Qualités d'une indexation portée par XML et une ontologie au regard d'un standard

Michel Crampes, Sylvie Ranwez, Michel Plantié, Christophe Vaudry

Qualités d'une indexation portée par XML et une ontologie au regard d'un standard

Michel Crampes - Sylvie Ranwez -Michel Plantié-Christophe Vaudry

Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P)
EMA - Site EERIE, Parc Scientifique Georges Besse
F-30 035 Nîmes cedex 1
{Michel.Crampes, Sylvie.Ranwez, Michel.Plantié}@ema.fr,
christophe.vaudry@up.univ-mrs.fr

RÉSUMÉ. L'indexation d'objets pédagogiques revêt des formes multiples. Nous posons le problème de la caractérisation des qualités d'une indexation dans le contexte de la conception de Documents Virtuels Personnalisables (DVP), et plus particulièrement la composition automatique de dispositifs pédagogiques. Nous présentons une grille d'analyse constituée des qualités attendues avec leurs conflits possibles. Nous montrons en théorie et sur une application concrète comment la compatibilité avec un standard (le LOM), et la prise en compte de XML complété de plusieurs constructions formelles (prédicats à deux termes proches de RDF, ontologies en DAML, sous-ensembles flous), peuvent contribuer à la recherche du meilleur compromis possible entre ces qualités.

ABSTRACT. Indexing pedagogical objects may have several and contradictory forms. We expose the problem of characterizing the qualities of indexing in the context of Adaptive Virtual Document (AVD), and more specifically, automatic composition of pedagogical curricula. The qualities are translated into an analytical grid with possible conflicts. Using theory and a practical example, we show how XML, along with the LOM standard and several formal structures in XML (RDF-like two term predicates, ontology in DAML, fuzzy sets) can be used to look for the best compromise between these qualities.

MOTS-CLÉS : Indexation, Tuteurs Intelligents, DVP, IMS, LOM, XML, RDF, DAML, Ontology.

KEY WORDS: Indexing, Intelligent Tutoring, AVD, IMS, LOM, XML, RDF, DAML, Ontology.

1 Introduction

Dans le présent article, l'indexation de documents numériques est vue comme une composante de la conception de Documents Virtuels Personnalisables (DVP). Un DVP est un ensemble d'éléments numériques, autrement appelés fragments d'information, associés à des mécanismes d'identification, sélection et assemblage sous contrainte(s) pour produire un document composite adapté aux besoins d'un utilisateur ou d'un groupe d'utilisateurs dans un contexte particulier [Ranwez et Crampes, 99 ; Iksal et Garlatti, 01]. Par exemple, un mécanisme capable de composer des résumés de match de football spécifiques pour le supporter d'une équipe à partir de vidéos de matchs est un DVP. Un système capable de produire des dispositifs pédagogiques personnalisés est un DVP. Nous englobons sous le terme "composition" ces mécanismes d'identification, sélection et assemblage. Notre équipe porte un regard transversal sur la conception de DVP, en cherchant des méthodes et des outils de composition communs à des domaines très variés : programmes de télévision, programmes de musique radio, interfaces multimédia adaptatives, construction de parcours pédagogiques personnalisés. Il est essentiel de noter que la composition ne se limite pas à la simple sélection des fragments, mais que les relations que ces fragments entretiennent entre eux dans le document composé jouent un rôle fondamental pour couvrir les différentes facettes du besoin et pour construire une entité cohérente et porteuse de sens.

Pour qu'un robot[1] puisse mettre en œuvre de manière efficace les mécanismes de composition qui donnent sa dynamique au DVP, il doit disposer d'une représentation interne des fragments d'information sur lesquels il travaille. En l'absence de capacité de reconnaissance d'image ou d'interprétation suffisante de textes, les fragments doivent être indexés. C'est dans ce contexte que l'article qui suit donne un cadre de réflexion sur les qualités attendues d'une indexation pour concevoir des DVP en général, et plus particulièrement dans le domaine de l'enseignement. L'enjeu pour l'enseignement se situe dans la construction de parcours individualisés et adaptatifs pour le E-Learning [Derycke, 02]. Nous appuyons notre réflexion sur les deux éclairages suivantes.

1) De 1997 à 1999 nous avons conduit un projet intitulé Karina, qui visait à concevoir un environnement de construction de parcours pédagogiques personnalisés [Crampes et al., 98a ; Crampes, 98c ; Crampes et Ranwez, 00]. A cette occasion, il nous a fallu imaginer un dispositif d'indexation qui par ailleurs devait pouvoir être utilisé dans d'autres domaines d'application des DVP (télévision, radio, interfaces multimédia). Il en est résulté un langage formalisé par une DTD (Document Type Definition) [Crampes et al., 99] pour l'indexation de fragments multimédia en XML, et des outils d'aide à l'indexation qui reposent sur l'utilisation d'une ontologie du domaine. Ce projet s'est accompagné de nombreuses interrogations sur la forme et le fond de l'indexation.

2) La DTD Karina nous est personnelle pour des raisons historiques. En effet il n'existait pas à l'époque une DTD standard. Cependant elle a été dès le début conçue pour être compatible avec un standard naissant autour d'une recommandation

[1] Nous appelons "robot" un ordinateur et ses programmes.

issue de l'industrie américaine : IMS[2] (Instructional Management Systems). A l'heure présente, nous suivons de près les travaux de standardisation autour du LOM [LOMd6.4 02] menés par le Groupe de Travail LTSC de l'IEEE. Historiquement, nos travaux de recherche sur l'indexation de documents pédagogiques lors du projet Karina ont donc précédé ceux autour du LOM. Cependant, ce standard en cours de validation est l'héritier d'une part des recommandations du consortium IMS, ce dernier restant très actif outre-atlantique, et d'autre part du projet ARIADNE en Europe. Nous espérons ainsi, grâce à nos choix antérieurs de compatibilité ascendante avec IMS pouvoir réutiliser nos résultats et bénéficier des nouveaux acquis autour du LOM.

Nous nous proposons dans la suite de cet article de donner une grille d'analyse de nos interrogations, ainsi que quelques éléments de réponse, à partir de notre propre vécu durant le projet Karina, et à partir de la lecture que nous faisons du LOM. Cette lecture nous est propre car notre analyse est portée par la finalité de notre projet, à savoir la composition automatique de documents numériques personnalisés par un robot à partir de fragments documentaires indexés.

2 Principe d'économie et principe d'usage.

Si une structure standard d'indexation telle que le LOM vient à maturité, il reste cependant que l'indexation soulève de multiples interrogations dont certaines sont déjà perceptibles à la lecture de la version [LOMd6.4 02]. Nous en voyons en particulier deux grands types :

(i) le modèle est complexe et on peut déjà supposer que l'indexation d'un document pédagogique demandera beaucoup de travail et de compétences. Or la pratique actuelle de l'enseignement à distance, et plus généralement d'un environnement à large diffusion, montre que ce sont les solutions les plus économiques et les plus éprouvées qui s'imposent dans un contexte technico-économique très évolutif[3]. Nous parlerons ici d'un "principe d'économie". Que l'on regrette ou que l'on nie ce principe, il semble être têtu.

(ii) le LOM ne précise pas la forme et le contenu de l'expression du besoin lors d'une requête pour rechercher une ou plusieurs ressources pédagogiques. On peut supposer bien entendu que le standard découle, implicitement du moins, du vécu des auteurs en matière de manipulation de ressources pédagogiques. En effet, cette difficulté n'échappe pas à ces derniers car à terme l'objectif est "… de faciliter la recherche, l'évaluation, l'acquisition et l'utilisation des objets pédagogiques par les apprenants, les enseignants ou les processus logiciels automatisés" [LOMfr, 02][4]. L'indexation ne joue plus alors simplement un rôle d'aide pour l'échange de ressources pédagogiques, mais celui plus général de support pour la construction de dispositifs pédagogiques personnalisés soit par des humains, soit par des robots. Ce standard en cours d'élaboration n'est qu'une étape et son projet rejoint quelque part le

[2] http://www.imsproject.org/metadata/index.cfm
[3] Voir l'éclairage qui est fait sur ce sujet dans [Derycke, 02]
[4] [LOMfr, 2002] : Nous utilisons ici la traduction française du document [LOMd6.4 2002].

nôtre, la conception de Documents Virtuels Personnalisables, avec son lot d'interrogations sur ce que peuvent être les qualités d'une indexation dans ce domaine, en particulier quand le processus de traitement est un robot. Au cœur de ce projet, on voit donc poindre le problème de la valeur d'usage d'une indexation pour un traitement automatique. Nous parlerons d'un "principe d'usage", ou d'utilité.

A partir de ces principes d'usage et d'économie, il nous est possible de décliner un ensemble de qualités attendues d'une indexation. Une qualité relève plutôt de l'usage ou de l'économie, sans que sa contribution soit clairement tranchée pour l'un ou pour l'autre de ces principes.

3 Qualités d'une indexation

Nous proposons une classification en trois catégories des qualités qui peuvent être attendues d'une indexation. Nous définissons dans un premier temps ces trois catégories, puis nous justifions notre classification en soulignant les points communs des qualités appartenant à une même catégorie. Nous soulevons également certaines contradictions qui émergent de la volonté de concilier ces qualités entre elles.

3.1 Expressivité, technicité, exploitabilité

Catégorie 1. **Expressivité**. Cette catégorie englobe les qualités qui rendent compte d'une part de la capacité d'expression de l'indexation en regard du document indexé, et d'autre part de sa cohérence sémantique. Il s'agit ici de rendre compte des qualités intrinsèques du document dont l'indexation se fait le témoin.

- Fidélité : l'indexation doit représenter au plus près l'information qu'elle résume et éviter les interprétations subjectives.
- Complétude : elle doit représenter l'information contenue dans le document dans toutes les dimensions utiles au demandeur. On voit ici que la difficulté est de ne pas s'attacher à une demande, mais plutôt d'être capable de répondre à toutes les demandes possibles des utilisateurs potentiels.
- Objectivité : l'indexation ne doit pas refléter le point de vue de l'indexeur mais le document lui-même. L'objectivité demande une attention délicate de la part de l'indexeur. Par contre, une méta-interprétation du document peut aussi faire l'objet d'une indexation. Dans ce cas, cette interprétation doit pouvoir être tracée (qui est l'indexeur ? comment justifie-t-il son interprétation ? etc.). En effet, cette interprétation devient un document en tant que tel puisqu'elle rajoute de l'information originale sur le document.
- Consistance : l'indexation ne doit pas contenir de contradictions logiques ni de contradictions sémantiques. D'un point de vue logique, une affirmation et son contraire ne doivent pas coexister. Sur le plan sémantique, l'indexation doit respecter les définitions (règles) du domaine auquel le document se réfère. Ceci bien sûr suppose que le document lui-même respecte ces règles.

- Précision : elle doit éviter les ambiguïtés qui peuvent entraîner des mauvaises interprétations sur l'utilisation attendue du document et le sens que l'utilisateur peut être amené à extraire.
- Evaluabilité : une indexation doit pouvoir donner ses propres limites afin qu'un utilisateur ou un outil puisse juger de ce qu'il peut attendre du document, mais aussi de l'indexation elle-même (est-elle fidèle ? complète ? etc.).

Catégorie 2. **Technicité** : Cette catégorie contient les qualités indispensables pour un traitement automatique efficace des indexations afin de composer des dispositifs pédagogiques personnalisés.

- Accessibilité : l'indexation doit pouvoir être facilement accessible par des outils (indexeur automatique, robots, etc.).
- Calculabilité : l'organisation et la forme de l'indexation doivent permettre d'effectuer des calculs (recherche d'information, rapprochement conceptuel, éloignement conceptuel, inférences, etc.).
- Flexibilité : elle doit s'intégrer à un dispositif contextualisé du point de vue du contenu et du point de vue de l'activité de l'apprenant.
- Interopérabilité : une indexation doit pouvoir être traitée par des robots différents dans des contextes différents pour permettre la réutilisabilité requise par l'exploitabilité (voir la catégorie 3).
- Rigueur sémantique : elle rejoint les attentes de la catégorie 1. En effet, la qualité des calculs de composition dépend de la qualité sémantique de l'expressivité de l'indexation.

Catégorie 3. **Exploitabilité**. Cette catégorie contient les qualités qui relèvent essentiellement du principe d'économie. Il s'agit ici de mettre en valeur le document indexé pour une réutilisation maximale et opportune à moindre effort d'indexation.

- Concision ou automatisation : en l'absence d'outil entièrement automatique, l'indexation est lourde en mobilisation de ressources humaines. La forme de l'indexation et les outils d'aide à l'indexation ont pour but de rechercher cette qualité. Sans outils, l'indexation la plus économique est la plus concise.
- Lisibilité : cet aspect doit s'entendre comme la capacité pour un humain à lire l'indexation et à la comprendre. La lisibilité peut être améliorée à l'aide d'outils de représentation textuelle ou graphique.
- Réutilisabilité : Si le document pédagogique peut être réutilisé dans différents contextes (ce qui est le but recherché), l'indexation doit en témoigner et favoriser une utilisation optimale en fonction d'un besoin utilisateur précis ou d'une application précise. Cette qualité justifie l'investissement que l'indexation a nécessité. Le langage d'indexation doit donc être le plus standard possible, et l'expression sémantique la plus partagée possible. La réutilisabilité peut aller jusqu'à la "banalité", c'est-à-dire une capacité à s'intégrer dans des contextes les plus variés et contrastés.
- Granularité : dans la mesure où le document peut n'être utilisé qu'en partie, l'indexation doit permettre d'extraire la partie utile.
- Valorisation sémantique : la recherche de documents ou d'extraits par les robots exploite l'expressivité de l'indexation. La catégorie 1 est donc indirectement partie prenante dans l'exploitabilité du document.

– Evolutivité : Dans la mesure où l'état de l'art sur l'indexation et sur la construction de dispositifs pédagogiques personnalisés n'en est qu'à ses débuts, il convient de disposer de documents indexés qui pourront évoluer avec la technologie. Dans le cas contraire, ces ressources seront perdues.

3.2 Analyse des catégories

La première catégorie (expressivité) semble évidente quand on considère le rôle de l'indexation. Nous y trouvons des qualités qui réclament une indexation riche en contenu et en structuration sémantique. Elles conduisent à la complexité et s'opposent à celles de la troisième catégorie qui réclament à l'inverse de la concision. Toutes les qualités qui y figurent relèvent, dans l'absolu, de l'inaccessible, et l'indexeur doit savoir limiter son ambition. En effet, une indexation ne sera, par exemple, jamais complète car il est toujours possible de rajouter des détails contenus dans le document, ainsi que des compléments de lecture et d'interprétation du document. De même, la fidélité la plus poussée consisterait à reproduire le document en l'état, et donc à ne pas l'indexer. Certaines qualités peuvent aussi s'avérer contradictoires. Par exemple il n'est pas possible d'avoir plusieurs points de vue représentés et de conserver l'objectivité. La fidélité qui suppose le minimum de valeur ajoutée s'oppose à la complétude qui appelle à une information sans limite.

La deuxième catégorie (technicité) se rapporte à l'aspect technique de l'indexation. Elle renvoie aux possibilités de traitement de la méta-information : stockage, accès, échange, fusion, calculs sémantiques, etc. Ce sont ces capacités de traitement automatique qui permettent à un robot d'assembler un dispositif pédagogique personnalisé adapté au besoin de l'apprenant à partir des documents indexés ou d'extraits de ces documents.

La troisième catégorie (exploitabilité) regroupe les qualités qui privilégient la recherche du "minimum d'effort" lors du processus d'indexation pour le maximum de résultats en terme de mise en valeur et réutilisation du document indexé. Ici aussi émergent des contradictions et en conséquence certains compromis sont incontournables. Le minimum d'effort d'indexation, si l'indexation n'est pas automatique, invite à une indexation légère, à l'aide de quelques mots-clefs. Le maximum de réutilisation suppose des mots-clefs les plus génériques possibles. Mais alors l'indexation manque de précision et d'expressivité.

Les différentes méthodes d'indexation qui sont proposées dans la communauté scientifique privilégient souvent un axe au détriment des autres. Celles qui s'intéressent à la simplicité proposent plutôt l'usage de mots-clefs et une représentation dont la lecture est aisée. D'autres mettent en avant la représentativité et proposent des modes d'expression profonds et conceptuellement fondés [Prié, 99 ; Motta et al., 99]. Il s'agit d'un contexte d'usage savant et connaisseur du domaine, comme par exemple chez [Auffret, 99 ; Auffret, 00]. Les sections suivantes introduisent une démarche possible pour trouver un compromis entre ces différentes qualités dans le contexte de notre projet DVP.

4 L'apport d'XML

XML [W3C 98] est tout à la fois un langage permettant de créer des "balises" susceptibles d'indexer un document, un langage permettant de construire un autre langage (ceci est traduit par le terme "eXtensible") et un langage de structuration d'un type de documents avec le principe de la DTD[5] (Document Type Definition). Ses capacités pour indexer des documents et en particulier des documents pédagogiques lui donnent un rôle croissant dans la communauté [de La Passardière et Giroire, 01] et dans l'industrie.

Nous nous proposons d'étudier à partir de l'exemple du projet Karina quels apports et quels freins présentent XML et les langages qui en sont dérivés pour favoriser les qualités de l'indexation de documents pédagogiques dans le contexte de la construction de dispositifs pédagogiques personnalisés, ou, plus généralement, pour favoriser la composition de Documents Virtuels Personnalisables.

4.1 Apports d' XML pour la construction de dispositifs pédagogiques personnalisés. Etude de cas

L'objectif du projet Karina est la construction de dispositifs pédagogiques personnalisés pour l'enseignement à distance [Crampes, 98c]. Le principe de personnalisation consiste à fournir un dispositif qui est fonction des objectifs pédagogiques de chaque apprenant et de ses acquis. Chaque objet pédagogique entrant dans le dispositif répond soit à la recherche d'un ou plusieurs objectifs pédagogiques, soit à la recherche d'objectifs nouveaux liés à des pré-requis d'autres objets pédagogiques. Il s'agit donc d'une personnalisation dynamique du contenu.

Dès le début de ce projet, les soucis de "qualités" comme celles décrites au chapitre précédent ont été pris en compte. Le projet ayant débuté en 1997 et s'étant terminé en 1999, certains choix techniques qui paraissent assez évidents maintenant ne l'étaient pas au début du projet. Ils concernent en particulier l'utilisation d'XML pour indexer les objets pédagogiques et pour représenter la connaissance [Crampes, 98c]. Ces choix ont été effectués pour répondre au cahier des charges du projet qui peut se résumer ainsi.

Un apprenant ou un enseignant exprime un besoin de formation auprès d'un serveur de dispositifs pédagogiques personnalisables. Ce dernier prend en compte ce besoin pour rechercher des matériaux pédagogiques disponibles en ligne qui sont susceptibles de répondre aux objectifs, identifier les plus pertinents et les assembler en tenant compte des pré-requis de chaque ressource et éventuellement d'une contrainte de durée de formation imposée par le demandeur. On notera que le souci de pertinence des matériaux pédagogiques repose d'une part sur l'adéquation aux objectifs, et sur la recherche d'un temps imparti de formation. Cette définition large du cahier des charges renvoie à trois problématiques complémentaires : l'expression du besoin, l'indexation de matériaux (à laquelle nous nous intéressons ici), et la construction d'un dispositif sous contrainte de temps par un robot.

---

[5] Même si dans sa forme et dans sa précision (typage), la syntaxe d'une DTD est remise en question par XML Schema, le fond reste inchangé.

Sur un plan plus général, l'exploitabilité des matériaux pose comme exigence la recherche d'un langage d'indexation le plus standard, capable d'évoluer tout en restant pérenne.

4.1.1 La filiation IMS, pour répondre à certaines exigences d'exploitabilité

En l'absence d'une structure d'indexation déjà établie et reconnue en 1998, nous avons décidé d'en établir une qui réponde au mieux aux diverses exigences d'*expressivité*, de *technicité* et d'*exploitabilité*[6].

Cependant le consortium IMS[7] qui rassemble nombre d'acteurs internationaux importants dans le domaine des NTIC proposait une recommandation d'indexation de ressources pédagogiques sous forme de spécifications. Cette recommandation intègre déjà la structure du "Dublin Core" utilisée comme un standard aux USA pour l'indexation de documents en général. Nous avons alors conçu une structure d'indexation qui reproduit au mieux IMS en pariant sur la pérennité de cette recommandation. Nos choix se sont avérés heureux puisque par la suite IMS est devenue la base essentielle du LOM proposé comme standard au sein du IEEE. Cependant, en 1998, nous ne pouvions pas nous satisfaire d'un format en devenir alors que nos objectifs étaient d'aboutir rapidement à un environnement exécutable. Il nous a fallu en conséquence tout à la fois simplifier IMS sur certains points, et le préciser sur d'autres, en particulier en ce qui concerne la gestion de la granularité et la représentation de la connaissance.

4.1.2 Formalisation d'une structure d'indexation en XML pour répondre aux exigences de technicité et d'exploitabilité. La DTD Karina

Les exigences de "technicité" invitent à construire une structure d'indexation qui soit formalisée pour être calculable. Nous ne pouvions nous satisfaire d'IMS sous la forme d'une spécification. Le choix d'un langage formel s'imposait. Les exigences d'exploitabilité de leur côté nous ont conduits à prendre en compte XML dont la recommandation se stabilise et qui hérite du savoir-faire acquis sur SGML. Nous avons pu construire une structure d'indexation formelle en XML, une DTD, qui est représentative d'IMS tout en offrant un support de calcul suffisant pour être opératoire [Crampes et al., 99].

Le choix ne s'est pas fait sans regret pour des langages comme Scheme ou Prolog qui proposent des capacités de calcul et d'expression plus intéressantes, là où XML n'offre que des capacités de structuration et de balisage. Mais l'exigence d'exploitabilité était à ce prix. Ce choix qui peut paraître évident avec le recul ne l'était pas au moment où il a été fait.

4.1.3 Expressivité et exploitabilité de la DTD Karina

---

[6] Ces catégories d'exigence et leur déclinaison en "qualités" n'étaient pas explicites dans nos travaux antérieurs. L'analyse et la structuration de notre démarche sont postérieures et nous permettent à la fois de valider le travail effectué et d'en voir les limites pour mieux les dépasser ensuite.

[7] http://www.imsproject.org/

Compatibilité avec IMS-LOM[8] *:* L'essentiel des catégories IMS-LOM se retrouve dans la DTD Karina, et est explicitement reconnu comme tel dans des commentaires XML [Crampes et al., 99]. Le fait de calquer en partie la DTD Karina sur la structure IMS permet de bénéficier des réflexions et des résultats de cette communauté. De plus, le fait de corréler explicitement notre DTD avec IMS devait nous permettre d'envisager un portage éventuel d'une indexation Karina vers une indexation IMS (maintenant IMS-LOM) et réciproquement, quand IMS aurait une DTD, ce qui est maintenant le cas. A l'époque, nous ne savions pas par quel procédé cela était possible. Mais nous prenions date. Depuis lors, plusieurs solutions sont apparues possibles. Nous les détaillons dans la section suivante.

La structure de la DTD Karina présente d'autres aspects qui méritent d'être analysés en relation avec la recherche d'expressivité et d'exploitabilité d'une indexation.

La granularité du document est explicitement prise en compte là où IMS, puis le LOM ne font qu'indiquer la nature composite du document ("aggregation level"). En effet la racine de la DTD Karina présente quatre niveaux de qualification :

```
<!ELEMENT document_qualifie (description_editorielle,
                             global,
                             local*,
                             segment* ) >
```

La "description éditorielle" contient la trace des propriétés et droits attachés à l'objet. Elle réutilise au maximum les catégories IMS-LOM qui touchent à ces domaines. Les trois éléments suivants (global, local, segment) permettent d'indexer le document à trois niveaux différents, tout en utilisant une structure d'indexation identique. Exemple pour l'élément global :

```
<!ELEMENT global   (description_contenu,
                    description_pedagogique,
                    description_economique,
                    description_presentation,
                    description_karina) >
```

L'élément "global" permet d'indexer l'ensemble de l'objet pédagogique. Il correspond au niveau unique d'indexation dans IMS-LOM.

L'élément "segment" permet d'indexer des fragments à l'intérieur du document assurant ainsi un premier niveau de granularité. Cet élément répond bien au souci d'exploitabilité puisqu'il permet d'identifier et d'extraire l'essentiel d'un document par rapport à un contexte, autrement dit d'extraire le(s) fragment(s) de document le(s) plus pertinent(s). Nous avons fait largement usage de cet élément dans nos différentes applications tant pédagogiques que de télévision interactive afin d'extraire les parties les plus pertinentes d'un document.

Si l'on veut rechercher un niveau de granularité plus poussé, il est intéressant de considérer qu'un segment peut aussi contenir des segments. La DTD Karina ne le permet pas même si XML ne l'interdit pas. Il existe deux raisons à cela. La première

[8] Nous parlerons fréquemment d'IMS-LOM pour citer les spécifications actuelles d'IMS telles qu'elles sont prises en compte dans le LOM.

rejoint encore une fois le critère de “lisibilité”. Une indexation comportant des emboîtements est complexe. Nous avons préféré l’exclure. La seconde raison est plus fondamentale. La pratique dans les projets nous a montré que la segmentation a priori était à la fois difficile dans beaucoup de cas (comment déterminer le début et la fin d’un segment), et même malvenue dans certains cas. Nous avons analysé ce point dans [Crampes et al., 99]. On peut résumer le problème comme suit. Si nous nous situons dans le cas ou il faut composer un document en fonction d’une contrainte, par exemple une limite en temps, la segmentation a priori pose des problèmes car on ne peut jouer sur une certaine flexibilité des composants en fonction de leur intérêt sémantique. Il devient alors plus intéressant de considérer une segmentation dynamique où la longueur des segments est fonction de leur pertinence et du contexte. C’est dans ce but que nous avons envisagé l’élément “local”.

Celui-ci permet de repérer des évènements ponctuels à l’aide de balises sans extension (durée, …). En toute rigueur, ce type d’élément devrait être “vide” (“EMPTY”) pour témoigner de son caractère ponctuel. Mais l’homogénéité avec les autres éléments (global et segment) nous oblige à lui adjoindre des éléments “enfants”. Dans la pratique, il a été peu utilisé en l’absence d’un algorithme de segmentation dynamique sur lequel nous travaillons [Lemoisson, 01]. Il nous paraît toujours intéressant, même si son exploitation reste à définir.

4.1.4 Banalisation de l’objet pédagogique et balisage externe

Dès le début, il nous paraissait que, contrairement à une démarche très répandue, un objet pédagogique doit être vu comme un objet banal utilisé à des fins pédagogiques. Cette vision large a d’importantes conséquences car elle considère qu’il faut éviter au maximum de spécialiser des objets à des fins pédagogiques. L’exploitabilité doit s’entendre dans un sens très large. Finalement, des objets comme des reportages, des images, des reproductions d’ouvrages d’art, etc. peuvent être perçus comme des objets pédagogiques et peuvent être en totalité ou partiellement intégrés dans un dispositif. En conséquence, la DTD Karina a été définie pour indexer aussi bien des objets pédagogiques que des documents multimédias conçus à des fins autres que pédagogiques. En particulier, nous l’avons plusieurs fois utilisée sans difficulté pour des prototypes de télévision à la carte [Crampes et al., 99].

La conséquence de ce parti pris a été que nous avons considéré un document comme une entité non altérable dont l’indexation doit être disponible dans un fichier à part, voire une base de données. En fait, il est même possible de voir un document selon plusieurs indexations et d’utiliser à un moment précis celle qui paraît la plus adéquate.

4.1.5 Importation du LOM

Nous avons indiqué combien il serait précieux de pouvoir tout à la fois bénéficier d’une DTD standard comme IMS-LOM, et de disposer de sa propre DTD. En fait se pose ici la question récurrente de se conformer à un standard pour des raisons d’exploitabilité, ou bien de développer son propre format pour de multiples raisons

de simplification, de particularisation, etc. Ainsi [Michard, 99] conseille-t-il d'éviter des DTD trop générales.

Au moment où nous avons défini la DTD Karina, il n'existait pas de DTD IMS, mais seulement des spécifications. Le problème du choix ne se posait donc pas. Cependant, pour préserver l'avenir, notre modèle a calqué au plus près celui d'IMS, jusqu'à présenter explicitement des commentaires de références à ces spécifications. Si le choix se posait maintenant, le dilemme serait entier car il existe une proposition de DTD IMS-LOM[9]. En effet, IMS-LOM nous apparaît trop général, trop tourné vers la réutilisation de documents entiers, insuffisamment expressif, et peu susceptible de prendre en compte des matériaux non pédagogiques, comme par exemple des films ou des articles de journaux. Doit-on malgré tout le prendre tel quel pour des exigences d'exploitabilité, ou bien peut-on le prendre en compte en partie pour mieux cibler des objectifs particuliers de technicité et d'expressivité ? De nombreuses équipes feront bientôt sans doute face à ce dilemme qui n'est d'ailleurs pas propre à la communauté "ingénierie pédagogique".

Sur ce point, XML présente heureusement plusieurs facettes intéressantes. S'il existe déjà une DTD, il est possible de l'importer (section externe) en totalité dans une DTD spécifique (section interne) pourvu que l'on ne réécrive pas des noms d'éléments ou d'attributs. Cette solution est donc intéressante mais fort contraignante. La seconde solution consiste à l'introduire partiellement à partir d'une section externe faite d'entités paramètre ("parameter entities") pourvu que la section externe soit structurée en conséquence. Il est alors possible de réécrire les éléments importés. Mais la DTD IMS[10] n'est pas structurée en ce sens. Cette solution reste donc dépendante de choix stratégiques au niveau du LTSC.

La troisième solution, trop récente à l'époque pour être mise en oeuvre, consiste à utiliser des domaines de noms ("namespaces") pour distinguer ce qui relève de la DTD standard et de la DTD spécifique. Nous pourrions utiliser ce procédé dans le futur et revoir la DTD Karina dans cette direction.

Finalement, il est une quatrième solution élégante qui mérite d'être mentionnée. Dans la solution précédente, un document IMS-LOM reste dans un format différent de celui de la DTD spécifique, et il n'est donc pas possible de l'intégrer directement. A l'inverse, un objet pédagogique indexé avec la DTD spécifique n'est pas au format IMS-LOM et perd en exploitabilité (en particulier en réutilisabilité). Il est alors possible de définir une transformation XSLT [W3C, 99b] pour traduire une indexation en l'autre et une autre transformation pour la traduction inverse. La combinaison des solutions trois et quatre nous paraît actuellement le meilleur compromis. Mais cette recherche de compromis se justifie seulement s'il s'avère nécessaire de disposer d'une DTD plus "expressive" qu'IMS-LOM.

En effet, le choix d'XML et de IMS-LOM peut se résumer ainsi. L'accent n'est pas mis sur les capacités d'expression et de raisonnement, mais sur l'exploitabilité via la standardisation. Or un environnement d'auto-composition de documents pédagogiques suppose une forte capacité de représentation de la connaissance et de sa manipulation. La section suivante explore les limites d'XML et du LOM dans ce domaine et introduit notre représentation de la connaissance compatible avec RDF [W3C, 99a] toujours en partant de l'exemple du projet Karina. Le but est de

[9] http://www.imsproject.org/metadata/imsmdv1p2p1/imsmd_bindv1p2p1.html
[10] http://www.imsproject.org/metadata/imsmdv1p2p1/imsmd_bindv1p2p1.html#1208264

disposer d'un mode de représentation de la connaissance qui permette à un robot d'effectuer différentes opérations afin de composer un document personnalisé.

5 Un modèle de représentation de la connaissance en XML-RDF

Pour le projet Karina, un certain nombre d'opérations sur la connaissance avaient été spécifiées [Crampes, 98c ; Crampes et Ranwez, 00] à partir de l'étude d'un cycle de construction d'un dispositif pédagogique personnalisé basé sur le contenu. Il fallait pouvoir formaliser une opération comme le *calcul de la proximité conceptuelle* entre deux indexations représentant deux documents, *une opération de fusion conceptuelle* pour représenter un nouvel acquis pour un apprenant après consultation d'un document, *une opération de soustraction conceptuelle* pour actualiser les objectifs de formation, une opération de *réduction conceptuelle* qui consiste à supprimer les doublons dans une indexation, etc.[11]

Il nous faut disposer d'un mode de représentation de la connaissance qui permette à un robot de mettre en œuvre les opérations conceptuelles voulues (exigences de technicité), tout en respectant au mieux les exigences d'exploitabilité pour assurer la dimension économique de l'indexation, et d'expressivité pour représenter au plus près la connaissance. Nous introduisons progressivement ici le formalisme de représentation de la connaissance que nous avons élaboré. Un exemple simple d'indexation est donné à la section 5.2.

5.1 Représentation de la connaissance en XML adaptée aux opérations conceptuelles. Prise en Compte de RDF, RDFS, DAML+OIL et de l'imprécision

Les deux modes les plus courants d'indexation du contenu sont les mots-clefs, et le langage naturel. Le second mode est difficilement compatible avec un traitement de la connaissance par ordinateur du moins compte tenu de l'état de l'art. Nous l'avons dès le début écarté, bien que l'élément "description" permette ce type d'annotation conformément encore une fois à IMS. Nous avons alors exploré l'indexation du contenu à l'aide de mots-clefs (ici aussi conformément à IMS), puis des modes de description plus complexes que nous verrons par la suite, pour finalement en arriver à un mode de représentation à l'aide de triplets (dans le style RDF), doublés d'une pondération. Dans tous les cas, la présence d'une ontologie paraît intéressante, voire indispensable.

5.1.1 Représentation par mots-clefs

Comme il est coutumier dans la communauté "information retrieval", IMS puis le LOM proposent une indexation du contenu par mots-clefs. Le risque d'ambiguïté est levé par une référence à une (des) taxonomie(s). Cette approche présente l'avantage de la simplicité tant du point de vue des indexeurs, que du point de vue de l'ordinateur. En effet, les opérations décrites ci-dessus deviennent des opérations

[11] Nous ne détaillons pas ici ces opérations, l'indexation restant le centre de notre propos.

ensemblistes. Cependant l'expressivité est très limitée à deux titres. (i) Tous les mots-clefs ne sont pas identiquement représentatifs du contenu. Ceci est explicitement reconnu par le LOM puisque l'explication sur les mots-clefs conseille de placer en premier le mot-clef le plus "pertinent" ("most relevant first "). Ce degré de représentativité ne peut pas être pris en compte par un moteur sans autre indication de calcul. (ii) Se contenter de simples mots-clefs limite considérablement l'expressivité de la représentation. Par exemple, si un document pédagogique cite Verlaine, introduire le mot-clef 'Verlaine' apporte beaucoup d'ambiguïté (est-ce une citation de Verlaine ? parle-t-on de Verlaine ? Verlaine est-il le thème central ? etc.). Une indexation lexicale limite fortement la portée expressive. Nous nous sommes alors tournés vers une représentation plus riche, à savoir les graphes conceptuels.

5.1.2 Représentation à l'aide de Graphes Conceptuels

Les Graphes Conceptuels [Sowa, 84] présentent tout à la fois un fort degré d'expressivité conceptuelle et des capacités de calcul importantes eu égard aux opérations souhaitées. En effet, ils peuvent être assimilés à une représentation en logique des prédicats d'ordre 1 en particulier avec leur prise en compte du lambda calcul[12]. Plusieurs équipes de recherche utilisent ce formalisme pour indexer des matériaux sémantiquement riches comme des films.

Pour l'exemple, la phrase tirée d'Alice au pays des merveilles de Lewis Carol "Alice looks at the Caterpillar for some time in silence" peut être exprimée à l'aide du graphe conceptuel suivant[13] :

```
CG0: [LOOK_AT] -
      (AGNT)->[LITTLE_GIRL:Alice]
      (OBJ)->[CATERPILLAR:#]
      (MANR)->[SILENTLY]
      (DUR)->[TIME-PERIOD:#]
```

Dans ce graphe conceptuel, le concept "regarder" ([LOOK_AT]) indiqué entre crochets est en relation avec l'agent "Alice" de type "petite_fille" ([LITTLE_GIRL:Alice]) avec l'objet de type "chenille" dont le référent est indéterminé ([CATERPILLAR:#]), de manière "silencieuse" et durant une "période de temps" indéterminée.

Ce type de graphe peut servir à décrire tout autre élément d'information repéré par un indexeur, y compris en matière de meta-meta-connaissance. Il est par exemple possible de représenter l'information : « le document [Crampes, 97] prend pour exemple la phrase "Alice looks at the Caterpillar for some time in silence" tirée de l'ouvrage de Lewis Carol "Alice in Wonderland" ».

5.1.3 Difficultés d'une représentation avec les Graphes Conceptuels

---

[12] Nous ne détaillerons pas ici des représentation possibles en KIF ou en SCHEME que nous avons aussi envisagées, mais qui présentent encore plus de problèmes qu'une représentation à l'aide de Graphes Conceptuels [Ranwez et al., 00a] [Ranwez, 00b].

[13] Selon notre approche, tout document peut avoir valeur pédagogique dans un dispositif. L'exemple choisi ici peut être considéré comme un exemple d'indexation en vue de l'insertion d'un fragment d'*Alice au Pays des Merveilles* dans un dispositif pédagogique.

Les graphes conceptuels ont un fort pouvoir d'expressivité, mais ils pêchent sur plusieurs points, ce qui ne surprendra pas puisque l'on sait que la recherche d'une qualité met souvent à mal une autre ou plusieurs autres qualités. Nous citons quatre difficultés qui nous sont parus les plus pénalisantes.

Ils sont relativement compliqués à mettre en œuvre et sont peu sortis d'un cercle d'initiés. Ceci va à l'encontre de l'exploitabilité.

Des outils ont été construits, mais ils portent pour l'essentiel sur le dessin des graphes et leur manipulation logique. Nos opérations ne relèvent pas de l'inférence logique et ces outils ne sont donc pas susceptibles de les mettre en œuvre.

La troisième difficulté provient de l'absence d'un formalisme XML des graphes conceptuels au moment où nous avons dû définir un mode de représentation de la connaissance.

Finalement, un graphe conceptuel n'introduit pas des degrés de représentation de la connaissance même si certains travaux ont exploré le domaine [Ho, 94].

5.1.4 Des graphes conceptuels à RDF

Le besoin de simplifier la saisie et la manipulation de graphes conceptuels nous a conduit à les décomposer en un ensemble de triplets correspondant à des prédicats à deux termes. Ainsi, le graphe présenté dans l'exemple ci-dessus peut se réécrire :

```
[LOOK_AT] -
      (AGNT)->[LITTLE_GIRL:Alice]
      (OBJ)->[CATERPILLAR:#]
[LOOK_AT] -
      (AGNT)->[LITTLE_GIRL:Alice]
      (MANR)->[SILENTLY]
[LOOK_AT] -
      (AGNT)->[LITTLE_GIRL:Alice]
      (DUR)->[TIME-PERIOD:#]
```

Nous avons effectivement mis en œuvre cette décomposition qui s'est révélée très efficace pour l'indexeur et donc conforme à l'exigence d'exploitabilité[14]. Elle permet par ailleurs de retrouver le graphe conceptuel d'origine au travers des opérations sur ces graphes ("join", etc. [Sowa, 84]), ce qui assure qu'il n'y a pas de perte d'expressivité par rapport à des graphes conceptuels complets.

Mais les avantages de cette décomposition ne se limitent pas là. Elle peut facilement s'exprimer en XML sans faire appel à une structure formelle complexe ou une structure informelle du type PCDATA. Nous décrivons un "élément" triplet comme un composé de trois "éléments enfants" : sujet, verbe, objet.

Elle autorise de manipuler des prédicats simples comme ceux que l'on trouve souvent dans une ontologie. Ceci nous permet d'utiliser une ontologie du domaine

[14] Nous ne détaillons pas ici certains détails techniques comme la redondance des triplets visible sur l'exemple qui peut être gérée par l'utilisation de référents communs comme dans les graphes conceptuels ou en RDF.

comme aide à l'indexation pour peu que nous disposions d'un outil présentant cette fonctionnalité (voir section ci-dessous).

Finalement, le plus intéressant a été l'apparition de RDF [W3C 99a] et RDFS [W3C 02] pour exprimer la connaissance pour le Web Sémantique. Alors que notre principe de décomposition en triplets était antérieur à RDF, les auteurs de ce formalisme proposent aussi l'usage de triplets. Ainsi notre mode de représentation de la connaissance rejoint-il en grande partie le courant principal en vigueur au W3C. Ceci peut être le fruit du hasard, ou bien simplement une convergence de points de vue eu égard à une vision commune des qualités attendues d'une indexation de contenus.

5.1.5 Introduction d'un support ontologique

Le principe d'utilisation des ontologies pour décrire des fragments de connaissance n'est pas nouveau [Gruber, 93 ; UTE]. L'apport d'une ontologie ou d'un thésaurus pour une tâche d'indexation est le plus souvent associé au maintien de la conformité à un vocabulaire et à des relations. L'objectif est alors de conserver la consistance globale de l'indexation et la cohérence avec d'autres documents indexés dans une perspective d'échange de documents [Weinstein et Alloway, 97 ; Weinstein, 98 ; Motta et al., 99 ; Domingue et Motta, 99].

Pourtant, le rôle d'une ontologie peut être vu de manière plus vaste. L'ontologie fixe un certain nombre de règles sémantiques générales d'hyponymie (relation "est une sous-classe de"), d'hyperonymie ("est une super-classe de"), de méréonymie (relation "se compose de"), et de relations spécifiques au domaine (causalité, temporalité, etc.). La disponibilité de ces règles joue le rôle de contraintes qui favorisent un certain nombre de qualités. [Memzies, 99] présente quatre avantages apportés par l'utilisation d'une ontologie. L'interopérabilité et la réutilisabilité sont à l'évidence des retombées importantes pour des documents indexés sur la base d'une même ontologie. La structuration rejoint notre critère d'économie. Celle-ci est favorisée parce que l'indexeur dispose déjà d'une connaissance formalisée du domaine sur lequel porte le document. Sa tâche se résume à instancier des concepts et des relations sans avoir à reconstruire un corpus de connaissances sur le domaine. La disponibilité d'un outil d'indexation utilisant l'ontologie renforce l'économie d'indexation. Le quatrième avantage présenté concerne la navigation et la recherche. Plus généralement, une ontologie permet de renforcer la calculabilité de l'indexation. Des agents intelligents peuvent utiliser les relations proposées par l'ontologie pour effectuer des inférences conceptuelles afin de ne pas limiter les capacités des moteurs à la seule lecture des annotations disponibles à l'intérieur d'un document.

L'ontologie permet aussi de favoriser l'évaluation d'un document et de l'indexation associée. Un moteur peut calculer les trous conceptuels, le niveau de redondance, le niveau de généralité d'un document. Ces calculs sont liés à l'analyse de la couverture ontologique, c'est-à-dire au nombre de concepts de l'ontologie instanciés dans le document.

Finalement, l'avantage de disposer d'une ontologie tant pour l'indexation que pour la composition est de pouvoir inférer des rôles pour les objets à composer à partir du contenu de leur indexation repéré dans le contexte d'une ontologie [Asselborn et al., 97]. Le problème des rôles est également soulevé dans [Kabel et al., 99] qui propose un système d'indexation de fragments de documents

électroniques en fonction de différents points de vue, à l'aide d'un ensemble d'ontologies; cependant on peut regretter le fait que les rôles attribués aux fragments soient figés. Dans notre approche nous étudions la possibilité d'automatiser l'attribution de rôles en fonction du contexte d'utilisation [Ranwez et al., 00a ; Ranwez, 00b].

Cependant l'utilisation d'une ontologie pour indexer un document présente aussi certains inconvénients. En premier lieu, une évidence s'impose : il faut disposer d'une ontologie. Deux solutions sont possibles. La première consiste à réutiliser une ontologie disponible en ligne[15] mais cela entraîne plusieurs difficultés [Motta et al., 99]. Il faut se plier au formalisme imposé par l'ontologie en ligne comme par exemple Ontolingua, ou KIF, et adapter ses outils à ces formalismes. De plus les ontologies importées sont souvent trop générales et trop vastes. L'information utile est noyée dans une masse d'informations annexes. La deuxième solution consiste à construire sa propre ontologie. Il est alors possible de cibler l'information ontologique, mais les qualités d'économie et de réutilisabilité ne sont plus respectées puisque l'ontologie et sa représentation sont propriétaires. Par ailleurs, les ressources et compétences humaines nécessaires à cette tâche sont importantes. De plus construire une ontologie représentant fidèlement le domaine concerné est toujours difficile. Comment en effet représenter des concepts qui peuvent être interprétés différemment entre les utilisateurs, ou quelle est la hiérarchie de concepts la plus objective ?

L'autre difficulté que présente l'indexation supportée par une ontologie est que l'approche est fortement conceptuelle et il n'est pas toujours facile de décrire des situations, des lieux ou des évènements avec des modèles conceptuels parfois complexes. Nous en voulons pour preuve la complexité des ontologies du sens commun que l'on peut trouver sur les sites mentionnés ci-dessus.

Notre réponse à ces difficultés à été de (i) réutiliser une ontologie si elle existe, (ii) construire notre propre ontologie dans le cas inverse mais en utilisant un langage le plus standard possible. Nous avons finalement construit plusieurs ontologies en utilisant dans un premier temps un formalisme XML spécifique basé sur les mêmes triplets que ceux utilisés pour l'indexation. Dès que cela a été possible, nous avons migré vers un début de standard, DAML+OIL [DAML 01], qui présente de plus l'intérêt d'utiliser RDF et RDFS.

5.1.6 Introduction d'une pondération

Ainsi doté d'une représentation équivalente aux graphes conceptuels et donc à la logique d'ordre 1, d'un langage proche d'un standard (XML/RDF) qui favorise l'exploitabilité, et d'un support ontologique à l'intérêt multiple, nous pourrions voir notre dispositif suffisant. Il lui manque cependant de pouvoir intégrer une part d'imprécision.

Un fragment de texte, une image, un fragment sonore, sont autant d'unités d'information qui peuvent être interprétées de multiples manières. Cette interprétation résulte d'une mise en situation d'un récepteur[16] face à un document, dans une circonstance particulière. Un élément d'information se trouve donc dans un

[15] http://www.cyc.com/
http://logic.stanford.edu/kif/kif.html

[16] nous appelons récepteur la personne qui prend connaissance d'un document multimédia.

état potentiel de significations multiples. La rencontre d'une requête dans un contexte particulier détermine un sens privilégié qui peut ne pas être le même que celui provoqué par une autre requête dans un autre contexte. Un indexeur ne peut prévoir toutes les situations, a priori infinies, dans lesquelles l'unité d'information va être utilisée. Vouloir tout couvrir est non seulement intrinsèquement impossible, mais de plus va à l'encontre de certaines qualités comme l'exploitabilité et l'objectivité. Il revient à une méthode d'indexation soit de reconnaître l'unicité d'interprétation d'une annotation, soit de favoriser la multiplicité des points de vue, quitte à gérer l'augmentation de complexité qui peut en résulter.

L'approche classique dans la communauté "Information Retrieval" consiste à affecter des poids à des mots-clefs représentatifs de leur fréquence d'apparition dans un document [Kantor, 94 ; Salton et al., 96 ; Zobel et Moffat, 98]. On peut voir cette méthode d'expression de l'imprécision de deux manières. Soit il s'agit d'une probabilité de présence d'un mot dans le document, soit il s'agit d'un point de vue subjectif qui du point de vue théorique peut être rattaché à la théorie des sous-ensemble flous. C'est la seconde vision que nous avons privilégiée étant donné la dimension conceptuelle de notre méthode d'indexation.

La théorie des sous-ensembles flous se compose d'un corpus d'outils mathématiques particulièrement appropriés pour le traitement d'informations imprécises [Kaufman, 75 ; Kaufman, 77 ; Dubois et Prades, 00]. [Nkambou et al., 97] dans son modèle CREAM (Curriculum REpresentation and Acquisition Model) utilise cette théorie pour la construction d'un curriculum. Dans le domaine de l'interprétation d'images [Morton et Popham, 87] propose une méthode d'intégration du flou dans les graphes conceptuels et un algorithme pour faire des opérations sur ces graphes. Ho fait aussi un parallèle entre les graphes conceptuels et la théorie des sous-ensembles flous [Ho, 94]. Dans son approche il précise que le meilleur moyen de déterminer les caractéristiques communes de deux concepts est de traiter chaque paire conjonctive (concept – relation – concept) séparément. Cela ne permet cependant pas de donner le support adéquat pour le processus de classification. Notre approche est voisine : nous utilisons également des paires conjonctives, car elles sont compatibles avec le format RDF. Nous y ajoutons une pondération des concepts afin d'introduire certains principes de la logique floue dans le traitement des connaissances.

Concrètement nous introduisons pour chaque graphe élémentaire des poids dans l'intervalle [0,1] que l'indexeur est libre d'instancier. L'ensemble constitue ce que nous appelons un Vecteur d'Etat Conceptuel (CSV) dont la justification a été donnée dans [Crampes,97]. L'indexeur peut faire apparaître autant de phrases élémentaires qu'il le souhaite avec des points de vue correspondant à des valeurs de pondération. Le principe des poids associés aux paires conjonctives présente d'autres intérêts. Il permet de mettre en œuvre des algorithmes d'apprentissage qui doivent personnaliser l'indexation selon le type d'utilisateur. Nous travaillons actuellement sur ce sujet. Il permet aussi d'utiliser les opérateurs présentés ci-dessus dans le cadre de stratégies d'optimisation quand il s'agit de construire un document sous contraintes, comme par exemple un cours en didactique ou un résumé d'un programme de télévision sous contrainte de temps [Crampes et al., 98a ; Crampes, 98b].

Ainsi la partie indexation de la connaissance avec des CSV consiste finalement, au vu d'un document ou de fragments d'un document (textuel, vidéo, son) à

construire un ensemble de paires conjonctives (appelées aussi triplets, ou prédicats d'arité deux) dans un formalisme XML proche de RDF. La section suivante montre comment l'indexation du contenu à l'aide de CSV écrits en XML peut être faite concrètement à l'aide d'un outil autour d'une ontologie du domaine.

5.2 Outil d'aide à l'indexation

L'indexation dans Karina consiste à donner des valeurs à des attributs qui représentent les différentes composantes de l'indexation telles qu'elles sont définies dans la DTD. Ce travail serait fastidieux, et donc le mode d'indexation ne répondrait pas à l'exigence d'exploitabilité, s'il n'était pas assisté par l'ordinateur. Nous avons en conséquence conçu et implanté un outil d'aide à l'indexation qui dissimule totalement les complexités structurelle, lexicale et syntaxique de la DTD et de la formalisation en XML. Par une sélection d'onglets, l'indexeur se positionne sur "l'élément de portée" (global, segment ou local) qui l'intéresse à un moment donné. Un menu lui permet ensuite de choisir les éléments (au sens XML) qu'il doit ou qu'il souhaite renseigner selon qu'ils sont obligatoires ou optionnels.

5.2.1 Indexation de la connaissance dans un contexte pédagogique

Parmi les différents aspects sur lesquels porte l'indexation, nous nous intéressons ici plus particulièrement à la représentation de la connaissance à l'aide de CSV. En effet ce point est le plus délicat étant donné le risque de complexité qui peut lui être associé. La figure 1 représente l'interface de l'outil dans le contexte du remplissage d'un CSV pour un cours sur le langage C. Le document à indexer est présent dans une autre fenêtre (non représentée sur la figure 1).

Après avoir sélectionné en haut à gauche l'ontologie du domaine dont traite le document à indexer, l'indexeur saisit les paires conjonctives de la description. Au milieu de l'écran, trois champs permettent de sélectionner (i) le premier concept de la paire conjonctive, (ii) la relation possible dans l'ontologie avec le concept précédent, (iii) le concept qui peut être associé au premier via l'ontologie par l'intermédiaire de cette relation. L'indexeur est ainsi assuré de ne pas faire d'erreur lexicale, syntaxique et même sémantique, puisque c'est l'ontologie qui fixe les contraintes sémantique, le cadre syntaxique et le vocabulaire. Des poids peuvent être affectés à chaque paire conjonctive, ainsi qu'à chaque élément de la paire. Cependant cette affectation manuelle de poids est lourde. Nous étudions la possibilité de l'automatiser en partie grâce à des calculs de couverture ontologique.

Les paires conjonctives sont ensuite traduites en langage naturel dans la fenêtre inférieure assurant ainsi une meilleure lisibilité. L'outil remplit bien là les rôles de facilitateur pour les qualités "consistance" (grâce aux contraintes ontologiques) et "lisibilité" (XML non visible, phrases en langage naturel)

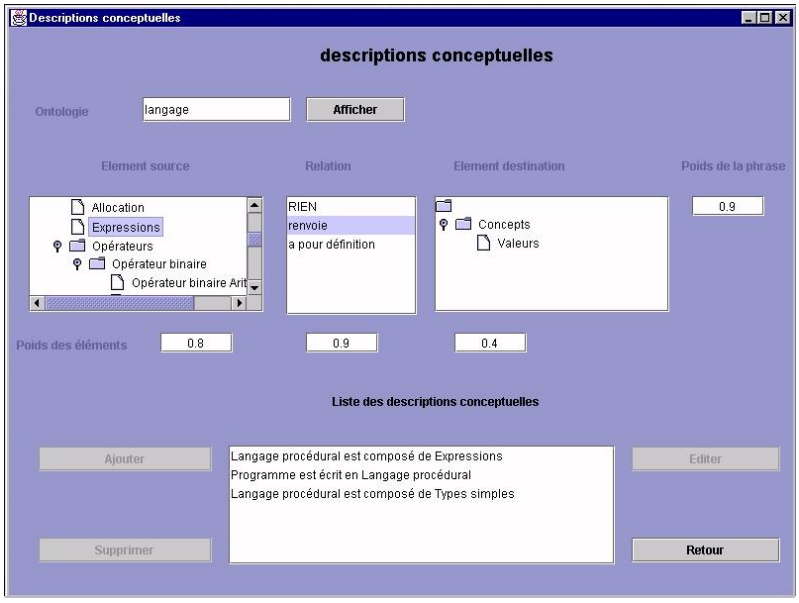

Figure 1 : *Interface d'indexation Karina*

L'exemple ci-dessus correspond à l'écriture d'un CSV pour un fragment de cours sur le langage C. On peut lire dans la fenêtre en bas au centre que ce fragment de cours parle de trois thèmes tandis qu'une quatrième assertion est en cours de construction et n'apparaît pas encore en bas : "un langage procédural est composé d'expressions" + "un programme est écrit dans un langage procédural" + "un langage procédural est composé de types simples". La pondération de chacune de ces trois assertions est par défaut de poids 1, contrairement à la quatrième assertion en cours de construction qui sera de poids 0.9 (en haut à droite) et sera visible à côté de la phrase lorsqu'elle sera validée. Nous analysons plus loin la qualité ergonomique de cet écran d'indexation au regard de la nature de la connaissance modélisée comparée à d'autres domaines comme la musique ou le sport.

5.2.2 Généricité de l'approche: indexation de fragments TV et de la musique

Plusieurs fois nous avons insisté sur le fait que les langages, méthodes et outils de conception de DVP devaient être autant que possible transversaux, c'est-à-dire applicables dans différents domaines. Nos travaux nous ont permis récemment de construire deux applications prototypes avec un industriel dans le domaine de la télévision et de la radio personnalisées. Il s'agit par exemple de pouvoir fournir des résumés de rencontres sportives selon les préférences des téléspectateurs ou bien un programme radio musical selon les préférences des auditeurs. On peut aussi imaginer qu'une séquence vidéo intervienne dans un dispositif de formation, ce qui justifie d'employer certaines méthodes transversales pour indéxer des documents pédagogiques et des documents vidéo.

Dans ce but nous avons utilisé les mêmes formalismes que pour la conception de dispositifs pédagogiques, avec pour base la construction de CSV et l'usage d'ontologies. Le moteur de composition opère cependant avec des algorithmes différents. En ce qui concerne la spécificité de l'indexation pour des fragments TV ou pour de la musique, il est apparu une difficulté majeure. Il nous faut pouvoir proposer une interface d'indexation pour un indexeur humain, ou *une interface d'expression des préférences pour un téléspectateur ou un auditeur*[17], qui soit ludique, tout en respectant le principe de la construction de CSV à partir d'une ontologie. Nous avons étudié plusieurs solutions. Nous présentons ici un exemple de l'interface utilisateur pour l'expression des préférences de l'utilisateur final. L'ontologie sélectionnée ("Domaine") est celle du football (en haut, à gauche). Les termes de l'ontologie sont le mieux possible représentés par de petites icônes (en haut) qui, sélectionnées, renvoient à des sous-arbres de l'ontologie (fenêtre en bas à gauche) et à des contraintes sémantiques exprimées par des triplets. Ici, la phrase "le capitaine contre attaque" a été sélectionnée par navigation dans les icônes et les sous-arbres (fenêtre en bas à droite). Le poids n'est pas pour l'instant pris en compte. Il fera l'objet d'une fonction d'apprentissage en liaison avec des stéréotypes d'utilisateurs.

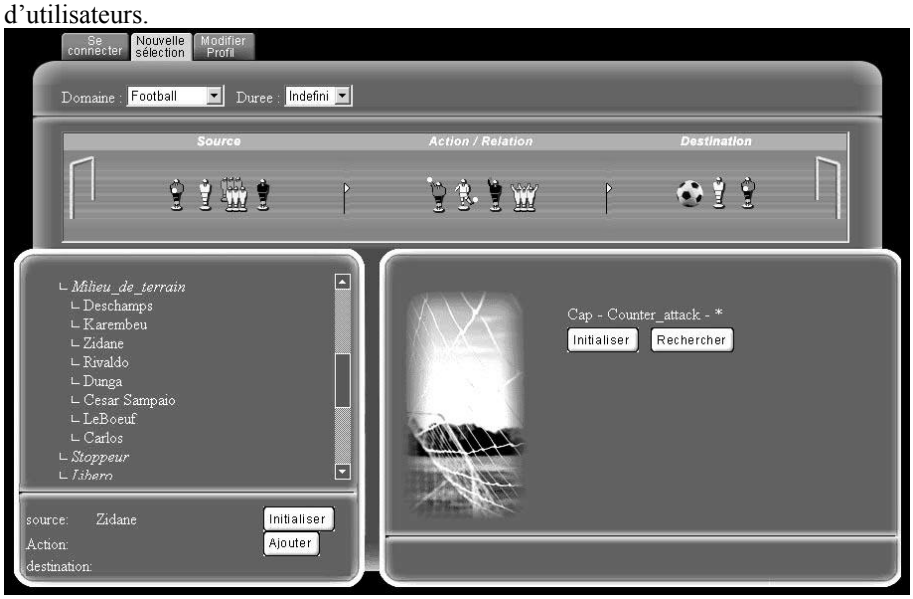

Figure 2 : *Interface pour l'indexation ou l'expression des préférences TV*

L'application radio musicale a aussi été développée sur les mêmes bases. La généralisation de l'approche semble pour l'instant en bonne voie. On notera cependant la difficulté de construire des interfaces d'indexation ou d'expression du besoin qui soient faciles d'usage et ludiques alors que l'on oblige l'utilisateur

[17] En effet nous demandons aussi que l'interface d'indexation puisse servir à un utilisateur final pour exprimer ses préférences à l'aide d'un CSV. Nous ne détaillons pas ici les raisons de ce choix exigeant. Le lecteur peut cependant imaginer qu'on a là les bases d'une navigation dans des DVD familiaux pour des applications futures.

(indexeur ou utilisateur final) à naviguer et faire des choix dans des ontologies, espaces abstraits par excellence. La section suivante discute ce point.

6 Conclusion : Qualités des méthodes, langages, et outils d'indexation au regard de la nature de la connaissance et du type d'application

6.1 L'approche qualité de l'indexation pour les DVP

Résumons notre démarche. Nous avons proposé une grille d'analyse de la qualité d'une indexation de documents numériques dans le cadre de l'indexation de fragments documentaires pour la conception de Documents Virtuels Personnalisables. Cette grille met en relief trois catégories de qualités. L'expressivité concerne les qualités de l'indexation pour rendre compte au mieux du contenu du document. La technicité concerne les qualités qui rendent compte de la capacité d'un moteur de composition à effectuer au mieux son travail en prenant pour base les indexations produites. L'exploitabilité concerne la dimension économique au sens large d'un processus d'indexation. On peut résumer cette dernière ainsi. A défaut d'une indexation automatique, indexer est une tâche fastidieuse, consommatrice de temps, et très subjective. Il faut en conséquence pouvoir indexer facilement, réutiliser au maximum les indexations déjà produites ou à défaut des outils d'indexation, et échanger au maximum le fruit de son travail en s'assurant qu'on parle le même langage.

Pour répondre à ces spécifications de qualité, tout en sachant qu'il est nécessaire de faire des compromis entre des qualités contradictoires, nous avons mené nos recherches selon les axes suivants.

6.2 A la poursuite de standards : XML, IMS, LOM

La technicité nous a conduit à construire un modèle d'indexation qui nous permet d'obtenir les informations nécessaires à un moteur pour composer des documents à partir de fragments indexés. Deux grandes classes d'information peuvent être identifiées : les informations de service (nature du document, auteur, média, rôle pédagogique pour un document pédagogique, granularité, etc.) qu'il est possible de fortement structurer à l'aide d'un schéma, et les informations de contenu qu'il est difficile de structurer, problème classique de représentation de la connaissance. Les soucis d'exploitabilité nous ont fait choisir la recommandation XML pour structurer l'ensemble de l'indexation. En poussant plus loin, ce même souci nous a amenés pour une application pédagogique à construire une structure d'indexation la plus proche possible d'une recommandation en devenir, IMS. Ce dernier choix s'est avéré opportun puisque le LOM, le standard en devenir, est l'héritier direct d'IMS. Il est important de noter que les technologies autour de XML permettent maintenant de basculer assez facilement d'une structure à une autre à l'aide de feuilles de style XSLT, assurant ainsi l'interopérabilité d'objets dont les structures sont différentes pour des raisons historiques, mais malgré tout assez voisines.

6.3 La représentation de la connaissance

La partie de l'indexation qui porte sur la représentation de la connaissance pose des problèmes d'une toute autre nature. Le souci d'expressivité nous a amené à retenir dans un premier temps un mode de représentation puissant en terme de logique de premier ordre, à savoir les Graphes Conceptuels de Sowa [Sowa, 84]. La capacité à les traduire éventuellement en langage Prolog, ou KIF, et à pouvoir effectuer diverses opérations conceptuelles, permet aussi de répondre au mieux au souci de technicité.

Cependant, la prise en compte du souci d'exploitabilité nous a amenés à découper une représentation de type GC en un ensemble de triplets plus faciles à manipuler par un indexeur. Ce choix s'est avéré heureux (effet de chance ou d'anticipation) puisque par la suite le langage RDF proposé par le W3C pour le Web Sémantique repose sur le même principe, assurant du coup une compatibilité a posteriori, et donc l'interopérabilité de nos modèles avec ceux qui devraient apparaître dans le futur sur Internet.

Les soucis à la fois d'expressivité, de technicité et d'exploitabilité nous avaient aussi fait asseoir une indexation du contenu sur une ontologie du domaine. Ce choix s'est avéré également judicieux puisque nous anticipions la tendance actuelle au sein du W3C à faire reposer une représentation de la connaissance en RDF sur la définition d'un vocabulaire (mots et sémantique) à partir d'un schéma (RDFS), puis sur une ontologie écrite en DAML (et bientôt OWL, dérivé immédiat de DAML).

Mais les besoins de calcul pour la composition (technicité) nous ont amenés à étendre cette représentation sous forme de triplets en une représentation "floue" qui associe des poids aux assertions. Nous avons ainsi obtenu une représentation du contenu sous forme de Vecteurs d'Etats Conceptuels (CSV) qui nous est propre tout en étant totalement transposable en d'autres représentations basées sur RDF, ou sur DAML. Les calculs effectués ensuite par un moteur de composition sont spécifiques à chaque domaine de DVP et ne sont pas traités ici.

6.4 Aide à l'indexation

Reste cependant le problème de l'aide à l'indexation sachant qu'il n'est pas aisé malgré tout pour n'importe qui de construire des CSV à partir d'une ontologie. Le premier outil d'aide à l'indexation conçu pour Karina propose une représentation abstraite. L'indexeur est aidé à deux niveaux.

Lors de la construction d'un triplet, le choix d'un premier concept se fait sur l'arbre entier des concepts. Ce choix entraîne ensuite la présentation d'un sous-arbre pour choisir un relation, qui entraîne aussi la sélection d'un sous-arbre pour le choix du second concept entrant dans la relation. Ces choix de sous-arbres sont guidés par les contraintes dans l'ontologie qui se présentent elles-mêmes sous la forme de triplets de type RDF. Ce principe assure tout à la fois une aide à l'indexeur dans un souci d'exploitabilité, et une cohérence sémantique de l'indexation dans un souci d'expressivité.

La seconde aide porte sur la présentation à l'écran de l'ontologie du domaine et des assertions construites. Plusieurs approches ont été explorées et deux,

complémentaires, ont été mises en œuvre. La première, dans le domaine de l'enseignement, consiste à traduire les triplets en des phrases en langage naturel. La seconde, explorée pour les applications de programmes TV et radio personnalisés, consiste à concrétiser au maximum les concepts et les relations sur l'écran à l'aide de petites icônes. Nous discutons ci-dessous ces différents choix.

6.5 Limites et perspectives sur l'indexation de la connaissance

Le LOM propose une structure intéressante pour l'indexation d'objets pédagogiques, mais le problème de la représentation et de l'indexation de la connaissance reste ouvert. Nous pensons avoir trouvé avec les CSV un compromis intéressant entre les différentes qualités attendues d'une indexation de la connaissance pour les DVP. Mais on voit que nous avons pour l'instant peu de solutions pour faciliter le travail de l'indexeur en l'absence d'une indexation automatique. La traduction d'un triplet en langage naturel est intéressante pour l'enseignement, mais la construction d'un triplet se fait toujours à la main sur des mots abstraits. En fait, dans beaucoup de domaines pédagogiques comme "la programmation en C", le caractère abstrait de la connaissance oblige à naviguer sur des abstractions. Dans des applications très visuelles et concrètes, comme une discipline sportive pour la TV, il est possible de proposer des écrans plus métaphoriques, voire esthétiques et ludiques. La tentative que nous avons faite dans ce sens est assez réussie, quoique fortement perfectible. En particulier, il est difficile de représenter des actions par des icônes (par exemple "joueur drible joueur"). Dans le domaine de la musique, nous nous sommes aperçus que peu de relations pouvaient être construites et que l'ontologie du domaine était très conceptuelle (par exemple : "rock") ou à l'inverse très concrète (par exemple "Johnny Haliday"). Dans ce dernier cas, on peut proposer des images des personnages, mais leur nombre constitue une difficulté.

Dans l'attente d'outils d'indexation automatique, nous voyons dans les techniques d'indexation vocale un moyen important pour faciliter le travail de l'indexeur. C'est aussi cette perspective qui nous avait guidés dans le choix des triplets et le support d'une ontologie. Les techniques de reconnaissance vocale sont encore peu efficaces. L'utilisation d'une ontologie permet de limiter le vocabulaire, et la représentation par des triplets permet de limiter la complexité des phrases et leur polysémie. Nous menons des essais en ce sens. Ainsi on voit combien la nature non professionnelle de l'indexeur et le souci de faciliter sa tâche ont pu guider les choix techniques en tenant compte non seulement des choix du moment, mais encore des perspectives futures.

Reste au bilan que nous voyons l'indexation comme une tâche indispensable, difficile, et qui constitue en conséquence un enjeu majeur. A terme, il sera sans doute possible de largement l'automatiser. Pour l'heure, elle relève d'un travail d'ingénierie cognitive qui mérite une analyse plus poussée et dont l'outillage reste largement à inventer.

Remerciements : Les auteurs de cet article remercient leurs partenaires, et plus particulièrement Florent Barbare de la société Netia, et Nicole Chanal de la société ActiMédia Systems, pour leur contribution aux projets présentés.

Références

ASSELBORN J.C., JANS J.M., BERTRAND A., SCHANET C. « Implémentation en Java de connaissances légales aux niveaux opérationnel et explicatif ». *Génie Logiciel,* N°46, Actes GL 97, 1997.

AUFFRET G. « Indexation Documentaire de Documents Virtuels : Vers un Nouveau Mode de Lecture des Documents Audiovisuels ». *Atelier sur les Documents Virtuels Personnalisables : De la Définition à l'Utilisation, 11ème Conférence Francophone sur l'Interaction Homme-Machine IHM'99*, Montpellier, novembre 1999. http://www.site-eerie.ema.fr/~multimedia/ihm99/

AUFFRET G. « Structuration de documents audiovisuels et publication électronique. Constitution d'une chaîne éditoriale numérique pour la mise en ligne de collections audiovisuelles ». *Thèse de l'Université de Compiègne*. Décembre 2000.

CRAMPES M. « Auto-Adaptative Illustration through Conceptual Evocation » *in Proceeding of the conference on Digital Library DL'97.* ACM '97, Philadelphia, PA., USA, ACM Press, pp. 247-253, July 1997.

CRAMPES M., VEUILLEZ J.P., RANWEZ S., « Adaptive Narrative Abstraction » *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia HT98*, Pittsburgh., PA., USA, ACM Press, pp. 97-105, June 1998.

CRAMPES M., « An Agent-Based Adaptive Program Composer for the Home TV of the Future », *ECAI-98*,Workshop on AI/Alife and Entertainment, 24 August 1998.

CRAMPES M., « Karina : Spécifications Fonctionnelles ». *Document interne Ecole des Mines d'Ales - Ecole Supérieur des Ingénieurs de Marseille*. Septembre 1998.

CRAMPES M., BAYART L., GELLY A., UNY P. « Spécification et proposition d'une DTD pour la qualification de matériaux pédagogiques adaptatifs *», Sciences et Techniques Educatives*, volume 6, Hermes 1999.

CRAMPES M., RANWEZ S. « From Adaptive Narrative Abstraction to Coherent Hypermedia Navigation », *Eleventh ACM Conference on Hypertext and Hypermedia, HT2000*, San Antonio, Texas, ACM, June 2000.

CRAMPES M. «Auto-Composition Active et émergence du sens dans l'Initeraction Homme-Machine sous contrainte ». *Mémoire d'Habilitation à Diriger des Rercherches*, Université de Montpellier II, Mai 2002.

DAML, Agent Markup Language Committee http://www.daml.org/2001/03/daml+oil-index.html, March 2001.

DERYCKE, A. « Sept questions sur le E-Learning : vers une problématique nouvelle pour la recherche ? ». Les technologies en éducation. Perspectives de recherche et questions vives. Actes du Symposium international francophone, Paris, 2002. Edités sous la direction de G. L. Baron et E. Bruillard. INRP 2002, pp. 29-39.

DOMINGUE J., MOTTA E., « A Knowledge-Based News Server Supporting Ontology-Driven Story Enrichment and Knowledge Retrieval ». *11th European Workshop on Knowledge Acquistion, Modelling, and Management EKAW '99, 1999.*

DUBOIS D., PRADES H., « Les ensembles flous en science et ingénierie de l'information ». Dans *Informatiques enjeux, tendances et évolution*, Techniques et science informatiques, Volume 19, N°1,2 et 3/2000, Hermès science, pp. 203-215, Janvier-Mars, 2000.

CRAMPES M., VERCOUSTRE A.M., NANARD M., RANWEZ S. Atelier « Documents Virtuels Personnalisables : de la Définition à l'Utilisation », *11ème conférence francophone sur l'Interaction Homme-Machine IHM'99,* Montpellier, France, 22-26 Novembre 1999. http:// www.site-eerie.ema.fr/~multimedia/ihm99

Garey, M.R., Johnson, D.S., « Computer and intractability : a guide to the theory of NP-completeness », W.H. Freeman and Company, New York, 1979.

GORDON A., KEDAR S., DOMESHEK E. « Interfaces for Managing Access to a Video Archive ». *Proc. of the Computer-Human Interaction Conference CHI'96*, Vancouver, BC, Canada, 1996.

GREEN S.J. « Automated link generation: can we do better than term repetition? » *Proceedings of the Seventh World Wide Web Conference WWW7*, Brisbane, also in a special issue of the *journal Computer Networks and ISDN Systems*, Volume 30, issues 1-7. http://decweb.ethz.ch/WWW7/1834/com1834.htm

GRUBER T.R. « A Translation Approach to Portable Ontology Specifications » *Knowledge Acquisition*, Vol.5 No. 2, pp.199-220, 1993.

HAUPTMANN A.G., WITBROCK M.J., CHRISTEL M.G. « Artificial Intelligence Techniques in the Interface to a Digital Video Library ». *Proceedings of the Computer-Human Interface Conference CHI-97,* New Orleans LA, March 1997.
http://www.cs.cmu.edu/afs/cs.cmu.edu/user/alex/www/HomePage.html

HO K.H.L. « Learning Fuzzy Concepts By Examples with Fuzzy Conceptual Graphs », *Proceedings of the 1st.Australian Conceptual Structures Workshop*, Armidale N.S.W. Australia, 1994.

IKSAL S., GARLATTI, S., « Documents virtuels et composition sémantique : Une architecture fondée sur les ontologies. ». Actes du Congrès Scientifique Nîmes TIC2001, Ecole des Mines d'Ales, pp. 91-96.

KABEL S.C., WIELINGA B.J., DE HOOG R. « Ontologies for indexing Technical Manuals for Instruction ». *Workshop on Ontologies for Intelligent Educational Systems, Ninth International Conference on Artificial Intelligence in Education, AI-ED'99*, Le Mans, France, July 19-23, 1999.

KANTOR P.B. « Information Retrieval Techniques », *Annual Review of Information Science and Technology (ARIST)*, Volume 29, Martha E. Williams Editor, Learned Information Inc., Medford, N.J., 1994.

KAUFMAN L.A. « Introduction à la théorie des sous-ensembles flous à l'usage des ingénieurs. Tome 1 Eléments théoriques de base », Deuxième édition, Masson 1977.

KAUFMAN L.A. « Introduction à la théorie des sous-ensembles flous à l'usage des ingénieurs. Tome 2 Applications à la linguistique, à la logique et à la sémantique », , Masson 1975.

LEMOISSON, P. Auto-composition d'un résumé vidéo sous deux approches : optimisation par contraintes ... négociation entre agents. Rapport de stage de DEA LGI2P/LIRMM, 2001.

LTSC, Learning Technology Standards Committee of the IEEE, Draft Standard for Learning Object Metadata, IEEE P1484.12.1/D6.4, 4 March 2002.

MEMZIES T. « Cost Benefits of Ontologies », *Intelligence*, vol.6, Number 3, ACM Fall 1999.

MICHARD A., « XML, Langage et application », Eyrolles 1999.

MORTON S.K., POPHAM, S.J. « Algorithm Design Specification For Interpreting Segmented Image Data Using Schemas And Support Logic ». *Image and Vision Computing* (5), Butterworth & Co publishers, pp. 206-216, 1987.

MOTTA E., BUCKINGHAM-SHUM S., DOMINGUE J. « Ontology-Driven Document Enrichment: Principles and Case Studies ». *Proceedings of the 12th Workshop on Knowledge Acquisition, Modeling and Management KAW '99*. Banff, Canada, October 16th - 21th, 1999

NKAMBOU R., GAUTHIER G., FRASSON C. « Un modèle de Représentation de Connaissances relatives au contenu dans un Système Tutoriel Intelligent ». *Sciences et techniques éducatives*. Volume 4, N°3/1997, p. 299-330, 1997.

DE LA PASSARDIERE, B., GIROIRE H. « XML au service des applications pédagogiques », *Sciences et Techniques Educatives*, volume 8, Hermes 2001.

PRIE Y. « Modélisation de documents audiovisuels en Strates Interconnectées par les Annotations pour l'exploitation contextuelle ». *Thèse de doctorat en informatique*, INSA de Lyon, décembre 1999.

RANWEZ S., CRAMPES, M., « Conceptual Documents and Hypertext Documents are two Different Forms of Virtual Document ». *Proceedings of the Workshop on Virtual Documents, Hypertext Functionality and the Web,* Eight International World Wide Web Conference, 1999, Toronto (Canada), pp. 21-27.

RANWEZ S., LEIDIG T, CRAMPES M. « Pedagogical Ontology and Teaching Strategies: A New Formalization to Improve Life-Long Learning » in *International Journal of Continuing Engineering Education and Life-Long Learning (IJCEELLL),* Inderscience Entreprises Ltd.

RANWEZ, S. « Construction automatique de dispositifs hypermedia adaptatifs à partir d'ontologies narratives et pédagogiques et de modèles conceptuels intentionnels de l'utilisateur ». *Thèse de doctorat, Montpellier II*, Décembre 2000.

SALTON G., SINGHALT A., BUCKLEY C., and MITRA M. « Automatic Text Decomposition Using Text Segments and Text Themes », *Proceedings of the Seventh ACM Conference on Hypertext HT96*, Washington DC, pp 53-65, March 1996.

SOWA J.F. « Conceptual Structures : Information Processing in Mind and Machine », Addison-Wesley, 1984.

UTE, « Some Ongoing KBS/Ontology Projects and Groups », http://www.cs.utexas.edu/users/mfkb/related.html.

W3C, « Extensible Markup Language (XML) 1.0 », eds. Tim Bray, Jean Paoli, and C. M. Sperberg-McQueen., W3C REC-xml-19980210, http://www.w3.org/TR/REC-xml, 10 February 1998.

W3C, « Resource Description Framework (RDF) Model and Syntax Specification) », W3C PR-rdf-syntax-19990105, http://www.w3.org/RDF/, 5 Janvier 1999.

W3C, « XSL Transformations (XSLT), Version 1.0 », *W3C Recommendation* , http://www.w3.org/TR/1999/REC-xslt-19991116, 16 November 1999.

W3C, « RDF Vocabulary Description Language 1.0: RDF Schema », W3C Working Draft, http://www.w3.org/TR/2002/WD-rdf-schema-20020430/,30 April 2002

WEINSTEIN, P., « Ontology-Based Metadata: Transforming the MARC Legacy ». *Actes Third ACM Digital Library conference*, Pittsburgh, PA, USA, June 1998.

WEINSTEIN, P. et ALLOWAY, G., « Seed Ontologies: growing digital libraries as distributed, intelligent systems ». *Proceedings of the Second ACM Digital Library conference*, Philadelphia, PA, USA, July 1997.

ZOBEL J., MOFFAT A. « Exploring the Similarity Space », *SIGIR Forum*, vol. 32, N° 1, ACM Spring 1998.

Michel Crampes *est ingénieur (IDN 1974) et docteur en informatique (Montpellier II, 1995). Après avoir dirigé à partir de 1981 le groupe de recherche sur les Nouvelles Technologies Educatives(NTE) au sein de la société SYSECA Temps Réel, filiale logicielle de Thomson Temps Réel, il rejoint en 1991 l'Ecole des Mines d'Ales (EMA) pour prendre la responsabilité des recherches sur le multimédia appliqué à la formation. Enseignant chercheur depuis 1995 au laboratoire de recherche en Intelligence Artificielle (LGI2P) de l'EMA, il obtient l'Habilitation à Diriger des Recherches en 2002 à l'Université Montpellier II. Ses travaux portent sur la composition d'informations et l'émergence du sens. Les domaines d'application concernent les interfaces adaptatives, les Documents Virtuels Personnalisables, l'ingénierie de la connaissance et le Web Sémantique.*

Sylvie Ranwez *est ingénieur EERIE et docteur en informatique de l'Université Montpellier II. Elle a effectué sa thèse au sein du LGI2P (Laboratoire de Génie Informatique et d'ingénierie de Production - Ecole des Mines d'Alès) sous la direction de Michel Crampes sur le thème de la composition de documents hypermédia adaptatifs. Elle a alors participé à deux projets dans le domaine des EIAH adaptatifs : Sibyl et Karina. Elle a travaillé ensuite sur la composition de programmes radio et TV personnalisables. Ses activités de recherche concernent les modes de composition de documents structurés basés sur des ontologies et les Documents Virtuels Personnalisables (DVP).*

Ingénieur EERIE 1998, ***Christophe Vaudry*** *rejoint le LGI2P en 1999 et présente une thèse à l'Université de Montpellier II en 2002, sous la direction de Marc Nanard et Michel Crampes, sur la composition des informations et les interfaces adaptatives. Il est actuellement en poste d'ATER à l'IUT de l'université de Provence, site d'Arles. Ses travaux portent sur la composition des informations et l'adaptation des interfaces. Les domaines d'application concernent les hypermédias et les documents virtuels, la prise en compte du contexte, la recherche d'information et le Web Sémantique.*

Ingénieur ENSEM (Nancy), ***Michel Plantié*** *rejoint l'Ecole des Mines d'Ales (EMA) dans laquelle Il soutient son DEA en informatique à l'Université Montpellier II en 2000 sur l'auto-composition d'information, et intègre le laboratoire de recherches en Intelligence Artificielle (LGI2P) de l'EMA. Il prépare actuellement sa thèse de doctorat dans le domaine de l'extraction d'information, et de l'aide à la décision. Ses travaux portent sur l'extraction de connaissance et l'émergence du sens. Les domaines d'application concernent l'ingénierie et les référentiels de connaissances, les moteurs d'extraction, et le Web Sémantique.*