Médiation documentaire et visée éditoriale :
Analyse des principes de mise en ligne de ressources pédagogiques

Richard Walter, Modyco, Université de Paris 10
Stéphane Chaudiron, Geriico, Université de Lille 3

1. Introduction

La mise en ligne de ressources documentaires implique une médiation avec des publics qui est effectuée à travers un artefact technique. Celui-ci est composé d'un ensemble de fonctionnalités de recherche et de lecture d'informations, formant un système éditorial qui instaure un contrat de lecture entre les « éditeurs » et les lecteurs/usagers de ces systèmes d'informations. Or une analyse des contextes réels d'usage montre que l'anticipation sur les besoins informationnels n'est très pas aisée avec la diffusion de mémoires d'étudiants.

Cet article examine deux approches différentes de mises à disposition de documents numériques réalisés dans un contexte pédagogique à travers les projets *Terminalf* et *Codex* qui sont deux expérimentations sur des corpus de mémoires d'étudiants :

- *Codex* : plate-forme de visualisation d'un corpus constitué de mémoires d'étudiants en sciences de l'information et de la communication (université Paris X), prenant en compte différents usages ;
- *Terminalf* : base de données rassemblant des mémoires d'étudiants en terminologie (universités Paris III et Paris VII), correspondant à des inventaires de domaines spécialisés.

Notre communication se propose de comparer les approches tant conceptuelles que méthodologiques, utilisées pour ces deux projets. Pour cela, nous examinerons comment ces deux projets ont traité les questions des formats, de la granularité documentaire, du guidage dans les ressources, des métadonnées et enfin de la chaîne de production.

Les deux projets ayant pour objectif de proposer en ligne un ensemble cohérent de mémoires, il convient au préalable de préciser les notions de « corpus » et de « ressources » utilisées dans ces projets. Chaque mémoire peut en effet être considéré comme un document autonome mais est également intégré dans un ensemble. Nous discuterons ensuite la notion de médiation documentaire qui est en jeu ici et plus précisément sur les finalités pédagogiques *et/ou* scientifiques de ce type de projet.

2. Présentation des projets *Codex* et *Terminalf*

2.1 Réflexions préalables sur les notions de corpus, d'archives et de ressources

La construction et la mise en accès d'un ensemble qui se veut « raisonné » de documents textuels impliquent une démarche de mise en cohérence à la fois interne (structuration, homogénéité…) et externe (visée éditoriale, usages prévus, publics ciblés…). En ce sens, la construction de « ressources » à visée pédagogique et de corpus renvoie à une démarche qui est très différente de celle des « entrepôts de données » que sont par exemple les archives ouvertes. De fait, comme l'indique F. Rastier (Rastier, 2002), une archive se définit par l'ensemble des documents qu'elle contient et qui sont accessibles. Elle n'est donc pas constituée pour une recherche particulière ni pour un usage *déterminé*. Les archives ouvertes sont un ensemble de documents déposés et rendus accessibles par leurs auteurs sans détermination *a priori* des usages qui en seront faits. Cette caractéristique n'enlève rien à l'intérêt des archives ouvertes, notamment du point de vue de l'accélération de la diffusion des résultats de la recherche ou du modèle économique alternatif mis en œuvre.

Une autre question concerne la différence entre la notion de « corpus » et celle de « ressources ». Dans la typologie des corpus qu'il propose dans une approche linguistique, F. Rastier distingue trois types principaux de corpus. Le *corpus de référence* qui est constitué par l'ensemble de textes sur lequel le chercheur va faire des contrastes entre les différents corpus d'étude. Ensuite, le *corpus d'étude* qui est

délimité par les besoins spécifiques de la recherche et correspond à un sous-ensemble du corpus de référence. Et enfin le *sous-corpus de travail en cours* qui varie selon les phases de l'étude et qui contient des passages pertinents du texte ou des textes étudiés par rapport à l'objet de l'étude.

Pour illustrer cette typologie, Rastier donne l'exemple de l'analyse thématique des données textuelles, et plus précisément l'exemple de l'analyse des sentiments. Dans un corpus littéraire, le corpus de référence sera constitué des romans, le corpus d'étude est constitué des passages contenant des noms de sentiments et les corpus de travail sont les corpus propres à tel ou tel sentiment. Cette approche souligne bien que tout corpus se constitue dans le cadre d'un objectif de recherche ; il n'y a donc de corpus que *par intention*. De même, le développement de certaines applications informatiques nécessite des corpus d'apprentissage, des corpus de test ou des corpus de validation. C'est le cas en particulier des logiciels de traitement automatique des langues et des méthodes pour les évaluer.

De son côté, la notion de « ressources », qui connaît actuellement un emploi inflationniste, en particulier dans le domaine des sciences humaines et sociales, est empreinte d'un flou conceptuel qu'il convient également d'examiner. Selon M. Pothier (Develotte & Pothier, 2004), on peut en particulier différencier d'une part les ressources entendues comme des données de toutes formes (textuelles, sonores, visuelles) et d'autre part les ressources informatiques qui sont mises à la disposition des pédagogues pour concevoir leurs apprentissages. La notion de ressources est également utilisée en sciences du langage pour désigner des données linguistiques de différentes natures (lexicales, morpho-syntaxiques, sémantiques…) et qui seront utilisées dans le traitement automatique des langues, non pour leur développement (il s'agit dans ce cas des corpus) mais pour leur fonctionnement (ainsi, la ressource peut être le dictionnaire bilingue de spécialité utilisé dans un traducteur automatique ou les règles de grammaires implémentées dans l'analyseur).

Les deux projets sont au départ une mise à disposition de *ressources* produites dans un contexte donné. Nous allons voir si maintenant elles forment un *corpus* grâce au système éditorial utilisé.

2.3 *Terminalf*

Depuis 1998, le projet *Terminalf* a pour objectif de diffuser et de faire connaître des ressources terminologiques en langue française. Ces ressources sont constituées de mémoires de terminologie présentant et structurant le vocabulaire technique ou scientifique d'un micro-domaine spécialisé. Ces inventaires sont susceptibles d'aider à la compréhension de la langue française utilisée dans les milieux de spécialistes.

Ce projet se concrétise à travers un site internet qui diffuse déjà plus d'une centaine de mémoires de terminologie et une autre centaine sont en cours d'introduction. Ce sont des mémoires d'étudiants de deuxième cycle (maîtrise ou DEA) en langue étrangère et appliquée (LEA) ou en ingénierie linguistique, provenant de deux universités (Paris III et VII) ; ils étaient « dormants » sur des étagères, l'objectif est donc de les valoriser et de les diffuser.

Comme en témoignent les derniers mémoires mis en ligne, ils sont d'une grande diversité thématique : la résistance du VIH aux médicaments antirétroviraux, le solfège et la théorie musicale, le tennis, l'écologie générale, l'homoparentalité, le Stade de France, le cheval, la composition et structure de l'inflorescence, la télévision numérique terrestre.

Chaque mémoire est composée d'environ 70/80 concepts traités de façon homogène (définition, contexte d'usage, statut grammatical, traduction en une ou plusieurs langues – le français étant la langue pivot). Ces concepts sont présentés sous forme de fiches individuelles mais sont reliés les uns aux autres par l'indication de leur liaison (homonymie, synonymie, tout ou partie, etc.). Il s'agit donc de nomenclatures qui proposent une arborescence pouvant être facilement gérée sous forme de base de données ou d'hypertexte.

Les mémoires se composent de deux parties. Une partie textuelle est ainsi enregistrée au format Word (introduction, méthodologie, bibliographie, etc.) et une partie structurée avec des fiches est réalisée sous forme d'une base de données. La partie textuelle a été relativement simple à traiter car elle se présente comme du texte linéaire. Par contre, les fiches terminologiques sont réalisées et stockées dans un fichier créé avec le logiciel propriétaire *Access* qui pose des problèmes en termes d'accès simultanés et de gestion de la volumétrie.

Terminalf est donc appelé à se développer tant sur le plan terminologique que sur le plan informatique, notamment en considérant l'apport des outils automatiques d'indexation des données. Quatre traitements demandent actuellement une attention particulière :

- Différenciation entre terme vedette et terme associé.
- Fabrication et pérennité des webographies ;
- Visualisation des relations d'un domaine et des arborescences ;
- Critères de création d'index automatique et semi-automatique ;

À travers cette construction, *Terminalf* propose une réflexion méthodologique et épistémologique sur le traitement numérique de données structurées, ici pédagogiques et terminologique. Cela est d'autant plus possible que nous disposons d'une expérience sur la longue durée pour mesurer les évolutions tant techniques que conceptuelles.

2.4 *Codex*

Le projet *CodeX* (Consultation et organisation de documents électroniques en XML) (Chaudiron, 2000) est constitué de versions électroniques de mémoires d'étudiants en deuxième cycle (maîtrise) en sciences de l'information et de la documentation de l'université Paris X. Ce fonds a été rassemblé depuis 1999 et représente environ 50 documents. L'accroissement de ce corpus est d'environ 20 documents par an.

Contrairement à *Terminalf*, les mémoires relèvent d'un même domaine, les sciences de l'information et la communication même si ils couvrent des sujets très divers comme en témoignent quelques-uns des premiers mémoires mis en ligne :

- Le Louvre, Orsay et le Centre Georges Pompidou sur cd-rom : vers des Néo-musées ?
- La création et la diffusion de l'art sur Internet
- La valorisation des fonds bibliographiques et des fonds d'archives du service historique de la marine
- Comparaison des regards portés sur les révolutions de l'imprimerie et du numérique
- Activité d'un veilleur au sein d'un système d'information
- Filtrage de l'information : ses techniques, ses usages et ses outils
- Intégration organisationnelle des nouvelles technologies de l'information et de la communication dans l'industrie cinématographique

CodeX est une plate-forme expérimentale qui vise à rendre accessible un ensemble de documents électroniques structurés selon des modes de consultation adaptés aux différents contextes d'usage recensés. Alors que *Terminalf* diffuse essentiellement du matériau, *CodeX* expérimente différents modes de consultation déterminés à partir d'une enquête d'identification des besoins menée auprès des futurs usagers, à savoir les étudiants.

Les documents traités sont des mémoires universitaires assez volumineux (au moins une centaine de pages). Compte tenu de ce caractère volumineux, plutôt que de donner accès aux documents dans leur entier, l'approche retenue a été de donner à l'utilisateur la possibilité de consulter au préalable ces documents à travers des modes de visualisation (des *vues*) qui ne retiennent que les aspects pertinents pour un usage précis. Chaque vue vise à répondre à un besoin informationnel précis et identifié.

Actuellement, quatre vues possibles du même document, sont identifiées et correspondent donc à quatre modes de consultation :

- la visualisation de la table des matières ;
- la visualisation de l'introduction, de la conclusion et de la table des matières ;
- la visualisation du résumé d'auteur ;
- la visualisation du résumé d'auteur et de la bibliographie.

Schématiquement, on peut définir *le système de vues multiples* utilisé dans *CodeX* comme un mode de consultation qui permet à l'utilisateur de visualiser des facettes différentes du même document. Héritée du monde des bases de données, la notion de « vue » a ensuite été utilisée dans le domaine de la documentation structurée par certains éditeurs pour désigner l'extraction d'un ou plusieurs blocs d'information à partir de la structuration logique du document. Dans cette approche, le premier objectif a été de construire le système de vues seulement sur la structure logique du document.

Ces vues sont statiques car elles sont prédéfinies par les concepteurs de l'interface d'accès à la base et les utilisateurs ne peuvent les modifier. D'un point de vue technique, ces vues structurelles s'appuient sur un balisage XML des documents. Par exemple, une vue peut chercher à partir de la structure balisée la partie ou le chapitre demandé. Le système de vues multiples ainsi défini s'inscrit dans une logique de gestion des connaissances et se présente comme un outil de filtrage et de visualisation d'une base de documents.

3. Une comparaison des deux approches

Entre un projet avec système de vues multiples et un autre avec extraction de concepts, des différences existent bien sûr, mais aussi des similitudes. Celles-ci sont d'abord factuelles. Ce sont des projets universitaires lancées à la même époque. Tous les deux ont des visées pédagogiques (fournir des indications de qualité pour les mémoires suivants) mais aussi des visées scientifiques (fournir un matériau de recherche par la mise à disposition de mémoires « dormants »). Par contre, la mise en pratique s'est avérée différente à cause de la structure différente des mémoires et donc d'unités documentaires à manipuler différemment.

3.1 Format

Les formats électroniques actuellement les plus utilisés sont conçus pour un usage prédéfini (affichage à l'écran pour HTML, impression pour les formats bureautiques comme Word). S'ils s'acquittent généralement bien de cette fonction, ils se prêtent cependant mal à d'autres types d'exploitation pour lesquels ils n'ont pas été conçus (indexation, création de vues sur les documents, stockage en bases de données, etc.).

CodeX se base sur une approche de la représentation structurée des documents électroniques. L'idée est de représenter un document électronique en utilisant un codage XML ayant pour caractéristiques d'être à la fois rigoureusement validé selon un schéma bien précis et neutre par rapport aux utilisations possibles.

Cette dernière caractéristique fait du document structuré un support idéal pour l'implémentation de vues multiples. L'approche adoptée au sein du projet est donc de convertir le corpus en XML à partir des fichiers Word produits par les étudiants puis d'utiliser le mécanisme des feuilles de style XSL (*Extensible Stylesheet Language*) pour générer des vues adaptées à chaque usage. Le terme « feuille de style » est d'ailleurs réducteur dans la mesure où une spécification XSL consiste schématiquement en un ensemble de règles de traduction permettant non seulement d'associer des attributs de présentation physique (par exemple une police, une taille, une couleur, etc.) aux éléments d'un fichier XML mais également de filtrer et de réordonner ces derniers. Cette méthode s'adapte parfaitement à la typologie des contenus présents dans *CodeX*.

Pour *Terminalf*, l'extraction est plus complexe car le contenu est fait d'un mélange hybride de contenu linéaire (les textes en Word) et de contenu déjà structuré (les fiches terminologiques en Access). Les bases de données sont stockées dans un répertoire spécifique ; une spécification des métadonnées est faite à l'introduction d'un mémoire dans *Terminalf* et elles stockées dans une base générale. Les fichiers Word sont transformés en HTML et intégrés dans cette même base générale. Celle-ci rassemble donc les métadonnées et le contenu générique d'un mémoire, ainsi la recherche des informations demandées par l'utilisateur peut s'effectuer sur l'ensemble de celui-ci. Un des chantiers informatiques en cours est la simplification de cette liaison entre base de données, métadonnées et contenu générique d'un même mémoire. Cette simplification passera par la généralisation de XML comme langage de stockage et d'exploitation de l'ensemble des données.

Enfin, pour diffuser un des éléments essentiels aux mémoires de domaine, l'arborescence du domaine, un outil est en train d'être développer permettant d'automatiser une mise en ligne aisée et graphique de celle-ci. La généralisation de formats comme PowerPoint ou PDF a permit de récupérer des arborescences réellement sous forme d'arbre structuré et non plus sous forme de simple pavé textuel. Le format des arborescences va devenir à terme homogène, ce qui permettra un traitement graphique systématique et donc une meilleure visualisation des relations dans le domaine concerné.

3.2 Granularité

Pour accéder à un document dans *CodeX*, l'utilisateur doit sélectionner une vue puis cliquer sur le lien correspondant au document. Une feuille de style XSL correspondant à la vue est alors appliquée dynamiquement, côté serveur, à ce document. Les règles de cette feuille de style extraient les éléments pertinents pour la vue et génèrent une page HTML présentant ces éléments. Cette page est alors transmise au navigateur de l'utilisateur. Dans *CodeX*, la granularité de l'information pertinente est donc déterminée par la vue que choisit l'utilisateur.

Dans *Terminalf*, l'approche est multiple car elle est basée sur une granularité variable. L'unité de base peut être un terme (concept ou non) de l'index général, une fiche terminologique de la base de données d'un domaine, un concept de l'arborescence. Chacun de ces éléments répond à un besoin spécifique et nécessite un traitement particulier pour la lecture et l'exploitation

3.3 Guidage dans les ressources

Dans *CodeX*, la conception de modes diversifiés d'interaction lors de la consultation de la base de documents est basée sur la notion de *vue multiple* qui présente l'intérêt de permettre à l'utilisateur d'évaluer rapidement la pertinence des documents dont il a retrouvé les références. Cette vue s'appuie exclusivement sur la structure logique des documents. Pour élaborer des vues exploitant d'autres sources d'information, il conviendrait de se fonder sur les index ou d'autres notations créés par les auteurs ou par des outils d'ingénierie linguistique.

Dans *Terminalf*, la possibilité de naviguer entre les mémoires a été instaurée dès l'origine. Compte tenu de la discipline, l'approche a été très lexicographique : chaque nouveau mémoire ajoute des unités à l'index général. Contrairement au moteur de recherche qui permet de chercher à l'intérieur des fiches terminologiques, l'index général cumule tous les concepts présents ainsi que leurs équivalents dans une langue étrangère. On arrive ainsi à près de 8000 formes dans 130 mémoires, sachant que les doublons ne sont pas éliminés car les concepts n'ont pas la même application selon les domaines ou les langues. Ainsi, par exemple, le concept d'« infrastructure » en télévision numérique terrestre ne peut être rassemblé avec celui d'« infrastructure » en volley-ball.

Pour la navigation, nous sommes en présence de deux dispositifs différents : l'un extrait des vues d'un contenu linéaire et l'autre se base sur une arborescence, un contenu non linéaire mais structuré.

3.4 Métadonnées

De nombreux dispositifs de navigation ont été imaginés dans le contexte du livre imprimé : table des matières, index, glossaire, etc. La table des matières correspond généralement aux divisions de l'ouvrage (chapitre, section) dont elle reproduit séquentiellement les titres. Par contraste, l'index se présente comme un raccourci significatif du contenu, qui peut exister indépendamment de la structure.

La plupart des documents traités dans *CodeX* sont accompagnés d'index créés par les auteurs. Ces index sont en général assez riches et pertinents. On peut en effet penser que les index auteurs ont une certaine valeur puisqu'ils contiennent des termes et des concepts jugés importants par les créateurs des documents. Dans *CodeX*, l'exploitation de cette source d'information a permis de créer des vues basées non plus sur la structure logique mais sur le contenu des documents.

Schématiquement, la structure logique du document pourrait se concrétiser dans la table des matières et son contenu dans l'index. Lors du processus de conversion d'un document pour son introduction dans *CodeX*, chaque occurrence d'une entrée d'index Word est traduite en notation de sa position dans la structure XML. On dispose ainsi du moyen de relier un terme d'indexation à ses différents contextes d'apparition.

Terminalf présente lui deux niveaux de métadonnées bien différents :

- Le mémoires a comme étiquette nom de l'auteur, titre du mémoire, année de soutenance, enseignant responsable, expert garant, institution, langue prise en charge, etc.
- Pour la fiche terminologique, le concept pivot et toutes les catégories associées peuvent être considérés comme des métadonnées. On peut faire la comparaison avec la structure d'une notice de dictionnaire : y sont considérées comme éléments structurants le mot-vignette et les différentes parties de la définition (genre, catégorie grammaticale, niveau de langue, exemple d'usage, etc.). Ces éléments structurants sont donc des métadonnées sur lesquels on peut appliquer différents outils de recherche.

Le protocole de recherche et la présentation du domaine sont importants pour valider les fiches des concepts ; il y a enfin nécessité de relier ces deux types de contenus par des métadonnées.

3.5 Chaîne de production

La chaîne de production n'est pas forcément très différente entre les deux projets. Ils partent du même objectif initial de diffusion des mémoires et ont adopté le même critère de sélection (note minimum de 14/20). Intégrer un mémoire à un ensemble a permis d'harmoniser les méthodes et les procédures de réalisation d'un mémoire. Dans *Terminalf*, le plus difficile fut sans doute d'obtenir une structure de document satisfaisante aussi bien pour les besoins pédagogiques (et donc terminologiques) que pour une exploitation à grande échelle.

Pour *CodeX*, si le corps de ces documents est assez simple à analyser (la mise en page est proche du modèle par défaut de l'outil bureautique), il n'en va pas de même des pages initiales (comme la page de titre) qui contiennent les métadonnées des documents mais ont une structure irrégulière. Une procédure a été mise au point pour pouvoir recueillir des métadonnées normalisées. Dans un premier temps, un formulaire très simple permet une saisie rapide et systématique des métadonnées dont on souhaite avoir une représentation XML. Une fois ce formulaire validé, une macro Word en récupère les données et génère automatiquement un fichier XML contenant un en-tête documentaire. Dans un second temps, le corps du document est soumis à un convertisseur qui en produit une représentation XML. Les deux fichiers XML ainsi obtenus sont ensuite combinés en utilisant les méthodes du DOM (Rôle, 1999) pour produire un document XML unique avec métadonnées et contenu. Le fichier XML contenant les métadonnées sert à alimenter une base de données classique permettant un premier

niveau de requêtes portant sur le titre, l'auteur, les mots-clés, etc. Le fichier XML complet est lui stocké dans l'arborescence de fichiers du serveur du projet.

Dans *Terminalf*, le contenu est plus structuré ; chaque mémoire a déjà sa propre base de données. Ces bases de données ont été facilement cumulables, car elles ont la même structure informatique. On est d'ailleurs assez vite arrivé à pouvoir faire des recherches sur l'ensemble du corpus. Pour l'introduction d'un mémoire dans *Terminalf*, une interface en ligne donne accès à des formulaires qui permettent la saisie et la modification des métadonnées et du contenu générique, et le dépôt de la base de données dans un répertoire spécifique.

Dans les deux projets, on notera qu'il y a une certaine redondance puisque les métadonnées sont à la fois stockées dans un fichier spécifique et dans une base de donnée générale. Cette redondance fut au départ sans grande conséquence vu le nombre assez faible de documents mais, avec l'accroissement des corpus, elle doit maintenant permettre de pouvoir migrer vers d'autres solutions de stockage basées entièrement sur XML.

4. Discussion

Comme on a pu le constater, les deux projets proposent des opérations différentes sur le contenu : *CodeX* produit une vue du texte alors que *Terminalf* extrait une donnée d'un lexique. Ces deux dispositifs forment un système éditorial de nature différente : pour *CodeX*, le système se fonde sur la structure du document et donne des vues éclatées de celui-ci ; pour *Terminalf*, le système opère à partir des termes du domaine et en donne une vue complexe.

Les deux projets proposent ainsi une visée éditoriale des documents mis en ligne, ce qui souligne le fait qu'un document numérique n'est pas seulement du contenu informationnel mais constitue une *hybridation* entre le contenu et les outils d'accès (outils de lecture comme outils de manipulation).

L'analyse de ces deux systèmes permet d'observer trois phénomènes particulièrement intéressants pour statuer sur la pertinence de diffuser des ressources sorties de leur contexte pédagogique d'origine.

4.1 Effet d'entraînement

Rendre accessible l'intégralité des travaux des étudiants et les valoriser via un support ouvert de consultation a eu un « effet d'entraînement » sur les pratiques de production. La généralisation d'internet et de la bureautique et l'acculturation technique des étudiants qui en a résulté ont provoqué une modification des pratiques d'écriture. L'appropriation de ces outils n'a sans doute pas modifié la qualité du contenu mais plutôt sa présentation et son volume. Nous avons ainsi constaté que les mémoires deviennent plus complexes (avec l'insertion d'images et de graphiques, l'utilisation d'index et de table des matières). Dans *Terminalf*, les bibliographies « s'internetisent » et les arborescences deviennent graphiques.

Cela a aussi entraîné une évolution pédagogique qui s'est traduite par une moindre indulgence dans la notation des mémoires incomplets ou insatisfaisants. Par contre, il y a toujours le risque de juger le mémoire non sur sa qualité intrinsèque mais sur son intégration possible au corpus. Pour l'étudiant, savoir que son mémoire peut se retrouver en ligne reste motivant. Cela a produit un saut qualitatif : au début de *Terminalf*, entre 1998 et 2001, il y a eu une augmentation de 60% du nombre de mémoires « diffusables » sur une promotion. Enfin, le corpus ainsi constitué est devenu une aide pédagogique pour les étudiants qui peuvent désormais consulter d'autres mémoires et avoir ainsi des modèles pour réaliser les leurs.

4.2 « Effet imprimante »

L'« effet imprimante » désigne une pratique qui consiste à utiliser le nouvel outil de publication qu'est la Toile avec un protocole cognitif ancien datant du temps où le seul outil de publication était l'imprimante. Cette pratique est celle de la plupart des archives ouvertes où les documents sont déposés en format PDF et sont figés ainsi dans la linéarité ancestrale du document, sans tenir compte de la possibilité actuelle d'offrir à l'usager d'extraire des « granules » d'information du document ou d'avoir un accès non-linéaire à l'information.

Le langage XML permet de s'affranchir de cet « effet imprimante » pour répondre à d'autres besoins. Il permet de considérer le document comme un ensemble d'informations dont l'usager peut avoir des accès parcellaires par la vue ou circulaires par l'hypertexte. Tout en proposant une vue complète et donc entièrement imprimable du mémoire, *CodeX* propose une médiation plus enrichie avec son système des vues. *Terminalf* n'est que peu concerné par l'« effet imprimante » à cause de l'hybridation originelle des données manipulées. D'ailleurs, extraire un mémoire à partir de ce qui a été intégré dans *Terminalf* et le rendre à l'identique de la version papier déposée à l'université est impossible.

Pour dépasser l'« effet imprimante », nous avons comme projet pour *CodeX* d'aller vers un hypertexte multivalué permettant de mettre plusieurs liens typés sur une unité (différentes définitions d'un terme, accès à des synonymes…). Ces parcours seraient toujours déterminés *a priori* mais devraient pouvoir être personnalisables si une navigation par profil peut être implémentée.

4.3 Statut des données

Le dernier point concerne le statut scientifique des données accessibles qui sont le reflet d'une pratique pédagogique à un moment donné. Peuvent-elles alors prétendre au statut d'information scientifique et technique (IST), au même titre qu'un article de revue scientifique filtré et validé par les pairs ?

Dans les deux projets, l'accès à la ressource est immédiat et la seule évaluation est celle de l'enseignant qui a noté le mémoire en fonction de ses propres objectifs pédagogiques. Cette garantie se montre insuffisante dès qu'il s'agit de diffuser une information pertinente pour un besoin informationnel *autre*. Dans *Terminalf*, une fiche terminologique n'est valide que si on connaît d'autres éléments comme le protocole de recherche, la présentation du domaine, l'arborescence de celui-ci, la date du mémoire ; dans certains domaines spécialisés, le vocabulaire peut changer rapidement. Les mémoires diffusés ont donc toujours été présentés comme des données de travail exploratoires.

Se pose alors la question de la modification éventuelle d'un document afin de l'améliorer ou le mettre à jour. La réponse commune aux deux projets a été de n'accepter aucune correction postérieure à la mise en ligne. Il aurait été en effet illusoire de penser obtenir la même qualité de correction sur l'ensemble du corpus compte tenu de sa croissance et de son évolution thématique. La seconde raison concerne le souci déontologique de respecter l'intégrité du travail remis par l'étudiant et noté par l'enseignant.

À cet égard, le cas de la webographie présente dans les mémoires est significatif : les liens indiqués peuvent devenir obsolètes. Si on enlève les liens morts, on modifie le travail de l'étudiant ; au final, on peut même le décrédibiliser si on en arrive à une webographie inexistante alors que l'étudiant, à l'époque, avait fait sa recherche correctement. Pourtant on ne peut renvoyer l'internaute vers une mauvaise adresse. Il donc été choisi de ne rien modifier en précisant que les données sont datées et sont donc à prendre avec recul par l'utilisateur.

Cette utilisation de travaux d'étudiant soulève de nombreuses questions sur la fiabilité ou la subjectivité des sources d'informations. Pour les deux projets, un lien de confiance est à établir pour permettre d'utiliser « scientifiquement » les données et il doit être instauré vers l'utilisateur mais aussi vers l'étudiant. Ce lien est renforcé par un contexte cumulatif et institutionnel. L'institution donne une

crédibilité de « label » mais cela n'est pas suffisant. On constate en effet que l'« effet de masse » est plus marquant : plus la base est importante et visible, plus elle est prise au sérieux.

5. Conclusion et perspectives

À partir des analyses précédentes, trois constats peuvent être établis :

- Il y a rupture avec la linéarité classique du document ;
- Il y a hybridation des documents et des outils d'accès, d'où une difficulté à définir une granularité du document qui est de plus en plus variable ou adaptable au contexte ou à l'utilisateur ;
- Il y a prédominance des contextes d'usage, variables et évolutifs, dans la définition manipulatoire du document.

Ces contextes d'usage font que le document est manipulé différemment : il se transforme en donnée manipulable plutôt qu'en document fixé *ad vitam aeternam*… On passe aussi d'un ensemble de ressources disponibles à un corpus *par intention*, dans lequel l'utilisateur va pouvoir faire trois types de corpus (*corpus de référence / corpus d'étude / sous-corpus de travail en cours*) grâce au système éditorial utilisé. Ainsi, le simple et nécessaire archivage de pratiques pédagogiques est orienté vers la mise à disposition d'un matériau de travail susceptible de répondre à différents besoins informationnels.

Les documents considérés ici sont des *traces* d'une pratique pédagogique et d'un état de l'art à un instant donné mais, dans un contexte cumulatif et institutionnel (de recherche), ils peuvent devenir des *signes* (une information scientifique et technique). Un des prochains chantiers consistera à perfectionner les outils de manipulation des ressources permettant de renforcer ces *signes* : cartographie, hypertexte dynamique…

6. Bibliographie

Aubry Christine, Janik Joanna (sous la dir. de), *Les Archives ouvertes : Enjeux et pratiques*, Paris, Ed. de l'ADBS, 2005.

Chaudiron S., Role F., Ihadjadene M., « CodeX : un système pour la définition de vues multiples guidées par les usages ». In *Document Électronique Dynamique - Actes du troisième Colloque International sur le Document Électronique*, CIDE 2000, 4-6 juillet 2000, Université de Lyon III, p. 71-81.

Develotte Christine, Pothier Maguy, *La notion de ressources à l'heure du numérique*, Lyon, ENS Editions, 2004.

Jeanneret Yves (sous la dir. de), *Métamorphoses médiatiques, pratiques d'écriture et médiation des savoirs*, Paris, rapport de recherche, février 2005, programme ACI Cognitique.

Pédauque, Roget T., *Document et modernité*, 16 mars 2006, document de synthèse du RTP-DOC (http://rtp-doc.enssib.fr/article.php3?id_article=255)

Rastier François, « Enjeux épistémologiques de la linguistique de corpus », In *Actes des deuxièmes Journées de Linguistique de Corpus*, Geoffrey W. (sous la dir. de), Lorient, septembre 2002, Presses universitaires de Rennes.

Role François, Verdret Philippe, « Le Document Object Model», In *Actes de GUT'99*, Goossens M. (sous la dir. de), Lyon, INPL, p. 155-171.