Sur la piste de l'indexation conceptuelle de documents

Une approche par l'annotation

Yannick Prié

LISI - UFR Informatique
Université Claude Bernard Lyon 1
F-69622 Villeurbanne Cedex
yprie@lisi.univ-lyon1.fr

RÉSUMÉ. *Dans une première partie, nous justifions l'engouement actuel pour l'indexation conceptuelle de documents que nous définissons comme ensemble de connaissances symboliques en permettant l'exploitation, ce que nous illustrons en présentant des modèles récents de descriptions de documents fondées sur les connaissances. Nous proposons également de considérer toute structure documentaire comme structure sémantique ayant forme de graphe. Dans une deuxième partie, nous proposons de voir toute mise en place d'indexation conceptuelle comme se faisant lors d'une annotation, laquelle correspond à une documentation de sa propre pratique d'exploitation du document. Nous présentons quelques exemples récents liés à la notion d'annotation, et discutons les perspectives ouvertes tant par le concept d'indexation conceptuelle que celui d'annotation.*

ABSTRACT. *In the first part of this article, we justify current interest for conceptual indexing of documents, which we define as symbolic knowledge allowing their exploitation. We illustrate this notion with recent knowledge-based document description models. We also propose to consider every documentary structure as a semantic structure in a general graph model. In the second part, we propose that every setting of conceptual indexing be seen as an annotation task corresponding to one's documentation of his/her document exploitation practice. We present recent examples related to the annotation notion, and we discuss the open perspectives both for conceptual indexing and annotation concepts.*

MOTS-CLÉS : *annotation, connaissances, exploitation et inférences documentaires, indexation conceptuelle, structures, transmission de connaissances.*

KEY WORDS: *annotation, knowledge, documentary exploitation and inferences, indexing, structures, knowledge transmission.*

1. Introduction

Tandis que la Recherche d'Information « classique » à base de mots clés atteint une certaine maturation, ses limitations intrinsèques s'en trouvent soulignées. Du fait d'une part de son application *automatique* aux textes seuls, alors que de nombreux autres médias sont disponibles numériquement, d'autre part de la polysémie des unités langagières que sont les mots, ainsi que de la surcharge cognitive afférente dans la gestion des résultats de requêtes. Le besoin se fait alors sentir d'un passage à un niveau supérieur d'indexation, d'une véritable indexation sémantique des documents de tous types. Il s'agit de développer des outils et des méthodes permettant la mise en place d'une sémantique explicite des contenus documentaires autorisant en premier lieu leur recherche, mais également – de façon plus générale – leur exploitation, afin de faciliter certaines applications, tels que les agents intelligents de recherche, de construction automatique de résumés, de catalogues, etc. Si les thesaurii constituent par exemple une première façon de normaliser le sens des termes d'indexation, l'enjeu est désormais d'exprimer de véritables *connaissances* documentaires utilisables comme telles. La notion de « web sémantique » (Semantic Web) nous semble à ce titre suffisamment générique pour prendre en compte de façon intégrée l'ensemble des enjeux liés à l'exploitation des documents numériques fondée sur les connaissances documentaires (pages web, documents de formation configurables, documents médicaux ou de conception, catalogues, etc.).

La nouveauté de cette prise de conscience tranche avec la relative ancienneté des travaux sur la sémantique documentaire et sa représentation. Il apparaît qu'un certain nombre de facteurs expliquent ce contraste.

Tout d'abord, on assiste à un double mouvement d'extension des notions de document et de système documentaire. Ainsi, au document simplement numérisé (instrumentation numérique) s'ajoutent des métadonnées (auteur, date de création), et de la valeur ajoutée (e.g. des liens vers d'autres documents en rapport), qui peut devenir contenu supplémentaire faisant de droit partie du document (un film sur support DVD comprend désormais, de façon établie, des chapitres). Le document est considéré dans son *genre*, résultant d'un contexte de production donné, et apparaît comme connecté à d'autres documents, et il s'agit de reconnaître et de décrire ces caractéristiques. Dans la thématique prometteuse des rapports entre documents et connaissances, les notions de systèmes documentaires et de bases de connaissances ont tendance à se conjuguer. Ainsi, les documents et les corpus sont désormais reconnus comme sources principales de connaissances, qu'il y a lieu d'organiser en systèmes utiles. Le Web comme système documentaire global[1] peut même être considéré comme une base de connaissances. Il convient alors de partager des vues sur ces systèmes, des façons de les utiliser : mise en place collaborative de réseaux documentaires dans les entreprises et nécessité de discipline d'utilisation de modèles prescriptifs de description des documents ; possibilité de partager des bases de

[1] Somme toute guère plus anarchique que le réseau virtuel de tous les documents « matériels » du monde extérieur.

données sur le web, c'est à dire d'assurer une interopérabilité des modèles ; partage de connaissances entre agents, partage de traces d'exploitation des réseaux, etc.

Ensuite, la prise de conscience « massive » de l'enjeu des représentations sémantiques de document est facilitée voire engendrée, techniquement par l'arrivée du langage XML[2], dont l'utilisation prochaine généralisée sur le Web ainsi que dans un grand nombre de systèmes d'information[3] permet de penser l'ensemble des données et des documents disponibles dans un format unique. La plus lente maturation de RDF[4], liée à la notion de Web Sémantique [BER 99][5] permet de définir un langage fondé sur XML, et offrant la possibilité de parler *à propos* de ressources quelconques. Les relations entre structures documentaires et structures de description s'étudient naturellement dans ce cadre.

S'ajoutent à ces facteurs les changements de pratiques dus à la numérisation et à l'indexation des documents (tels qu'ils ont par exemple été décrits pour les documents audiovisuels en Sciences Humaines dans [AUF 99]), qui induisent un besoin nouveau de documents ; ainsi que le facteur « psychologique », lié à l'explosion du Web et des marchés connexes, dû à une volonté nouvelle, sinon de tout connaître, du moins d'avoir accès à tout *personnellement*. Si une majorité des informations (des connaissances) disponibles sur le Web n'est pas nouvelle, sa mise à disposition quasi- instantanée l'est, et implique qu'il y a lieu de répondre *obligatoirement* à toute question, parce que la connaissance existe, et ne peut qu'être proche. Ce dernier facteur, lié à tous les autres, peut-être le moins négligeable, tient au développement économique du Web et aux enjeux financiers énormes qui y sont liés : l'indexation sémantique des documents est nécessaire, au titre du développement de nouveaux services, car il y a beaucoup d'argent à gagner.

Cet article, à vocation exploratoire, poursuit plusieurs objectifs.

Le premier objectif est de mettre en exergue un certain nombre d'idées et de réflexions liées aux développements des rapports entre connaissances et documents au travers de la notion d'indexation conceptuelle comme connaissance *ajoutée* à un document, mais également comme connaissance présente dans toutes ses structures. A cet effet, nous proposerons de réévaluer la notion d'index, nous définirons comment nous envisageons l'indexation conceptuelle, avant de décrire quelques

---

[2] Extensible Markup Language, http://www.w 3c.org/XML

[3] La notion de document bien formé, éventuellement en dehors de toute DTD, permet de représenter à l'aide de XML des données telles qu'elles se trouvent dans les bases de données (perdant tout lien avec une structure documentaire), car il est possible dans le langage de traiter des instances comme telles, sans référence à un modèle.

[4] Resource Description Framework, http://www.w 3c.org/RDF (pour un tutorial : http://www710.univ-lyon1.fr/~champin/rdf-tutorial/)

[5] http://www.w3.org/Talks/1999/05/www8-tbl : il s'agit de considérer un espace universel dans lequel tout peut référer à tout, et de mettre en place des « espaces inter-créatifs » (if you notice a connection, make a link). Ces sujets prennent de plus en plus d'ampleur dans les conférences, et des workshops leur sont consacrés, tel http://www.ics.forth.gr/proj/isst/SemWeb/call.html.

travaux récents de description documentaire fondée sur les connaissances s'intégrant dans ce cadre. Nous argumenterons alors afin de considérer toute structure documentaire comme structure sémantique sous la forme d'un graphe. Le second objectif de l'article est de proposer d'envisager la notion d'annotation comme plus appropriée que celle d'indexation à la mise en place d'indexation conceptuelle. En effet, nous considérons que l'annotation permet de considérer toute indexation comme documentation de sa tâche d'exploitation d'un document, et nous présenterons quelques travaux récents dans ce domaine. Une discussion sur quelques problèmes et perspectives ouverts par la perspective d'annotation et l'évolution des documents numériques viendra conclure l'article. Un troisième objectif est de fournir au lecteur quelques pointeurs vers des recherches en cours dans certains domaines de l'indexation conceptuelle.

2. Indexation conceptuelle de documents

Dans cette partie, après quelques définitions rapides, nous discutons particulièrement la notion d'index documentaire, que nous ramenons à un concept opératoire dans le moment où il est utile pour une certaine tâche, manipulant à un moment ou à un autre un document ou bien une partie de document. Nous nous intéressons alors à l'indexation conceptuelle en soi, et proposons un cadre permettant de penser les systèmes d'information documentaire fondés sur les connaissances. Les évolutions des structures documentaires notamment arborescentes sont ensuite évoquées.

2.1. *Documents*

Un document est une trace de l'activité humaine, créée par un auteur et mise à disposition de lecteurs. L'intentionnalité de l'auteur se retrouve plus ou moins dans les *structures documentaires* (au sens général). Celles-ci ne sont pas obligatoirement explicites (la manière dont est interprété un document est loin d'être connue) ou explicitées dans un codage (par exemple, les moments d'un raisonnement ne se retrouveront pas obligatoirement dans la mise en page d'un texte). Les structures documentaires conditionnent – prescrivent – l'interprétation du document. En retour, interpréter un document peut également consister en la mise au jour de structures documentaires concernant sa description sémantique et son organisation.

Un document a été créé dans un certain contexte de production, en vue d'un certain contexte de réception, il appartient également à un genre, et réfère de façon implicite ou non à d'autres documents, qui en tant qu'intertexte en prescrivent également l'interprétation.

2.2. *Documents numériques*

Un document numérique [BAC 99] mobilise un support et une forme d'enregistrement numériques (fichiers), qui permettent de construire une forme d'appropriation sur un support d'appropriation (écran par exemple), répondant à une ou plusieurs modalités d'appropriation (texte, image). Il est « nécessairement ce qui est consulté dans le cadre d'une forme d'appropriation sur un support d'appropriation. Il ne correspond pas à l'enregistrement interne, contrairement à ce que la locution de « document numérique » pourrait laisser entendre ». « Le document numérique, considéré sur le support d'enregistrement, n'est pas un document, mais une ressource à partir de laquelle peuvent être calculés autant de documents, c'est à dire de formes d'appropriation sur un support d'appropriation. » Un document numérique autorise la navigation, c'est à dire le passage d'une partie de document à une autre.

Les documents numériques peuvent être variés et nombreux, mobilisant diverses modalités d'appropriation, et correspondant à des pratiques d'interprétation et d'utilisation diverses. Certains genres ont passé sans heurts le cap du passage au numérique, tandis que d'autres sont nés avec celui-ci (courrier électronique). Pour illustrer cette diversité, on trouvera entre autres les textes (dépêches, romans, manuscrits anciens, manuels techniques, etc.), les images et les schémas, les sons (musique, bruit, dialogues), l'audiovisuel, le multimédia interactif (pages web, cédéroms), les documents de conception assistée par ordinateur, les documents de travail collaboratif (courriers électroniques, messages de groupes de discussion), etc.

Avec les documents numériques (principalement textuel) sont apparus les formats documentaires (mimant plus ou moins les formes d'appropriation) permettant leur manipulation. Des normes telles que SGML[6], HyTime[7], HTML et plus récemment XML fournissent des langages permettant de décrire les documents et les structures documentaires (désormais : dans le sens de structures décrites dans des langages formels). La TEI définit des DTD (Définition de Type de Document) correspondant à divers genres de documents littéraires, dont les structures, issues d'un usage ancien, peuvent être normalisées. Du fait de leur composante non structurée *a priori* (puisqu'il n'y a pas eu de normalisation de structure liée à un usage) les images et documents audiovisuels sont le plus souvent codés en tant qu'entités dans un format binaire ne mimant pas de structure d'analyse ou d'interprétation.

Le codage des documents structurés dans des langages de balises se révèle donc être une donnée qu'il y a lieu de prendre en compte dans l'indexation. Conséquence de la possibilité de XML de se détacher de tout modèle documentaire, il est également désormais possible de considérer n'importe quelle instance XML comme un document, y compris le codage d'un brin d'ADN !

2.2. *Index documentaires*

[6] Standard Generalized Markup Language, http://www.w3.org/MarkUp/SGML/
[7] Hypermedia / Time-based Structuring Language, http://www.ornl.gov/sgml/wg8/document/1920.html

De façon générale, un index est « quelque chose » permettant d'accéder à « autre chose ». L'indexation est alors mise en place d'objets permettant d'accéder à d'autres objets, ceux-ci pouvant être, dans le cas qui nous intéresse, des documents, des parties de documents, voire des ensembles de documents. L'indexation comme processus vise à mettre en place des index (qui correspondront à l'indexation comme résultat), afin d'être à même de retrouver et d'avoir accès à des documents.

L'accès est permis par l'*identifiant* de ce qui est indexé (numéro, page). Remarquons que celui-ci représente déjà en soi un index à partir du moment où il est exprimé dans une forme sémiotique interprétable par un utilisateur, pour peu que celui-ci en ait la clé, c'est à dire les règles et le contexte de production (un numéro représentant un ordre d'arrivée, une page de livre ou un moment temporel peuvent être indicatifs de ce qui est indexé). En ce qui concerne le véritable index, associé à l'identifiant, [BAC 99] le définit dans une bibliothèque comme *«* la paraphrase d'un contenu en une forme sémiotique interprétable permettant l'exploitation du contenu indexé dans le cadre d'une pratique donnée. » Il s'agit alors de mettre en place des *documents* de description qui sont documents *per se*, parce qu'ayant contenu et forme matérielle et qui réciproquement pourront être indexés.

De manière générale, les index sont valables dans une pratique donnée, c'est à dire que leur création se fait dans un contexte précis, en vue d'au moins une utilisation prévue explicitement. Cependant, il est possible de dire *a maxima,* qu'est index de contenu documentaire tout ce qui est utilisé comme tel, *que cela ait été mis en place dans cette volonté ou non*. Cela est très important dans le domaine numérique. En effet, « tout système numérique est documentaire et mobilise une indexation » [BAC 99] (exemple limite, un code binaire est indexé par son source s'il est disponible, et de toute façon par la table des symboles). Dès lors, n'importe quelle partie du support d'enregistrement d'un document est susceptible d'une part d'être accédée comme contenu documentaire, d'autre part de servir d'index pour ce contenu.

En conséquence de quoi il nous apparaît que tout système d'information documentaire est composé d'autant d'index potentiels qu'il y a de possibilités de calcul de liens document/index faisant sens dans le cadre d'une pratique. L'enjeu est alors celui de la *compréhensibilité* des index, de la possibilité de leur interprétation, de leur validité, qui en font des index véritables.

2.3. Indexation conceptuelle

L'objectif d'un système d'information documentaire, quel qu'il soit, est de gérer un ensemble de documents en vue de leur utilisation, au cours d'une collaboration entre l'homme et le système. Il s'agit donc de disposer d'un environnement riche permettant non seulement de rechercher (ce qui correspond à la définition classique),

mais également d'exploiter les documents. Alors, pour l'utilisateur, la recherche d'information n'est qu'une phase parmi d'autres de l'utilisation des documents. Inversement, on peut dire que l'exploitation de documents passera, au moins pour le système, par une recherche d'information permanente, puisqu'elle suppose un accès aux documents et à leur index pertinents. Les connaissances du domaine, de structuration documentaire, sur l'utilisateur, si elles sont disponibles, seront par exemple en permanence mobilisables en tant qu'index des documents.

Il s'ensuit qu'il y a lieu, au moins au niveau des documents, d'*instrumenter* ceux-ci en vue de leur exploitation – plus ou moins automatique pour les documents numérique – ce qui s'apparente dans tous les cas à de l'indexation (*cf.* ci-dessus).

Dans le cas d'un système informatique, les index devant pouvoir être exploités comme connaissances dans des processus d'inférences automatiques par la machine, on se rapproche de ce que nous avons appelé [PRIE 99b] *indexation intelligente* comme une indexation offrant dans son mode de fonctionnement/représentation même la possibilité de l'interroger et de la manipuler elle-même, en tant qu'indexation, et plus seulement en tant qu'index à traverser vers une information « brute ». Un document ainsi indexé offre par la description de son contenu structuré et indexé la possibilité de manipulation, de calcul, voire d'inférences sur ses index, que ceux-ci soient décrits en dehors de lui ou bien en son sein. Toute utilisation du document passe par conséquent par l'un de ses index, en tant que connaissance en soi.

Ainsi, la notion de *connaissances* devient importante, et partant la notion d'*indexation conceptuelle*, comme « explicitation de structures et de concepts contenus dans les documents numériques ou qui leur sont associés, pour mieux les exploiter » (appel à proposition du présent numéro).

Nous définissons l'indexation conceptuelle comme recouvrant toute connaissance ajoutée à un document pouvant servir dans le cadre de « calculs » sous-tendus par l'exploitation de ces documents, pourvu que cette connaissance soit utilisable aussi bien par l'homme que par la machine[8]. La possibilité d'*inférence* sur les index conceptuels fait de droit partie des possibilités d'exploitation documentaire. L'indexation conceptuelle consiste en toute explicitation symbolique de connaissances contenues dans les documents ou bien à leur propos, en permettant la recherche et la manipulation. Ceci dépasse à notre sens la simple explicitation des concepts contenus dans les documents, mais recouvre également tout ajout de connaissances pouvant servir d'une manière ou d'une autre, par exemple la position géographique d'une caméra au moment où un plan audiovisuel a été tourné ou bien l'âge du réalisateur. Nous verrons plus loin que même les structures documentaires (au sens des langages de balises) relèvent de l'indexation conceptuelle.

Comme nous l'avons souligné dans l'introduction, la possibilité de représenter dans un même langage à la fois des documents et des assertions sur ceux-ci qui en

[8] Cela élimine *a priori* les traitements automatiques sur des données purement non symboliques.

serviront d'index se révèle ici fondamentale, et les travaux autour de RDF semblent promis à un avenir certain.

2.4. Descriptions de documents orientées connaissances

Nous nous restreignons dans cette partie à l'indexation conceptuelle consistant en la description des connaissances contenues dans un document dans des formalismes de représentation de connaissances permettant de mener des inférences sur celles-ci. De nombreux travaux récents s'attaquent en effet à cette question, et nous nous proposons de les décrire et de les comparer (2.4.5) au travers d'une grille composée des quatre critères suivants : les documents ou parties de documents décrits (2.4.1) ; les connaissances de description utilisées pour mettre en place les descriptions (2.4.2) ; les descriptions elles-mêmes (l'indexation conceptuelle) et leur mise en place (2.4.3) ; enfin la manière dont les descriptions sont utilisées (2.4.4).

2.4.1. *Documents, parties de documents*

Nous l'avons vu, le support documentaire de l'indexation conceptuelle peut être composé d'un document entier ou bien d'une partie du document. Dans ce dernier cas, les parties de documents décrites le sont le plus souvent à l'aide de balises SGML/XML qui en donnent la portée. Pour les images, il est d'usage de déterminer des zones (par exemple décrire des images médicales), et la future norme JPEG2000[9] fournira ces possibilités. Pour le son, les zones sont en général temporelles, et peuvent prendre en compte une seule ou bien l'ensemble des pistes (par exemple un instrument). Pour les documents audiovisuels, il est possible de se limiter à des segments temporels, ou bien passer à d'autres niveaux, par exemple des zones d'images, des objets temporels tels que ceux disponibles dans la norme MPEG 4[10].

Deux questions restent ouvertes quant aux parties de document faisant l'objet d'une indexation. La première concerne la compréhensibilité minimale de la partie visée, et son indépendance par rapport au reste du document (par exemple, dans [NAN 95], un contexte minimal de compréhension est défini autour d'une ancre n'englobant qu'un seul terme du document textuel). La seconde question concerne la capacité propre à avoir sens de la partie de document considérée (par exemple, un paragraphe n'est pas une partie quelconque d'un texte, pas plus qu'un plan audiovisuel).

2.4.2. *Connaissances de description*

L'indexation fondée sur les connaissances d'un document ne peut se faire sans un ensemble de connaissances de description qui sont définies avant l'indexation, le plus souvent *hors* du document. Celles-ci, clairement identifiées et disponibles, permettront le partage de l'utilisation de l'indexation, c'est à dire la continuité

[9] Joint Picture Expert Group, http://www.jpeg.org
[10] Moving Pictures Expert Group, http://www.cselt.it/mpeg/standards/mpeg-4/mpeg-4.htm

sémantique entre la mise en place et l'utilisation de l'indexation. Elles seront *a minima* de simples vocabulaires, et prendront *a maxima* la forme d'ontologies, c'est à dire de définition des concepts du domaine visé et de leurs relations, ainsi que de règles exprimant les possibilités de description. Certains concepts pourront prendre la forme d'objets (attributs/valeurs) et une relation d'héritage pourra être mise en place. D'autres règles permettront éventuellement de servir de support à des inférences non liées à la relation de spécialisation (est-un) de l'ontologie. La couverture de ces connaissances pourra être limitée à un domaine (connaissances techniques par exemple), associer différents types de descripteurs (e.g. de structure et sémantiques), voire correspondre à un vocabulaire général (type Wordnet).

2.4.3. *Descriptions*

On peut distinguer au moins deux degrés dans les descriptions fondées sur les connaissances.

Il s'agit au premier degré d'éclairer un ensemble de documents à l'aide d'une base de connaissances (le plus souvent une ou plusieurs ontologies, sur lesquelles il est possible de mener des inférences minimales). Certains concepts sont alors repérés comme s'instanciant dans des documents, qu'il est alors possible de retrouver : un monde est décrit, et ses concepts sont illustrés par des documents. La plupart des systèmes de gestion des connaissances de l'entreprise (*knowledge management*) suivent une telle approche. La base de connaissances est alors construite en fonction des documents qu'il y a lieu d'intégrer, et se limite le plus souvent à un domaine relativement précis. [MOT 00] propose une méthodologie de mise en place de tels systèmes, en insistant sur la difficulté qu'il y a à construire une ontologie et sur les coûts afférents, qu'il s'agit d'évaluer.

Le second degré, plus ambitieux, vise à décrire des faits relatifs aux contenus des documents, le plus souvent sous la forme d'assertions logiques. Un monde est décrit *dans* les documents. Ces assertions peuvent être stockées dans les documents et/ou dans des bases externes. Les systèmes fondés sur cette approche, qui donne plus de possibilités de finesse et de précision dans les descriptions relèvent pour l'instant du domaine de la recherche.

La mise en place des descriptions peut avoir lieu de façon manuelle, semi-automatique ou même automatique. Un domaine de recherche actif consiste en effet dans l'étude de la possibilité d'aller chercher les connaissances dans les documents (par exemple en recherchant des motifs dans des sous-parties identifiées de documents structurés).

Remarquons qu'à partir du moment où l'on va considérer une indexation conceptuelle *interne* au document se pose la question de sa valeur ajoutée, ainsi que celle de son cycle de vie. En effet, il devient possible d'envisager de façon intégrée par exemple la documentation du document au cours de son cycle de vie, y compris lors de sa création (lorsqu'il y a production directe en numérique). On conçoit alors les enjeux en termes d'évolution de la notion de document même, puisque les

structures d'indexation/documentation font alors de droit partie de la structure documentaire. Nous reviendrons sur ces questions plus loin dans l'article.

2.4.4. *Exploitations*

L'exploitation du système d'information documentaire fondé sur les connaissances consiste le plus souvent en la recherche de documents guidée par les connaissances mises en place, permettant d'atteindre les connaissances virtuelles qui y sont présentes. Celle-ci peut se faire par navigation à partir de la base (cas du premièr degré de description), mais aussi en spécifiant des faits dans une requête, à charge pour le système de trouver des documents dont les assertions correspondent[11]. Les inférences sont le plus souvent menées dans une base de connaissance externe aux documents, construite à partir des connaissances de description et des descriptions conceptuelles documentaires. Quant à l'utilisation « non documentaire », il s'agira par exemple de répondre à des questions sur la base en déduisant de nouvelles connaissances. Les exploitations automatiques des descriptions consistent également dans la mise en place de tables de matières, dans la création de liens entre documents (éventuellement à la volée), dans les résumés sur des sujets particuliers, etc.

2.4.5. *Quelques systèmes récents*

Le lecteur trouvera ici les descriptions succinctes de quelques systèmes récents, illustrant la vitalité et la diversité de la recherche dans le domaine de la description de documents fondée sur les connaissances.

[MOT 00] propose une approche d'enrichissement de bases documentaires « dirigé par les ontologies », ce qui correspond à notre premier degré de description. Il ne s'agit pas de décrire formellement ce qui est informellement dit dans les documents, mais plutôt d'y associer une valeur ajoutée pour l'exploitation et la recherche tout en fournissant des services de raisonnement (principalement liés à la relation de spécialisation). Ainsi, pour un ensemble de documents, on construit une ontologie, externe, qui servira de point d'accès, et permettra de rechercher des documents ou de construire de nouveaux documents à partir du corpus. Un certain nombre d'outils et une méthodologie ont été développés pour faciliter le travail du concepteur, et l'architecture a permis de construire des systèmes adaptés aux informations, à la veille scientifique, mais également pour l'organisation de documents pédagogiques.

[MAR 00] propose de décrire la sémantique de parties de documents du Web à l'aide de graphes conceptuels qui y seront stockés dans un langage de balises, car ceux-ci ont l'avantage d'être exploitables et lisibles à la fois par l'homme et la

[11] Si l'indexation de la « Recherche d'Information historique » telle que présentée dans [ROU 99] perd un peu de pertinence au niveau recherche, alors qu'elle reste la plus utilisée par les moteurs, ceci ne présage pas d'adaptations du *modèle logique* pour prendre en compte l'indexation conceptuelle et les requêtes inférentielles liées : on considère alors que la description conceptuelle du document visé doit *impliquer* logiquement la requête.

machine. Les connaissances de description consistent en un treillis des types de concepts et de relations, organisés suivant la relation *est-un*. L'intérêt principal est ici la volonté de disposer d'une base de connaissances générale, c'est pourquoi les concepts sont extraits de Wordnet, tandis qu'une base générale de relations (près de 200) a été mise en place. Les inférences liées à la recherche consistent en des projections des graphes requêtes dans les graphes description. Un *shell* permet de poser des requêtes et d'organiser les réponses dans des nouveaux documents.

[GUA 99] propose d'utiliser les ontologies pour la recherche d'information. Le système OntoSeek vise des corpus correspondant à des catalogues (pages jaunes, catalogues commerciaux) dont la structure est peu ou prou la même pour toutes les pages. Sur la constatation de la difficulté à mettre en place et à faire évoluer des ontologies strictes, l'originalité consiste à utiliser une ressource linguistique telle que Wordnet, qui offre un vocabulaire (des sens) et une ontologie (des relations sémantiques). L'utilisateur décrit alors les documents à l'aide de « graphes conceptuels lexicaux »[12] (LCG) que le système, en se basant sur Wordnet et un certain nombre de règles lui propose de désambiguiser. La recherche de documents consiste en de l'appariement de graphes dirigé par les ontologies, en se basant notamment sur la subsomption. On remarquera avec intérêt que l'ensemble des données du système se trouve dans un graphe unique.

[ZAR 99] présente le projet CONCERTO d'annotation conceptuelle de documents Web : il s'agit d'associer au document un texte descriptif en langue naturelle, qui peut être extrait ou non du document, lequel sera ensuite traduit dans le langage de haut niveau NKRL (Narrative Knowledge Representation Language). Ce langage est fondé sur un certain nombre d'ontologies, et dispose de patrons qui permettent de décrire la manière d'utiliser celles-ci pour décrire. Le système fournit un environnement semi-automatique de mise en place des connaissances, et un module de recherche intelligente d'information, qui permet notamment de gérer des extensions de requêtes.

[ERD 00] présente les derniers développements d'Ontobroker. Ce système est construit autour d'une ontologie centrale composée d'un vocabulaire organisé en taxonomie et de règles d'inférences. Des documents Web sont annotés en HTML (éventuellement automatiquement lorsqu'ils sont assez réguliers dans leur structure), et les connaissances sont accumulées dans la base de connaissances sous la forme d'assertions F-logic [KIF 90] (dont toute la puissance est alors disponible). Le système permet alors de répondre aux requêtes en menant des inférences sur la base de connaissances et de retrouver des documents décrits. Les développements récents (DTDMaker) s'intéressent à la possibilité de représenter des parties d'ontologies

[12] Ces graphes sont considérés comme étant « modérément expressifs », et se basent uniquement sur des relations langagières, de noms relationnels issus de Wordnet. Par exemple, il n'existera pas de relation ad-hoc abstraite du type *has-part*, mais une relation étiquetée à l'aide de *part*.

sous la forme de DTD XML, afin de pouvoir mener des contrôles sémantiques sur des faits cohérents avec l'ontologie[13].

[HEF 99] choisit de rompre radicalement avec les approches fondées sur des ontologies centralisées, considérant le Web dans son ensemble comme base de connaissances, ce qui nécessite de repenser les outils de représentation de connaissances. La taille du Web par exemple nécessite un relâchement de l'expressivité des langages, tandis que l'hypothèse du monde clos doit être abandonnée, ainsi que la consistance de toutes les assertions entre elles. De plus, il s'agit dès le départ de prendre en compte l'évolution du Web et de la connaissance de description. Les ontologies utilisées devront alors être assez larges pour se charger du maximum de domaines, mais point trop, afin de pouvoir évoluer. Le langage SHOE est un langage de représentation de connaissances avec des ontologies identifiées, disponibles sur le Web, qui peuvent être étendues à tout moment. Une ontologie est une taxonomie ainsi que des possibilités de relations entre catégories. Les extensions d'ontologies et les assertions, toujours uniques, rajoutées et positives (la négation n'est pas exprimable dans le langage), voire éventuellement inconsistantes, sont stockées dans les pages Web sous la forme de balises supplémentaire. Un moteur recherche ces pages, stocke les faits et permet de répondre aux requêtes. Un schéma d'utilisation des *claims* (ce qui a été dit à un moment donné, par une personne donnée), est également proposé afin de s'accommoder des inévitables inconsistances dans les assertions. Les études en cours de ces travaux originaux portent sur l'évolution et le partage d'ontologies sur le Web.

2.5. Indexation conceptuelle et structuration documentaire

Certains des modèles que nous venons de présenter ont pour objectif de stocker les descriptions mêmes (l'indexation conceptuelle) dans les documents structurés. L'intégration de XML dans la majorité des systèmes d'information nous incite à penser que de plus en plus de connaissances seront ainsi décrites dans les documents, en tant qu'indexations conceptuelles, qui pourront certes se trouver à part de la structure documentaire, mais aussi sans doute de plus en plus naturellement

[13] Pour [EUZ 99] il y a bien convergence des outils du Web, des documents et des connaissances, notamment convergence lente entre langages de représentation de connaissances et langages de description de documents. Une DTD permet par exemple d'échanger des informations sur la manière dont est codé, au niveau connaissances, un document. Cependant, il n'existe pas (encore ?) de sémantique permettant la validation des opérations appliquées aux documents. Cela signifie qu'il est nécessaire d'interpréter une DTD en tant que connaissance de description pour savoir comment exploiter les documents décrits suivant celle-ci. [EUZ 99] suggère par ailleurs qu'une des voies intéressantes pourrait consister à joindre des informations sémantiques aux DTD permettant de valider des manipulations sur les documents décrits.

directement en XML, dans la structure documentaire. Nous sommes alors enclins à penser 1/ que les différentes structures deviendront uniquement sémantiques, et que l'importance de la structure logique des documents pourrait s'atténuer 2/ qu'il y a également lieu de considérer la description des documents en termes de graphes et non plus d'arbres.

2.5.1. *Des structures sémantiques*

Considérons rapidement et grossièrement l'évolution des structures de présentation, que nous définissons comme structures documentaires mimant la manière dont un document est nativement présenté au lecteur.

Les premiers modèles documentaires ont été fondés sur des structures de présentation liées à la présentation du document à l'écran, ensuite s'est mise en place – avec les langages de balise tels que SGML – la différenciation entre structure logique et structure physique, dont un calcul permet de passer de l'une à l'autre. D'autres travaux ont alors défini la notion de structure sémantique comme structure de connaissances non liée à la présentation seule des documents (par exemple les balises META dans HTML, ou encore des liens entre ressources dans RDF). De plus, il est maintenant reconnu que les balises liées à la présentation – même les plus banales – décrivent en fait des connaissances qui peuvent être utilisées dans d'autres tâches que celle de présentation. Par exemple, une balise *Titre* exprime deux choses : <titre> comme dénomination qui entraînera une certaine présentation et un arrangement de ce qui est encadré par la balise, mais aussi le fait que ce qui est encadré est un titre, c'est à dire une information sémantique utilisable en tant que telle, par exemple pour en extraire des mots clés *a priori* plus pertinents que d'autres sur le document. On constate alors que la frontière entre connaissances de présentation et connaissances autres est en fait relativement ténue, voire s'estompe. Les structures de présentation et les structures d'indexation conceptuelle décrivent toutes des connaissances sur le document, et dans le même temps fondent celui-ci. Ainsi un document accompagné de sa documentation est encore un document, désormais adapté à des utilisations autres que la tâche classique de présentation pour l'appréhension. XML, n'étant pas *a priori* fondé sur des considérations de présentation participe de cette évolution.

Si la structure de présentation telle que nous l'avons définie pouvait s'apparenter dans un premier temps à la structure logique définie dans la terminologie documentaire classique, laquelle est bien différenciée de la structure sémantique, nous préférons considérer que toute structure est sémantique et fait partie d'une structure d'indexation conceptuelle. Nous considérons donc que toute structure interne ou externe à un document est structure sémantique, laquelle prend son statut lorsqu'elle est utilisée dans le cadre d'une tâche qui peut être une tâche de présentation[14]. Le terme structure de présentation nous semble à cet égard devenir

[14] L'approche des Strates-IA [PRI 99c], sur laquelle nous reviendrons, est un pas dans cette direction : annoter une partie de document audiovisuel comme < Plan > revient à ne pas

une notion dynamique et opératoire, dont la définition est finalement : « fait partie de la structure de présentation tout élément de structure de connaissance documentaire (sémantique) utilisé pour mettre en place une présentation du document ». Dans ces conditions la structure logique pourrait perdre de son importance initiale pour n'être plus qu'une structure de présentation « canonique » extraite d'une structure générale de connaissances. Toute structure utilisable mise en place dans un document ou sur un document est donc une structure sémantique de connaissances documentaires, qui peut être utilisée (manipulée) en tant que telle dans le cadre d'une tâche. Toute connaissance peut dès lors servir d'index dès qu'elle permet d'accéder au document

Il convient pour terminer cette courte étude de se poser la question des structures de connaissances documentaires minimales attachées à un contenu documentaire. Il nous semble qu'il existe minimalement une structure de présentation canonique correspondant au genre du document. Ces connaissances sont mises en place (implicitement ou explicitement) par l'auteur du document, on pourrait donc les appeler connaissances « auctoriales » premières. A ces connaissances minimales il est possible d'ajouter toutes les connaissances possibles, afin de compléter l'indexation conceptuelle.

Pour résumer, nous proposons de considérer toute structure documentaire comme sémantique, l'utilisation même du document révélant finalement la structure documentaire considérée par l'utilisateur. Indexation conceptuelle et structurelle nous apparaissent donc relever du même niveau.

2.5.2. *Arbres et graphes*

Conséquence logique de la structure documentaire considérée comme une, il nous semble qu'il y a désormais lieu de considérer celle-ci plutôt comme un graphe que comme un arbre. En effet, il est reconnu que l'organisation des connaissances de description ne doit pas *a priori* suivre un schéma hiérarchique[15], ce qui signifie que les possibilités de description doivent être suffisamment riches. Le graphe comme moyen le plus général de représentation des connaissances s'impose alors[16]. En d'autres termes, les schémas de descriptions ne doivent pas être plaqués sur les outils de description « technologiques ». Même si la description dans un langage documentaire tel que XML suppose une ossature de description arborescente primaire, il est possible de définir des liens entre les balises. Par contre, il est

donner de primat à l'information « Plan » comme ce serait le cas si était définie une unité « syntaxique » ou logique *Plan.*

[15] Un livre peut être décrit en chapitre et section, ou bien suivant l'organisation des idées ou des schémas ; un document audiovisuel pourra être décrit suivant les flux visuel, audio, ou bien audiovisuel, en notant les interrelations entre les descripteurs ; un match de tennis peut être décrit suivant ses plans, ou bien en fonction des points du jeu, *etc.*

[16] C'est par exemple la conclusion à laquelle arrive également [BIR 00] pour la modélisation d'annotation de corpus linguistiques, à la suite d'un travail d'abstraction des modèles d'annotation les plus utilisés : seuls des modèles génériques sous forme de graphe se révèlent assez riches pour les besoins des chercheurs.

possible que la présentation à l'utilisateur se base de façon privilégiée sur la notion d'arbre, en attendant que des interfaces simples et généralisées de visualisation de graphes existent. Se pose alors la question de ce qu'il y a lieu de présenter de la structure de connaissances à l'utilisateur dans le cadre de sa pratique.

3. Une perspective générale d'annotation

La partie précédente, consacrée à l'indexation conceptuelle, nous a permis de définir celle-ci comme s'incarnant dans toute structure symbolique disponible de description du document. En effet, nous avons touché du doigt le fait que toute connaissance explicite interne ou externe au document pouvait servir d'index conceptuel pour celui-ci. Nous avons aussi plaidé pour la prise en compte de toute structure documentaire comme structure sémantique, qu'il s'agit de considérer globalement comme graphe unique. Dans la présente partie, nous proposons de considérer la notion d'annotation comme processus général de mise en place d'indexation conceptuelle.

3.1. *Ecrire son interprétation*

Dans le cas le plus général, annoter un document, c'est attacher à l'une de ses parties une description qui correspond à un usage que l'on – ou que toute autre personne – souhaitera en faire plus tard. L'annotation savante est nécessaire au travail intellectuel sur les textes, et ressort souvent du commentaire, de la mise en relation, de la construction d'un réseau d'intertextes. Cette tradition a par exemple été prise en compte dans les travaux sur la mise en place de la station de lecture de la Bibliothèque Nationale de France [VIR 95].

Au niveau des résultats, les notions d'annotation et d'indexation semblent équivalentes : on se retrouve dans tous les cas avec une indexation conceptuelle du document, qui sera utilisée dans le cadre d'une tâche d'exploitation.

Cependant, il nous semble que c'est dans le *processus* et sa *dynamique* qu'il y a différence. Indexer, c'est avant tout décrire un document pour le retrouver. Annoter, c'est décrire son interprétation du document, en vue de n'importe quelle tâche d'exploitation de ce document. On indexe pour rechercher plus tard, on annote pour donner des traces de son interprétation, pour documenter la tâche que l'on est en train d'accomplir. Ces traces pourront alors être destinées à soi-même, ou bien partagées, et leur mise en place peut se faire de façon plus ou moins automatique, et plus ou moins collaborative.

Par exemple, un collecticiel assurera le partage d'annotation de documents sur lesquels plusieurs utilisateurs travaillent, chacun documentant la partie qu'il écrit, et ses choix de rédaction, éventuellement en vue d'une réécriture [CHI 99]. L'annotation de ressources linguistiques est également un domaine qui prend de l'importance, puisqu'on est en train de passer de l'étiquetage (*tagging*) à l'annotation suivant des schémas d'annotation [ISA 00]. L'annotation personnelle de

pages Web devient de plus en plus nécessaire afin d'une part d'être capable de maîtriser la prolifération des signets, et d'autre part d'améliorer sa productivité d'utilisation des pages (annoter la partie importante d'une page permet par exemple de décider de son intérêt sans la charger) [DEN 00]. Enfin, l'annotation permet de considérer et d'échanger par exemple l'interprétation qu'un chercheur en génétique fera d'un brin d'ADN, interprétation qui n'est pas figée [LIB 99].

Le concept d'annotation au cours d'une certaine tâche, en vue d'une certaine tâche, menant à une indexation exploitée au cours d'une autre tâche nous semble dès lors pouvoir être considéré comme unificateur, qui dépasse celui d'indexation en tant que processus. Ainsi, une personne indexant un document documente en fait la tâche qu'elle est en train d'accomplir, annote celui-ci par son interprétation en vue d'une tâche de recherche menée par d'autres, au cours d'une autre tâche qu'elle connaît plus ou moins (par exemple, écrire un article).

Nous proposons dès lors d'envisager la mise en place d'indexation conceptuelle pour un document sous la perspective unificatrice de l'annotation, liée à l'écriture sur le document, par définition non figée, évolutive, permettant l'attachement de ressources à d'autres et la documentation de sa propre tâche. Point positif, remarquons à ce titre que l'annotation permet de penser la mise en place de balises documentaires, quelles qu'elles soient, y compris liées à la structure de présentation (par exemple une mise en gras), puisque celles-ci correspondent à une documentation directe de la tâche d'écriture. L'indexation mise en place du fait des balises devient alors réellement conceptuelle.

3.2. *Auteurs et lecteurs, partager des annotations*

Faire rentrer l'indexation dans le cadre de l'annotation nous permet de penser dans un cadre unificateur l'ensemble des opérations consistant à ajouter – d'une manière ou d'une autre – des connaissances à un document, au cours d'une certaine tâche, en vue d'une autre tâche.

Il est possible de considérer l'acte d'annotation comme acte de communication entre un auteur et un lecteur (voire de transmission de connaissances), dans le cadre d'un document structuré. L'auteur est par définition le premier « documenteur » de son texte, éventuellement de façon inconsciente lorsqu'il utilise un traitement de texte, par exemple en sélectionnant un niveau de titre, ou de façon consciente par exemple s'il indique en note quelle version d'un titre il a écartée.

Concernant la mise en place d'annotations dans un second temps, il faut remarquer que le plus souvent, l'annotation est annotation de parties de documents qui sont déjà désignées, c'est à dire que la structure « auctoriale » est utilisée comme support pour une description annotative qui lui est seconde. Ceci signifie qu'est généralement accepté le fait qu'il existe une analyse primordiale, qui est celle du concepteur du document, qui servira de support à l'analyse. Ceci est faux dans le cas

général, où l'auteur d'un document ne maîtrise pas toutes les utilisations et exploitations qui en seront faites, hormis dans des cas très particuliers. Il n'empêche que le point de vue de l'auteur, bien que non définitif, n'en est pas moins très important, et est entièrement porté par la documentation et les prescriptions *explicites* du document. La liberté de définition des annotations dépend en fait de la tâche que ces dernières doivent pouvoir supporter. Dans le cas de l'audiovisuel par exemple, on peut imaginer que dans un futur proche un auteur mettra en place une structure documentant son travail créatif, sur laquelle s'appuyer pour annoter, par exemple dans un objectif d'analyse de l'œuvre et de sa création. Il est cependant nécessaire de pouvoir définir d'autres unités que celles qu'il aura explicitement spécifiées, la liberté d'analyse en dépendant. Dans le cas d'un travail plus précisément défini, comme celui présenté dans [CHI 99], où des parties de documents sont clairement documentées et partagées dans le cadre d'une méthode de conception, la liberté d'annotation est plus contrainte.

Nous résumerons le cas général en spécifiant que *l'auteur du document n'est pas coauteur de toutes les annotations* au sens où il en définirait les zones de portée. De l'exploitation dépend ensuite le rapport auteur/lecteur : il y aura par exemple une différence entre le savant qui annote et est son propre lecteur, donc relativement libre au niveau vocabulaire et méthodes, et d'autres cas plus généraux (par exemple quand un documentaliste annote pour une communauté d'utilisateurs).

La notion de *schéma d'annotation* s'impose alors, afin d'assurer la continuité sémantique entre l'auteur et l'utilisateur, éventuellement machine, de l'annotation, et de pouvoir envisager une utilisation autre qu'exploratoire, par exemple inférentielle. Ces schémas d'annotation s'inscrivent bien entendu dans le cadre de l'indexation conceptuelle défini en 2.4., comme connaissances de description. Ils sont plus ou moins formellement définis et sont plus ou moins partagés entre les utilisateurs, qu'ils soient DTD, simples thesaurii, ontologies ou autres.

3.3. *Quelques exemples*

Après avoir justifié l'intérêt à notre sens de la notion d'annotation plutôt qu'indexation dans la mise en place d'indexation conceptuelle, nous présentons à titre d'illustration quelques approches de description de documents qui rentrent dans ce cadre.

Tout d'abord, remarquons que quelques produits généraux d'annotation commencent à être disponibles sur la marché, par exemple Annotation SDK[17] qui permet d'annoter et de stocker des annotations associées à des documents quelconques dans l'environnement Windows, autorisant un utilisateur à documenter son exploitation générale de l'environnement informatique auquel il fait face. Des

[17] http://www.blackice.com/annot1.htm

produits commerciaux d'annotation de pages Web existent également[18]. Les systèmes que nous évoquons maintenant appartiennent au domaine de la recherche.

Egalement dans le domaine des annotations Web, [DEN 00] considèrent des signets améliorés comme annotant les documents qu'ils désignent. Il est également possible d'annoter des sous-parties de document. Une annotation est considérée comme l'association d'une ancre et d'attributs, et certains attributs normalisés sont proposés, tels que *sujet* et *type de document*, *type du texte surligné*, *commentaire libre* et *style* (d'accord, pas d'accord), mais l'utilisateur peut considérer ses propres attributs. Le système permet de visualiser les annotations et d'y mener des recherches. On constate que le schéma d'annotation est ici relativement libre, peut être plus ou moins partagé, ainsi que les annotations elles-mêmes.

Avec le développement de bases de documents linguistiques annotés (textes étiquetés, transcription de discours, etc.) mis librement à disposition des chercheurs, il devient nécessaire de penser l'annotation de corpus linguistiques[19], voire de documents traces de communication (par exemple des dialogues multimodaux comprenant des gestes de désignation). Un exemple récent est entre autres celui de MATE[20], projet dans lequel ont été étudiés tous les schémas d'annotation disponibles, afin de permettre d'annoter à l'aide des meilleurs. A ce titre, un outil d'annotation de corpus codés en XML a été mis en place : un éditeur générique est spécialisé pour une certaine tâche d'annotation, gère la présentation des documents et les actions d'annotation utilisées suivant les schémas. On notera que si les schémas d'annotation conduisent actuellement à des descriptions conceptuelles arborescentes, les auteurs reconnaissent qu'il y a lieu de passer aux graphes (et non à des forêts), l'annotation par exemple de corpus de dialogues ne pouvant s'accommoder d'une simple description hiérarchique [ISA 00].

Dans le cadre du projet SESAME consacré à l'exploitation de documents audiovisuels, nous avons proposé le modèle des Strates Interconnectées par les Annotations (Strates-IA) pour l'instrumentation de ceux-ci [PRI 99a, PRI 99c]. Un document audiovisuel, en effet ne dispose d'aucune autre structure de présentation que la superposition synchrone de flux audio et vidéo. Pour toute manipulation autre que la lecture, il est donc nécessaire de mettre en place une indexation conceptuelle.

---

[18] Par exemple http://www.thirdvoice.com/ ou http://www.imarkup.com/.

[19] Le workshop récent http://www.mpi.nl/world/ISLE/ (First EAGLES/ISLE Workshop on Meta-Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources and Data Architectures and Software Support for Large Corpora LREC 2000 Pre-Conference Workshops) illustre ces développements.

[20] *Multilevel Annotation Tools Engineering*: « The MATE project aims to specify, develop and test an internet-based set of tools for the annotation of spoken dialogue corpora. Annotated dialogues are a central resource for the efficient development of speech and language technologies, including spoken language dialogue systems. The market for such systems is currently expanding and there is an increasing need for re-use of annotated material and for making the process of annotation more efficient. » (voir http://www.ims.uni-stuttgart.de/projekte/mate/).

Dans ce modèle explicitement fondé sur l'annotation, les parties de documents sont des *unités audiovisuelles*, les descriptions sont composées de réseaux d'*éléments d'annotation* (tels que <Plan>, <Scène> ou <Chirac>, tous sémantiques, quelque soit l'analyse menée). Les connaissances de description consistent en une « base de connaissances » (*a minima* un thesaurus) regroupant les *éléments d'annotation abstraits*, ainsi que des *schémas de description* spécifiant comment il y a lieu d'utiliser les éléments d'annotation abstraits pour *écrire* sur le flux audiovisuel des réseaux d'éléments d'annotation. Nous considérons en effet – et c'est là l'originalité essentielle du projet – qu'il faut mettre en place l'annotation de façon à l'expliquer en la contextualisant (au moment de l'écriture), et que de la même manière, cela en permettra la compréhension par contextualisation au moment de la lecture [PRI 99b]. C'est la tâche de l'utilisateur qui seule permet l'interprétation de l'annotation et l'exploitation subséquente du document. L'indexation conceptuelle est alors également contextuelle, écriture sur le document à l'aide des éléments disponibles qu'il s'agira de structurer en vue d'utilisations variées : recherche, navigation, génération, analyse, etc. Les inférences sont également contextuelles, puisqu'on recherche toujours dans le graphe des Strates-IA des éléments à partir d'éléments connus, sous la forme d'isomorphisme de sous-graphe partiel [PRI 00a]. Les travaux en cours concernent 1/ l'extension des Strates-IA à d'autres types de documents, 2/ la description des Strates-IA à l'aide de RDF 3/ la gestion de la base de connaissances et des schémas de description comme pouvant évoluer, les éléments d'annotation n'ayant pas *a priori* de sémantique fixée, celle-ci découlant plutôt de leur exploitation contextuelle.

4. Discussion

La dernière partie de cet article est consacrée à la discussion de quelques questions liées à l'indexation conceptuelle, entre autres dans sa perpective d'annotation.

Une première remarque concernera l'objectivité supposée des descriptions que l'on retrouve le plus souvent dans les travaux sur l'indexation conceptuelle : il s'agirait de décrire *le* contenu du document. Ces travaux font peu de cas de l'interprétation et du caractère situé de la tâche d'annotation. Pourtant, il apparaît que toute annotation, comme interprétation, consiste justement en la caractérisation des traits de signification du document, c'est à dire en la définition des unités signifiantes et en la détermination des traits sémantiques attachés, et donc relève le plus souvent de choix liés à la pratique en cours.

Il nous semble par conséquent qu'il est nécessaire d'accorder dans un modèle d'annotation une grande liberté générale, tout en disposant de possibilités de contrôles particuliers, partageables, qui éclaireront des parties de celles-ci. Par exemple, en supposant un modèle général d'indexation conceptuelle en XML, des DTD pourront représenter des explications éclairant certains domaines de la

description. D'autres parties ne pourront qu'être explorées en se basant sur le fond sémantique commun que constitue la langue, tandis que des ontologies formelles pourront permettre de mener des inférences sur les descriptions, correspondant à des domaines de réalité modélisés. A la suite de nos travaux sur les Strates-IA, nous considérons les « conteneurs de connaissances » (thesaurii, terminologies, ontologies) comme des organisations de termes dont la sémantique est plus ou moins fixée. Alors, nous proposons de voir les connaissances comme de simples réseaux de termes (éléments d'annotation), tandis que la conceptualisation, c'est à dire la réalisation des concepts se fait de façon contextuelle, par intégration dans des réseaux sémantiques, la notion d'ontologie ou de sémantique formelle ne se concevant alors que dans un rapport de localité par rapport au réseau global [PRI 00b]. Tous les schémas de description s'expriment de la même manière, sont normatifs à divers degrés, et peuvent être mis en relation librement. C'est le principe de contextualisation (guidée par des schémas partagés, mais qui peut également être libre) qui permet de préciser le sens des termes en contexte au cours de l'annotation comme à la lecture/exploitation des annotations. On peut remarquer qu'on se situe entre une approche formelle stricte contraignant les descriptions possibles (ce qui est cependant autorisé) et une approche libre dans laquelle toute annotation est possible.

Une seconde remarque concerne la compréhension des documents. Celle-ci ne peut se faire qu'en reconnaissant ceux-ci comme appartenant à des genres – plus ou moins liés à des genres existants hors de la machine, ou bien apparus avec les documents numériques – liés à d'autres documents par des liens explicites (ce qui permet de définir des corpus, par proximité dans le graphe des documents). Un des rôles de l'annotation est de contribuer, par la mise en place de relations intertextuelles à l'inscription formelle des documents dans un réseau de co-textes[21], à l'éclairage de ceux-ci, gage d'une compréhension accrue qui se rapproche et remet au goût du jour une certaine lecture savante liée a l'herméneutique.

Notre troisième remarque prend appui sur nos travaux dans le domaine audiovisuel. En effet, il nous semble que l'étude des documents non textuels pourrait peut-être réussir à mettre en place ce qui a été *évité* jusque là pour les documents à forte composante textuelle, à savoir remettre en cause la linéarité de la description et l'arbre sous-jacent hérité des méthodes d'écriture sur papier. Etudier des documents n'ayant pas de structures de connaissance fixées *a priori* autorise en effet à mettre en place des descriptions qui ne leur sont pas obligatoirement redevables. Alors, le statut de connaissance de n'importe quelle annotation, y compris « simplement » structurelle prend tout son sens. Enfin, par un retour de situation, on peut imaginer que les documents à forte composante textuelle seront étudiés à nouveau au regard de descriptions « libérées » de l'arborescence (c'est à dire en graphes). Si pour de

[21] Ce que [KAN 99] appelle la *structure fonctionnelle* du document, qui est liée au calcul (par exemple les liens hypertextuels), et articule le mode et le type de la filiation intertextuelle, permettant au document électronique de faire sens au sein d'une communauté de documents, dans un corpus (en dépassant les structures internes du documents qui ne peuvent rendre compte de l'interprétation en vertu de seules lois de compositionalité de sens arborescentes).

nombreux types de documents on se rend compte que les descriptions hiérarchiques ne sont pas suffisantes, on peut de droit se poser la question de leur dépassement pour les textes eux-mêmes.

Notre dernière remarque concernera la notion d'*unité documentaire* : si les annotations font partie d'un réseau documentaire, dans un graphe semi-structuré de connaissances pouvant transcender les type intérieur/extérieur au document, elles font aussi partie de celui-ci. Alors l'unité documentaire devient aléatoire et liée à la tâche d'utilisation du document, c'est à dire à la façon de considérer le réseau à un moment donné. On peut ici trouver des liens avec la génération automatique de documents, par exemple dans des systèmes d'EIAO de deuxième génération, pour lesquels il s'agit de fournir des parcours personnalisés à l'apprenant dans des documents générés automatiquement en fonction de ses objectifs et de ses connaissances (hypermédias adaptatifs). Il devient alors nécessaire de normaliser les briques élémentaires, ou unités d'information [CRO 00] [HER 00] autosuffisantes, de gérer ces dernières ainsi que leurs organisations possibles dans des systèmes fondés sur les connaissances. La problématique des éléments documentaires, de leur unité documentaire, de leurs relations qui se pose dans le domaine des hypermédias éducatifs est du même ordre que celui que nous évoquions. Le statut effectif de document lui étant donné par la cohérence d'ensemble de sa forme d'appropriation, laquelle découlait en général de l'origine humaine directe du document, il y a lieu désormais d'évaluer celle-ci en cours d'exploitation. L'indexation conceptuelle des documents peut en effet conduire ceux-ci à changer de type, et à la création de nouveaux genres documentaires qui trouveront ou non stabilisation sociale.

Dans le même ordre d'idée, on peut finalement se poser, comme [KAS 00] la question de la généralisation et de l'intégration de toute donnée dans la notion de document pourvu qu'elle soit descriptible dans un langage de balises. Sans trop nous engager sur cette voie, nous proposerons simplement de considérer tout élément documentaire autosuffisant (document, partie de document, ou encore « extrait » de connaissances) comme trace de connaissance, qu'il est possible d'organiser dans des systèmes plus vastes, à l'aide de liens de cohérence documentaire, d'intertextualité ou d'interopérabilité dans le cadre de tâches d'exploitation (par exemple une communication). Toute connaissance s'inscrit alors sous une forme donnée dans un support informatique, sous la règle d'un genre, de conventions formelles et d'indices en permettant l'interprétation, qu'il s'agit de transmettre d'une manière ou d'une autre. Dans ces conditions, alors il nous semble possible que la notion de document même revienne à sa définition originelle de *trace* (désormais numérique) *d'une action humaine.*

5. Conclusion

Dans cet article exploratoire, nous nous sommes d'abord intéressés à l'indexation conceptuelle, dont nous avons justifié l'intérêt actuel, que nous avons définie comme connaissance symbolique permettant l'exploitation de documents, ce que nous avons

illustré par la présentation de travaux récents sur la description de documents fondée sur les connaissances. Nous avons constaté que la notion d'index était finalement opératoire (tout ce qui peut servir d'index en est un de droit dès qu'il est utilisé en tant que tel), et que les structures documentaires fournissaient une indexation des documents (ce qui permet de droit de les utiliser). Nous avons alors pu argumenter en faveur de structures documentaires toutes sémantiques (dont l'éclairage dépend de l'exploitation au cours d'une tâche d'utilisation du document) et défendu la cause des graphes contre des structures arborescentes qui nous semblent trop limitatives. Dans une deuxième partie, nous avons proposé de considérer qu'une indexation conceptuelle était mise en place dans le cadre d'une annotation, concept qui nous semble plus adapté à une tâche *de documentation de sa pratique d'exploitation d'un document* et en fourni un cadre général. Après une discussion de l'annotation comme transmission de connaissances, nous avons présenté quelques exemples représentatifs de la recherche actuelle au niveau de l'annotation, et donné quelques pistes concernant les possibles évolutions des documents dans le contexte d'indexation conceptuelles en autorisant de nouvelles utilisations.

Nous conclurons cet article en évoquant les perspectives de recherche ouvertes autour de la notion d'annotation. Il nous semble en premier lieu que les communautés de recherche concernées par l'annotation sont variées, et que leur travail interdisciplinaire, impératif d'un point de vue scientifique, le devient encore plus au regard des enjeux économiques. Ainsi, au moins la sociologie, l'ergonomie cognitive, l'informatique documentaire, l'ingénierie des connaissances, la recherche d'information sont concernées par la mise en place d'annotations fondées sur les connaissances et partageables entre communautés. Il s'agit par exemple d'étudier comment placer une annotation, la localiser, la diffuser, l'exploiter, dans le cadre de tâches et de domaines variés (tout le monde manipule, utilise des documents, met en relation des documents). Les aspects dynamiques de l'annotation, notamment les rapports entre connaissances d'annotation (vocabulaire, schémas d'annotation) et appréhension du document (notamment sa séquentialité, et son éventuelle temporalité), sont par exemple à étudier, et certains projets sont déjà démarrés[22].

Sans aller jusqu'à adopter la vue de [GAN 99][23] pour lequel beaucoup des problèmes qui se posent désormais ne dépendent plus des informaticiens mais des chercheurs des Sciences Cognitives (modélisation des phénomènes collectifs, vision, mémoire et modes de représentation, linguistique, ergonomie cognitive), reconnaissons que l'interdisciplinarité est désormais inévitable. L'annotation

[22] Citons à titre d'exemple le projet KDI lancé à l'Université de San Diego : A Distributed Cognition Approach To Designing Digital Work Materials For Collaborative Workplaces (http://hci.ucsd.edu/lab/projects.htm) « …Work materials become elements of the cognitive system itself, and cognition becomes an emergent property of the interactions among people and work materials. We will apply our integrated approach to developing a distributed cognition based theory of annotation and explore a range of application domains …»

[23] « L'informaticien qui a remplacé l'architecte dans la conception des supports matériels de connaissances, doit, à son tour, laisser la première place au cogniticien. »

conceptuelle de documents participe de ce mouvement, de même que leur probable évolution, ainsi que celle du Web sémantique vers une approche réticulaire totale qui mime numériquement le monde tel que nous l'interprétons. Nous espérons que cette courte et lacunaire étude – qui mérite désormais des extensions théoriques nombreuses – en aura convaincu le lecteur.

6. Bibliographie

[AUF 99] AUFFRET G., PRIE Y., « Managing Full-indexed Audiovisual Documents: a New Perspective for the Humanities », *Computer and the Humanities, special issue on Digital Images*, vol. 33, n° 4, 1999, pp. 319-344, Kluwer Academic Publishers.

[BAC 99] BACHIMONT B., « Bibliothèques numériques audiovisuelles : des enjeux scientifiques et techniques », *Document Numérique, Numéro spécial Les bibliothèques numériques*, vol. 2, n° 3-4, 1999, pp.219-242.

[BER 99] BERNERS-LEE T. « Challenges of the second decade », in World Wide Web Conférence WWW8, Toronto, may 1999.

[BIR 00] BIRD S., LIBERMAN, M. « A formal framework for linguistisc annotation », A paraître dans *Speech communication*, 2000, 31 p.

[CHI 99] CHIKH A., TAZI S. « Annotations structurées d'aide à la reconstruction de documents, le modèle ASARD », in *Conférence Internationale sur le Document Electronique CIDE'99*, Damas, Juillet 1999, pp. 205-218.

[CRO 00] CROZAT S., TRIGANO P., « Une démarche de conception pour les hypermédias pédagogiques : l'enjeu d'une approche centrée sur l'information », in *Ingénierie des Connaissances IC'2000*, Toulouse, Mai 2000, pp. 25-34.

[DEN 00] DENOUE L., VIGNOLLET F., « An annotation tool for Web browsers and its applications to information retrieval », in *RIAO'2000 Content-based Multimedia Information Access*, Paris, apr. 2000, pp. 180-195.

[ERD 00] ERDMANN M., STUDER R., « How to Structure and Access XML Documents With Ontologies », à paraître in *Data and Knowledge Engineering, Special Issue on Intelligent Information Integration*. 21 p.

[EUZ 99] EUZENAT J., « La représentation de connaissances est-elle soluble dans le web ? », in *Document numérique, numéro spécial Gestion des documents et gestion des connaissances*, vol. 3, n°3-4, pp.151-167.

[GAN 99] GANASCIA J.-G., « L'architecte, l'ingénieur et le cogniticien », in *Document numérique, numéro spécial Gestion des documents et gestion des connaissances*, vol. 3, n°3-4, pp. 337-348

[GUA 99] GUARINO N., MASOLO C., VETERE G., « OntoSeek : Content-Based Access to the Web », in *IEEE Intelligent Systems*, May/June 1999, pp.70-80.

[HEF 99] HEFLIN J., HENDLER J., LUKE S., « SHOE : A Knowledge Representation Language for Internet Applications », Technical Report CS-TR-4078 (UMIACS TR-99-71), Dept. of Computer Science, University of Maryland at College Park. 1999, 30 p., http://www.cs.umd.edu/projects/plus/SHOE/pubs/techrpt99.pdf

[HER 00] HERAUD J.-M., MILLE A., « PIXED : vers le partage et la réutilisation d'expérience pour assister l'apprentissage », A paraître *In Conférence sur le Technologies de l'Information et de la Communication dans les Enseignements d'ingénieurs et dans l'industrie TICE'2000*, 2000.

[ISA 00] ISARD A., MCKELVIE D., MENGEL A., BAUN MØLLER M. « The MATE Workbench – A tool for annotating XML corpora », *in RIAO'2000 Content-based Multimedia Information Access*, Paris, apr. 2000, pp. 411-425

[LIB 99] LIBOUREL, MOUGENOT I., SALLANTIN J., SPERY L. « Meta-data and biological sequence annotation », in *META-DATA'99 : 3rd IEEE META-DATA Conference* , Maryland, USA, April 6-7, 1999.

[KAN 99] KANELLOS I., « A propos de l'héritage critique du document électronique : multimodalité sémiotique, stratégies de lecture et intertextualité », in *2e Colloque Int. sur le Document Electronique CIDE'99*, Damas, Syrie, jul. 1999, pp. 97-109.

[KAS 00] KASSEL G., « Construction et exploitation d'une ontologie pour la gestion des connaissances d'une équipe de recherche », in *Ingénierie des Connaissances IC'2000*, Toulouse, Mai 2000, pp. 251-259.

[KIF 90] KIFER M., LAUSEN G., "FLogic: A Higher-Order Language for Reasoning about Objects, Inheritance, and Scheme," in SIGMOD RECORD, Vol. 18, N° 6, Juin 1990, pp. 134-146.

[MAR 00] MARTIN P., EKLUND P., « Knowledge Retrieval and the Word Wide Web ». In *IEEE Intelligent Systems, special Issue on "Knowledge Management and the Internet"*, May/June 2000.

[MOT 00] MOTTA E., BUCKINGHAM SHUM S., DOMINGUE J. « Ontology-Driven Document Enrichment:Principles, Tools and Applications », in *International Journal of Human Computer Studies*, vol. 52, num. 5, pp. 1071-1109.

[NAN 95] NANARD J., NANARD M., « Adding macroscopic semantics to anchors in knowledge-based hypertext », in *International Journal on Human-Computerf, 43 :* 363-382, 1995.

[PRI 99a] PRIE Y., MILLE A., PINON J.-M., « A Context-Based Audiovisual Representation Model for Audiovisual Information Systems », in *Context'99, Second International and Interdisciplinary Conference on Modelling and using Context*, sept. 1999, Trento, Italy, Lecture Notes in Artificial Intelligence, vol. 1688, pp. 296-309.

[PRI 99b] PRIE Y., « Modélisation de documents audiovisuels en Strates Interconnectées par les annotations pour l'exploitation », Thèse de doctorat, Institut National des Sciences Appliquées de Lyon, dec. 1999, 270 p.

[PRI 99c] PRIE Y., MILLE A., PINON J.-M., « Connaissances et documents audiovisuels : un modèle pour l'exploitation contextuelle des annotations », *Document numérique, numéro spécial Gestion des documents et gestion des connaissances*, vol. 3, n°3-4, pp. 241-262.

[PRI 00a] PRIE Y., LIMANE T., MILLE A., « Isomorphisme de sous-graphe pour la recherche d'information audiovisuelle contextuelle», in *12ème congrès Reconnaissance de Formes et Intelligence Artificielle, RFIA2000*, feb. 2000, Paris, vol. I, pp. 277-286.

[PRI 00b] PRIE Y., MILLE A., « Reuse of knowledge containers: a "local semantics" approach in *Workshop on Flexible Strategies for Maintaining Knowledge Containers, 14th European Conference on Artificial Intelligence ECAI 2000*, Mirjam Minor (Ed.), Berlin, Aug. 2000, pp. 38-45.

[ROU 99] ROUSSEY C., CALABRETTO S., PINON J.-M. « Etat de l'art en indexation et recherche d'information », *Document numérique, numéro spécial Gestion des documents et gestion des connaissances*, vol. 3, n°3-4, pp.121-147.

[VIR 95] VIRBEL J., « Annotation dynamique et lecture expérimentale : vers une nouvelle glose ? », in *Littérature*, n° 96, 1995, pp. 91-105.

[ZAR 99] ZARRI G.-P., BERTINO E., BLACK B. et al. « CONCERTO, An Environment for the "Intelligent" Indexing, Querying and Retrieval of Digital Documents », in *Proc. of the 11th Int. Symp. ISMIS'99, Foundations of Intelligen*t *Systems*, LNAI 1609, Warsaw, Jun 1999, pp. 226—234.