction avec les attributs de légitimité. Ainsi, à l'image de ce qu'évoque Morand [MORA04] concernant l'utilisation des diagrammes en conception logicielle, le DopA peut être vu comme un support d'espace de résolution de problème : les instances stables ou même finales sont d'une importance secondaire par rapport aux évolutions entre chacune d'elles qui reflètent bien l'activité associée à la construction du signifiant.

Au travers des théories que nous venons de présenter, nous proposons d'aborder le Système de Gestion de Connaissances en le considérant comme un document numérique. Pour cela nous allons analyser le SGC en le traitant successivement comme un medium, comme un signe et enfin comme une forme.

34. APPROCHE SOCIALE DU SYSTÈME DE GESTION DE CONNAISSANCES : LE SGC COMME MEDIUM

Un Système de Gestion de Connaissances quel qu'il soit doit nécessairement être vu et analysé sous une approche sociale de prime abord. En effet, un

projet de gestion de connaissances nécessite la mise en place d'un dispositif particulier monopolisant un réseau d'acteurs qui vont devoir collaborer· Classiquement, ce réseau sera composé de 4 types d'acteurs :

11. L'analyste : Il s'agit de la personne qui va mener la démarche de capitalisation des connaissances. Il doit cultiver la dynamique de cette démarche et nous le verrons par la suite y joue un rôle de médiateur. Il conduit le processus de capitalisation en tenant des réunions avec les experts afin de les accompagner dans la formalisation des connaissances. L'analyste mène une activité de co-construction des modèles, de la formalisation et éventuellement des algorithmes.

12. L'expert : Il s'agit de celui qui possède ce que l'on pourrait appeler le savoir métier. Il a acquis cette qualité d'expert du fait d'une longue expérience de terrain ou encore du fait d'un savoir théorique du phénomène ou de l'objet auquel on s'intéresse. Dans le cadre d'un vaste projet de gestion des connaissances, on tentera classiquement de faire appel à plusieurs experts, chacun apportant un point de vue différent sur le ou les domaines que l'on cherche à formaliser.

13. L'opérationnel : Que la finalité soit la production d'un document ou bien un système d'information, un projet d'Ingénierie des Connaissances fait référence à un ou plusieurs opérationnels qui exploiteront le résultat du projet dans des pratiques métiers.

14. Le commanditaire du projet : Il s'agit du porteur de la problématique à traiter. Il peut être à l'origine de cette problématique, mais pas toujours. C'est bien évidemment un acteur tout à fait important, avec qui il convient de savoir négocier. En effet, le commanditaire peut avoir une vision à priori de ce qu'il voit, ce qu'il est capable d'imaginer qui peut se révéler en contradiction avec une réalité de terrain identifiée par l'analyste.

Le développement du guide et des SGC en général fait le plus souvent référence à un contexte difficile. En effet, au-delà de la " simple " multiplicité des acteurs, l'analyse des pratiques de diagnostic sur de grands ouvrages implique nécessairement de devoir solliciter plusieurs domaines de compétence. Ainsi, dans les cas d'application du guide, il s'agit de consulter des experts en hydro mécanique, en génie civil ou encore en contrôle commande. Nous sommes alors confrontés à un contexte d'étude multi domaines et multi points de vue.

34.1. Un contexte multi domaines et multi points de vue : l'analyste en médiateur

L'objectif du guide est d'aboutir à une description consensuelle et généraliste du diagnostic où chaque compétence sollicitée doit pouvoir se retrouver et s'identifier. L'analyste dans sa démarche joue un rôle de médiateur entre des

compétences qui amènent chacunes un vocabulaire et des points de vue différents. A cela s'ajoutent des contraintes liées aux aspects social et organisationnel. Les compétences sollicitées peuvent ainsi introduire des clivages : le guide, en sa valeur de media qui tente de véhiculer une représentation consensuelle, doit alors être un point d'entrée à la construction d'un terrain commun [NICO03] pour faciliter le dialogue et la coopération entre ces domaines d'expertise.

Nous avons évoqué le fait que chaque domaine d'expertise introduit un point de vue différent. De même le commanditaire peut apporter un autre point de vue et se révéler intrusif dans le dispositif mis en place par l'analyste pour mener la capitalisation. De plus, il peut avoir une vision à priori du SGC à produire. Par exemple dans une démarche de capitalisation de connaissances en rapport avec l'identification de bonnes pratiques, il pourra s'attendre à ce que l'on produise un outil prescripteur [LEPL04]. Ceci pourra aller à l'encontre des attentes de l'opérationnel qui n'est pas nécessairement demandeur d'une procédure imposée, mais plutôt d'un outil support à son activité, se portant en aide mémoire. Il s'agit donc pour l'analyste de s'adapter à la demande du commanditaire tout en trouvant parfois les arguments pour faire " évoluer " sa vision.

Les opérationnels représentent des utilisateurs qui exploiteront une restitution des connaissances dans une pratique métier qui leur est propre. Ils sont les seuls en mesure de valider le système, et particulièrement le mode de restitution, en termes d'usage et d'utilisabilité [MALL02]. Leur point de vue est donc essentiel et doit être pris en compte dans le travail de médiation de l'analyste puisqu'il influencera nécessairement la définition du processus de restitution des connaissances.

Notons que les visions des différents acteurs sont bien évidemment changeantes à mesure que le projet évolue. Cette fluctuation est liée à l'interaction entre les différentes compétences, à la confrontation des points de vue ainsi qu'à l'impact des résultats intermédiaires du projet - pouvant faire référence aux connaissances formalisées ou encore le moyen technologique ou non qui sont mis en œuvre pour les présenter - lorsqu'ils sont proposés aux acteurs.

34.2. Des acteurs en interaction médiatisée : le SGC comme media à la coopération

Le contexte multi domaines et multi points de vue inhérent à la réalisation du guide, introduit bien le fait que le développement d'un Système de Gestion de Connaissances représente une activité collective nécessitant une coopération forte entre les acteurs. Il est essentiel qu'un dialogue s'instaure entre eux pour converger vers un résultat consensuel. Aboutir à ce résultat peut d'ailleurs être appréhendé comme une situation coopérative de résolution de problème relative à la recherche d'un équilibre négocié entre les différents

acteurs en rapport avec la représentation des connaissances en terme de formalisation mais également de restitution.

La section précédente a permis de dégager l'activité de médiation que doit jouer l'analyste. Il doit mener une réelle négociation entre expert(s), commanditaire(s), opérationnel(s). Il est intéressant d'évoquer un instant le processus mis en œuvre dans la problématique du guide. En effet, durant la phase de capitalisation, l'analyste va rencontrer dans un premier temps les acteurs de manière indépendante. A l'issue de cette première étape, il va formaliser les différentes visions qu'il a pu réunir durant les différents entretiens et tenter d'en proposer une représentation qu'il proposera aux différents acteurs impliqués dans le processus, afin que chacun puisse intervenir, critiquer, annoter ce qui lui est proposé. Ce processus se poursuit jusqu'à aboutir à une stabilisation de la formalisation : une représentation consensuelle. Cette démarche illustre bien le développement du guide comme une activité collective d'un réseau d'acteurs coopérant sur la durée de manière asynchrone et délocalisée.

En outre, cette activité se traduit par de nombreuses interactions et échanges entre les acteurs, l'analyste se plaçant en pivot, en médiateur. Les échanges sont alors largement médiatisés. Les medias touchent à plusieurs modalités puisqu'on peut observer des échanges vocaux (en présentiel ou non), des échanges de courriels ou plus généralement de documents qui seront autant de productions soumises à un processus " annotatif ". Dans ce contexte, le guide, dans sa globalité, représente lui-même un media puisqu'à mesure de son évolution, il véhicule la formalisation et les façons de la restituer entre les acteurs. Nous sommes alors tentés de définir l'ensemble de ces medias porteurs d'informations et en constante évolution, à mesure que l'on avance dans le cycle de vie du SGC, comme des documents, des documents pour l'Action. En ce sens, le guide peut alors être vu comme un document multi sources, dynamique, et devant offrir une interactivité, médiatisant une représentation des connaissances : il constitue alors un Document pour l'Action à part entière. Ainsi, les acteurs prendront tantôt le rôle d'auteur, d'éditeur ou de lecteur face à la représentation des connaissances.

35. LE SYSTÈME DE GESTION DE CONNAISSANCES COMME SIGNE : LA FORMALISATION DES CONNAISSANCES

Nous venons d'argumenter que le SGC pouvait être considéré comme un media de la coopération, par la représentation des connaissances qu'il véhicule auprès des acteurs. Cette représentation doit alors constituer un objet signifiant interprétable par ces acteurs. La construction de cet objet signifiant résulte en partie du processus de formalisation des connaissances où l'analyste joue un rôle prépondérant.

L'analyste apporte un savoir faire de modélisation et conduit largement la définition de l'objet signifiant en question. Notons bien qu'il s'agit d'une

co-construction entre les acteurs. En effet, l'analyste ne peut être lui-même un expert de l'ensemble des domaines auxquels il est confronté. Ceci nous amène à introduire, au-delà de son rôle de médiateur, qu'il est accompagnateur dans la formalisation. Ainsi, plutôt que d'analyser les acteurs et les informations qu'ils fournissent, il se situe plutôt dans une activité de " mise en analyse ". Les acteurs impliqués dans la capitalisation, et notamment les experts, sont alors incités à s'interroger sur leur propre savoir et savoir faire.

35.1. Formalisation multi domaines et multi points de vue

Les aspects multi domaines et multi points de vue sont un enjeu de premier plan dans la définition de l'objet signifiant que doit représenter le SGC. En effet, il incombe à l'analyste d'être capable d'interpréter les informations recueillies au sein des différents domaines de compétence consultés - ce qui va lui demander de se constituer une certaine expertise sur ces domaines pour être en mesure de communiquer avec les experts - puis tenter d'en formaliser, en collaboration avec les acteurs, une représentation pivot où chacun puisse d'une part, s'identifier dans le contenu et d'autre part, identifier les autres compétences, tout en se positionnant par rapport à elles. Il s'agit là d'un exercice difficile qui souligne encore une fois le rôle de médiateur de l'analyste mais également le rôle de media du SGC, en tant que support de l'information, dans la coopération entre les acteurs.

Dans la réalisation du guide, nous sommes confrontés à ce contexte d'étude. Il s'agit en effet de réaliser une base de connaissances offrant un panorama général sur l'ouvrage auquel on s'intéresse qui pourra à posteriori être particularisé en fonction des spécificités de chaque ouvrage. Le guide doit donc refléter les informations de chaque domaine, les structurer, en identifiant les interfaces entre chacun d'eux pour souligner leur complémentarité et soutenir leur coopération. La complexité inhérente à l'aspect multi domaines est augmentée par le caractère multi points de vue de la problématique. Les différents points de vue sont en relation avec l'aspect multi domaine - il peut exister des divergences inter domaine mais également intra domaine - mais pas uniquement. En effet, dans l'exemple du guide, il s'agit de traiter et mettre en relation 2 modes de représentation des connaissances. Le premier est lié à une formalisation de la pratique métier du diagnostic qui renvoie à l'usage quotidien des experts tel qu'ils appréhendent leur activité. Il en résulte une représentation séquentielle, parfois descriptive, d'étapes imbriquées dans des phases de diagnostic. Le second type correspond à la définition d'une Analyse des Modes de Défaillance de leurs Effets et de leur Criticité (AMDEC). Cette AMDEC est une représentation fonctionnelle de l'ouvrage qui est liée à sa décomposition matérielle. La constitution de l'AMDEC représente un effort intellectuel important pour les experts puisque différente de leur pratique usuelle. Ce changement de point de vue, introduit par l'AMDEC, est alors tout à fait intéressant car il oblige les experts à beaucoup s'interroger sur leur savoir.
Pour répondre à ce contexte d'étude, nous appuyons que l'utilisation de documents numériques offre à l'analyste un environnement propice à la

formalisation multi domaines et multi points de vue. Ainsi dans la réalisation du guide, plusieurs documents numériques ont été construits pour rendre compte des différents points de vue. Nous avons ainsi dissocié formalisation de la pratique métier et représentation fonctionnelle dans des documents différents. De plus, nous avons exploité la structuration logique interne des documents pour organiser les connaissances relatives aux différents domaines de compétences. Ces différents documents constituent des objets signifiants disjoints, en fonction des points de vue, pouvant être filtrés selon les domaines, permettant de médiatiser une représentation aux acteurs.

De là, ces documents numériques peuvent être mis en relation, associés, fusionnés de manière intégrale ou partielle. Ainsi, le guide dans sa version globale donne lieu à la fusion du document relatif à la formalisation métier et celui en rapport avec le point de vue fonctionnel. De plus, dans le calcul de ce document, on enrichit un point de vue par des fragments documentaires du second en fonction " d'informations communes ". Les informations partagées permettent également de définir de nombreux hyperliens entre ces documents. Les différents documents qui sont produits dynamiquement sont alors autant de Documents pour l'Action se focalisant tantôt sur un point de vue, tantôt sur un domaine ou encore permettant de confronter et d'associer points de vue et domaines permettant ainsi de traduire les interfaces entre eux. Ainsi, dans le guide, la représentation fonctionnelle dans son association avec la formalisation métier vient appuyer et justifier les pratiques de terrain. Il en résulte une vue du diagnostic " argumentée " par des éléments d'analyse fonctionnelle issus de l'AMDEC.

Nous venons d'aborder les possibilités offertes par l'utilisation de documents numériques pour une formalisation multi domaines et multi points de vue. Il convient maintenant d'aborder la problématique de la rédaction de ces documents.

35.2. Formalisation déclarative des connaissances au sein de documents numériques

La capitalisation des connaissances a déjà été évoquée comme une activité peu déterministe. En fait, elle relève d'une démarche pouvant être qualifiée de créative. Il semble alors difficile de prévoir en dehors de la phase de capitalisation elle-même les informations qui seront identifiées, de même que leur structuration effective, ou encore l'ordre selon lequel elles seront saisies et structurées par la personne en charge de cette tâche. Tout ceci est influencé par les aspects multi domaines et multi points de vue de la problématique, ainsi que l'aspect social décrit dans la section 3.

Dans ce contexte, il semble alors peu adéquat de proposer un espace de formalisation contraint imposant une logique de saisie. L'utilisation de documents numériques dans le cadre de la formalisation permet de pallier cette difficulté. Pour cela, nous nous appuyons sur l'utilisation de l'eXtended Markup Language (XML). En tant que meta langage, XML permet une

formalisation très souple des connaissances puisque les balises, leur nom et leur organisation logique peuvent être définis au fur et à mesure de la saisie. Nous avons ainsi pu observer de manière expérimentale, au travers de l'utilisation d'un éditeur XML, qu'une saisie déclarative offre un confort dans l'activité de l'analyste lorsque les structures de données, les modèles, sont peu identifiés. La formalisation, au travers de ce système auteur, se fait selon un processus constructiviste, en extension, faisant évoluer structuration et instanciation de concert, jusqu'à converger vers une certaine stabilité.

Dans notre expérience, le système auteur correspond à un éditeur de code source XML qui offre la possibilité, à l'issue d'une phase de saisie, d'extraire automatiquement la DTD (Document Type Definition) ou le Xschema qui correspondent à la structure logique canonique du document (Modèle de Document). Cette structure logique est essentielle puisqu'elle traduit, tout au moins en partie, une certaine image du modèle de connaissances inhérent à la formalisation produite. L'éditeur XML peut être alors vu comme un espace propice à l'activité créative que nous tentons d'instrumenter. En effet, il est possible de saisir balises et contenu, faire évoluer la structuration ou encore laisser certains éléments vides pour y revenir à un moment plus adéquat en fonction de sa réflexion et ce sans contraintes fortes, si ce n'est rédiger un document bien formé au sens de la norme XML. De plus, l'extraction de la structure logique du document permet d'observer l'évolution du " modèle de connaissances " à intervalles réguliers. Le document XML devient alors un DopA, dont on peut suivre l'évolution au gré de l'activité de l'analyste et pouvant s'intégrer dans un contexte multi auteurs. Cependant, cette méthode de saisie n'est pas sans poser des difficultés.

En effet, la saisie des informations et du balisage - bien qu'elle soit facilitée par l'éditeur- demande une certaine rigueur et un temps d'adaptation. De plus, le balisage peut se révéler difficile à faire accepter car éloigné des interfaces " Wysiwyg " auxquelles la majeure partie des utilisateurs sont habitués. Ainsi, le retour d'expérience concernant le guide nous montre que l'analyste en charge de la formalisation a été confronté à certains phénomènes de désorientation au sein du contenu lorsque le document numérique s'est étoffé et a pris de l'ampleur. Pour pallier ce problème, nous avons introduit un second éditeur qui, lui, est centré sur l'édition du contenu au détriment de la structure. Il offre un espace de saisie qui occulte le balisage en le transformant par des cadres de saisie, étiquetés par le nom de balise correspondante, qu'il est possible de colorer pour faciliter le repérage dans la structure du document. L'éditeur offre une mise en forme qui traduit le balisage par une sémantique graphique accompagnant le processus d'édition.

Dans notre expérience, l'analyste méthode reconnaît l'intérêt avec le premier éditeur de pouvoir faire évoluer structure et données de concert. Cependant, il est également important de fournir un espace d'édition, particulièrement

lorsque la structure se stabilise, même pendant une période temporaire, plus orienté sur le contenu des balises. Les 2 éditeurs véhiculent en fait 2 modes de restitution du contenu qui se révèlent plus complémentaires qu'antagonistes, adaptés à différentes tâches de l'analyste. Ici apparaît le lien entre la "construction " de l'objet de signifiant que doit constituer le SGC et la problématique de la forme.

36. LE SYSTÈME DE GESTION DE CONNAISSANCES COMME FORME : RESTITUER LE CONTENU AUX UTILISATEURS

La définition de la structure et la définition des modalités de restitution constituent des opérations qui sont lourdes de sens et de signification [ECO88][PEDA03] du point de vue de la responsabilité " autoriale " ou plutôt " éditoriale ". En effet, les aspects multi domaines et multi points de vue introduisent différents systèmes de signes [ECO88] qu'il faut tenter de faire cohabiter dans la restitution afin que les acteurs puissent l'interpréter correctement. Le SGC doit donc être construit comme un objet de communication, une forme au sens de [PEDA03].

Les éléments que nous avons évoqués précédemment apportent déjà des réponses à cette problématique. En effet, il est possible de calculer des restitutions selon différentes stratégies à partir des documents numériques produits durant la construction de l'objet signifiant : chaque restitution correspond alors à un DopA en relation avec un point de vue, un domaine, pouvant être adapté, en terme de contenu, à l'action qu'il doit supporter. Nous proposons ici de nous concentrer sur les modalités de restitution.

36.1. Différentes formes de restitution : traduire la sémantique des modèles dans des systèmes de représentation

La définition de la forme de la restitution est en relation avec le contexte d'utilisation où elle doit s'insérer et particulièrement l'utilisateur qui va devoir l'exploiter. Ceci a déjà été illustré par l'utilisation de 2 types d'éditeurs XML qui fournissent une forme de restitution de l'information s'adaptant à une saisie tantôt focalisée sur l'évolution de la structure et des données, tantôt sur les données seules. La problématique qui nous apparaît est alors d'exploiter les possibilités de différents systèmes de représentation pour produire une représentation de l'objet signifiant traduisant les traits sémantiques du modèle de connaissances. Pour répondre à cette problématique, nous implémentons un mécanisme supervisé permettant de traduire une formalisation, se matérialisant par un document numérique, vers différents systèmes de représentation (cf. figure 1 : textuel, graphique…).

Notre mécanisme, dans sa définition, s'appuie largement sur la théorie de l'image [BERT73].

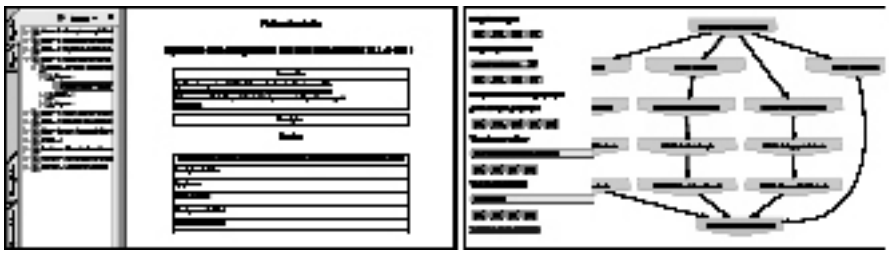

Figure 1. Restitution textuelle et Restitution graphique du guide

Ainsi, nous classifions les éléments constituant la structure logique des documents - qui reflète le modèle de connaissances - selon une typologie qui emprunte les niveaux de perception visuelles (Associative, Sélective, Ordonnée, Quantitative) introduits par Bertin auxquels nous avons ajouté quelques catégories (Entité, Relation, Identifiant, Description). Nous sommes conscient que les catégories ajoutées sont discutables car elles peuvent se révéler redondantes face à certain niveau de perception visuelle ou d'autres notions liées à la théorie de l'image. Par exemple, une information classifiée comme un " Identifiant " peut être vu comme une " Variable Sélective ". De même, les notions " d'Entité " - correspondant à une instance d'une composante au sens de [BERT73] - et de " Relation " renvoient à la construction graphique des réseaux. Ceci étant dans un premier temps, nous avons décidé d'introduire ces catégories pour faciliter nos premières expérimentations et leur compréhension.

De cette classification, il est possible de traduire la formalisation des connaissances dans différents systèmes de représentation (textuelle, graphique). Les différents types d'informations sont ainsi mis en forme de manière " standardiser " et ce en fonction des possibilités (variables) du système de représentation choisi. Le tableau 1 présente la façon dont sont représentées les informations dans le guide. Cette mise en correspondance n'est pas arbitraire et là encore s'appuie sur la sémiologie graphique [BERT73]. L'objectif est bien de contraindre le calcul des restitutions en intégrant la notion d'efficacité [BERT73] d'une part, mais également avoir la possibilité de se rapprocher des systèmes de signes [ECO88] auxquels les acteurs sont habitués si cela est jugé pertinent. De cette manière, il est possible de produire des représentations graphiques exploitant une symbolique de représentation en rapport avec le domaine de compétence où sera exploitée la restitution. Inversement, il pourra parfois s'agir de s'éloigner de systèmes de signes connus. Par exemple dans la capitalisation, s'éloigner d'une sémantique visuelle à laquelle les experts sont habituées pour stimuler leur réflexion.
Le mécanisme présenté ici peut offrir la possibilité à l'utilisateur de se placer

Typologie des Informations	Système de Représentation textuelle	Système de Représentation Graphique
Entité	Saut de page + Titre Titre	Nœud
Relation	Hiérarchie : Partie, Section, Sous-section…	Arc
Identifiant	Titre	Libellé Nœud ou Arc
Description	Tableau + Titre	Lien vers une zone textuelle
Associative / Sélective	Mise en forme des titres : taille, gras, italique…	Couleur / Forme
Ordonnée	Dimension du plan (linéarité du document) et taille des titres	Dimension du plan
Quantitative	Ecriture textuelle de la valeur	Ecriture textuelle de la valeur

Tableau 1. Spécification d'une correspondance entre les types d'informations et des variables de mise en forme textuelles et graphiques

en position d'éditeur de la forme à condition qu'il ait un droit de " regard " etde modification sur la mise en correspondance présentée dans le tableau 1. A cela, nous ajoutons que dans certaines situations il serait bon que l'utilisateur puisse prendre une position d'auteur dans l'utilisation de la restitution. Nous entendons par cela non pas qu'il puisse modifier les données structurées, mais qu'il soit en mesure de construire la restitution avec les informations qu'il veut voir apparaître selon une organisation en rapport avec ses préférences.

36.2. Une restitution multimodale, interactive et dynamique

Dans la section précédente, nous avons décrit un mécanisme permettant à partir d'une formalisation, de produire une restitution textuelle et une restitution graphique de l'information. De là, nous proposons de prolonger le processus de restitution en associant les différents systèmes de représentation. Il en résulte une restitution " multimodale ", se concrétisant par du texte et des graphiques dans le cas du guide. La représentation multimodale tend ainsi à doubler les modalités d'expression d'une même formalisation pour en facilité la compréhension et l'analyse : chaque modalité sera " valorisée " de manière différente en fonction des acteurs et par exemple leur niveau d'expertise [STRA03]. L'objectif est bien d'accompagner le passage entre les différentes modalités d'expression en instrumentant la restitution via l'introduction d'expressions référentielles intermodales par exemple [BEST03] - pouvant être

de type icône, indice ou encore étiquette qui marquent explicitement les relations entre les modalités - ainsi que la définition d'hyperliens.

La définition d'hyperliens et de représentations multimodales, initiée dans le guide, illustrent assez bien le rapprochement entre SGC et hypermédia. L'intégration d'une interactivité forte doit permettre à l'utilisateur de construire son parcours de lecture dans le contenu, tels un processus de résolution de problème ou une enquête [DEWE93]. L'utilisateur devient acteur dans la restitution qu'il exploite, en prenant le rôle d'auteur de sa démarche de lecture. Pour être en mesure de répondre à ce mode d'interaction, en complément de l'utilisation de lien classique (ex : liens HTML ou PDF), nous avons initié des expérimentations, en nous appuyant sur la spécification XML Linking Language (Xlink). Cette norme propose une conception étendue de la notion de lien (liens multi ressources, bi directionnels, ou encore multi-directionnels…) qui nous permet d'entrevoir le calcul de restitutions interactives intégrant une logique de navigation guidée par les points de vue pouvant être adaptés de manière dynamique.

En complément d'un mode de navigation classique, cette spécification permet également à partir d'un document numérique initial d'enrichir des fragments documentaires pointés par les liens. L'utilisateur peut alors d'enrichir la restitution (l'augmenter en information) en fonction de ses préférences et surtout de ses besoins en terme de contenu informationnel et d'organisation de ce contenu. Il devient auteur de celle-ci : la restitution constitue alors un DopA. Il s'agit d'un mécanisme d'incorporation des informations pointées, ce qui n'opère pas de déplacement dans l'espace informationnel : ceci n'induit pas de changement de contexte, si ce n'est une modification d'une restitution que l'utilisateur est en train de modeler à sa convenance. Il s'agit à travers ce mécanisme de soutenir la compréhension en lui offrant la possibilité d'enrichir une restitution initiale par des informations connexes, pouvant être issues d'un autre point de vue, véhiculant par exemple un complément d'information ou des éléments de définition.

Notre analyse du SGC comme forme nous a conduit à produire une restitution des connaissances multimodales, interactive et dynamique. L'utilisateur est placé au centre du calcul de cette restitution pouvant prendre les rôles d'auteur, d'éditeur et de lecteur, face à une succession d'instances de restitution construit dans notre approche comme des documents numériques. Les différentes instances selon qu'elle sont produites " automatiquement " par le système - en fait selon une stratégie définie par les concepteurs - ou personnalisées par l'utilisateur peuvent alors être vus comme un Document pour l'Action à part entière du fait de leur évolution au gré des contributions de différents auteurs : les concepteurs comme auteurs de la restitution initiale qui peut être modelée par les utilisateurs de la restitution dans l'action au travers de mécanismes interactifs.

37. BILAN ET PERSPECTIVES

Dans cet article, nous avons proposé une étude de la notion de Système de Gestion de Connaissances et du contexte dans lequel il s'insère au travers d'une analogie avec la notion de document. L'analyse du SGC comme medium, argumentée par l'exemple du guide, appuie notre approche centrée sur le CSCW. Ainsi, le développement d'un SGC renvoie à une activité collective d'un réseau d'acteurs coopérant sur la durée de manière asynchrone et délocalisée. Le SGC véhicule une représentation des connaissances dans les interactions entre ces acteurs, et prend ainsi une dimension media à la coopération.

En considérant le SGC comme un signe, nous avons abordé son développement comme la définition d'un objet signifiant. Ceci se justifie particulièrement par les aspects multi domaines et multi points de vue inhérent à la gestion des connaissances. Nous avons ainsi pu observer, dans le cas du guide, que la mise en œuvre de documents numériques au travers d'une formalisation déclarative des connaissances permettait de répondre au contexte peu déterministe auquel nous sommes confrontés.

Enfin, l'étude du SGC comme une forme, nous a amené à proposer un mécanisme permettant de produire une restitution des connaissances multimodales interactive et dynamique. Ce mécanisme tire partie de l'approche documentaire qui de part la séparation entre les données structurées, la mise en forme (" style ") ainsi que les liens offre un environnement souple pour traduire une formalisation dans différents systèmes de représentation pouvant être associés.

L'ensemble de notre analyse permet de dégager une parenté étroite entre le développement du SGC et la notion de document dans sa rédaction, son édition et sa lecture. Ainsi, le SGC s'insère dans un réseau d'acteurs qui passent successivement dans un statut d'auteur, d'éditeur et de lecteur participant à la définition de la forme et du " fond ". Le SGC peut alors être appréhendé comme un agrégat de multiples fragments documentaires portés par des auteurs multiples. Ce contexte " hyper rédactionnel " nous renvoie au concept de Document pour l'Action : le DopA devient alors un support à la Gestion des Connaissances.

Cette conception documentaire du SGC, comme media de la coopération, dans un contexte informatique, nous rapproche de la vision de l'Interaction Homme Machine (IHM) introduite par Dourish [DOUR99] : " Human Computer Interaction can be thought of as a form of mediated communication between the end user and the system designer... ". Dans la même orientation, Bourguin et al [BOUR05] introduisent que les système de TCAO (Travail Coopératif Assisté par Ordinateur) doivent être développés comme une

intermédiation entre concepteur(s) et utilisateur(s) afin d'en faciliter le développement et l'évolution dans un contexte peu déterministe et mouvant. Nous pensons que les SGC, et particulièrement leur surface visibles (leurs interfaces homme machine), doivent être développés en ce sens.

Ainsi, nous pensons que l'approche documentaire présentée dans cet article et implémentée sur le guide va dans cette direction. Ceci étant, elle n'est pas sans poser de difficultés. On évoquera par exemple que si la formalisation déclarative répond à un besoin de souplesse, il convient de s'attarder sur les problématiques de gestion et de maintien de la cohérence des informations. Afin de répondre à ces difficultés, il s'agit certainement de se tourner vers des solutions hybrides associant la souplesse de l'approche documentaire et le cadrage des bases de données. Ainsi, nous avons engagé une étude en rapport avec un Système à Base de Connaissances qui s'appuie sur une base de données et met en œuvre des algorithmes de simulation reflétant un contexte beaucoup plus " formel ". L'idée est d'instrumenter ce système en appliquant l'approche documentaire présentée dans cet article au niveau de ses interfaces. L'ambition est alors de proposer une interface devenant un Document pour l'Action présentant les données contenues dans le système ainsi que les résultats qu'il produit. En prolongement, nous allons tenter d'implanter un mécanisme interactif permettant d'enrichir cette restitution par des informations connexes extraites par exemple de documents gravitant autour de ce système (document de conception, de capitalisation…).

38. RÉFÉRENCES BIBLIOGRAPHIQUES

[BERT73] Bertin, Jacques, "Sémiologie graphique - Les diagrammes - Les réseaux - Les cartes", Les réimpressions des Éditions de l'École des Hautes Études en Sciences Sociales, 1973.

[BEST03] Bestgen, Yves et Dupont, Vincent, "Impact des références intermodales sur la lecture et l'apprentissage d'un document multimédia", Actes de CIDE6, 11-24 (2003)

[BOUR05] Bourguin, Gregory, Derycke, Alain, et Tarby, Jean Claude, "Systèmes interactifs en co évolution réflexions sur les apports de la théorie de l'activité au support des pratiques collectives distribuées". RIHM, Vol 6 N°1, 2005.

[DEWE93] Dewey, John. "Logique, la théorie de l'enquête", PUF, 1993.

[DIEN01] Dieng-Kuntz, Rose, Corby, Olivier, Gandon, Fabien, Giboin, Alain, Golebiowska, Joana, Matta, Nada et Ribière Myriam. "Méthodes et outils pour la gestion des connaissances : Une approche pluridisciplinaire du Knowledge Management (2ème édition)". Dunod, 2001.

[DOUR99] Paul Dourish. "Where the action is - The Foundations of Embodied Interaction", MIT Press, 1999.
[ECO88] Eco, Uberto. "Le Signe", 1988.

[LEPL04] Leplat, Jacques, "Eléments pour l'étude des documents prescripteurs", revue électronique activités (http://www.activites.org), Vol 1 N°2, 2004.

[MALL02] Mallein, Philippe et Tarozzi, Sylvie, "Des signaux d'usage pertinents pour la conception des objets communiquants", Les cahier du numérique, Vol 3 N°4, 2002.

[MORA04] Morand, Bernard. "Logique de la conception - Figures de sémiotique générale d'après Charles S. Pierce". L'Harmattan, 2004.

[NICO03] Nicolle, Anne, Interaction langagière personnes / machines, Variation, construction et instrumentation du sens, 251-285, 2003.

[PEDA03] Roger T. Pédauque. "Document : forme, signe, médium, les reformulations du numérique". Technical report, STIC-CNRS, 2003.

[SCHI96] Schmidt, Kjeld, et Simone, Carla "Coordination mechanisms : Towards a conceptual foundation of CSCW systems design". CSCW Journal, 5, 2-3, 155-200, 1996.

[STRA03] Strahm, Maeva et Baccino, Thierry, "L'intermodalité dans la lecture de documents électroniques : investigations oculométriques", Actes de CIDE6, 79-104, 2003.

[ZACK04] Manuel, Zacklad. "Processus de documentation dans les documents pour l'action (dopa) : statut des annotations et technologies de la coopération associées". In Le numérique : Impact sur le cycle de vie du document pour une analyse interdisciplinaire, EBSI, Montréal, 2004.

Session 04

Gestion et accès à des collections de documents

PFC 12
Un outil d'aide à la découverte du contenu des documents et à la création de dossiers

André ALUSSE
Jean-Charles LAMIREL
Abdel BELAÏD

LORIA-Université Nancy 2, Campus Scientifique, B.P. 236, Vandoeuvre-Lès-Nancy,
LORIA, Campus Scientifique, B.P. 236, Vandoeuvre-Lès-Nancy, France
{alusse,abelaid,lamirel}@loria.fr

RÉSUMÉ

Cet article traite de la construction automatique et dynamique de dossiers consolidés. La construction de dossiers utilise plusieurs étapes : recherche des documents les plus significatifs à partir d'une requête par mots-clés, classification dynamique du résultat de la requête en utilisant plusieurs classifieurs aux comportements différentiés, combinaison des résultats de ces classifieurs pour mieux faire ressortir les thématiques extraites, et enfin personnalisation de l'organisation en introduisant les choix de l'utilisateur. Une évaluation statistique des paramètres utilisés par les classifieurs a permis de mesurer leur intérêt et surtout leurs incidences sur la constitution finale des classes thématiques. En sus de l'utilisateur, d'autres opérateurs de type plus large : groupes ou communautés peuvent intéragir avec le système pour l'enrichir. Le prototype présenté dans cet article est une plate-forme expérimentale d'observation sur l'organisation de documents et sur les méthodes de classification. L'application pilote concerne la consolidation des textes de loi de la Commission Européenne.

Mots-clés : Recherche d'information, classification automatique, combinaison de classification, dossier, partage d'informations, appropriation des documents par l'utilisateur.

INTRODUCTION :

Cet article traite de la construction automatique et dynamique de dossiers consolidés. Nous entendons par dossier consolidé, un ensemble bien

12 PFC : Personal File Consolidation

organisé de documents, et de parties de documents, traitant d'un sujet précis répondant à un besoin de l'utilisateur. Un tel type de synthèse et de consolidation d'information s'avère souvent indispensable. En effet, il intervient aussi bien dans la création d'un nouveau cours que dans la constitution d'un état de l'art sur une thématique, ou encore, dans des cas plus précis, comme la réalisation d'un dossier de législation à partir de différents textes de lois.

La motivation de ce projet émane principalement de demandes de SSII collaborant avec la CEE et cherchant à offrir, notamment aux juristes, la possibilité d'établir des synthèses ou des dossiers de consolidation à partir de l'ensemble des publications de la Communauté Européenne. Ces dossiers sont susceptibles de couvrir des sujets variés comme : quelles sont les dernières réglementations en matière de transport d'animaux ? Quels textes concernent plus particulièrement les douaniers en matière d'importation d'animaux et quelles sont les dernières règles en vigueur ? Quels textes concernent plus précisément les vétérinaires sur le sujet et quelle était la législation en vigueur l'année précédente ?

Au vu de ces motivations, la consolidation se traduit dans ce projet par l'intégration de plusieurs aspects du domaine de la recherche d'information : la recherche en elle même, la classification pour le rapprochement et l'organisation des documents et l'intégration du point de vue de l'utilisateur final.

A l'heure actuelle, il n'existe pas de système intégrant l'ensemble de ces fonctionnalités. De fait, les outils existants sont plutôt spécialisés sur une tâche précise : par ex. sur la recherche d'information, ou sur l'extraction d'informations et de résumés, ou sur la classification, statique ou dynamique ou bien encore, sur la catégorisation.

La catégorisation consiste à organiser les documents en se basant sur une hiérarchie thématique pré-existante. La constitution de cette hiérarchie, aussi bien que l'analyse des documents à catégoriser, sont en général des opérations manuelles laissées à la responsabilité de spécialistes.

L'objectif de la classification non supervisée est de découvrir les groupes (clusters) de documents similaires et de faire émerger des classes "latentes" d'un ensemble de documents. De nombreux types de méthodes de classification non supervisées sont proposés dans la littérature, parmi lesquels on peut citer :

les méthodes basées sur un calcul préalable d'une matrice de similarité entre les documents, comme HAC (Hierarchical Agglomerative Clustering) [Voorhees86], Suffix Tree Clustering [Zamir98], Semantic Online Hierarchical Clustering [Zhang01] ; les méthodes de nuées dynamiques de type : K-means [Rocchio66] ; les méthodes neuronales comme les cartes auto-

organisatrices de Kohonen (SOM) [Roussinov98] ou les gaz neuronaux simple (NG) [Martinetz91] ou évolutifs (GNG) [Friedske94].

Cependant, ces approches ne permettent pas d'intégrer l'expérience de l'utilisateur, ce qui est insuffisant pour la constitution de dossiers personnalisés. Il faut donc une stratégie plus centrée autour de l'utilisateur qui prenne en compte davantage ses habitudes de travail et sa vision sur les documents. L'utilisateur doit ainsi pouvoir créer sa propre catégorisation et pouvoir la confronter à la fois à celles des autres utilisateurs et aux classifications calculées par le système.

Pour mesurer l'intérêt porté aux documents par les utilisateurs, la littérature propose la notion de "filtrage collaboratif" dont le but est de faire ressortir les avis majoritaires sur des documents qui seront ensuite utilisés comme des recommandations personnelles. Cette approche n'a cependant pas été exploitée dans le contexte plus large de l'organisation de l'information.

Parallèlement, beaucoup de travaux ont également été effectués sur la combinaison de classifications hétérogènes dans le domaine de la recherche d'information, et ceci avec des objectifs très variés. Ces combinaisons offrent des possibilités intéressantes pour fournir des résultats classifiés précis, intuitifs et complets. [Lam01] et [Bennett02] ont suggéré et généralisé de telles méthodes. Leur principale limite est cependant de fonctionner hors-ligne.

D'un autre côté, certains moteurs de recherche, comme Exalead ou Vivissimo, mettent en œuvre une classification en ligne. Mais, ils n'offrent que des possibilités réduites d'organisation qui ne répondent pas à une problématique de dossiers.

L'objectif du projet Paploo est de fournir à l'utilisateur une série d'outils pour rechercher et trier les documents en fonction de ses besoins (dossier, consolidation, synthèse, cours, etc.). Ce qui revient donc à fédérer l'ensemble des approches précédemment décrites.

L'article est organisé de la façon suivante : la solution retenue est présentée dans la deuxième section, la troisième section décrit le fonctionnement du système, la quatrième section est consacrée à la présentation des premiers résultats et à la discussion, alors que la conclusion et les perspectives sont données dans la dernière section.

LE SYSTÈME DÉVELOPPÉ

LA DÉMARCHE

La solution de constitution de dossier que nous proposons repose sur les trois principes directeurs suivants :

1) opérer plusieurs classifications thématiques des documents, autorisant ainsi différentes visions organisationnelles des résultats,

2) exploiter l'intérêt porté par les utilisateurs sur les documents pour réaliser un filtrage beaucoup plus pertinent, ciblé sur les appréciations et besoins de ceux-ci, et de faire bénéficier l'expérience de la communauté à l'ensemble,

3) offrir différents outils d'analyse permettant à l'utilisateur de comprendre les relations entre documents et de faire ressortir les propriétés communes en conformité avec ses besoins.

L'ARCHITECTURE

La figure 2.1 illustre le fonctionnement de la plate-forme expérimentale PFC que nous proposons. Celui-ci se décompose en quatre étapes : la recherche (requête), les classifications, la visualisation, qui constitue la partie interactive, et enfin, la construction du dossier.

D'un point de vue pratique, la plate-forme PFC est composée de plusieurs modules :

1) Pour la recherche : un moteur de recherche générique permettant à partir d'une requête sur un sujet de restituer les documents pertinents.

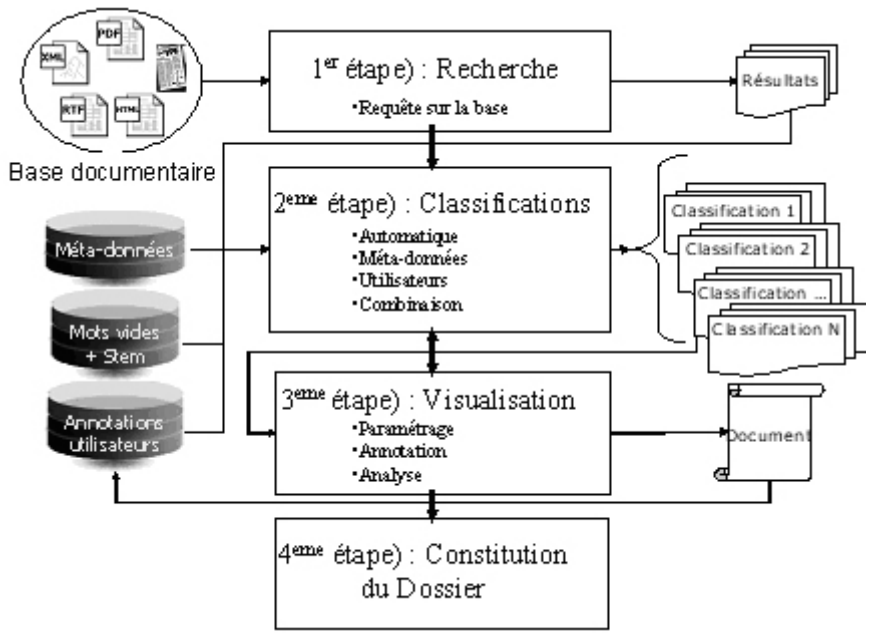

Figure 21 : Enchaînement des tâches dans PFC

2) Pour la classification : un ensemble d'outils de classification. Ces outils sont choisis de manière à réaliser différents types de classification : agglomérative, hiérarchique, syntaxique, sémantique, ou par méta-données, afin d'obtenir plusieurs organisations de documents et un module de combinaison et d'intégration de l'ensemble de ces classifications. Chacun de ces classifieurs est interactif, permettant à l'utilisateur d'adapter ses paramètres de calcul.

3) Pour la partie interactive :
- un module d'annotation permettant à l'utilisateur de donner son propre point de vue sur les documents,
- un module d'analyse du comportement des classifieurs,
- un module de visualisation des documents.

4) Pour la constitution de dossier : un module qui réorganise les documents suivant un schéma hiérarchique choisi par l'utilisateur.

DESCRIPTION DES TÂCHES

LA RECHERCHE

La requête est composée d'une suite de mots clés. En réponse à cette requête, le moteur de recherche retourne une liste de documents (D). Nous retenons le titre (T) et un extrait du document (R) en lien avec la requête, de chacun d'entre eux.

LES CLASSIFICATIONS

Classification automatique

L'objectif de la classification automatique est de regrouper les documents les plus proches à partir de la proximité des mots qu'ils contiennent. La première étape du processus se focalise sur l'extraction des mots. Chaque document est analysé de manière à constituer la liste des mots qui lui est associée. Les mots vides sont exclus et seul le radical des mots retenus (algorithme de Porter) est conservé. Pour chacun des mots retenus (radicaux), deux fréquences sont calculées : la fréquence du mot dans le document (TF pour Term Frequency) et la fréquence du mot dans l'ensemble des documents (DF pour Document Frequency). Ces valeurs permettent ensuite de calculer l'"Inverse Document Frequency" (IDF), afin de pondérer l'impact des mots sur-représentés dans les documents et qui finissent ainsi par perdre de leur valeur discriminante : si un mot est présent dans tous les documents, il ne permet pas de les différentier.
Etant donné un mot mi, on a : IDF(mi) = log((1+N) / (1+DF(mi))).

Le tableau 3-1 présente un extrait de la liste des racines de mots avec le

nombre d'occurrences trouvées, leur valeur de TF*IDF, ainsi que la liste des mots dont elles sont déduites.

Mot (radical)	occurrence	tf*idf	Mot entier
communaut	22	24,97	communauté communautés
fievr	15	33,50	fièvre
animal	72	61,01	animaux animale
utilis	5	13,20	utilisation utilise utilisés

Tableau 31 : mots extraits du vocabulaires des documents

L'objectif de l'étape suivante est d'établir une liste des séquences de mots consécutifs qui se répètent dans plusieurs documents pour obtenir des expressions (ou motifs) les décrivant. Pour cela, on utilise la technique des "Suffix Array", introduite par [Manber93]. Le tableau 3-2 donne un extrait de cette liste.

Expression	Occurrence	Tf*idf
Protection des Animaux	4	8,93
Salaires et Rémunérations des Ouvriers d Abattoir	4	7,78

Tableau 32 : expressions extraites des documents

Puis, trois types de classifications sont finalement construites :

1) La première classification s'inspire de la méthode AHC (Agglomerative Hierarchical Clustering). A partir de ces extractions, qu'il s'agisse des mots ou des motifs, le système construit une représentation vectorielle des documents (Vector Space Model ou VSM [Salton75]). A l'issue de cette étape, chaque document sera alors décrit par les mots ou motifs associés à un coefficient de pondération égal au produit des coefficients TF et IDF. Une matrice de similarité est construite en se basant sur un calcul de distance de type corrélation cosinus [Salton75] entre les vecteurs documents. Chaque vecteur est assigné à une classe, puis une hiérarchie est établie en regroupant deux à deux les classes les plus proches, mais dont la distance est néanmoins inférieure à un certain seuil (technique du dendogramme). La classification résultante est donc un arbre dont les documents représentent les feuilles et dont les nœuds représentent les clusters. Les intitulés décrivant chaque cluster sont définis à partir des mots ou expressions majoritairement partagées par l'ensemble des documents du cluster. On peut agir sur le seuil minimal permettant le regroupement des documents et des clusters en classes.

2) La seconde classification dérive de la méthode "Suffix Tree Clustering" (STC). Les clusters sont construits à partir des mots ou des expressions les plus fréquemment retrouvés dans les documents. Un même document peut donc se retrouver dans plusieurs clusters différents. La classification n'est donc pas recouvrante. Chaque expression constitue un cluster de base. Son score dépend du nombre de mots qu'elle contient, ainsi que du nombre de documents qui la contienne. Ces clusters sont ensuite regroupés hiérarchiquement suivant leur similarité calculée à partir d'une fonction de distance.

3) La troisième classification est établie d'après la technique de Lingo [Osinski03]. Il s'agit de découvrir d'abord les labels les plus représentatifs et les plus discriminants, tout en englobant un maximum de documents. La détection de ces labels s'opère à l'aide de la technique du "Latent Semantic Indexing" [Deerwester90]. Comme pour AHC, le seuil de regroupement peut être ajusté.

Pour compléter ces classifications, nous avons implémenté trois autres techniques, dont deux sont issues du domaine neuronal. Ces techniques sont plus adaptées à la découverte de relations entre les données que les précédentes si la connaissance sur la nature de ces relations est limitée.

La première méthode est basée sur le partitionnement de type K-Means.

La deuxième méthode utilise les cartes auto-organisatrices SOM de Kohonen, dont le principe est le suivant : les données d'entrée sont des vecteurs, issus du VSM dans notre cas, et la carte SOM représente l'espace dans lequel elles doivent se ranger. Nous utilisons la phase d'apprentissage comme phase de découverte des relations entre documents. Les neurones de la carte (dont le nombre est fixé au départ) représentent au final les clusters de documents et les relations de voisinage entre neurones traduisent la proximité entre les classes de documents.

La troisième méthode repose sur la technique "Growing Neural Gas" (GNG) où le nombre de neurones n'est pas imposé à l'avance comme pour les cartes SOM, ce qui permet une plus grande souplesse dans la découverte des relations possibles.

Pour ces trois méthodes, nous avons considéré que les mots et expressions représentaient les propriétés d'entrée, et que leur fréquence d'apparition dans les documents représentaient les poids des vecteurs d'entrée. Afin de restreindre l'espace d'entrée et ainsi de mieux cibler les propriétés, nous avons défini un seuil minimal de fréquence pour retenir les mots simples (Sm = 10 %), et, un seuil minimal inférieur pour retenir les expressions (Se = 5%) afin de les valoriser.

Soit P={p1,p2,…,pj} l'ensemble des propriétés, mi un mot et Ei une expression extraite précédemment. Nous avons donc la relation suivante :
mi P DF(mi)) > Sm et Ei P DF(Ei)) > Se
Soit D = {d1,d2,…dn} l'ensemble des documents résultats, C le corpus. Nous avons la relation suivante :
C = D P / Ci,j = TF(di,pj)*IDF(pj)
Ce corpus ainsi constitué servira d'entrée aux trois autres méthodes de classification. A l'issue de l'application de ces méthodes, les expressions apparaissant le plus fréquemment dans les clusters serviront à constituer les étiquettes de ceux-ci.

Catégorisation par méta-données

Les vues sur l'organisation des documents obtenues précédemment sont complétées par des informations provenant des indexations thématiques établies par un expert du domaine. Chaque document est décrit avec des méta-données, comme cela est par exemple défini par la norme "Dublin Core", ou bien par le thesaurus Eurovoc, dans le cadre de notre étude. Les clusters de documents sont déduits de ces informations. A titre d'exemple, le contenu des méta-données Eurovoc suivant, associé en parallèle à un document :
<EUROVOC_DOM CODE="56"> Agriculture, Sylviculture </EUROVOC_DOM>
<EUROVOC_MTH CODE="5641">5641 pêche</EUROVOC_MTH>
produira les deux clusters imbriqués "Agriculture, Sylviculture" et "pêche".

Catégorisation utilisateur

Au cours de la consultation, l'utilisateur est invité à donner son avis sur le document, ceci en en introduisant ses propres mot-clés pour le décrire. L'utilisateur s'approprie ainsi les documents, et, comme ses annotations sont stockées, cela permet d'enrichir la base d'informations au fur et à mesure.
Pour construire la classification découlant de ces annotations, la liste des mot-clés décrivant chacun des documents résultats est dressée. Ce ?sac de mots? sert à établir le VSM, point de départ d'une classification type ?AHC?. Comme chaque utilisateur est associé à un groupe partageant le même profil et les mêmes besoins, une classification peut être déduite sur le même principe avec l'ensemble des mot-clés utilisé par le groupe. L'utilisateur peut donc enrichir ses points de vue sur les documents en découvrant les remarques d'usagers proche de lui. Une troisième classification est construite à partir des annotations de l'ensemble des intervenants. Cette approche permet à l'utilisateur de partager et confronter son avis et au système d'offrir plus de point de vue pour avoir une vision plus globale des documents.

Combinaison des classifications

La figure 3-1 illustre le mécanisme de combinaison des classifications. Chaque classification prise individuellement n'apporte pas de solution idéale.

Pour bénéficier des avantages de chacune des classifications et minimiser leurs défauts, une combinaison en est donc opérée. Nous avons choisi d'utiliser à nouveau l'algorithme AHC pour effectuer cette combinaison car : 1) il est le plus simple à mettre en œuvre et 2) on peut choisir la granularité (indice) de classification recherchée pour l'application.

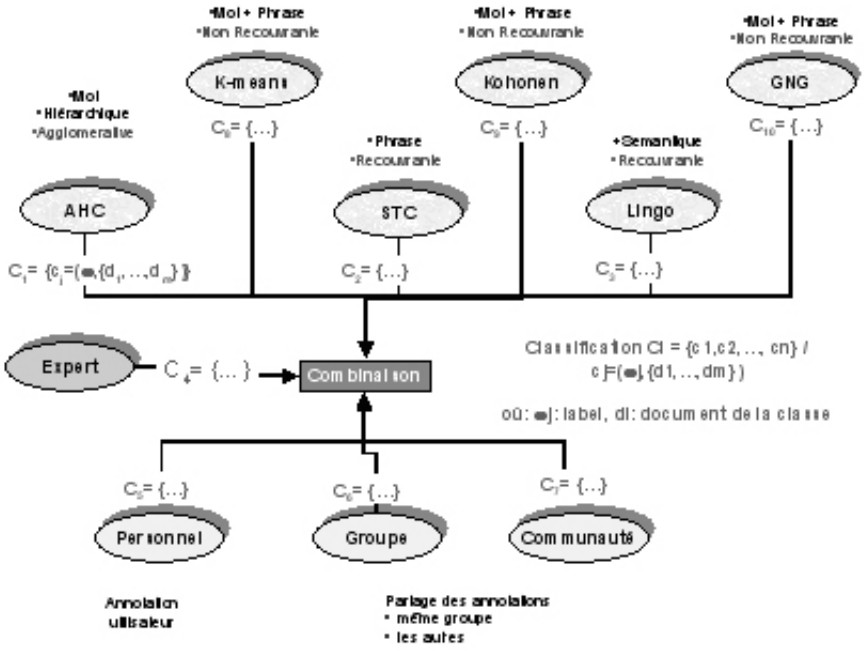

Figure 31 : combinaison des classifications

Soit Ci une classification, avec Ci={c1,c2,…, cn} où cj=(i,{d1,…,dm}); est le label de la classe et di un document de la classe. Tous les labels i sont extraits des classifications. Ces labels servent ensuite à construire le VSM (M). Le poids d'un label est fonction du poids accordé au classifieur et du nombre de fois où ce label est mentionné. Le poids d'un classifieur est attribué en fonction de l'importance que l'utilisateur lui donne et de son opportunité à répondre à ses besoins.

$$M_{i,j} = \sum_{k=1}^{n} (d_i \in C_k(\omega_j)) \times p_k$$

où pk est le poids de la classification k,
et

$$d_i \in C_k(w_j)$$

une fonction binaire (la classe j contient ou non le document di).

Visualisation

Il s'agit de la phase interactive du système qui permet à l'utilisateur de consulter les documents, de prendre connaissance des relations inter-documents (classification), de modifier les paramètres pour affiner le comportement des classifieurs, et enfin, d'annoter les documents.

Paramétrage des classifieurs

Comme nous l'avons préalablement mentionné, le paramétrage des classifieurs permet à l'utilisateur d'agir sur le comportement de ceux-ci et de découvrir les relations de différents niveaux de généralité entre documents. De manière pratique, ces modifications de paramètres sont prises en compte, d'abord au niveau du classifieur concerné, et ensuite, au niveau de la combinaison. Cette opération peut être itérée autant de fois que l'utilisateur le souhaite.

Annotation

En cours de consultation, l'utilisateur a la possibilité d'annoter les documents consultés, d'apporter sa propre vision sur un document, et donc d'influer sur les classifications utilisateurs, et ainsi d'enrichir le système au cours de son exploitation. De manière pratique, la note saisie (avec une interface appropriée) pour un document donné est stockée dans la base et le processus de classification est relancé pour les classifications utilisateur. Le module est en phase de réflexion et d'élaboration.

Analyse des résultats

Quatre mesures objectives ont été définies pour permettre à l'utilisateur de mesurer la qualité des classifications et de juger de l'opportunité de celles-ci pour atteindre son but. Ces mesures servent aussi à comparer le comportement des classifieurs et à découvrir les plus aptes à détecter les propriétés communes :

la couverture permet de calculer le nombre de documents classés,
C = (NbDocResultat - NbNonClasse) / NbDocResultat)
la dispersion permet de mesurer la répartition en classes,
D = NbClasse / NbDocResultat
la précision permet de mesurer l'homogénéité d'une classe,

$$\text{Précision } P = \frac{\sum_{i}^{n} Pc_i}{|C|} \Big/ Pc_i = \frac{\sum_{i}^{n} Pc(tj)}{|Ci|} \Big/ P(ci(tj)) = \frac{nbDoc(tj \in d)}{nbDoc(Ci)}$$

le rappel permet de mesurer l'indépendance des classes.

$$\text{Rappel } R = \frac{\sum_{i}^{n} Rc_i}{|C|} \Big/ Rc_i = \frac{\sum_{i}^{n} Rc(tj)}{|Ci|} \Big/ R(c(tj)) = \frac{nbDoc(tj \in d, d \in Ci)}{nbDoc(tj \in d, d \in \{Ci\})}$$

Les mesures de précision et de rappel que nous avons choisies représentent une adaptation, à l'évaluation des classifieurs, des mesures de précision et de rappel utilisées en ingénierie documentaire. Cette approche, proposée par [Lamirel04], présente l'avantage majeur d'être totalement indépendante de la méthode de classification utilisée, contrairement aux méthodes d'évaluation classiques basées sur l'inertie inter et intra classe. Elle permet donc de comparer objectivement les résultats de plusieurs classifieurs différents, ce qui s'avère important dans notre contexte.

Dossier

A l'issue de ces étapes, le dossier est créé à partir des nœuds de classification sélectionnés par l'utilisateur. Ces nœuds sont transmis à un utilitaire qui se charge de la recomposition et de la reformulation. Des considérations de recomposition et de reformulation automatique basée sur une annotation sémantique sont l'objet d'un autre travail.

Expérimentations et discussions

La méthode que nous avons proposée a été expérimentée avec une petite partie des documents de la Communauté Européenne, à savoir 2000 documents incluant 453 règlements, 368 questions écrites, 242 traités, etc. Ce corpus reste cependant suffisamment représentatif pour valider notre approche.

Le tableau 4.1 présente les classifications établies par les différents algorithmes pour la requête ? transport d'animaux?. Ces premiers résultats amènent les remarques suivantes :

1. pour AHC, le nombre de clusters est important ; il est difficile de juger de la qualité des clusters et de l'opportunité des termes pour l'identification des propriétés communes,

2. pour STC, les documents ne sont pas assez discriminants (beaucoup d'introductions communes à plusieurs documents) pour en déduire des regroupements intéressants. Le recouvrement est trop important,

3. Pour Lingo, moins de clusters sont détectés, les intitulés sont plus opportuns mais beaucoup de documents ne sont pas classés,

4. Pour K-means, l'équilibre entre la précision et le rappel n'est pas optimum, la taille des clusters est hétérogène,

5. Pour l'approche SOM, l'équilibre est meilleur, mais au détriment d'un cluster poubelle qui regroupe 10 documents,

6. Pour l'approche GNG, les clusters sont plus équilibrés,$

7. La classification issue des méta-données (non présente dans le tableau) construit beaucoup de clusters avec un fort recouvrement du fait que les documents contiennent beaucoup de méta-données.

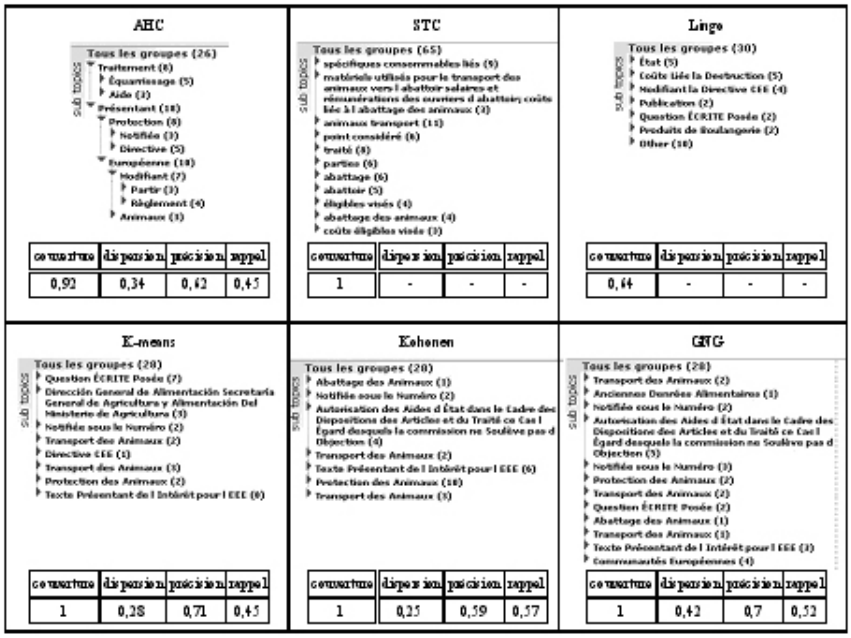

Tableau 41 : Tableaux des résultats de classifications

Ces différentes classifications permettent aussi à l'utilisateur de découvrir les mots et les concepts partagés par les documents. Dans le cadre des documents de la Commission Européenne, d'autres phénomènes peuvent être détectés comme par exemple : les références croisées (par ex. vu l'article n°…), les termes de mise à jour (par ex. modifié le …), les "questions écrites", les "directives", etc. Cela induit différents regroupements possibles pour l'utilisateur.

Classification utilisateurs

Afin de simuler les contributions utilisateurs, nous avons créé 8 utilisateurs virtuels, répartis en 3 groupes. Un groupe douanier avec deux profils, l'un orienté "règlement", l'autre "frontière", un second groupe vétérinaire avec trois profils, sanitaire, maladie et abattoir ; enfin un troisième groupe agriculteur ayant pour profils, agriculture, pisciculture et élevage de poulets. Pour chacun des utilisateurs, nous avons attaché des mots-clés aux documents en fonction de leurs préoccupations supposées et d'une analyse subjective des documents, par exemple "grippe aviaire" pour le vétérinaire ayant en charge les maladies. Les résultats montrent que les clusters construits à partir de ces contributions sont en forte connections avec les préoccupations utilisateurs et donc, que l'enrichissement des données à partir des annotations utilisateurs constitue un apport primordial à notre problématique.

Analyse d'un paramètre

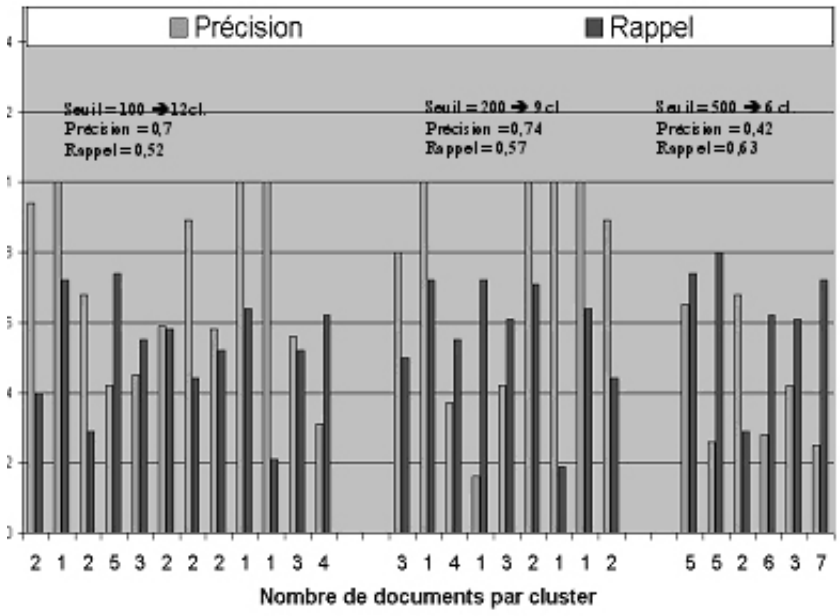

Figure 4 1 : analyse d'un paramètre pour la méthode "Growing Neural Gas"

La Figure ?4 1 montre l'impact du suivi d'un paramètre dans le comportement d'une classification. L'accroissement de la valeur du seuil pour créer un nouveau nœud illustre le comportement de l'algorithme et permet d'analyser le type des clusters créés. Plus le seuil augmente, moins le nombre de clusters est élevé. Corrélativement, la taille des clusters augmente, mais la précision moyenne recule. Toutefois, les valeurs moyennes de précision et

de rappel ne représentent que des valeurs indicatives. En effet, sur l'exemple de la figure 3-1, que la classification de précision moyenne la plus faible (seuil 500) a permit d'isoler un cluster de 5 documents, où il existe un équilibre, s'opérant à des valeurs élevées, entre la précision et le rappel. Ceci permet d'en déduire une forte proximité entre les documents du cluster (ici, 2 traitant de l'éradication de la peste porcine, 1 sur la fièvre aphteuse et 2 sur l'influenza aviaire). Dans la classification la plus précise en moyenne (seuil 100), ces 5 documents se trouvent répartis dans 3 clusters différents. Ce cluster potentiel de 5 documents n'a donc pas été identifié dans ce dernier cas. Cet exemple démontre bien que la possibilité pour l'utilisateur d'intervenir à tout moment sur les seuils de classification s'avère nécessaire pour lui permettre de mieux comprendre l'organisation des documents.

Combinaison des classifications

La figure suivante analyse l'impact de chaque classifieur dans la combinaison. Pour le cas "égalité" chaque classifieur est affecté du même poids alors que dans les autres cas l'impact du classifieur analysé est multiplié par 10 afin de majorer clairement son importance. Cela nous permet de mieux comprendre la contribution de chacun des classifieurs dans la composition. La dispersion est favorisée par la contribution de l'ensemble des utilisateurs, cela étant dû à la multitude d'avis sur les documents. Elle est minimisée par les classifieurs non recouvrants, étant donné que ceux-ci discriminent moins. Le rappel est faible pour les classifieurs non hiérarchiques car ils conservent moins d'informations, à la différence de la technique AHC. Ces analyses sont primordiales pour s'approcher de la construction du dossier souhaité. Toutefois, elles sont moins évidentes à appréhender que l'impact des paramètres sur le comportement d'un classifieur donné. Elles demandent en effet une plus grande contribution de l'utilisateur, mais lui offre, en contrepartie, des outils pour une meilleure compréhension des liens possibles entre les documents, ainsi que la possibilité de contrôler l'organisation de ses dossiers.

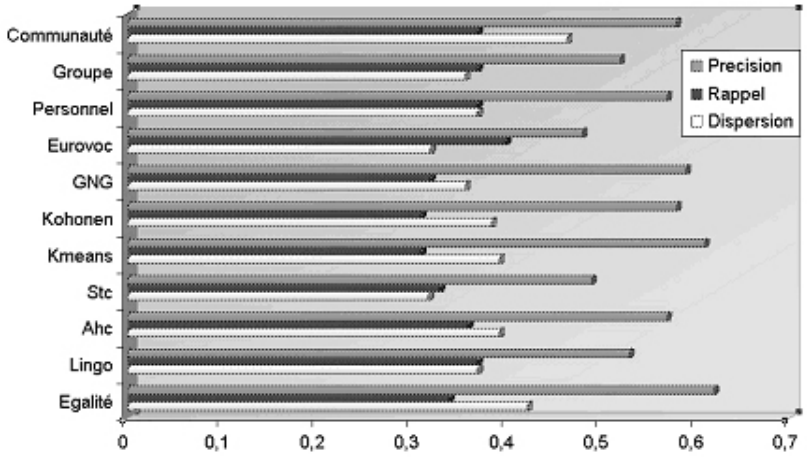

Figure 4 2 : Analyse de l'impact des classifieurs

Le tableau suivant montre trois classifications élaborées en combinant les résultats des différents algorithmes. Pour la première, le poids de chacun est le même, pour la seconde, les classifications utilisateurs et les classifications non hiérarchiques ont été privilégiées, et, dans la troisième, le poids de la classification hiérarchique a été renforcé, ainsi que les paramètres impliqués dans la hiérarchisation. Cette troisième solution permet donc d'obtenir des dossiers plus hiérarchisés alors que la seconde s'attache plus à un découpage à plat, par thème.
Ces observations permettent donc à l'utilisateur de comprendre le comportement des classifieurs et d'adapter et valider la stratégie en fonction de ses besoins.

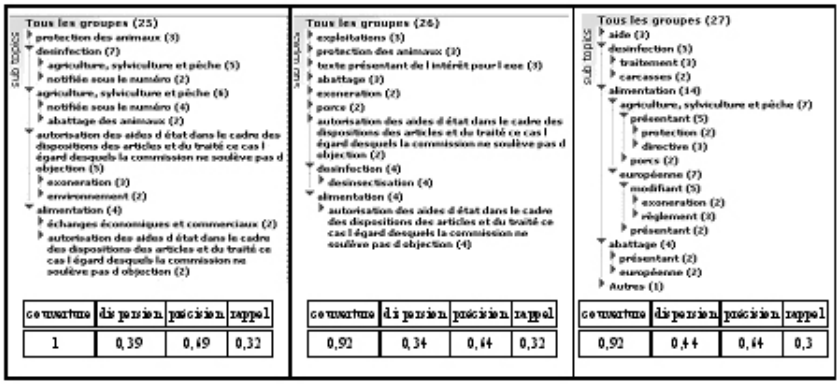

Tableau 4 2 : Exemple de résultats de combinaison de classifications

CONCLUSION ET PERSPECTIVES

Nous avons mis en place un nouveau système pour guider l'utilisateur dans la construction de dossier. Il en est encore au stade expérimental mais doit être progressivement intégré à la plateforme de la société leader du projet. Ce système repose sur les techniques et les idées suivantes : recherche d'information et classifications de documents, appropriation d'information et partage de connaissances, analyse d'outils et aide à la compréhension des liens et des partages de concepts entre documents. La construction de dossier passe tout d'abord par la recherche des documents les plus significatifs, puis par leur réorganisation en clusters et en hiérarchie. Ce processus interactif fait ressortir les mots, ou groupes de mots, partagés par un ensemble de documents ce qui permet à l'utilisateur de découvrir les liens existants entre documents. D'un autre côté, l'analyse des paramètres de travail et leur ajustement possible lui offre des clés pour contrôler l'organisation de son dossier. Afin de tenir compte de son profil, de sa compréhension sur les textes et de ses besoins, il est également sollicité pour enrichir la connaissance sur les documents, connaissance qui est ensuite partagée par l'ensemble de la communauté. Tout cela contribue à fournir un outil facilitant le regroupement de documents tout en tenant compte des nécessités de chaque utilisateur.

Pour améliorer le système, il est envisageable d'associer à l'utilisateur un ensemble de mots-clés ou expressions qui le caractérise, lui ou ses besoins, sous la forme d'un profil. Ces informations serviraient à renforcer l'influence des documents en adéquation avec ce profil, et par conséquent, à orienter le partitionnement vers des concepts plus en rapport avec les préoccupations de l'utilisateur. Un enrichissement de la qualification de l'impact des classifieurs et une analyse des comportements utilisateurs dans la réalisation de l'objectif fixé pourraient permettre d'établir des stratégies et de proposer plus facilement des solutions pré-établies adaptées aux besoins particuliers. Ce travail constitue néanmoins une première approche dans l'aide à la découverte de relation entre documents et de réalisation de dossiers.

BIBLIOGRAPHIE

[Bennett02] Bennett P.N., Dumais S.T. et Horvitz E., "Probabilistic combination of text classifiers using reliability indicators: Models and results", In Proceedings of SIGIR-02, Tampere, Finland, 2002.

[Deerwester90] Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. et Harshman R., "Indexing by Latent Semantic Analysis", J. American Society for Information Science, 1990.

[Fritzke94] Fritzke, B., Growing Cell Structures - A Self-Organising Network for Unsupervised and Supervised Learning, Neural Networks, 7(9):1441-1460, 1994.

[Lam01] Lam W. et Lai K.Y., "A meta-learning approach for text categorization", Proceedings of SIGIR-01, New Orleans, US, 2001.

[Lamirel04] Lamirel J.C., Francois C., Al Shehadi S., Hoffman M., "Multi-Topographic new classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping".
In Scientometrics international Journal, Vol. 60, No. 3 445-462, 2004.

[Manber93] Manber U. et Myers G.,. "Suffix arrays: a new method for on-line string searches", SIAM Journal of Computing, 22(5), pp. 953-948, 1993.

[Martinetz91]. Martinetz T. et Schulten K., "A "neural gas" network learns topologies". In Kohonen, T., Makisara, K., Simula, O., and Kangas, J., editors, Articial Neural Networks, pages 397-402. Elsevier Amsterdam, 1991.

[Osinski03] Osinski S., "An Algoritm for Clustering of Web Search Results", Master thesis, Poznan Universitv of technology, 2003.

[Rocchio66] Rocchio J.J., "Document retrieval systems - optimization and evaluation", Ph.D. Thesis, Harvard University, 1966.

[Roussinov98] Roussinov D. et Ramsey M., "Information forage through adaptive visualization", In Proc. ACM Conf. on Digital Libraries 98 (DL98), Pittsburgh, PA, USA, 1998.

[Salton75] Salton G., Wong A., Yang C.S., "A Vector Space Model for Automatic Indexing ", Communications of the ACM, 18 (11): 613-620, 1975.

[Voorhees86] Voorhees E.M., "Implementing agglomerative hierarchical clustering algorithms for use in document retrieval", vol. 22, 465-476, Information Processing and Management, 1986.

[Zamir98] Zamir O. et Etzioni O., "Web document clustering: a feasibility demonstration", In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), 1998.

[Zhang01] Zhang D. et Dong Y., "Semantic, Hierarchical, Online Clustering of Web Search Results", 3rd International Workshop on Web information and data management, Atlanta, Georgia, 2001.

L'ARCHITECTURE CoMED
pour la gestion collective de documents électroniques dans l'organisation

Guillaume CABANAC
Max CHEVALIER
Claude CHRISMENT
Christine JULIEN

IRIT/SIG - Unit´e Mixte de Recherche 5505 CNRS
118 route de Narbonne
F-31062 Toulouse cedex 9
LGC - Equipe d'Accueil 2043
IUT Paul Sabatier Toulouse III
129 avenue de Rangueil - BP 67701
F-31077 Toulouse cedex 4

{cabanac,chevalier,chrisment,julien}@irit.fr

RÉSUMÉ

A l'heure actuelle, de nombreuses organisations souvrent d'une overdose informationnelle : les individus accèdent et conservent quotidiennement une quantitè croissante de documents électroniques. De plus, les efforts individuels de recherche mis en œuvre ne sont pas rentabilisés et valorisés par une difusion adaptée. Pourtant, la gestion rationnelle des documents est une condition nécessaire pour être performant dans les diférentes activitès liées au cycle de vie du document. C'est pourquoi nous présentons dans cet article une architecture intégrée pour la gestion collective des documents électroniques de l'organisation baptisée CoMED (Collective Management of Electronic Documents). Notre proposition, fondée sur l'activité d'annotation, exploite l'interdépendance des activités documentaires pour valoriser l'information introduite dans l'organisation. Son objectif consiste à aider l'individu dans ses activités documentaires, tout en faisant progresser l'organisation dans son ensemble.

Mots-clés: gestion collective de documents, architecture intègrèe, activitès documentaires, document électronique, annotation.

1. CONTEXTE ET PROBLÉMATIQUES DE NOS TRAVAUX

Comme le proposait [Seletzky, 2002] les organisations modernes peuvent être qualifées d' "orgaNETisées" car elles reposent de plus en plus sur les Systèmes d'Information (SI) disponibles, qu'ils soient internes (un réseau de l'entreprise, un ERP. . .) ou externes (Extranet, Internet. . .). Ces SI sont

d'importants vecteurs d'information pertinents pour l'organisation. Dans le même temps, ils provoquent une overdose informationnelle : l'organisation est le plus souvent incapable de traiter de façon optimale toute l'information collectée. Ce problème peut être identifié à deux principaux niveaux : individuel et collectif. Au niveau individuel, les membres de l'organisation peuvent avoir du mal à identifier, à trouver et à stocker l'information pertinente pour leurs activitès par exemple. Au niveau collectif, le problème consiste surtout à propager au mieux l'information introduite dans l'organisation (e.g. par chacun de ses membres) que les individus susceptibles d'être intéressès puissent l'exploiter au mieux. En efet, nous supposons que les besoins en information des diférents membres de l'organisation sont proches voire similaires, du moins au regard de leurs activités dans l'organisation. Une solution à ces problèmes permettrait de faire " vivre " l'information qui est trop souvent statique, stèrile et dispersée dans les méandres de l'organisation. Dans cet article, nous proposons d'étudier ces problèmes au travers des activités liées aux informations et plus particulièrement liées à leur support : les documents. Ces activités issues du cycle de vie d'un document sont présentées dans la deuxième section, selon leurs caractéristiques individuelles mais également collectives, couvrant ainsi les problèmatiques exposées. Enfin, nous proposons dans la troisième section une architecture de SI interne à l'organisation lui permettant d'exploiter au mieux ses documents. Cette architecture nommée CoMED (Collective Management of Electronic Documents) repose sur un élément central : l'activité d'annotation, ainsi que sur une approche collective - englobant les activités collaboratives et coopératives - permettant l'amélioration mutuelle des diférentes activités documentaires. Ainsi, le résultat d'une activité augmente automatiquement la performance des autres activités. Cette architecture est illustrée à partir de documents issus du média pouvant être vu aujourd'hui comme une source privilègiée d'informations : le Web.

2. LES ACTIVITÉS DOCUMENTAIRES DE L'ORGANISATION

Dans l'organisation, la gestion des documents peut être perçue comme un facteur de performance ; elle repose sur une optimisation de diférentes activités facilitant l'accès aux documents et, par conséquent, aux informations qu'ils contiennent. Ces activités forment le cycle de vie du document cf. figure 1 [Sellen et Harper, 2003, p. 203]. Les sections suivantes illustrent les activités identifiées (numérotées de ? à ?) qui peuvent être réalisées individuellement mais aussi collectivement, ce qui permet de tirer parti d'un groupe.

2.1. Recherche d'information et navigation ?

Le challenge quotidien de l'organisation moderne consiste à fournir à chacun de ses membres des informations utiles pour son activité. Pour ce faire, l'individu s'inscrit dans une activitéde Recherche d'Information (RI) ?
qui nécessite d'alterner sans réellement s'en rendre compte entre deux tâches distinctes : la recherche et la navigation [H¨olscher et Strube, 2000].

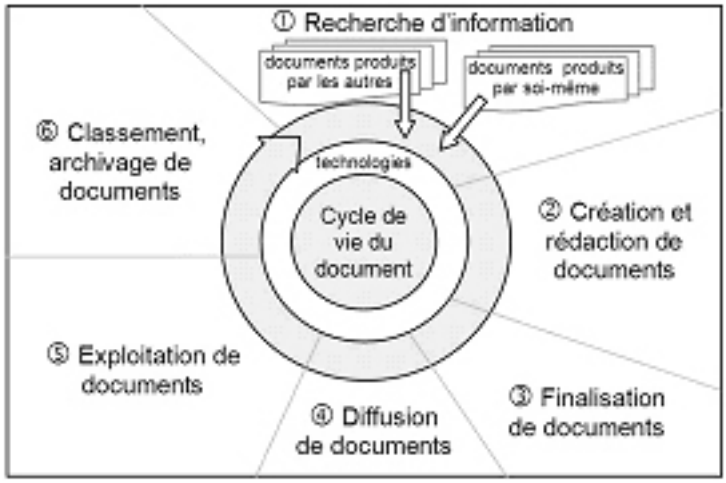

Fig. 1 - Cycle de vie du document [Sellen et Harper, 2003, p. 203]

Ces taches peuvent être réalisées individuellement ou collectivement. Activité individuelle. En phase de recherche, un individu explicite ses besoins sous la forme d'une requête qu'il soumet à un outil de recherche appelé moteur sur le Web. Ce dernier restitue alors à l'utilisateur une liste de documents triée par pertinence (système) d´écroissante. Il est a noter que d'autres types de visualisation ont été proposés au travers d'interfaces graphiques évoluées [Chen, 2006]. En parallèle, l'utilisateur navigue dans l'espace des documents disponibles sans avoir à formuler ses besoins mais également sans connaitre a priori ni le contenu ni l'organisation de l'espace des documents [Agosti et Smeaton, 1996]. Diférentes approches ont été proposées afin d'optimiser ces deux tâches complémentaires : la reformulation de requête, le réordonnancement de la liste des résultats voire la personnalisation de cette dernière pour améliorer la tâche de recherche. Concernant la navigation, les propositions suivent deux principaux courants. D'une part, des accélérateurs de navigation tels que Letizia [Lieberman, 1995] proposent à l'internaute les liens de l'hypertexte local qui sont jugés les plus pertinents pour la navigation courante. D'autre part, des systèmes proposent des documents intéressants par rapport à ceux consultés durant la navigation de l'utilisateur, en interrogeant un outil annexe e.g. WBI [Barrett et al., 1997].
Activité collective. On observe dans la vie courante que les individus se rassemblent en groupes sociaux ou organisationnels pour résoudre des problèmes de recherche [Karamuftuoglu, 1998]. Ainsi, des systèmes reposant sur une approche collective permettent ériences a un utilisateur de tirer profit des exés de recherche et de navigation d'autres individus qui possèdent des centres d'intérêts proches. Par exemple, pour la tâche de recherche, Cosydor [Jeribi et al., 2001] utilise les expériences de recherche des utilisateurs pour

aider un individu à formuler ses besoins, en amont de la RI. Dans IronWeb [Dussaux et Pécuchet, 2000] les individus peuvent bénéficier des connaissances d'experts en cherchant dans les hiérarchies de signets de ces derniers. De façon plus générale, le logiciel Human-Links permet la RI pair-à-pair dans le corpus constitué par les documents d'individus connectés sur le Web lors de la recherche. Par ailleurs, VR-Vibe [Benford et al., 1995] représente les résultats des recherches du groupe dans un environnement virtuel en 3D explorable par l'utilisateur. Enfin, l'aspect collectif dans la navigation est également présent dans Broadway [Jaczynski et Trousse, 1999] qui utilise des techniques de raisonnement par cas pour anticiper la navigation d'un utilisateur en fonction des navigations d'autres personnes pr´ec´edemment observ´ees.

2.2. Création et finalisation de documents

A partir des ressources trouvées ?, les membres organisationnels produisent une valeur ajoutée qu'ils communiquent en l'explicitant au moyen de rédactions individuelles et/ou collectives de documents à l'aide d'outils adaptés.

Activité individuelle. La rédaction ? de documents est facilitée par des logiciels de traitement de texte comme OpenO?ce Writer ou Microsoft Word qui permettent la mise en page de textes, schémas, graphiques. . . En phase de relecture, préalablement a la finalisation `? du document, l'individu peut ajouter des commentaires, reformuler des passages, identifier des erreurs typographiques, etc. grâce à la fonctionnalité correspondante du logiciel utilisé.

Activité collective. La phase de rédaction ? peut être réalisée collectivement grâce a un Wiki (un site Web entiérement modifiable par ses lecteurs), dont l'intérêt pour la création, la relecture et la difiusion des documents est montrée dans [Guzdial et al., 2000]. Similairement, des produits comme Microsoft Shared Point Services permettent éditer de façon synchrone

a plusieurs individus d'un même document : en phase de finalisation ?, chaque relecteur voit les modifications apportés par l'ensemble des collaborateurs.

2.3. Difiusion manuelle et automatique de documents

L'organisation, par le biais des activités précédentes, doit être capable de gérer un grand nombre de documents générés par ses activités ou bien issus de sources externes. Pour optimiser l'impact des informations contenues dans les documents, il est important que ceux-ci soient difiusés de manière adéquate dans l'organisation. Ces documents difiusés manuellement ou de façon automatique aux diférents membres susceptibles d'être intéressés permettront une meilleure exploitation de l'information tout en évitant sa perte dans les méandres de l'organisation.

Activité individuelle. Traditionnellement, la difiusion manuelle d'un document aux diférents membres de l'organisation est réalisée grâce au courrier électronique. Cette difusion est hautement cognitive pour l'individu qui la réalise car elle nécessite une connaissance de l'organisation et de ses membres, ce qui représente une limite non négligeable dans un contexte organisationnel de grande taille. En efet, un membre recevant une information ne pourra la diffuser correctement que s'il connaît les personnes susceptibles d'être intéressées, ce qui est peu probable dans une grande structure. La création de listes de difusion pointant vers un ensemble préétabli de courriels simplifie le problème sans toutefois le résoudre : l'individu doit connaître les centres d'intérêts de chaque liste. Une autre approche consiste étier et leurs a formaliser les processus m´caractéristiques (actions, délais associés et leur ordonnancement, intervenants et leurs rôles, données nécessaires et/ou produites) pour spécifier un work?ow [Marshak, 1994]. Ce dernier achemine automatiquement les documents en fonction d'événements qui déclenchent les règles de distribution définies a priori. Cependant, il est dificile de connaître l'ensemble des personnes intéressées par un sujet donné. C'est pourquoi une difusion plus large est parfois souhaitable au travers de portails organisationnels proposant des techniques de recherche d'information par exemple.

Activité collective. Afin de limiter léfort nécessité par une recherche d'information active, un SI de type Push propose automatiquement des documents aux individus (contrairement à l'approche Pull qui n´ecessite une démarche active de leur part cf. section 2.1.). Ainsi, un système defiltrage d'information " amène a l'utilisateur les documents qui vont lui permettre de satisfaire son besoin en information" [Belkin et Croft, 1992] en construisant et en faisant ´evoluer un mod`ele d'utilisateur appel´e " profi utilisateur ". Il existe plusieurs types de filtrages caractérisés par des modèles distincts d'appariement entre l'utilisateur et les documents. Nous pouvons citer Syskill & Webert [Pazzani et al., 1996] comme exemple de filtrage cognitif car il recommande un document à un individu lorsque son contenu est similaire au profil de l'utilisateur. Une autre approche qualifiée de filtrage collaboratif consiste à s'appuyer sur une communauté d'utilisateurs comme dans GroupLens [Konstan et al., 1997] qui exploite les jugements des individus vis-`a-vis des documents pour identifier des groupes d'opinion au sein desquels les documents sont difusés. Par conséquent, seuls les jugements des utilisateurs doivent être connus du système qui n'a ainsi pas besoin d'indexer les documents, contrairement au filtrage cognitif. Une évolution de ces approches consiste a tirer parti de l'organisation même de la communauté d'utilisateurs exploitée par ailleurs dans le filtrage collaboratif. Nous pouvons citer par exemple [Zhang et Ackerman, 2005] qui présentent des approches exploitant le réseau social (les relations entre individus) en complément des jugements émis par les utilisateurs. De tels systèmes privilégient les documents issus des contacts les plus proches d'un individu, en faisant l'hypothèse qu'il saura davantage apprécier la pertinence de ces sources.

2.4. Exploitation et classement de documents

Chaque membre de l'organisation est un " gestionnaire " potentiel des documents capitalisés collectivement. Il doit donc disposer d'outils adaptés pour gérer et exploiter au mieux les documents, qu'ils proviennent de sources internes ou externes à l'organisation e.g. du Web.

Activité individuelle. L'individu consulte les documents électroniques sur son écran ou bien sur papier après impression. Outre la lecture, les a?ordances1 du papier encouragent le lecteur à personnaliser le texte, a se l'approprier en formulant des annotations. Concrètement, la création d'annotations permet aux individus de matérialiser leur réflexion critique en reformulant, commentant, corrigeant, etc. des passages précis du document qu'ils lisent. Cette activité favorisant l'apprentissage est appelée " lecture active " [Adler et van Doren, 1972] par opposition à la lecture de loisir. En réponse au besoin d'annoter les documents électroniques (une exp´erience [Sellen et Harper, 2003, p. 95] montre que les individus se déclarent être frustrés de ne pas pouvoir le faire), de nombreux logiciels appelés Systèmes d'Annotation (SA) permettent l'annotation informelle 2 de tout ou partie d'un document sur le Web. Concernant le stockage ?, la hiérarchie thématique ad hoc de répertoires est une organisation trés répandue cf. les systèmes de gestion de fichiers, le classement hiérarchique pour les messageries électroniques. . . C'est également le cas sur le Web où la hiérarchie de signets 3 est un véritable espace d'information personnel pour l'utilisateur qui construit une représentation explicite et organisée des documents reflétant ses centres d'intérêt. D'après [Abrams et al., 1998] cette structure évolue de façon assez rapide car un utilisateur y ajoute trois à quatre signets en moyenne lors d'une session de navigation.

Activité collective. La facilité de partage est un avantage important en faveur des annotations électroniques par rapport a leurs homologues papier. Ainsi, le SA Amaya [Kahan et al., 2002] permet entre autres de consulter des documents électroniques accompagnés des annotations formulées par leurs lecteurs : l'individu n'est plus restreint au seul point de vue défendu par l'auteur du document car il bénéficie des analyses des lecteurs/annotateurs précédents. Par ailleurs, [Cabanac et al., 2006] propose à chaque lecteur de réagir a des passages d'un document grâce à une annotation de type donné d´ecrivant l'opinion exprimée (confirmation ou réfutation) ainsi que le contenu du commentaire (exemple, correction, question) par exemple. Par la suite, un individu peut réagir con réaction, ce a une annotation ou, de fa¸ecursive, a une autre r´1 Propriétés naturelles actionnables entre le monde et un individu e.g. les propriétés physiques du papier (?n, léger, poreux, opaque et ?exible) suggèrent les actions humaines de saisie, transport, pliage, écriture. . . [Sellen et Harper, 2003, p. 16]

2 Dans cet article, nous considérons des annotations informelles [Marshall, 1998] également qualifiées de cognitivement sémantiques [Zacklad et al., 2003] car leur contenu n'est pas contraint à un vocabulaire contrôlé e.g. issu d'une ontologie.

3 Le terme signet a pour synonymes " favori " en français et " bookmark " en anglais. qui forme une arborescence de réactions dont la racine est une annotation ancrée sur le document. Cette structuration hiérarchique et chronologique des réactions est appelée " fil de discussion ". Concernant le stockage, les membres de l'organisation peuvent centraliser ? leurs documents dans un entrepôt de documents [Khrouf et Soul´e-Dupuy, 2004] a?n d'établir une source commune d'information. Cette vision du partage peut être également remarquée au travers de systèmes de partage de hiérarchies de signets [Kanawati et Malek, 2000]. A contre-pied de la structuration hiérarchique, le phénomène récent de folksonomy c'est-`a-dire de " classification sociale " consiste à associer aux documents des termes choisis librement appelés tags. Ainsi, la maison d'édition Nature propose le système Connotea [Lund et al., 2005] qui permet de partager des documents " taggués " et de d´ecouvrir de nouvelles informations en explorant l'ensemble des documents collectivement constitué.

2.5. Synthèse sur les activités documentaires

Au travers des activités documentaires, nous pouvons constater que la prise en compte de l'expérience, des activités, des contacts, etc. d'un groupe permet de mettre en œuvre plusieurs systèmes distincts visant a exploiter de manière eficace les documents dans l'organisation. Ces systèmes sont basés sur le principe donnant/donnant car un individu œuvre pour un groupe et, réciproquement, un groupe œuvre pour un individu. Cependant, au regard des exemples mentionnés nous observons que les multiples systèmes se concentrent sur une, voire deux activités au maximum. Or, nous pensons que les approches actuelles ne prennent pas sufisamment en compte la réalité. En effet, l'être humain ne réalise pas ces diférentes activités linéairement et de façon cloisonnée comme le montre la figure 1 : il peut chercher de l'information ?, commencer à ecrire un document ? et chercher a nouveau pour approfondir un point précis ?. Comme conséquence à cette spécialisation, chaque système ne prend en compte qu'une partie de l'activité réelle de l'individu en délaissant les autres activités. Partant de cette constatation, nous proposons dans cet article une approche collective, globale et intégrée nommée CoMED (Collective Management of Electronic Documents) qui vise à couvrir l'ensemble des activités documentaires. Notre approche permet de faciliter mutuellement les diférentes activités réalisées. Notre proposition repose essentiellement sur l'activité d'annotation qui permet aux membres organisationnels de s'approprier et de m´emoriser les documents qu'ils jugent pertinents. Ainsi, l'activité de classement ? qui vise à conserver des documents pertinents au travers des annotations constitue l'élément central de l'approche CoMED.

3. COMED : GESTION COLLECTIVE DE DOCUMENTS

La gestion collective de documents doit recouvrir l'ensemble des activités documentaires de l'organisation pour optimiser son eficacité. Cependant, les diférents outils examinés dans la section précédente sont principalement

limités à une seule activité ; par conséquent, chaque système construit une vision distincte et partielle de l'organisation et de l'individu. C'est pourquoi nous proposons une intégration de toutes les activités documentaires par le biais d'une d´emarche principalement collective. Pour cela nous avons identifé l'élément fédérateur de notre approche : l'activité de stockage ?. En effet, nous constatons que toutes les tâches effectives de l'individu sont assujetties au stockage : une personne qui trouve des documents sur le Web ou qui crée des documents - en exploitant notamment des ressources déjà conservées - va vraisemblablement stocker tout ou partie de ces documents. Au regard de l'usage traditionnel fait des documents, nous pouvons souligner que cette activité traduit une appropriation personnelle de ces derniers, qui peuvent être stockés dans un système de fichiers au travers d'une arborescence de répertoires, par exemple.
Nous avons également souligné dans la section précédente que certaines approches se basent sur les hi´erarchies de signets pour conserver les documents suscitant un intérêt pour l'individu. Elles représentent de véritables espaces personnels d'informations [Abrams et al., 1998]. Toutefois, ces hiérarchies soufrent de certaines limites dont la sous-information car le contenu des signets est trés limité et même parfois peu pertinent. Par conséquent l'utilisateur peine parfois à se souvenir de la raison pour laquelle il a créé un signet vers un document sans avoir à en visualiser le contenu [Maarek et Ben-Shaul, 1996]. Nous pensons que l'individu gagnerait à créer des annotations à la place des signets parce qu'elles permettraient une mémorisation d'information plus riche : alors que le signet désigne obligatoirement un document dans son intégralité, le point d'ancrage de l'annotation est plus précis car défini sur tout ou partie du document (un paragraphe, une phrase, un mot, etc.). De plus, l'annotateur peut spécifier un commentaire et en donner un aperçu en y associant un type e.g. confirmation, réfutation, question. D'un point de vue collectif, les annotations permettent les débats au sein de fils de discussions. C'est pourquoi CoMED est basée sur des annotations qui sont exploitées comme vecteurs de stockage des documents pour l'individu. Notre approche est de type donnant/donnant car chaque individu, au travers de ses activités, participe automatiquement aux activités d'un groupe. En contrepartie, il bénéficie à son tour de l'activité des autres membres du groupe. Ainsi, même si l'activité au travers de cette application peut sembler importante, le gain que l'individu peut obtenir n'est pas négligeable. Afin de présenter la démarche adoptée dans CoMED, nous définissons tout d'abord la notion de hiérarchie d'annotations argumentatives, puis les diférents processus mis en œuvre dans notre architecture.

3.1. Hiérarchie d'annotations argumentatives

Dans CoMED, un utilisateur mémorise tout ou partie d'un document en créant une annotation. Afin de faciliter l'accès à ces documents qui ont fait l'objet d'un intérêt, nous proposons à chaque individu d'organiser thématiquement les annotations créées dans une Hiérarchie d'Annotations (HdA) personnelle ;

une HdA possède une structure arborescente similaire à celle bien connue d'une hiérarchie de signets, ou plus généralement d'un système de fichiers. Pour choisir le répertoire d'accueil de son annotation, l'individu doit réaliser un effort cognitif [Rucker et Polanco, 1997] qui consiste à sélectionner celui qui contient les annotations les plus similaires à l'annotation candidate. Cet effort cognitif individuel sera exploité pour améliorer les activités du groupe.

En termes de définition, les annotations formulées par les individus contiennent des données objectives DO ainsi que des informations subjectives IS. En effet, CoMED crée des DO pour mémoriser les attributs de l'annotation suivants : son identification, l'identité de son créateur qui donne accès à ses caractéristiques (identité, courriel, etc.) ; sa date de création qui permet d'organiser le ?l de discussion chronologiquement (cf. section 2.4.) ainsi que son point d'ancrage qui sp´ecifie de manière non ambiguë son emplacement au sein de la ressource annotée - diférentes techniques proposées sont applicables au contexte des documents semi-structurés e.g. XPointer [Kahan et al., 2002]. D'autre part, les IS sont formulées par les annotateurs ; elles peuvent être omises et comprennent : le contenu de l'annotation ainsi que sa visibilité (privée ou publique) qui permet de restreindre la portée de l'annotation, cette donnée renseigne CoMED quant à la difusabilitéde l'annotation.

De plus, nous proposons de conserver l'expertise de l'annotateur car les individus accordent en général davantage de crédit à l'opinion d'un expert qu'à celle d'un novice [Marshall, 1998]. Pour étayer ses remarques, l'annotateur peut également spécifier la liste des références sur lesquelles il s'appuie. Enfin, il peut qualifier la sémantique du contenu de son annotation en y associant diférents types qui sont, dans notre approche, liés au commentaire (correction, exemple ou question) ou à l'opinion de l'annotateur (confirmation ou réfutation) dans le cas d'une annotation argumentative (c'est une réaction dans le ?l de discussion). Par ailleurs, le type jugement (positif ou négatif) permet aux annotateurs de spécifier leur point de vue suggestif. Enfin, une annotation peut susciter des réactions (récursivement annotables) organisées chronologiquement au sein d'un ?l de discussion.

3.2. Aide `a la gestion collective et int´egr´ee de documents

Nous proposons dans cette section une vue synthétique de l'approche CoMED (cf. figure 2). Elle met en perspective le caractère global et intégré de notre architecture de SI qui couvre l'ensemble des activités documentaires. Notre architecture comporte quatre processus interdépendants car le résultat d'un processus est conjointement amélioré par celui des trois autres. Le premier processus (reco) diffuse ? les documents capitalisés par l'organisation sur la base des centres d'intérême (r´ets des membres ; le deuxièorg) permet la réorganisation ? thématique des espaces personnels d'annotations construits à l'aide de reco. Le troisième processus (navi) émet des recommandations en fonction de la navigation ? courante de l'utilisateur, en se basant sur les HdA

structurées par ré sociale eorg. Le dernier processus (valid) calcule la validité des annotations au travers des FdD, ce qui permet à l'utilisateur d'estimer le consensus global suscité par une annotation en phase de création ?, de finalisation ? ou de lecture active ? pour d´ecider de son traitement (et, de façon indirecte, de sa prise en compte éventuelle par reco). Nous détaillons ces processus dans les sections suivantes.

3.2.1. Processus de difiusion (reco) et de classement (réorg) de documents issus de la navigation collective

Au sein de l'organisation, les documents trouv&s ? par chacun des membres sont rarement diffusés ? car cela demande une démarche active hautement cognitive cf. section 2.3. Ainsi l'effort de recherche n'est pas factorisé, ce qui entraîne une perte d'eficacié et un coût non négligeables pour l'organisation. C'est pourquoi nous proposons le processus de recommandation reco qui automatise la diffusion des documents dans l'organisation. En entrée du processus, le corpus de recommandation est constitué de l'intégralité des documents visualisés dans l'organisation car ils ont suscité l'intérêt d'au moins un individu. En sortie du processus, des recommandations sont insérées dans les HdA des individus intéressés et plus précisément dans le répertoire le plus adéquat sur la base d'une comparaison thématique. Une telle recommandation est en réalité une annotation dont la granularité du point d'ancrage est maxi-male : il est d´efini sur l'intégralité du document. Concrètement, le processus reco réalise un parcours en profondeur d'abord de chaque HdA pour recommander les documents dans les répertoires les plus spécifiques tout en évitant plusieurs recommandations du même document dans un même chemin de la HdA. Le processus évalue alors une distance thématique entre le document à re-commander et chaque répertoire ri. Pour ce faire, un classifeur Ci = ?fi,ti? est associé à chaque ri, c'est un filtre qui accepte le document d lorsque fi(d) ? ti. Ce filtre est construit à partir des annotations de type " jugement positif " car nous sommes certains qu'elles représentent un intérêt pour l'utilisateur, contrairement à des annotations qualifiées avec des types différents. Les éléments clés du processus reco, évoqués sommairement ici, sont détaillés dans [Chevalier et Julien, 2003]. Déclenché à intervalle régulier, il permet à chaque individu de bénéficier de l'activité de recherche collective. Les documents recommandés sous forme d'annotation sont insérés directement dans les HdA des individus sous la forme d'une annotation. Une telle introduction de recommandations peut amener un utilisateur à vouloir réorganiser sa HdA. Comme cette tâche nécessite un effort cognitif important [Rucker et Polanco, 1997], nous proposons d'aider l'individu à réorganiser thématiquement tout ou partie de sa HdA grâce au processus ré sur l'al eorg. Ce dernier est bas´gorithme de classification ascendante hiérarchique [Jardine et van Rijsbergen, 1971] couplé à une technique d'étiquetage des classes obtenues et de seuillage permettant à l'utilisateur de spécifier la profondeur de la hiérarchie réorganisée [Chevalier et Julien, 2003].

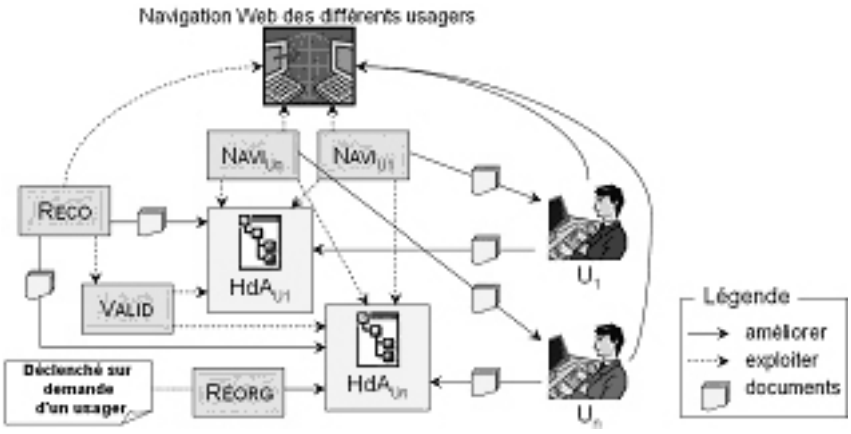

Fig. 2 - Vue synth´etique de l'architecture CoMED

3.2.2. Processus d'aide à la navigation (navi)

La navigation d'un individu profite au groupe (reco) mais aussi à l'individu lui-même : CoMED lui propose les annotations en rapport avec sa navigation courante (navi) qui proviennent des HdA de ses collègues. Ainsi, à contre-pied des approches basées sur l'hypertexte local ou sur des outils tiers cf. section 2.1., nous faisons l'hypothèse qu'un utilisateur peut tirer parti, lors de sa navigation, des ressources conservées par d'autres utilisateurs ayant des centres d'intérêt proches de son besoin momentané, caractérisé par sa navigation. C'est pourquoi le processus de recommandation navi se base en entrée sur la navigation courante d'un individu pour lui proposer en sortie un ensemble d'annotations pertinentes. Ces recommandations sont issues des HdA des membres organisationnels. Concrètement, le processus se base sur la structuration des HdA (résultant des efforts cognitifs individuels) pour recommander les annotations les plus proches du document en cours de navigation, tout en privilégiant les documents qui appartiennent au plus grand nombre de HdA. Dans cette approche, la proximité est calculée au travers des HdA ; elle est égale à la distance moyenne entre le document à recommander et le document visité. Notons que les annotations à la racine des HdA ne sont pas prises en compte car elles n'ont pas fait l'objet d'un effort cognitif de classement. L'intégralité de ce processus de filtrage est détaillée dans [Chevalier et Julien, 2003].
Les recommandations émises par CoMED lors de la navigation de l'individu lui permettent de découvrir des ressources jugées intéressantes par d'autres membres organisationnels. Ces derniers ont pu annoter les ressources et débattre de points de vue divergents au sein de fils de discussion. De telles

annotations argumentatives, véritables valeurs ajoutées aux documents, sont des sources d'informations complémentaires pour l'individu.

3.2.3. Processus de validation sociale d'annotations (valid)

L'individu engagé dans des activités collaboratives dans l'organisation a parfois besoin d'identifier les propositions validées socialement parmi toutes celles présentes sur un document, c'est-`a-dire celles qui font consensus. Par exemple, en phase de rédaction collective ? il est nécessaire d'identifier les idées qui font consensus au sein du groupe ; en phase de relecture ? il semble cohérent de prendre en compte les remarques à propos desquelles s'accordent de nombreuses personnes e.g. une réfutation confirmée par l'ensemble des relecteurs.

Enfin, lors de la lecture de documents ?, repérer les annotations qui font consensus permet d'obtenir une valeur ajoutée sensée alors que celles qui suscitent des débats permettent d'identifier les arguments de leurs détracteurs et défenseurs. En fait, évaluer le degré de consensus d'une annotation argumentative i.e. sa validité sociale nécessite d'extraire l'opinion (confirmation ou réfutation graduelles) de chaque réaction de son FdD et d'en faire une synthèse mentalement. Cette tâche entraîne une surcharge cognitive qui doit pourtant être évitée car elle perturbe l'activité principale de l'individu : sa lecture [O'Hara et Sellen, 1997]. C'est pourquoi le processus valid calcule la " validit´e sociale " de chaque annotation au travers du ?l de discussion associé. Cette valeur représente le degré de consensus des participants au débat. Concrètement, nous modélisons les différentes réactions (qui sont des annotations) sous la forme d'un graphe d'arguments reliés par des arcs étiquetés selon le type de chaque argument e.g. réfutation, confirmation, correction. Une valorisation de ce graphe d´etaillée dans [Cabanac et al., 2006] permet de connaître le degré d'accord entre les intervenants. Cette valeur peut être exploitée pour mettre en emphase les annotations qui font consensus et pour décider de leur traitement par reco. Cela permet aux individus de se focaliser sur les débats qui ont abouti à un accord durant l'exploitation des documents ? ou sur les re-marques les plus appuyées des relecteurs ? par exemple. éd´ A l'opposé, un procé indiquant les annotations qui ne font pas consensus est également possible.

3.3. Discussion

Notre architecture CoMED souffre de quelques limites que nous discutons dans cette section. Nous faisons l'hypothèse que les individus créent et font évoluer une hiérarchie thématique contenant leurs annotations. La structuration thématique est en effet couramment utilisée, bien que d'autres types d'organisations existent (e.g. par auteur, par date de lecture, par tâche) et, dans de telles configurations, notre architecture n'est pas adaptée. De plus,

en considérant que les annotations sont organisées hiérarchiquement, nous supposons qu'elles refletent les centres d'intérêts de l'utilisateur à l'image des hiérarchies de signets [Abrams et al., 1998]. Or, cette supposition devrait être vérifiée par des études comportementales avec des utilisateurs. Par ailleurs, nous ne prenons en compte actuellement que les annotations de type " jugement positif " car nous considérons que seul ce type traduit un intérêt explicite de l'utilisateur. Or, nous devrions nous demander si d'autres types peuvent également indiquer un intérêt. De même, nous n'exploitons pas pour le moment ce qui difiérencie les annotations des signets (point d'ancrage à granularité ajustable, présence de commentaire, de références sous la forme de liens hypertextes qui étayent un argument e.g. article scientifique, de types, notion d'expertise subjective de l'annotateur). L'architecture CoMED proposée dans cet article fait actuellement l'objet de réalisations logicielles qui sont au stade de prototypes. Nous désirons à terme évaluer nos contributions par des expérimentations en milieu écologique i.e. avec d'authentiques utilisateurs. Nous désirons notamment considérer certains résultats d'études comportementales [Beenen et al., 2004] qui montrent, dans des contextes applicatifs précis, qu'un individu participe davantage lorsqu'il sait que ses activités bénéficient à d'autres personnes.

4. CONCLUSION ET PERSPECTIVES

Cet article présente un état de l'art des systèmes proposés pour faire bénéficier l'individu des compétences collectives de l'organisation au sein de chaque activité documentaire. L'apport de ces systèmes est réel. Toutefois nous remarquons que leur spécialisation pour telle ou telle activité dans le cycle de vie du document est une limite importante. En effet, la caractérisation de l'utilisateur dans une activité ne profite qu'à l'amélioration de celle-ci alors que toutes les autres pourraient en bénéficier car elles sont interdépendantes. C'est pourquoi nous proposons l'architecture originale CoMED (Collective Management of Electronic Documents). Notre approche est constituée de processus intégrés et interdépendants qui couvrent les activités documentaires réalisées dans une organisation. L'objectif principal de CoMED vise à diffuser les documents introduits dans l'organisation par les individus afin que toutes les personnes intéressées en bénéficient automatiquement. Ainsi, l'information n'a pas être chercher à nouveau par les individus car elle leur est automatiquement

proposée en fonction de leurs centres d'int´erˆet. En termes de perspectives à notre approche, nous envisageons de construire le réseau social de l'organisation grâce à l'analyse des débats conduits au sein des fils de discussion. Cette connaissance permettrait d'améliorer l'apprentissage des thématiques d'intérêt de chaque membre organisationnel pour afiner la pertinence des recommandations. D'autre part, notre approche ne permet pas à l'utilisateur de consulter la représentation que CoMED a de lui. Pourtant, la transparence du système permettrait aux utilisateurs de mieux

comprendre les propositions émises par le système, ils pourraient également mettre à jour les informations qui les concernent ; cela ne peut qu'encourager les individus à participer activement dans CoMED. Enfin, pour que notre système apporte une réelle plus-value à l'organisation dans son ensemble, nous désirons encourager la participation des membres organisationnels en termes de débats constructifs argumentés, de notes de lectures pertinentes, etc.

RÉFÉRENCES

Abrams, D., Baecker, R., et Chignell, M. (1998). Information archiving with bookmarks : personal web space construction and organization. Dans CHI '98 : Proceedings of the SIGCHI conference on Human factors in computing systems, pages 41-48, New York, NY, USA. ACM Press/A-W. Adler, M. J. et van Doren, C. (1972). How to Read a Book. Simon & Shuster.

Agosti, M. et Smeaton, A. F. (1996). Information Retrieval and Hypertext. Kluwer Academic Publishers, Dordrecht.

Barrett, R., Maglio, P. P., et Kellem, D. C. (1997). How to Personalize the Web. Dans CHI '97 : Proceedings of the SIGCHI conference on Human factors in computing systems, pages 75-82, New York, NY, USA. ACM Press.

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., et Kraut, R. E. (2004). Using Social Psychology to Motivate Contributions to Online Communities. Dans CSCW '04 : Proceedings of the 2004 ACM conference on Computer supported cooperative work, pages 212-221, New York, NY, USA. ACM Press.

Belkin, N. J. et Croft, W. B. (1992). Information Filtering and Information Retrieval : Two sides of the Same Coin ? Communications of the ACM, 35(12) :29-38.

Benford, S., Snowdon, D., Greenhalgh, C., Ingram, R., Knox, I., et Brown, C. (1995). VR-VIBE : A Virtual Environment for Co-operative Information Retrieval. Computer Graphics Forum, 14(3) :349-360.

Cabanac, G., Chevalier, M., Chrisment, C., et Julien, C. (2006). Validation sociale d'annotations collectives : argumentation bipolaire graduelle pour la théorie sociale de l'information. Dans INFORSID'06 : 24e congr`es de l'INFormatique des Organisations et Systèmes d'Information et de Décision, pages 467-482. Editions Inforsid.
Chen, C. (2006). Information Visualization : Beyond the Horizon. Springer.

Chevalier, M. et Julien, C. (2003). Interface adaptative et coopérative pour l'aide à la Recherche d'Information sur le Web. Information -Interaction Intelligence, 3(2) :47-73. Dussaux, G. et Pécuchet, J.-P. (2000). Création collective de bases de connaissances sur le web : Indexation par l'usage des documents. Dans CIDE 2000 : 3e Colloque International sur le Document Electronique, pages 185-203. Europia.

Guzdial, M., Rick, J., et Kerimbaev, B. (2000). Recognizing and Supporting Roles in CSCW. Dans CSCW '00 : Proceedings of the 2000 ACM conference on Computer supported cooperative work, pages 261-268, New York, NY, USA. ACM Press.

Hölscher, C. et Strube, G. (2000). Web search behavior of Internet experts and newbies. Computer Networks, 33(1-6) :337-346.

Jaczynski, M. et Trousse, B. (1999). Broadway : A Case-Based System for Cooperative Information Browsing on the World-Wide-Web. Dans Padget, J. A., éditeur, Collaboration between Human and Artificial Societies, volume 1624 de Lecture Notes in Computer Science, pages 264-283. Springer.

Jardine, N. et van Rijsbergen, C. J. (1971). The Use of Hierarchic Clustering in Information Retrieval. Information Storage and Retrieval, 7(5) :217-240.

Jeribi, L., Rumpler,B., et Pinon, J.-M. (2001). Système d'aide à la recherche et à l'interrogation de bases documentaires, fondé sur la réutilisation d'expériences. Dans INFORSID'01 : 19e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision, pages 443-463.

Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E., et Swick, R. R. (2002). Annotea : an open rdf infrastructure for shared web annotations. Computer Networks, 32(5) :589-608.

Kanawati, R. et Malek, M. (2000). Informing the Design of Shared Bookmark Systems. Dans RIAO 2000 : actes de la conférence Recherche d'Information Assistée par Ordinateur, pages 170-179, Paris, France.

Karamuftuoglu, M. (1998). Collaborative Information Retrieval : Toward a Social Informatics View of IR Interaction. Journal of the American Society for Information Science, 49(12) :1070-1080.

Khrouf, K. et Soulé-Dupuy, C. (2004). A Textual Warehouse Approach : A Web Data Repository. Dans Mohammadian, M., éditeur, Intelligent Agents for Data Mining and Information Retrieval, chapitre 7, pages 101-124. Idea Publishing Group.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., et Riedl, J. (1997). GroupLens : Applying Collaborative Filtering to Usenet News. Commununications of the ACM, 40(3) :77-87.

Lieberman, H. (1995). Letizia : An Agent That Assists Web Browsing. Dans IJCAI'95 : International Joint Conference on AI, pages 924-929.

Lund, B., Hammond, T., Flack, M., et Hannay, T. (2005). Social Bookmarking Tools (II) : A Case Study -Connotea. D-Lib Magazine, 11(4).

Maarek, Y. S. et Ben-Shaul, I. (1996). Automatically Organizing Bookmarks per Contents. Computer Networks, 28(7-11) :1321-1334.

Marshak, R. T. (1994). Work?ow White Paper : an Overview of Workfow Software. Dans Bierman, B., éditeur, Workfow'94. The Conference Group.

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. Dans HYPERTEXT '98 : Proceedings of the 9th ACM conference on Hypertext and hypermedia, pages 40-49, New York, NY, USA. ACM Press.

O'Hara, K. et Sellen, A. (1997). A Comparison of Reading Paper and On-Line Documents. Dans CHI '97 : Proceedings of the SIGCHI conference on Human factors in computing systems, pages 335-342, New York, NY, USA. ACM Press.

Pazzani, M. J., Muramatsu, J., et Billsus, D. (1996). Syskill & Webert : Identifying Interesting Web Sites. Dans AAAI/IAAI, volume 1, pages 54-61.
Rucker, J. et Polanco, M. J. (1997). Siteseer : personalized navigation for the web. Communications of the ACM, 40(3) :73-76.

Seletzky, S. (2002). L'entreprise orgaNETis´ee. InfoPro. Dunod.
Sellen, A. J. et Harper, R. H. (2003). The Myth of the Paperless Ofice. MIT Press, Cambridge, MA, USA.

Zacklad, M., Lewkowicz, M., Boujut, J.-F., Darses, F., et Détienne, F. (2003). Formes et gestion des annotations numériques collectives en ingénierie collaborative. Dans Dieng, R., éditeur, IC2003, pages 207-224, France. PUG.

Zhang, J. et Ackerman, M. S. (2005). Searching For Expertise in Social Networks : A Simulation of Potential Strategies. Dans GROUP '05 : Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work, pages 71-80, New York, NY, USA. ACM Press.

UN MODÈLE POUR LA CONFRONTATION D'OPINIONS NUMÉRISÉES SOUS PORPHYRY

Samuel Gesche
Sylvie Calabretto
Guy Caplat

LIRIS, INSA de LYON. 69621 Villeurbanne Cedex - France
{Samuel.Gesche, Sylvie.Calabretto}@insa-lyon.fr
Département Informatique de l'INSA de Lyon
- 69621 Villeurbanne Cedex - France
Guy.Caplat@insa-lyon.fr

RÉSUMÉ

Le système Porphyry s'appuie sur un nouveau modèle pour les bibliothèques numériques, dans lequel des experts peuvent exprimer etreprésenter des structures sémantiques reliant des documents. Notre contribution à ce travail consiste à y ajouter la possibilité de confronter les points de vue de ces experts, par le biais du système Platon, qui sera à terme une extension de Porphyry.

MOTS-CLES : Bibliothèques numériques, confrontation, points de vue, opinions, multiples experts, sciences humaines

ABSTRACT

The porphyry system is based on a new system for digital libraries, in which experts can express and represent semantic structures between documents. Our contribution to this work is to allow the confrontation of the viewpoints given by them, and the subsequent system, Platon, is planned as an extension of Porphyry.

KEYWORDS: Digital Libraries, confrontation, viewpopints, opinions, multiple experts, humanities

1. INTRODUCTION

Notre travail se trouve à l'interface entre deux thématiques, celle des bibliothèques numériques et celle de la multi-expertise. Comme la première, nous avons à gérer des contenus documentaires vastes, et comme la seconde nous avons comme but d'outiller un travail précis faisant appel à la coopération de plusieurs spécialistes.

Cependant, contrairement aux bibliothèques numériques nous ne nous limitons pas à la simple gestion de contenu, et contrairement aux systèmes experts

nous n'avons pas la vocation de " remplacer " l'expert en le reléguant au statut de superviseur. L'objectif de Porphyry est en effet de fournir un instrument de travail pour la construction du sens.

Les bibliothèques considérées dans Porphyry sont des bibliothèques spécialisées, destinées à des experts. Dans un tel cadre, limiter la description des documents à une indexation unique, fixe et effectuée par un tiers, revient à nier leur expertise. Porphyry repose donc sur des structures qui sont construites par les experts en fonction de leurs problématiques et de leurs spécialisations.

Dans son état actuel, Porphyry offre un moyen de visualiser différents points de vue lorsqu'ils sont appliqués aux même cas expérimentaux. Cependant, la visualisation des points de vue n'est qu'une étape : Platon (PLateforme d'Analyse et de Traitement d'Opinions Numérisées), étend l'activité de construction de sens à la confrontation de points de vue, et ceci quel que soit le formalisme d'expression.

Le présent document est organisé comme suit : après une présentation de quelques travaux connexes aux nôtres en section 2, nous faisons, en section 3, une présentation de Porphyry, des réseaux de description et de l'évolution apportée à Porphyry par le module Platon. Puis, en section 4, nous définissons notre concept d'opinion, ainsi que d'autres concepts utiles à notre étude. Dans la section 5 nous présentons notre modèle de la confrontation, avant de conclure et de donner quelques perspectives de notre travail dans la dernière section.

2. TRAVAUX CONNEXES

Les bibliothèques numériques sont une technologie largement employée aujourd'hui, mais le défi d'en faire un système efficace -aussi bien au niveau des performances qu'au niveau de l'adéquation aux besoins- est toujours nouveau. Chaque projet a ses propres spécificités, ce qui fait la variété du monde des bibliothèques numériques. Dans ce contexte, le défi principal reste toujours de fournir aux utilisateurs un système adapté à leur mode de travail, suivant qu'ils soient simples internautes ou philologues à l'expertise pointue. Cependant, comme cette technologie est en train d'acquérir sa maturité, il existe un certain nombre de projets qui visent à développer des structures unifiées pour les bibliothèques numériques. Le réseau européen DELOS [DEL06] a pour but de réunir les équipes de recherche dans le domaine des bibliothèques numériques afin de mettre en commun leurs capacités. Ils cherchent donc à élaborer des théories et des structures complètes et unifiées, pour fournir une technologie générique et robuste pour le développement de bibliothèques numériques. Une autre méthodologie de conception, développée outre-atlantique [GON04], est basée sur cinq aspects, les flux, les structures, les espaces, les scénarios et les communautés.

Cette méthodologie a été instanciée notamment sur un projet lié à l'archéologie [SHE05]. Dans le cadre de l'acquisition de connaissances à partir de multiples experts, dans le but de perfectionner les systèmes experts, l'INRIA, dans le cadre du projet ACACIA, a développé une approche multi-point de vue basée sur les graphes conceptuels (comme le montre la thèse [RIB99]). [RIB97] présente un certain nombre d'approches antérieures (TROPES et ROME notamment). [GAM94] et [DIE98] posent les bases mathématiques et algébriques d'une confrontation entre plusieurs experts tandis que [DIE94] en donne une approche stratégique. [RIB02] résume la théorie des points de vue en utilisant le terme d'opinion pour décrire les connaissances non consensuelles (terme que nous reprenons dans un contexte plus large dans notre travail). D'autres approches multi-points de vue ont vu le jour dans le but de l'intégration de multiples compétences dans le cadre d'un projet transversal ([NAN01] par exemple). Elles sont basées pour leur part sur les technologies de l'Internet. Il est à noter également que deux approches basées sur UML, incluant pour chacune une extension multi-points de vue de ce formalisme, sont menée également (voir [NAS03] et [LAH05]).

Le projet Porphyry vise à offrir aux chercheurs en Sciences Humaines des assistants à la construction du sens dans les bibliothèques numériques. Il s'agit d'un domaine spécialisé qui n'est pas très représenté dans les publications. En effet, les contraintes métier de la recherche amènent souvent les chercheurs à développer eux-mêmes leurs outils, et les publications qu'ils font portent sur les résultats obtenus avec ces outils et non sur les outils eux-mêmes (généralement d'ailleurs l'effort est axé sur l'efficacité et non sur la réutilisabilité dans d'autres projets avec d'autres équipes). On trouve néanmoins des traces de tels outils, comme [ACH04] ou [ORI00].
Citons encore le projet européen Arkeotek [ARK06] qui vise à réformer les pratiques éditoriales en Sciences Humaines et Sociales, en dégageant une structure logique de la manière dont les chercheurs présentent leurs théories, et le projet HyperTopic ([CAH04], [CAH06]) qui, bien que tourné vers l'entreprise plus que vers la recherche en Sciences Humaines, développe une certaine intéropérabilité avec Porphyry [BEN06].

3. DE PORPHYRY À PLATON

3.1. Porphyry

Porphyry [POR06] propose l'instrumentation du travail des experts par l'enrichissement itératif du corpus par des structures hypermédias. Ces structures, comme nous l'avons dit dans l'introduction, sont construites par les experts en fonction de leurs problématiques et de leurs spécialisations. Elles sont exprimées dans le formalisme des réseaux de description, sur lequel nous reviendrons plus loin. Prenons l'exemple des collections de l'Ecole Française d'Athènes, initiatrice du projet : différents acteurs vont les organiser en fonction de leur spécialité, comme le montre la figure 1.

La première structuration est donnée par le maquettiste. Chaque page est désignée sans ambiguïté par le triplet " Collection/Volume/Folio ". La seconde est donnée par le bibliothécaire afin de faciliter l'accès au corpus pour les chercheurs d'où la structuration en titre, date, auteurs. La troisième structure est celle de l'archiviste. A chaque figure publiée dans les collections sera associée la référence et la description (auteur, date de prise de vue). Enfin, la quatrième structure ouvre sur beaucoup d'autres : celles des experts.

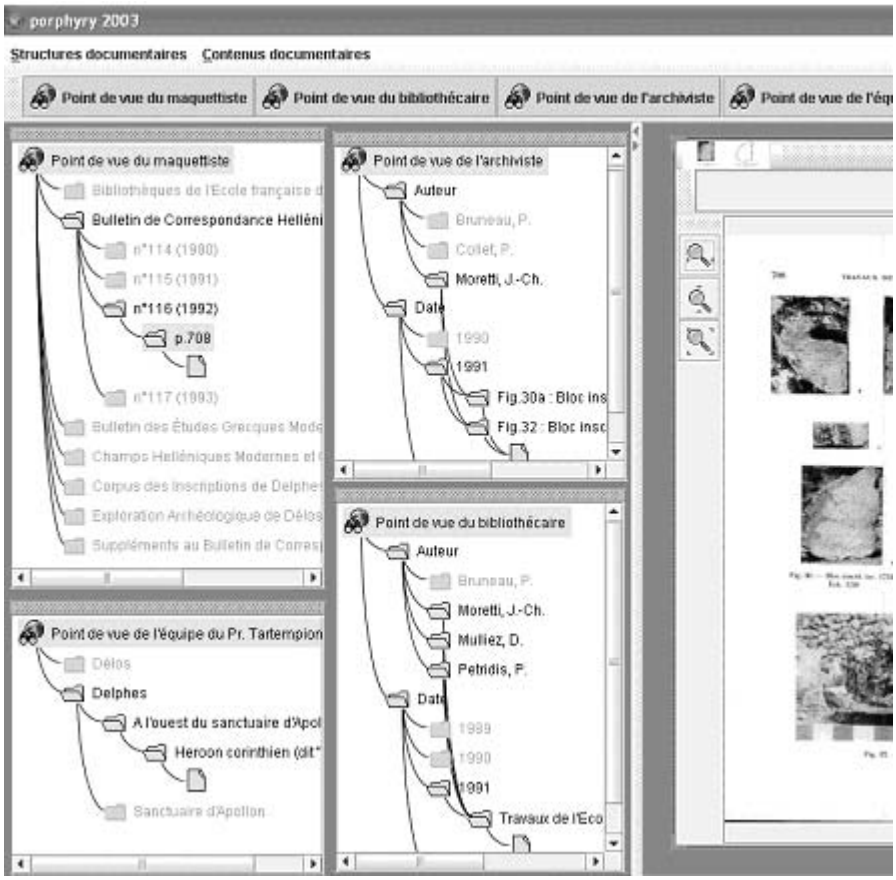

Fig.1 - Quatre points de vue dans le prototype Porphyry

Chacun des experts (ou communauté d'experts, du moment que leur structuration est unique, comme dans notre exemple) fournit sa propre structuration de la partie du corpus sur laquelle il travaille. Et chacune de ces structurations correspond à une opinion sur le corpus qui n'a pas à être (et ne peut pas être) comparé aux autres dans un esprit d'intégration ou de consensus.

Pour expliciter cela, prenons un deuxième exemple, toujours en archéologie. Philippe Bruneau [BRU76], en réponse aux premières " banques de données archéologiques ", faisait déjà remarquer l'impossibilité de décrire objectivement une photographie de mosaïque noire et blanche. Etait-on en présence de la représentation d'une mosaïque noire sur fond blanc ou blanche sur fond noir ? Dans un tel cas, nous devons disposer d'un modèle permettant d'exprimer qu'une première opinion affirme qu'il s'agit d'une mosaïque noire sur fond blanc et qu'une seconde affirme l'inverse. Ces deux opinions étant contradictoires, notre " modèle " doit être plus permissif que la normale. De fait, la réponse dépend de l'opinion de l'expert, de ses références, de sa grille d'interprétation des documents, donc d'une théorie qu'il applique au corpus. Ce qui est intéressant, ce sont donc bien les relations entre théories, et leurs points communs tout autant que leurs contradictions.

3.2. Les réseaux de description

Les réseaux de description sont le formalisme dans lequel sont exprimées les opinions sous Porphyry. Il s'agit d'une variante des réseaux sémantiques limitée aux relations de spécialisation et de généralisation ([BEN04], [POR06]).

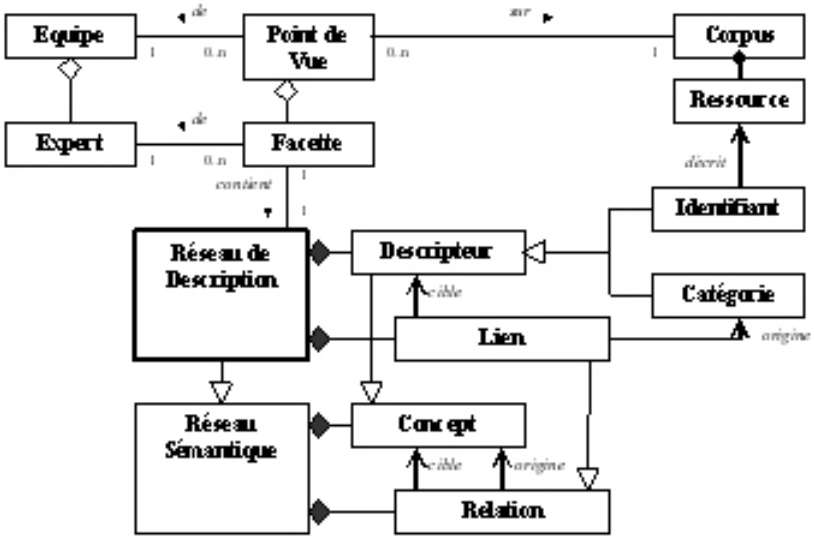

Fig.2 - Diagramme UML des réseaux de description

Proposer la structuration de corpus par les experts eux-mêmes nécessitait de trouver un formalisme aussi pauvre sémantiquement et peu contraint que possible. Pauvre sémantiquement, car la richesse sémantique ne résisterait pas à la variété des points de vue (par exemple, si on considère le formalisme

des graphes conceptuels et que chaque expert crée son treillis de relations entre concepts, à quoi cela sert-il ?). Peu contraint, car les experts doivent définir eux-mêmes les contraintes (si on propose un jeu de relations unifié, en quoi est-on plus performant qu'un système qui propose une indexation unique ?).
Il fallait de plus que ce formalisme soit à la portée d'experts en Sciences Humaines, dont les modèles ne sont pas ceux des informaticiens, mais néanmoins implantable dans l'interface du prototype Porphyry.

Les réseaux de description (présentés dans la figure 2) sont donc issus des réseaux sémantiques. Cependant, dans le cas des réseaux de description, les relations ne sont pas typées car le typage est jugé trop riche sémantiquement (on l'a expliqué plus haut). A la place, on a une relation unique, de type généralisation / spécialisation, qui peut signifier n'importe quoi, de la composition à l'antériorité. Les concepts sont remplacés par des descripteurs, qui soit pointent sur des documents (chaque document peut avoir un unique descripteur), soit sont reliés à d'autres descripteurs. Chaque descripteur a un nom qui lui est propre.

L'implantation visuelle des réseaux de description dans le prototype ressemble de fait à une arborescence de fichiers qui permettrait de placer un fichier ou un répertoire dans plusieurs répertoires, comme on peut le voir dans la figure 1.

3.3. Platon

Platon repose sur le principe d'un empilement de niveaux de modélisation (principe que l'on retrouve souvent en modélisation informatique, par exemple dans [UML06]), où chaque niveau modélise le niveau inférieur. Ainsi, les documents sur lesquels travaillent les chercheurs qui utilisent Porphyry traitent d'objets réels, objets que l'on place par convention au niveau 0.
Les documents eux-mêmes en sont des modèles, et sont par conséquent placés au niveau M1.
Les modèles construits dans Porphyry, qui sont les structurations que les chercheurs appliquent sur ces documents, sont situés au niveau supérieur M2.
Le système de confrontation de Platon, pour pouvoir fonctionner, a besoin d'un troisième niveau M3 : dans ce niveau on décrit les langages dans lesquels sont écrits les modèles placés en M2.
Afin de permettre l'interopérabilité au sein du système de confrontation, un dernier niveau (M4, donc) a pour objet de décrire les métamodèles du niveau M3. Les formalismes considérés ici doivent se définir eux-mêmes de manière à ne pas requérir de niveau supérieur. Ce niveau ne contient donc pas uniquement les formalismes de description de langages, mais également tout formalisme ou langage utile à la traduction d'un langage dans un autre. Sa nature réflexive l'exclut au moins en partie des traitements informatiques.
Les formalismes du niveau M4 sont donc soit des langages formels, soit des procédures qui devront être respectées par ceux qui décrivent les langages.

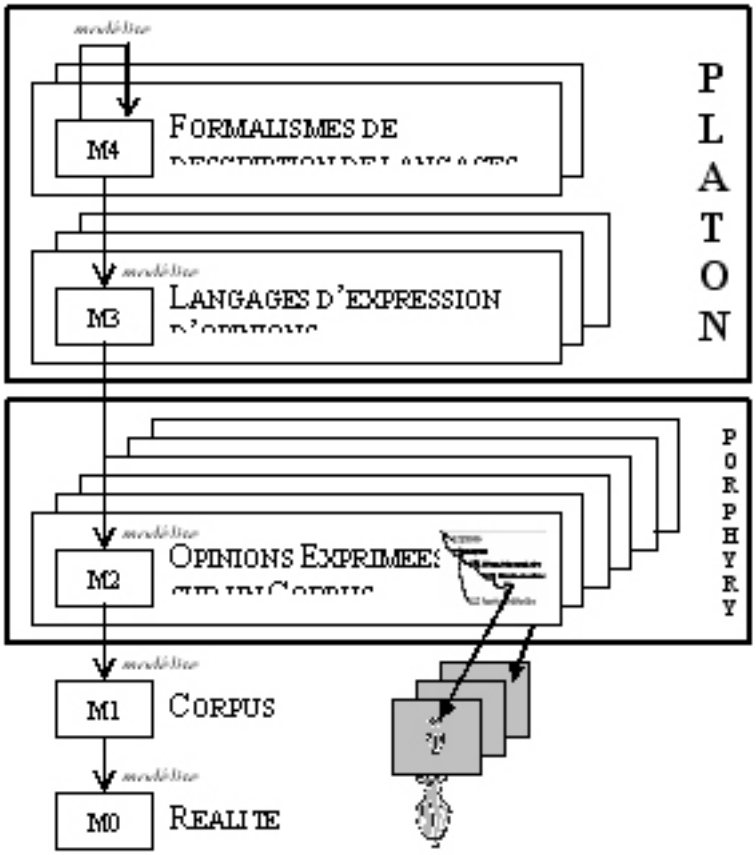

Fig.3 - Les cinq niveaux de modélisation

Les langages et leurs formalismes de description, tout comme les opinions ou les documents du corpus, font partie du vaste ensemble documentaire de Porphyry. Ainsi, on peut exprimer des opinions sur eux.

Il est à noter que aussi bien les opinions, que leurs langages d'expression, que les formalismes de la couche supérieure, sont fixés sur des supports documentaires qui font partie du vaste ensemble documentaire de Porphyry (en d'autres termes, de la base documentaire ou du corpus en M1). Ceci permet l'expression d'opinions sur tous les niveaux comme par exemple une analyse critique de la pertinence de tel langage de structuration.

Un même document pourra donc être, en tant que structuration d'un corpus, placé au niveau M2, et en tant que document publié pouvant prendre place dans une structure, placé au niveau M1.

4. MODÈLE DE L'OPINION

4.1. Le point de vue " Opinion "

Comme Porphyry est un outil à destination des Sciences Humaines, le désaccord entre chercheurs y est un sujet d'enrichissement et de réflexion [BEN01], ce qui va à l'encontre d'une intégration du savoir.
Les connaissances inscrites dans les réseaux de description de Porphyry ne sont donc pas consensuelles. De là vient le terme d'opinion qui fait référence à des points de vue non consensuels [RIB02]. Il est à remarquer cependant que, en Sciences Humaines plus qu'ailleurs, le terme d'opinion a une connotation péjorative . Il est donc important de noter que pour nous une opinion est le résultat d'un travail d'expertise, donc une construction scientifique à part entière. Simplement, il ne fait pas l'objet d'un consensus.

Une opinion est une théorie portant sur un domaine. La théorie en question est d'ordre abstrait, c'est une idée. La théorie qu'est l'opinion comporte naturellement des lacunes, des incohérences et en règle générale elle ne peut pas être connue de manière intégrale, pas même par son auteur. Bien entendu, cela étant, nous ne tiendrons compte que de l'expression d'une telle théorie.

L'expression d'une opinion est un modèle exprimant la théorie qu'est l'opinion. Le modèle est une construction concrète, c'est un document. En tant que tel, il est écrit dans un langage bien défini. Le langage peut être n'importe quel mode d'expression qui permet de sous-tendre une communication, cependant nous nous limitons aux formalismes (langages de modélisation ou de programmation).

Les opinions, de par l'absence de consensus, apportent au modèle un certain nombre de contraintes :

- Accepter des lacunes et des incertitudes dans les connaissances. En d'autres termes, pas de tiers exclu.
- Etre conciliable avec une autre opinion contradictoire. En effet, le principe de non-contradiction est fréquemment violé dans ce contexte.
- Supporter l'incohérence à l'intérieur même de sa structure.

Dans Porphyry, l'opinion est une structuration de corpus faite par un expert. Elle est exprimée par un modèle dans un langage permettant l'expression des opinions (les réseaux de description de Porphyry par exemple).

En philosophie, l'opinion est un avis que l'on considère comme vrai. La caractéristique principale de cet avis est son immédiateté : il n'y a aucun raisonnement derrière, et encore moins une démarche scientifique. Une opinion est donc affirmée et non élaborée. Elle provient souvent d'une source extérieure à la personne.
Ce n'est pas du tout la signification que nous rattachons au terme d'opinion.

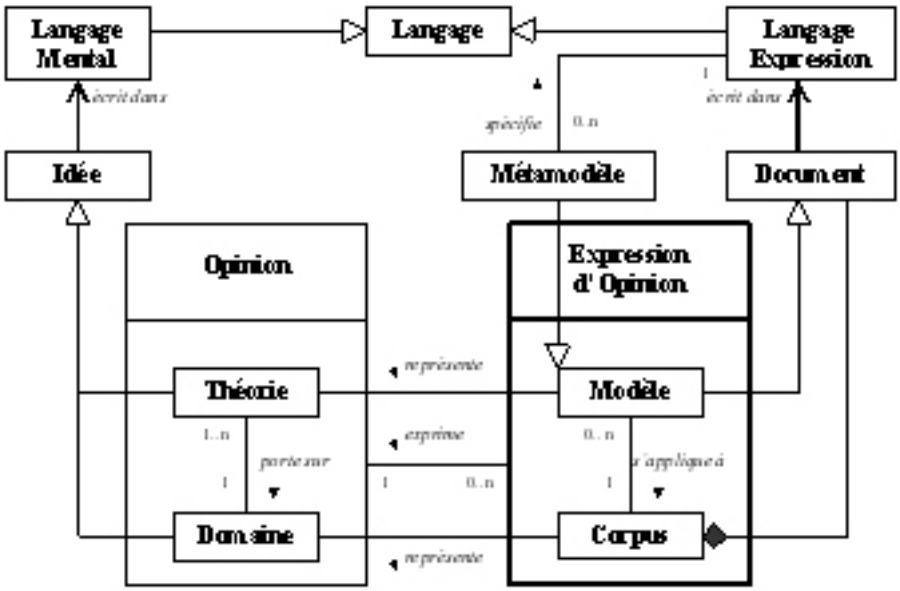

Fig. 4 - Modèle de l'opinion sous Porphyry

4.2. La notion de Domaine

De même que la théorie qui constitue l'opinion, le domaine sur lequel s'applique cette théorie est une construction abstraite qui a besoin d'être exprimée. Posphyry permet par conséquent l'expression du domaine, et cela grâce à la couche Steatite [POR06].

L'expression du domaine, que nous appelons corpus, est l'ensemble des documents que le chercheur a utilisés pour construire son modèle, et auxquels celui-ci s'applique. Ces documents sont aussi divers que des photos, des fragments de texte, des articles ou des notes personnelles (la liste n'est pas exhaustive) -le tout sous format numérique.

L'importance du domaine pour la confrontation des opinions ne doit pas être sous-estimée : en effet, deux théories contradictoires n'ont pas le même intérêt si elles ont été construites sur des éléments qui n'avaient rien à voir que si les documents sont communs. Le deuxième cas est beaucoup plus intéressant que le premier.

Il faut donc, dans le cadre de la confrontation d'opinions, être en mesure de savoir sur quoi s'appuient ces opinions.

4.3. La notion de langage

Notre approche du langage est d'abord une approche globale : nous incluons dans ce terme aussi bien un aspect informatif qu'un aspect éditorial.
En d'autres termes, nous regroupons sous le terme " langage ", en plus du langage au sens habituel (composé d'expressions agencées selon des règles lexicales, syntaxiques, sémantiques etc.), la mise en page et les contraintes éditoriales qui président à l'expression.

Supposons par exemple le cas de deux journaux. Ces deux journaux, supposons-le, commentent les mêmes nouvelles, et ce dans le même dialecte local. Cependant, les rédactions respectives imposent des rubriques différentes, ou les mêmes rubriques dans un ordre différent, ou encore les mêmes rubriques mais avec des contenus différents. Si ce sont des contraintes, nous considérons que c'est une partie intégrante du langage, et donc les deux journaux n'utilisent pas le même langage.

La raison de ceci est que nous parlons d'opinions, et la différence entre opinions peut se faire au niveau des contraintes éditoriales aussi bien qu'au niveau du message. En effet, supposons le cas d'une satire sous forme poétique visant un texte juridique. Il est important lors de la confrontation de considérer le fait que l'un soit poétique, et l'autre juridique, sous peine de perdre de précieuses indications (indépendamment de l'influence des styles sur le message lui-même).

Cependant, le langage d'expression de l'opinion est nécessairement, dans Porphyry comme dans les autres systèmes informatiques, un formalisme plus restreint que la langue naturelle. Pour passer de la langue naturelle au formalisme, on bénéficie du concours d'une interface homme-machine.

A ce propos, la puissance d'expression des formalismes utilisés dépend généralement de ce que les concepteurs ont voulu que l'on puisse exprimer. Par exemple, les réseaux de description de Porphyry, nous l'avons vu, permettent l'expression de concepts et de relations de spécialisation et de généralisation entre ces concepts, ce qui permet d'exprimer moins de choses, mais plus facilement. Pour exprimer des opinions, il sera donc utile de savoir si le formalisme que l'on utilise est prévu pour cela.

Notre approche de la confrontation est multi-langage. Ainsi, nous travaillons avec plusieurs formalismes. D'abord, les réseaux de description, qui sont à la base de Porphyry, et le formalisme des Topic Maps vers lequel Porphyry se dirige. Nous rajoutons à cela les graphes conceptuels [SOW00] qui sont souvent utilisés dans le contexte de l'acquisition de connaissances et des solutions multipoints de vue. Dans le contexte de la modélisation (et dans certaines solutions multi-points de vue comme VUML), c'est souvent le langage UML [UML06] qui revient, il a par ailleurs l'avantage de disposer d'une spécification très détaillée. Nous l'intégrons également.

5. CONFRONTATION D'OPINIONS

5.1. Objectif de la confrontation

Notre objectif est de permettre la confrontation d'opinions, donc des modèles dans lesquels elles sont exprimées.

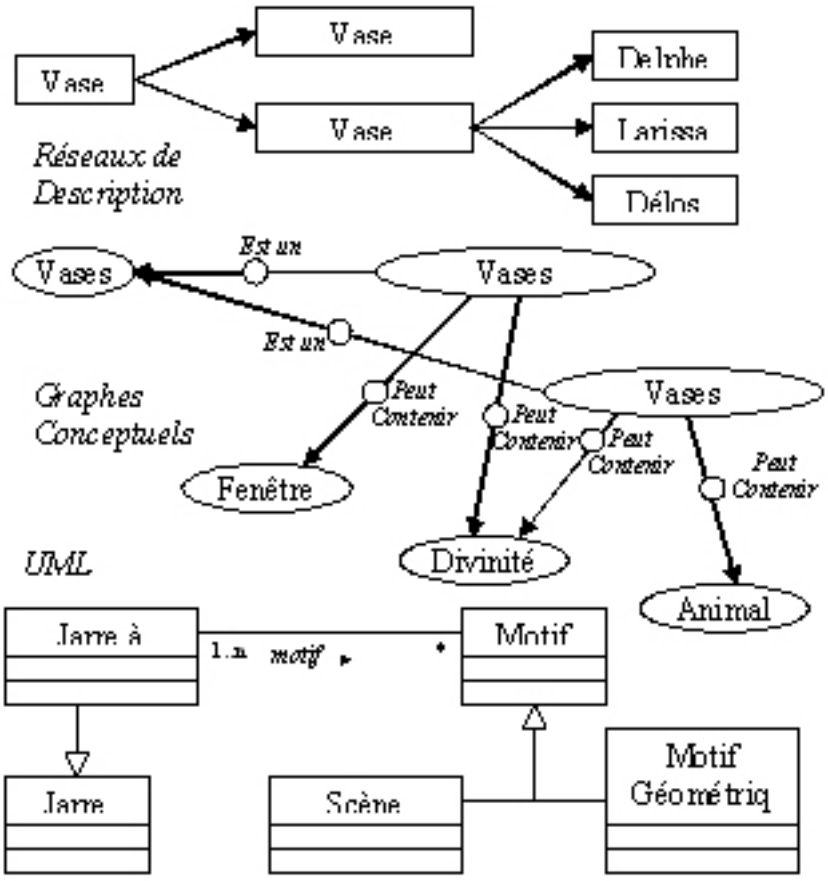

Fig. 5 - Trois opinions dans trois langages

Soient trois opinions distinctes sur un corpus, exprimées par les trois modèles de la figure 5, dans trois langages différents (les réseaux de description, les graphes conceptuels et le formalisme UML). Les experts à l'origine de ces modèles, s'ils connaissent les langages, sont capables d'effectuer la confrontation. Cependant, on peut délimiter un certain nombre de tâches dans cette confrontation qui auraient un grand intérêt à être

automatisées. Par exemple, le " vase " du premier, les " vases " du second et la " jarre " du troisième font probablement référence au même concept, qui est par exemple un récipient étanche et décoré en terre cuite. Peut-être faudrait-il remplacer les différents termes par une unique appellation. Ou alors, on peut vouloir évaluer la différence entre ces termes en s'appuyant sur la structure du graphe.

5.2. Modèle de confrontation

Le processus de confrontation d'opinions est constitué d'une succession d'actions. Décrire l'ensemble des confrontations nécessite donc de définir un langage d'expression de ces actions et des notions sur lesquelles ces actions agissent, donc un langage de confrontation. On peut alors définir un modèle de confrontation pour la confrontation d'un ensemble de modèles.

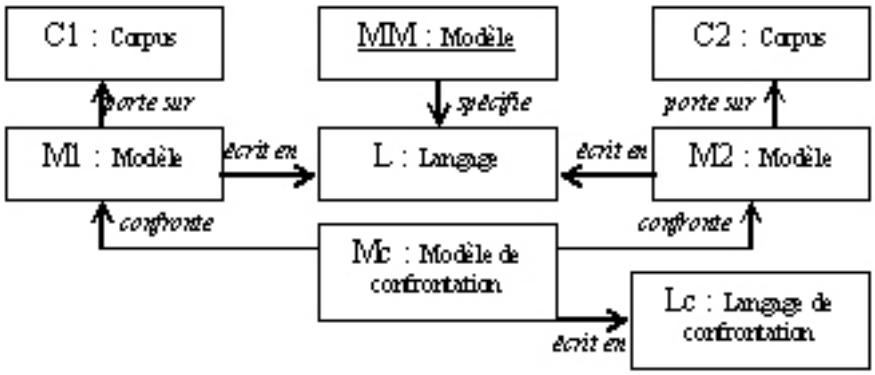

Fig. 6 - Modèle de la confrontation

Le langage de confrontation doit soit être indépendant des langages d'expression des modèles, soit être adapté à tous les langages utilisés.
Pour confronter deux modèles, nous avons donc besoin, comme c'est indiqué sur la figure 6:

- De deux modèles M1 et M2 à confronter, écrits dans le même langage L, spécifié par le modèle MM, et qui portent sur les corpus C1 et C2.

- D'un langage de confrontation Lc permettant la confrontation de modèles écrits en L ;

-D'un modèle de confrontation Mc, écrit en Lc, spécifiant toutes les actions de confrontation. C'est le déroulement de ce modèle qui donne le résultat de la confrontation.

Pour exprimer ce résultat, il est nécessaire d'avoir des critères de confrontation, ainsi que des métriques associées à ces critères. Voici quelques critères possibles :

- Complexité : le nombre de concepts et de relations entre eux ;
- Connectivité : indiquant dans quelle mesure ils sont liés (ce qui va de " rien en commun " à " inclusion ") ;
- Cohérence : la cohérence intrinsèque à chacun des modèles, ainsi que leur cohérence mutuelle ;
- Proximité culturelle : la proximité entre les auteurs, et entre leurs cultures respectives ;
- Proximité temporelle : savoir si les opinions sont contemporaines ;
- Proximité thématique : la proximité entre les corpus respectifs.

Les critères et leurs métriques doivent être définis dans le modèle du langage commun aux modèles, ou directement dans le langage de confrontation.

Comme nous l'avons dit, la confrontation est une analyse de théorie, elle se situe donc un niveau plus haut que la construction de ces théories. De fait, le travail est similaire entre l'analyse de faits réels, celle de faits réels et de représentations, et la confrontation qui est l'analyse de faits réels, de représentation et d'opinions, qui sont des représentations de représentations.
On peut rajouter, enfin, que le résultat d'une confrontation est une théorie, donc une opinion sujette à confrontation.

6. CONCLUSION ET PERSPECTIVES

Cet article présente les premières étapes de notre projet de recherche. Nous avons d'ores et déjà défini notre modèle de l'opinion, du domaine, du langage et de la confrontation d'opinions.

Nos travaux actuels portent sur l'élaboration du langage de confrontation, dans le cadre d'une confrontation multi-langages. En effet, comme le projet Porphyry n'est pas le seul projet à destination des chercheurs en Sciences Humaines, il est préférable de saisir au plus tôt les opportunités de développement d'une certaine interopérabilité entre les différentes applications. Le mécanisme d'import/export de données en est généralement le premier stade.

BIBLIOGRAPHIE

[ACH04] Projet Achemenet (2004), http://www.achemenet.com/

[ARK06] Projet Arkeotek (2002-2006),
http://www.epistemes.net/arkeotek/index.htm

[BEN06] Iacovella A., Bénel A., Calabretto S. : Porphyry & Steatite: Software layers for sense makers in humanities, In Workshop on Indexing and Knowledge in Human Sciences (IKHS 2006), SdC2006, Nantes, France, juin 2006.

[BEN04] Bénel A. : Expression du point de vue des lecteurs dans les bibliothèques numériques spécialisées, In : Actes du Colloque International sur le Document Numérique, "Approches sémantiques sur le document numérique", La Rochelle, 22-25 juin 2004.

[BEN01] Bénel A., Eyged-Zsigmond E., Prié Y., Calabretto S., Mille A., Iacovella A., Pinon J.M. : Truth in the Digital Library: From Ontological to Hermeneutical Systems. ECDL'2001. Lecture Notes in Computer Science, Vol. 2163, Springer Verlag. pp.366-377.

[BRU76] Ph. Bruneau : Quatre propos sur l'archéologie nouvelle [en ligne], In Bulletin de Correspondance Hellénique, n°100, Athènes : Ecole française d'Athènes, 1976. p.103-130.

[CAH06] Zaher L'H., Cahier J.P., Zacklad M. : The Agoræ Hypertopic approach, In Workshop on Indexing and Knowledge in Human Sciences (IKHS 2006), SdC2006, Nantes, France, juin 2006.

[CAH04] Cahier J.P., Zacklad M., Monceaux A. : Une application du Web socio-sémantique à la définition d'un annuaire métier en ingénierie, In Proc. of the Conference Ingénierie des Connaissances IC 2004, Lyon, Mai 2004.

[DEL06] The DELOS Network of Excellence on Digital Libraries (2004-2006), http://www.delos.info

[DIE98] Dieng R., Hug S. : MULTIKAT, a Tool for Comparing Knowledge of Multiple Experts. 6th International Conference on Conceptual Structures, ICCS'98, Montpellier, France, August 1998. Lecture Notes in Computer Science, volume 1453, p 139.

[DIE94] Dieng R., Labidi S., Lapalut S., Martin P. : Comparaison de graphes conceptuels dans le cadre de l'acquisition des connaissances à partir de multiples experts. In Actes des Journées ``Graphes Conceptuels'', LIRMM, Montpellier, Mars 1994. GC'94.

[GAM94] Gammoudi M.M., Labidi S. : An Automatic generation of Consensual Rules between Experts using Rectangular Decomposition of a Binary Relation. Proc. of the XI Brazilian Symposium on Artificial Intelligence (SBIA'94), pages 441-455, Fortaleza, Brazil, 17-20 Octobre 1994. SBIA'94.

[GON04] Gonçalves, M., Fox, E., Watson, L., and Kipp, N. : Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. ACM Transactions on Information Systems, vol. 22(2), pp. 270-312, April, 2004.

[LAH05] Lahna B., Roudies O., Giraudin J.P. : Une approche multivue pour la conception de systèmes d'information à composants. LSR-IMAG,

Grenoble et SIR, Rabat, In actes du XVIIIème congrès INFORSID, 2005.

[NAN01] Nanard M., Nanard J. : Cumulating and Sharing End Users Knowledge to Improve Video Indexing in a Video Digital Library. Proc. JCDL'2002 (Joint ACM/IEEE Conf. On Digital Libraries), ACM Press, 2001.

[NAS03] Nassar M. : VUML : a Viewpoint oriented UML Extension. ASE, p. 373, 18th IEEE International Conference on Automated Software Engineering (ASE'03), 2003.

[ORI00] "HyperNietzche", Paolo DiOrio, Presses Universitaires de France, Collection Ecritures Electroniques, 2000.

[POR06] Projet Porphyry (2001-2006), http://www.porphyry.org

[RIB02] Ribière M., Dieng R. : A Viewpoint Model for Cooperative Building of an Ontology. Proceedings of the 10th International Conference in Conceptual Structures (ICCS'2002), Springer-Verlag, LNCS 2393, editeur : U. Priss, D. Corbett, G. Angelova. p. 220-234, Borovetz, Bulgarie, 15-19 juillet, 2002.

[RIB99] Ribière M. : Representation et gestion de multiples points de vue dans le formalisme des graphes conceptuels. Thèse de doctorat : Université de Nice-Sophia Antipolis, 1999.

[RIB97] Ribière M., Dieng R. : Introduction of Viewpoints in Conceptual Graph Formalism. In Proceedings of the International Conference on Conceptual Structures ICCS'97, Aout 97, University of Washington, Seattle, USA.

[SHE05] Shen R., Gonçalves M.A., Fan W., Fox E. : Requirements Gathering and Modeling of Domain-Specific Digital Libraries with the 5S Framework: An Archaeological Case Study with ETANA. Research and Advanced Technology for Digital Libraries: 9th European Conference, ECDL 2005, pp 1-12.

[SOW00] Sowa J.F. : Knowledge Representation : Logical, Philosophical and Computational Fondations. Brooks/Cole, 2000, ISBN 0-534-94965-7.

[UML06] Unified Modeling Language, http://www.uml.org/

BIBLIOMÉTRIE ET LINGUISTIQUE :
Évaluation de la production scientifique et annotation sémantique

Marc BERTIN
Jean-Pierre DESCLES
Brahim DJIOUA
Yordan KRUSHKOV

Laboratoire LaLICC
(Langage, Logique, Informatique, Cognition et Communication),
28 rue Serpente, 75006 Paris, France
Université Paris-Sorbonne/CNRS UMR8139
marc.bertin@paris4.sorbonne.fr
jpdescles@paris4.sorbonne.fr
bdjioua@paris4.sorbonne.fr

RÉSUMÉ

L'identification et l'évaluation de la production scientifique sont un problème d'actualité. Nous nous sommes donc intéressés à étudier comment les auteurs sont cités à travers les publications scientifiques. Cette approche linguistique permet de catégoriser les renvois bibliographiques d'une publication. L'application informatique de cette étude s'appuiera sur la plateforme informatique EXCOM (EXploration COntexuel Multilingue).

ABSTARCT

The identification of the scientific production and the evaluation of the researchers is a problem of current events. It is the reason for which we suggest being interested in the way are quoted the authors in articles. This approach allows to categorize quotation and annotate automatically a text. The computer application of this study will be integrated into the platform EXCOM (Multilingue Contexual EXploration).

MOTS-CLÉS : Evaluation, EXCOM, exploration contextuelle, linguistique, bibliographie

KEYWORDS : Evaluation, EXCOM, contextual exploration, linguistic, bibliography

INTRODUCTION :

Identifier la production scientifique, tout comme l'évaluation de la science en générale, sont des exercices périeux. Sans rentrer dans un débat, je soulignerai une prise de conscience plus vive de cette problématique avec la classification de Shangaï. Aussi, si l'approche de Garfield est loin d'offrir à travers le Facteur d'Impact une solution très pertinente, elle n'en reste pas moins la solution la plus présente, et cela malgré les biais introduits et souvent discutés dans la littérature. Nous allons donc essayer de comprendre le pourquoi de cette situation et proposer une approche nouvelle sans nous appuyer sur l'approche statistique habituelle, c'est-à-dire celle de Zipf qui utilise les fréquences d'apparitions de termes dans le texte, mais en préconisant l'approche linguistique. En effet, les méthodes statistiques usuelles comme Okapi et autres "tf.idf" sont largement utilisées dans les systèmes de recherche d'informations et d'évaluation. Nous proposerons donc une classification des relations entre auteurs à base de critères qualitatifs basée sur une étude linguistique textuelle.

ORIGINE DE L'APPROCHE STATISTIQUE

Nous prendrons comme point de départ la bibliologie telle qu'elle est définie par Otlet. Elle décrit la science systématique et raisonnée du livre qui a pour objet l'histoire du livre et ses procédés de fabrication, de diffusion et de conservation. L'application de l'outil statistique à l'univers de la bibliologie conduit naturellement à la bibliosociométrie qui est la mesure de l'activité du livre et du document sur l'homme et la société, et à la bibliométrie, qui est l'ensemble des méthodes et techniques quantitatives, de type mathématique et statistique, susceptible d'aider à la gestion des bibliothèques et d'une manière très générale des divers organismes ayant à traiter de l'information. Il s'agit là de la définition proposée par A. Pritchard. En fait, il est clair que le livre dépend de par sa matière de l'économie et de par son texte, il relève de la linguistique et de la textologie.

SOCIÉTÉ DE L'INFORMATION

À la source des sciences de l'information, nous pouvons citer trois lois qui sont celles de Bradford, Lotka et Zipf. Nous connaissons aussi cette dernière sous le nom de Zif-Paretto, soulignant ainsi le lien avec le monde de l'économie. La loi de Lotka est relative à la production d'articles par les chercheurs. Celle de Zipf exprime la fréquence d'apparition des mots dans un texte. Quant à Bradford, elle exprime la répartition des articles dans les revues. C'est à partir de la loi de Lotka que Price proposera l'idée que le nombre d'auteurs les plus productifs est donné par la racine carrée du nombre total d'auteurs. Malgré les travaux de Glänzel et Schubert, il fut difficile de proposer une formulation mathématique en concordance avec les données empiriques. Cependant, la grande idée de Price a été de définir une propriété essentielle du champ

scientifique en proposant la Distribution sur les Avantages Cumulés. Cette théorie se propose comme générale et unificatrice des lois empiriques de la bibliométrie. L'idée sous-jacente est qu'une minorité de scientifiques se trouve être à l'origine de la majorité des publications dans un domaine. Nous citerons Price [PRI63] afin de mettre en évidence sa formation de physicien et donc son approche :

" On étudie le comportement d'un gaz à différentes conditions de température et de pression. On ne s'intéresse pas à une molécule appelée Georges, se déplaçant à une vitesse spécifique et située en un endroit spécifique à un instant donné; on considère seulement la moyenne de l'ensemble total des molécules où certaines sont plus rapides que d'autres, où elles sont situées au hasard et se déplaçant en différentes directions."

THÉORIE DE LA CITATION

Au delà de l'aspect purement quantitative de cette thermodynamique des auteurs, le questionnement sur les motivations des auteurs à citer et la signification de la citation a permis de mettre en évidence deux écoles de pensée. L'étude des citations nécessite de comprendre les normes utilisées, les différentes fonctions de la citation, de leurs qualités ainsi que les motivations et les raisons pour citer des travaux. On peut citer [CRO84], [KIN87], [LIU93] et [LEY98]. La fonction communicative de la citation se résume à deux courants. La première approche peut se définir par une citation de Wilson (1999):

" Document is cited in another document because it provides information relevant to the performance and presentation of the research, such as positioning the research problem in a broader context, describing the methods used, or providing supporting data and arguments. "

L'auteur qui cite est conditionné par les normes de la science en générale, et plus particulièrement par les normes de son domaine de recherche. Cela rejoint les points de vue de Garfield, Price et Cole [PRI63], [PRI65] et [COL92]. Il est admis dans cette théorie que les citations sont égales entre elles et qu'elles sont suffisantes à l'argumentation de l'auteur. Une des idées fortes de Price est qu'il a été le premier à souligner la possibilité de mettre en relation les auteurs afin de pouvoir cartographier la science ou au moins un domaine de celle-ci à travers une analyse des co-citations.

S'opposant à l'approche à la théorie normative, l'approche sociale constructiviste prône que les citations sont des instruments rhétoriques afin de persuader les lecteurs selon des critères autres que scientifiques. On peut renvoyer pour cela aux écrits de [LAT87]. Un travail de synthèse, en marge de cette problématique, détail plus en profondeur cette réflexion [SCH04] dans son introduction.

LES AUTRES TRAVAUX

Cependant, différentes études ont été menées à des fins plus applicatives comme le résumé automatique. Il n'est plus à démontrer l'importance des citations qui sont très présentes dans les systèmes informatiques comme le SCI de l'ISI ou Citeseer. Garfield a proposé le premier en 1955 un système d'indexation des citations. Les hyperliens offrent la possibilité de naviguer d'un article à l'autre avec facilité et permettent ainsi d'identifier plus facilement les travaux connexes au domaine de l'article. L'étude des citations n'est donc pas nouvelle et différentes catégorisations ont vu le jour. Celles-ci remontent aux travaux de Garfield [GAR65] ou de Lipetz [LIP65]. Nous pourrons donner comme exemple les catégories suivantes : conceptual or operational, organic or perfunctory, evolutionary or juxtapositional, and confirmatory vs. negational tel que proposées par [MOR75, MUR75]. White proposera une classification basée sur des études sociologiques et déterminera les raisons pour lesquelles certains travaux sont mis en avant ou plus cités que d'autres [WHI04].

Bradshaw construira une nouvelle métrique RDI (pour Reference Directed Indexing) [BRA03]. De façon succincte, la pondération variera en fonction des termes présents dans la " citance " qui est un néologisme anglophone indiquant le segment textuel entourant la citation. Cette méthode recherche dans l'article cité les termes trouvés dans le segment textuel. Cela permet de pondérer plus fortement les articles cités ayant les mêmes termes.

Nanba identifie les citations a des fins de résumé. Son hypothèse de travail est qu'une citation représente un résumé succinct selon le point de vue de l'auteur. [NAN00 et al.] repère des segments textuels (citations area) pour identifier les citations afin de générer des résumés. Nous soulignerons que dans ces travaux, il propose également une catégorisation des citations (citations types) afin d'organiser les " aires de citations ".

Plus proche de nos préoccupations, nous pouvons citer les travaux de Simone Teufel qui propose également une classification [TEU00, MOE00], [TEU01]. Sa catégorisation, en 13 points, couvre assez largement les citations et leurs motivations. De même, Bonzi s'est intéressée à la citation négative expliquant qu'une citation n'était pas forcement un signe d'acceptation de la part de celui qui cite [BON82]. Enfin, d'un point de vue de l'automatisation, les travaux de Mercer [MER04, MAR04] proposent d'utiliser ces classifications qui sont ignorées par les systèmes d'indexation de citation.

L'ensemble de ces travaux sur les catégorisations n'offrent pas de solutions de cartographies à grande échelle. De plus elles sont limités dans le cadre d'une application ou d'une méthode. Elles ne s'appuient pas sur une étude linguistique qui permettrait de traverser les domaines en s'intéressant non pas aux termes clés mais aux relations entre les termes. C'est l'approche linguistique que nous prônons ici.

LIMITATION DE L'APPROCHE QUANTITATIVE

Mesurer la qualité de la production est relativement difficile dans le sens où les indicateurs bibliométriques caractérisent le contenant et non le contenu. Elle apporte une valeur et des mesures, mais ils ne sont pas et ne doivent pas être des signes de la qualité de la recherche scientifique. On constatera ces dernières années une attitude du " publish or perish " conduisant à des pratiques d'écriture qui peuvent mettre en péril la qualité des articles. Cela peut provoquer des comportements antiscientifiques comme le plagiat, la publication dans une revue où le FI est élevé plutôt que dans une revue adéquate ou bien encore de diviser les données en parties ridiculement petites. L'un des risques encouru par cet état des faits est sans doute à court terme une production scientifique accrue, mais d'une qualité moindre, nécessitant de parcourir un certain nombre de publications pour couvrir une pensée ou un concept. À moyen terme un risque d'uniformisation de la recherche est présent. Cette homogénéité a déjà été soulignée et plusieurs articles présentent les biais introduits par cette méthode d'évaluation. Au-delà de l'aspect rédactionnel, il existe un nombre de limites intrinsèques à cette approche impliquant l'acceptation des biais ainsi introduits. Pour exemple, seul le premier auteur est pris en compte, de plus il faut considérer les problèmes d'homonymie ou de fautes de frappe présentes dans les bases de données. Les domaines sont inégalement représentés et les indicateurs bibliométriques s'appliquent très difficilement pour les sciences humaines et sociales. Toutes les revues ne sont pas recensées et pour celles qui le sont, il peut y avoir sur- ou sous-estimation de la revue et donc des travaux et des équipes. On notera que l'autocitation ou la citation d'un article controversé n'est pas abordée par l'approche statistique. De plus, les ouvrages ne sont pas pris en compte. Nous pouvons aussi constater que deux ans ne suffisent pas pour qu'un article se révèle or il s'agit de la durée retenue pour le calcul du facteur d'impact. Enfin, la citation négative n'est pas prise en compte. Pour le moment, il n'y a guère de solutions innovantes, seulement de nouvelles approches statistiques permettant de minimiser les biais introduits.

MÉTHODOLOGIE

Face à ce constat, il serait intéressant pour la communauté scientifique de disposer d'un outil plus qualitatif pour la conception de réseaux d'auteurs. Les outils de cartographie actuels s'appuient sur une approche quantitative et matricielle. Une nouvelle approche de cette problématique doit être envisagée. Sans prétendre fournir un traitement sémantique complet d'un article scientifique, nous pourrons dans un premier temps considérer les relations sémantiques entre l'auteur, les co-auteurs et les références bibliographiques. Il serait tout à fait pertinent de savoir si un article est cité de façon positive ou négative. Une référence bibliographique citée en contre-exemple est tout à fait révélatrice des relations entre les travaux des chercheurs. Il peut s'agir entre autres d'une référence par rapport à une définition, une hypothèse ou bien une méthode, mais également d'un point

de vue, d'une comparaison ou bien d'une appréciation. Suite à l'identification des appels bibliographiques, nous proposerons une annotation de ceux-ci avec une catégorie afin de définir comment l'auteur a été cité. Cette catégorisation est définie par l'étude d'indices que nous relèverons dans la phrase. Nous rechercherons les indices positifs/négatifs de citation d'un auteur, ainsi que les citations hypothèses/méthodes utilisées par un auteur. On caractérisera alors ce point de vue comme étant une catégorisation sémantique des références de citation d'auteur. Le renvoi bibliographique qui se trouve dans le texte permet de définir un segment textuel où se trouvera l'information de catégorisation de ce renvoi. L'implémentation informatique de cette approche utilise la plateforme EXCOM (Exploration Contextuelle Multilingue) développée au sein du laboratoire LaLICC. Nous pourrons nous référer à l'article de [DJI06] décrivant plus en détail la plateforme.

BIBLIOGRAPHIE ET RENVOIS BIBLIOGRAPHIQUES

Selon Malcles, bibliographe, il est vrai que l'étude bibliométrique passe par une analyse de la bibliographie, mais Palanco souligna que l'application des statistiques à la bibliographie est réducteur, évoquant l'idée de réductionnisme bibliométrique puisque cette approche élimine la diversité des thèmes au profit de l'unité des matières. Nous prendrons comme postulat de départ que la bibliographie est effectivement une donnée essentielle pour l'évaluation des publications. L'appel de citation dans un texte peut prendre différentes formes. Il peut s'agir principalement d'un renvoi numérique ou d'un renvoi par nom d'auteur. Pour cela, nous dresserons une classification des différentes familles numériques et alphanumériques des références bibliographiques.

Pour ce travail, nous avons utilisé les normes, mais également les " coutumes ". En effet, les renvois bibliographiques dans le texte sont plus ou moins normalisés selon les normes ISO 690-1 (Z 44-005) et ISO 690-2, mais il était nécessaire de prendre en compte des pratiques dépassant le simple renvoi numérique ou alphanumérique afin de pouvoir traiter exhaustivement l'ensemble des renvois bibliographiques. Afin de traiter automatiquement cette tâche d'identification et d'extraction, nous pourrons par exemple définir un alphabet adéquat permettant d'appliquer au corpus un automate fini déterministe. Pour identifier les renvois bibliographiques se trouvant présents dans le texte, nous nous appuierons sur les travaux déjà effectués [BER06], qui proposent un automate à états finis afin de localiser les renvois bibliographiques. Cependant, au lieu de considérer l'aspect numérique d'une référence bibliographique, nous utilisons les renvois dans le texte afin de catégoriser les relations entre auteurs. Nous avons émis l'hypothèse que la pensée de l'auteur par rapport aux travaux de ses confrères se trouve à proximité de la référence bibliographique. Aussi considérons-nous dans cette première approche que la prise de position d'un auteur vis-à-vis de ces confrères se trouve dans un espace proche d'un renvoi bibliographique.

15. Indicateurs et indices

Nous nous proposons donc d'utiliser les renvois bibliographiques identifiés par l'automate à états finis d'un article afin de déterminer segments textuels et déterminer. Les renvois bibliographiques seront alors considérés comme étant nos indicateurs. Les indices linguistiques, quant à eux, permettent de déterminer une information sémantique spécifique. Ils permettent de réduire l'indétermination et de spécifier la qualité du renvoi. Il s'agit du seul savoir dont nous avons besoin pour déterminer nos catégories et se trouvent présents autour de l'indicateur, dans le même segment textuel que celui-ci. La méthode de l'Exploration Contextuelle, développée par Mr Desclés [DES91, DES96], va permettre à l'aide des indices, de lever les indéterminations sémantiques de l'unité linguistique analysée et proposer une catégorisation qualitative des références bibliographiques.

16. Segments textuels et localisation

L'indicateur permet de déterminer le segment textuel nécessaire et suffisant à l'accomplissement de notre tâche. Dans cette étude, nous ferons coïncider ce segment textuel avec la phrase. Nous nous gardons la possibilité d'étendre nos recherches à des zones plus larges, comme la théorie nous le permet, si cela s'avérait nécessaire à lever certaines ambiguïtés. Une fois l'espace de recherche déterminé, il faut prendre en compte la localisation de l'indice par rapport à l'indicateur. Nous avons identifié cinq localisations possibles par rapport à l'indicateur :

" premier mot du segment textuel | avant le milieu | au milieu | après le milieu | à la fin du segment textuel ". D'un point de vue pratique, seul le contexte droit|gauche est implémenté et se révèle pour le moment suffisant dans le cadre de ce travail.

CATÉGORISATION

Les différentes catégories ont été identifiées par Krushkov Yordan [KRU05] dans son travail de mémoire de maîtrise sous la direction de Mr Desclés. Elles se trouvent à la base de ce travail, aussi allons nous détailler les différentes catégories sur lesquelles nous nous appuyons.
Le point de vue est la première catégorie que nous avons identifiée. Il est très présent dans les corpus étudiés. Les indices linguistiques suivants font partie de cette catégorie :

" Selon | d'après | pour | considérer que | nous y voyons |comme le dit |... ". Ils sont généralement localisés en amont de l'indicateur.
La seconde catégorie à laquelle nous nous sommes intéressés est la comparaison. En effet, nous comparons souvent le travail de nos confrères.

Dans ce cas précisément, nous pouvons trouver des similarités ou bien des dissimilarités :

" ressembler |comme dans les travaux de | le rapport avec |... ".

Pour la non-ressemblance, nous avons comme indices linguistiques :
" différer de | contraire l'approche de |contrairement ce qu'affirme |... "

La catégorie de l'information est vaste. Pour cela, elle est divisée en sous-catégories comme l'hypothèse, l'analyse et le résultat. Pour la sous-catégorie de l'analyse, nous pouvons donner comme exemple :

" a été analysé dans | l'analyse de | lors de son analyse | ... ".
Pour concevoir la sous-catégorie des résultats, nous avons considéré les indices linguistiques suivants :

" nous avons démontré | donner de nombreux exemples de | a publié ses résultats | a dégagé |... "

La catégorie de la définition est également importante avec pour indices :
" ils caractérisent | la notion ... introduite dans |... "

La catégorie de l'appréciation met en valeur le jugement d'un auteur sur un autre auteur ou plutôt sur un ou plusieurs travaux de celui-ci. Il peut s'agir d'un jugement positif ou négatif :
" ont rejeté | n'as pas répondu | en trahissant sérieusement notre proposition | ... ".
Cette catégorie est très importante dans le sens où elle apporte une réponse à l'un des biais introduits par l'approche statistique.

Quotation	Point de vue *Soi-même|Autrui*	Pris de position
	Comparaison *Soi-même|Autrui*	Similitude
		Dissimilitude
	Information *Soi-même|Autrui*	Hypothèse
		Analyse
		Résultat
		Méthode
		Citation
		Contre-exemple
	Definition *Soi-même|Autrui*	
	Appréciation *Autrui*	Accord
		Désaccord

Figure 1 : Catégorisation des renvois bibliographiques

Constitution du corpus

Pour cette étude, nous avons constitué un corpus d'articles issus du laboratoire LaLICC afin d'identifier les indices et de constituer notre base de connaissances. Afin de traiter le caractère pluridisciplinaire de notre approche, nous avons augmenté le corpus avec des publications extraites de HAL, la base de données de l'INRIA. Nous avons également choisi des articles de la revue INTELLECTICA. Ce petit corpus de test couvre les domaines de la linguistique, de l'informatique, et des sciences cognitives afin de démontrer la capacité du système à traiter une information multidisciplinaire. À la rédaction de cet article, le corpus est exclusivement constitué de textes en français. La couverture de l'anglais sera une prochaine étape dans le développement de ce système.

Plateforme informatique

L'architecture informatique de la machine à annoter automatiquement EXCOM, qui s'inspire de l'architecture modulaire GATE, est décrite dans la figure suivante. Les textes traités par EXCOM sont d'abord prétraités pour les préparer à une segmentation en phrases, paragraphes et sections en s'appuyant sur les travaux de [MOU99a, MOU99b].

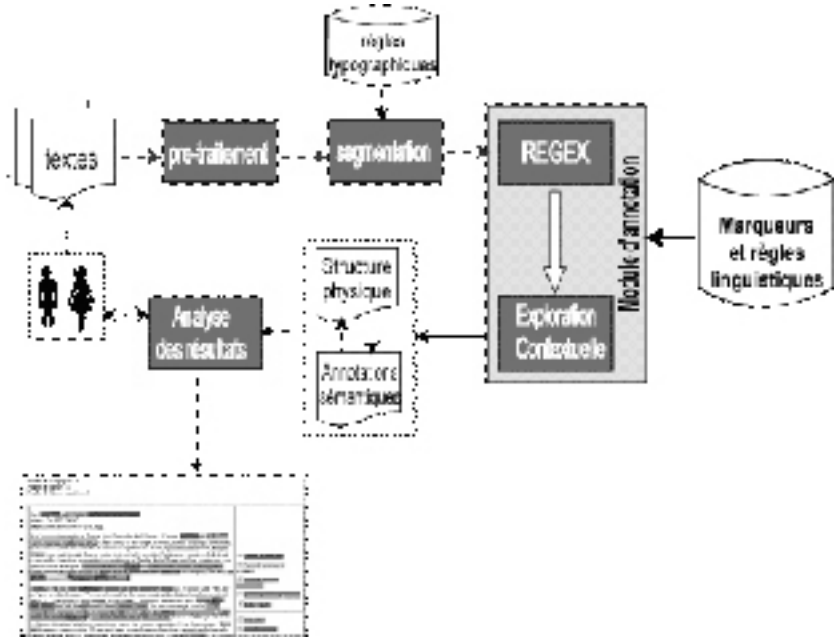

Figure 2 : Architecture informatique de la machine EXCOM

À chaque tâche d'annotation sémantique sont associés un ensemble de marqueurs linguistiques (listes d'indicateurs et d'indices) et un ensemble de règles applicables sont les textes segmentés. Les conditions de déclenchement de ces règles sont exprimées de différentes façons qui déclenchent certains niveaux du moteur d'annotation. Chaque niveau fait appel à un algorithme général de fonctionnement. Ce moteur est construit sous une forme multicouche où chaque brique répond à un besoin d'annotation particulier. Les briques les plus élémentaires sont définies pour répondre à un besoin d'annotation des références bibliographiques (indicateurs du point de vue bibliosémantique) sous forme d'expressions régulières et les plus externes prennent en charge les aspects d'une exploration contextuelle. Les modules les plus externes s'appuient sur les modules les plus internes.

Le module REGEX fait appel à un moteur d'expressions régulières. Les domaines d'utilisation des expressions régulières sont nombreux : elles interviennent dans le cadre de l'analyse de contenu des textes. Avec le support d'Unicode, l'extraction d'information peut se réaliser sur des documents multilingues.

Le module d'exploration contextuelle (EC) est composé :

8. d'un ensemble de marqueurs linguistiques (indicateurs et indices) ;

9. d'un ensemble de règles d'EC qui se présentent sous la forme de règles déclaratives (si certaines conditions sont vérifiées alors certaines actions sont appliquées).

10. d'un moteur d'EC qui applique les règles en respectant la primauté de l'indicateur sur les indices complémentaires.

Le résultat de l'application de ces règles est un texte annoté. Les annotations sont des marques sous forme d'éléments et attributs XML. La sémantique de ces annotations est liée à l'organisation de la catégorie du point de vue reconnue par le système EXCOM. L'objectif de cette plateforme est de proposer une exploration du texte afin de l'augmenter d'informations sémantiques sous forme d'annotations. Si la plupart des travaux menés dans ce domaine s'appuient sur une analyse morpho-syntaxique, la méthode préconisée pour cette plateforme est l'Exploration Contextuelle et utilise une base de connaissances, constituée de marqueurs linguistiques. Elle permet d'étiqueter automatiquement un texte à partir de ressource linguistique.

Déclaration de Règles

L'application informatique nécessite l'écriture de règles. Celles-ci se présentent sous la forme d'un fichier XML. Aussi allons-nous détailler une règle qui permet d'annoter la publication selon le point de vue de la méthode, qui est une sous-catégorie de information.

```
<regle nom_regle="RegIInfMet3" tache="bibliosemantique"
point_de_vue="information" type="EC">
 <conditions>
  <indicateur espace_de_recherche="phrase" type="annotation"
valeur="RenvBiblio" />
  <indice contexte="droite" espace_de_recherche="." type="liste"
valeur="IdAuteur" />
  <indice contexte="droite" espace_de_recherche="." type="liste"
valeur="IdMethode" />
  </conditions>
 <actions>
  <annotation type="ajout_attribut" espace="identique" annotation="methode"
/>
  </actions>
</regle>
```

Cette règle traite donc du point de vue de l'information : point_de_vue="information". L'indicateur a pour valeur : valeur="RenvBiblio" qui permet de retrouver les renvois bibliographiques et identifier l'espace de recherche qui est la phrase : espace_de_recherche="phrase". Les indices, de type liste, sont définis par leur contexte qui peut être droite ou gauche par rapport à l'indicateur, dans l'espace de recherche préalablement identifié. Dans le cas présent, les deux indices se trouvent à droite de l'indicateur. Si l'ensemble des conditions de cette règle est validé, alors EXCOM annote le segment textuel en ajoutant un attribut : <annotation type="ajout_attribut" espace="identique" annotation="methode" />

Résultats

Les résultats sont affichés sous la forme suivante : Le segment textuel est coloré en bleu. L'indicateur est en vert et les indices primaires et secondaires sont respectivement en vert clair et mauve.

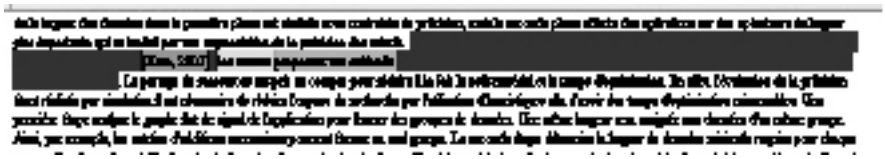

Figure 3 : Exemple du point de vue méthode

Discussion

Le premier point que nous discuterons est celui des renvois bibliographiques. L'étude des segments textuels repose principalement sur l'identification deces renvois. Aussi est-il très important dans cette approche que l'ensemble

des renvois soit reconnu. Si sur cet exemple, aucun problème d'identification n'a mis à défaut cette approche, il faudra cependant tenir compte, sur des corpus plus littéraires, de la notion de courant ou de personnes associées en tant que telles. Par exemple, " Selon Pottier, nous devons concevoir que ... ".
Le deuxième point est la nécessité de continuer le travail linguistique afin de pouvoir couvrir l'ensemble des catégories identifiées et de les implémenter au sein de la plateforme. Seule la volumétrie nous permettra de proposer à cette approche qualitative un protocole dévaluation.
Le troisième point est une remarque quantitative. Sur cet exemple, nous avons constaté que l'auteur se référait plusieurs fois à la même publication d'un de ces confrères selon le point de vue de la méthode. L'identification des renvois bibliographiques peut donc apporter une pondération à l'outil bibliométrique. Cependant, il faut bien souligner que notre approche, va au-delà d'une simple pondération puisqu'à une référence bibliographique, nous faisons correspondre une catégorisation sémantique.

Conclusion

À court terme, cette approche permettra de proposer un outil beaucoup plus fin et complémentaire à l'approche proposée actuellement. D'une part, la prise en compte de la bibliographie comme unité est loin d'être satisfaisante et de nombreux biais sont introduits. Le fait de pouvoir catégoriser la bibliographie par une analyse linguistique donc qualitative et automatisé via la plateforme informatique EXCOM offrira un outil plus pertinent et offrira à moyen terme de nouvelles possibilités d'exploration des textes scientifiques. D'autre part, cette approche permet de catégoriser sémantiquement les renvois bibliographiques. Aussi et contrairement à une approche statistique, pouvons-nous étudier et obtenir des résultats sur un très petit nombre de publications, à l'échelle d'un laboratoire par exemple, tout en conservant la possibilité de travailler à une plus grande échelle. À l'approche statistique de l'évaluation, l'approche linguistique permet de porter un regard qualitatif des relations entre les travaux des différents auteurs.

Perspectives

Cette étude servira de point de départ à une nouvelle façon de cartographier la science ou du moins un domaine. L'utilisation d'un logiciel comme Pajek [BAT01], [BAT02] et [BAT05] va nous permettre d'analyser des réseaux sous forme de graphes en permettant d'annoter les arcs non pas une pondération mais une catégorie sémantique. Nous serons alors à même de déterminer des ensembles de sous-graphes en fonction des catégories précédemment établies.

BIBLIOGRAPHIE

[BAT01] Batagelj, V., Pajek - program for large networks analysis and visualization Presented at Dagstuhl seminar Link Analysis and Visualization Dagstuhl 1-6. July 2001.

[BAT02] Batagelj V., Mrvar A., Zaveršnik M., "Network Analysis of Texts". Jezikovne tehnologije / Language Technologies, T. Erjavec, J. Gros eds., Ljubljana 2002, p. 143-148.

[BAT05] Batagelj, V. Brandes U., "Efficient generation of large random networks", 2005 Physical Review E 71, 036113.

[BER06] Bertin M., Desclés J.P., Djioua B.,Krushkov Y., "Automatic Annotation in Text for Bibliometrics Use", FLAIRS 2006, Floride, 11-13 mai

[BON82] Bonzi, S., "Characteristics of a literature as predictors of relatedness between cited and citing works". Journal of the American Society for Information Science, 1982, 33(4): 208-216

[BRA03] Bradshaw S., "Reference directed indexing: Redeeming relevance for subject search in citation indexes". In Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries, 2003.

[COL92] Cole, S., "Making Science. Between Nature and Society", 1992, Harvard University Press, Cambridge, MA.

[CRO84] Cronin, B., "The Citation Process: The Role and Significance of Citations in Scientific Communication",1984, Taylor Graham, London.

[DES91] Desclés, J. P., "Exploration contextuelle et sémantique: un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte.", 1991, Knowledge modeling and expertise transfert p.371-400.

[DES97] Desclés, J. P., "Système d'exploration contextuelle." Co-texte et calcul du sens p.215-232. 1997.

[DJI06] Brahim, D. , Flores, J.G., Blais, A., Desclés J-P., Gael, G., Jackiewicz, A., Le Priol, F., Leila,N.B., Sauzay B., "EXCOM: an automatic annotation engine for semantic information", FLAIRS 2006, Floride, 11-13 mai,

[GAR65] Garfield, E., "Can citation indexing be automated ?" National Bureau of Standards Miscellaneous.1965. Publication, 269:189-192

[KIN87] King, J., "A review of bibliometric and other science indicators and their role in research evaluation", Journal of Information Science, 1987. Vol. 13 No. 5, pp. 261-76.

[KRU05] Krushkov, Y. (2004-2005), "L'exploration contextuelle des appariements entre les références bibliographiques et les passages textuels dans un corpus de textes linguistiques." Mémoire de Maîtrise, Université Paris IV Sorbonne sous la dir. de Mr J.P Desclés.

[LAT87] Latour, B., Science in Action, Open University, Milton Keynes. 1987.

[LEY98] Leydesdorff, L., "Theories of citation ?", Scientometrics, 1988, Vol. 43 No. 1, pp. 5-25.

[LIP65] Lipetz, B. A., Improvements of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. American. Documentation, 16:81-90, 1965.

[LIU93] Liu, M., "Progress in documentation: the complexities of citation practice - a review of citation studies", Journal of Documentation, 1993. Vol. 49 No. 4, pp. 370-408.

[MER04, MAR04] Mercer, R. E. and Marco, C. D., "A design methodology for a biomedical literature indexing tool using the rhetoric of science". In BioLink workshop in conjunction with NAACL/HLT, pages 77-84. 2004.

[MOR75, MUR75] M. J. Moravcsik and P. Murugesan., "Some results on the function and quality of citations." Social Studies of Science, 5:86-92.1975

[MOU99a] Mourad Ghassan, "Rôle de la typographie dans la segmentation de textes", JILA'99 (Journées Internationales de Linguistique Appliquée), p.203-206.

[MOU99b] Mourad Ghassan, "La segmentation de textes par l'étude de la ponctuation", CIDE'99 (2e Colloque International sur le Document Électronique), p.155-171.

[NAN00 et al.] Nanba, H. Kando, N. and Okumura, M., "Classification of research papers using citation links and citation types: Towards automatic review article generation". In American Society for Information Science SIG Classification Research Workshop: Classification for User Support and Learning, pages 117-134, 2000.

[PRI63] Price, D., "Little Science, Big Science", p.IV-V. 1963.

[PRI65] Price, D. De Solla, "Networks of scientific papers", Science, 1965 Vol. 149, pp. 510-5.

[PRI86] Price, D. De Solla, "Little Science, Big Science . . . And Beyond", Columbia University Press, New York, NY.1986

[SCH04] Schneider J.W., "Introduction to biblimetrics for construction and maintenance of thesauri". Journal of documentation. 2004. Vol.60 N°5 p.524-549

[TEU00, MOE00] TEUFEL,S. MOENS, M., "What's yours and what's mine: Determining Intellectual Attribution in Scientific Text" In: Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong, Oct 2000.

[TEU01] TEUFEL S., "Task-Based Evaluation of Summary Quality: Describing Relationships Between Scientific Papers". Workshop Automatic Summarization', NAACL-2001.

[WHI04] White H. D., "Citation analysis and discourse analysis revisited". Applied Linguistics, 25(1):89-116. 2004.

[WIL99] Wilson, C.S., "Informetrics", in Williams, M.E. (Ed.), Annual Review of Information Science and Technology, Vol. 34, pp. 107-247.1999