La fragmentation et l'unité documentaire en question

Stéphane Chaudiron, Madjid Ihadjadene, Azzeddine Maredj

La fragmentation et l'unité documentaire en question

Stéphane Chaudiron
stephane.chaudiron@univ-lille3.fr
Université de Lille3-GERiiCO

Madjid Ihadjadene
ihadjade@u-paris10.fr
Université de Paris10-CRIS

Azzeddine Maredj
amaredj@mail.cerist.dz
CERIST-Alger

Références : Chaudiron S., Ihadjadene M., Maredj A., « La fragmentation et l'unité documentaire en question », Actes du 16ème Congrès de la SFSIC, 11-13 juin 2008, Compiègne.

Résumé :

Dans cette communication, nous nous intéressons au processus de déconstruction/reconstruction de l'unité documentaire à l'oeuvre sur internet, et plus spécifiquement dans le champ de l'information scientifique et technique. Après avoir illustré le processus de fragmentation du document, nous comparons ensuite trois approches qui visent à (re)construire une nouvelle unité : une approche fondée sur l'utilisation du langage de structuration XML, une approche fondée sur les pratiques sociales d'indexation et d'annotation et enfin une approche qui, à partir d'une schématisation a priori des constructions scientifiques, vise notamment à rendre accessibles les données brutes de la science dans le cadre d'une nouvelle unité informationnelle.

Abstract :

This paper presents the deconstruction/reconstruction process of the documentary units on the net in the field of scientific information. We first illustrate the fragmentation process, then we compare three approaches which aim at rebuilding a new documentary unit : an approach based on XML, another based on social indexing and a last one which uses *a priori* modelized frames for writing and editing scientific information in order to make available the scientific results in a e-science context.

Mots-clés : unité documentaire, fragmentation, information scientifique et technique, XML, approche collaborative, données brutes de la science

Keywords : documentary unit, fragmentation, scientific information, XML, collaborative approach, e-science

1. Introduction

Nous connaissons actuellement une mutation majeure du processus de production et de diffusion de l'information scientifique et technique (IST) qui touche l'ensemble du processus d'acquisition, d'indexation, de repérage, de traitement et de commercialisation de l'information. Cette profonde mutation peut s'analyser sur le plan de l'écriture et de la lecture ou encore sur le plan de l'évolution des modèles économiques et des stratégies d'acteurs. Parmi ces éléments de mutation, nous retiendrons ici le phénomène de « *fragmentation de*

l'unité documentaire » qui, d'une certaine manière, double un autre phénomène, largement décrit et commenté, qui est celui de la surcharge informationnelle.

Ces deux phénomènes induisent des effets de désorientation cognitive dans la recherche et l'accès à l'information. D'une part, il n'est plus possible, à l'échelle individuelle, de « consommer », c'est-à-dire de s'approprier le nombre croissant d'écrits touchant à nos propres domaines de recherche. D'autre part, ces écrits sont de plus en plus difficilement identifiables en tant que tels car fragmentés, déstructurés, sortis du contexte qui les légitime, difficilement localisables également à cause de la multiplicité des plates-formes d'hébergement (serveurs commerciaux, bibliothèques électroniques, bases de pré-publications, archives ouvertes, sites personnels…).

En réponse à la surcharge d'informations, trois approches ont été adoptées. À partir des travaux de Gerard Salton dans les années 1950, l'effort a d'abord porté sur la réalisation de systèmes de recherche d'information (SRI) optimisant les processus de recherche de documents. L'accent a été mis, en particulier, sur le développement d'outils performants pour une identification et une extraction optimale de contenus informationnels. Cette première piste reste très féconde comme en témoignent les travaux portant sur les algorithmes d'appariement des index, d'ordonnancement des résultats et le développement de nouvelles fonctionnalités (classification, catégorisation, production de résumés, visualisation de l'information…). Plus récemment, une deuxième approche s'est développée sous le nom générique de web sémantique. Il s'agit d'un ensemble de technologies qui visent à décrire les ressources, notamment documentaires, du *World Wide Web* afin de les rendre accessibles et utilisables par des programmes grâce à un ensemble de métadonnées (comme par exemple le *Dublin Core*). Enfin, et notamment dans le domaine des SHS, une troisième approche concerne des recherches qui ont été engagées sur la modélisation *a priori* des textes scientifiques proposant une écriture multimédia adaptée à internet qui tente d'assurer à la fois la manipulation aisée des constructions scientifiques et l'accès à l'ensemble des données qui les fondent.

Dans cette communication, nous nous intéressons à la question transversale de l'*unité du document* en montrant que, sur internet, il existe une réorganisation constante de l'espace documentaire qui se fonde sur un double processus, quasi-simultané, de déconstruction et de reconstruction de cette unité. Nous montrons d'abord comment elle se déconstruit à différents niveaux aboutissant à une *fragmentation documentaire* qui est l'un des deux phénomènes contribuant à la désorientation cognitive des usagers : au niveau de l'appartenance d'un document à une collection au sens bibliothéconomique du terme, au niveau du document lui-même qui s'est dilué en une multitude de « granules documentaires » d'où la structure d'ensemble est absente, au niveau des « granules documentaires » également qui sont de plus en plus souvent décontextualisées et qui ne correspondent plus nécessairement à des éléments traditionnels (sections, paragraphes…), enfin au niveau de sa (multi)localisation sur le réseau. Nous comparons ensuite trois approches qui visent à (re)construire l'unité du document : une approche fondée sur l'utilisation du langage de structuration XML, une approche fondée sur les pratiques sociales d'indexation et d'annotation et enfin une approche qui, à partir d'une schématisation *a priori* des constructions scientifiques, vise notamment à rendre accessibles les données brutes de la science dans le cadre d'une nouvelle unité informationnelle.

2. Les différents niveaux de fragmentation documentaire

Sur internet, la notion même de document telle que défini par l'AFNOR TC 46 comme « une information enregistrée sur un support » est dépassée. Les technologies numériques associées à internet ont progressivement conduit à une déstructuration complète de la notion originelle de document mais non pas à sa suppression ou son effacement. La permanence du document a notamment été attestée par les travaux menés par le collectif Roger T. Pédauque (Pédauque, 2007) autour du concept de « redocumentarisation du monde ». Notre réflexion nous porte à détailler un aspect particulier de ce processus qui est celui de la fragmentation ou de la déconstruction de l'unité documentaire (avant d'aborder dans les sections suivantes celui de la reconstruction). Certes, ce processus est indéniablement lié à un contexte technologique particulier, l'économie numérique, mais il est également induit par des pratiques informationnelles nouvelles (le dépôt des pré-publications dans les archives ouvertes par exemple) et une mutation des modèles économiques des grands éditeurs (la vente à l'article ou au chapitre). Différents niveaux de fragmentation documentaire caractérisent le document dans l'espace numérique dont nous présentons quelques exemples pour illustrer notre propos.

En premier lieu, se pose la question de l'appartenance d'un document à une collection au sens bibliothéconomique du terme. Nous ne nous appesantirons pas sur ce point qui a déjà été souligné à maintes reprises mais nous rappellerons seulement que, dans le champ de l'information scientifique et technique, cette déconstruction de l'« effet collection » est particulièrement forte. Pour illustrer ce phénomène, citons le cas du dépôt des *pre-prints*, des *post-prints* ou des communications à des colloques dans les entrepôts d'archives institutionnelles qui, ce faisant, rompt avec l'unité formée par la revue ou l'ensemble des actes. Ainsi, dans le domaine des SIC, le dépôt dans HAL d'un article paru dans la revue *Études de communication* ou sa mise en ligne sur un site personnel dénoue le « fil rouge » qui lie les textes et en assure la jonction inte-textuelle[1]. Un autre exemple est donné par les contributions aux colloques qui sont de plus en plus souvent déposées dans les mêmes archives. C'est le cas de plusieurs communications du dernier congrès du Chapitre français de l'ISKO (Régimbeau G., Couzinet V., 2007) qui sont accessibles directement en ligne comme des objets numériques isolés de leur contexte de communication scientifique. Aucun des éléments de mise en contexte (introduction, sommaire, intitulé des sessions…) n'est plus présent pour relier la communication aux autres.

Un deuxième phénomène de déconstruction s'observe au niveau du document lui-même qui s'est dilué en une multitude de « granules documentaires » d'où la structure d'ensemble est absente. L'unité-document est ainsi remise en cause pour des raisons de facilité d'accès et de lisibilité. Grâce à l'apport des langages structurés (SGML puis XML), la consultation en ligne de textes volumineux comme les thèses peut désormais s'effectuer par partie, par chapitre, par section, etc. en s'appuyant sur des DTD (*Document Type Definition*). Ainsi, en se fondant sur la structure originelle du texte, l'accès au contenu est désormais possible par « granule documentaire » de taille variable. Le balisage du document qui s'effectue grâce aux DTD en fonction d'une structure déterminée *a priori* induit une fragmentation de l'unité documentaire globale en une multitude de fragments, de « granules » qui sont désormais accessibles en tant que tels. Plutôt que d'accéder à la totalité d'une « œuvre », l'usager a ainsi la possibilité d'accéder à des fragments dont la nature est déterminée *a priori*, et selon des chemins d'accès qui peuvent être statiques, c'est-à-dire déterminés eux aussi *a priori*, ou dynamiques, c'est-à-dire construits interactivement.

[1] Cette revue se caractérise notamment par la présence de textes courts, appelés « fil rouge », qui assurent la transition entre les différents articles thématiques du numéro et qui sont physiquement situés sur les pages précédant les articles eux-mêmes.

Un troisième phénomène de fragmentation se passe au niveau des « granules documentaires » qui sont de plus en plus souvent décontextualisées et qui ne correspondent plus nécessairement à des éléments traditionnels (sections, paragraphes…) du document. La sophistication croissante des outils de recherche d'information permet d'identifier et d'extraire d'un corpus (collection raisonnée de documents) ou d'un document des éléments textuels (mais il peut aussi s'agir d'autres objets tels que des figures, des tableaux, des images…) qui sont ensuite compilés, réorganisés, reconstruits afin de produire un nouveau sens. C'est le cas par exemple des *items* qui sont obtenus à l'aide de logiciels d'acquisition terminologique et qui peuvent alors être considérés comme des candidats-termes pour construire une terminologie ou les agrégats d'une classification (*clusterisation*) ou des candidats-descripteurs pour élaborer un thésaurus. Les termes ou syntagmes ainsi identifiés sont des fragments documentaires presque minimaux produits (semi)automatiquement, nécessaires à la reconstruction d'une nouvelle unité ultérieure qui sera celle de la représentation d'un document (index, résumé automatique) ou de son analyse comme dans le cas de la détection de signaux faibles en situation de veille informationnelle.

Dans la même logique, et même si les ressources nécessaires sont différentes, on peut mentionner le repérage et l'extraction des entités nommées (noms de personnes, de lieux, de marques…) ainsi que la détection et le suivi de thèmes ou d'évènements (opération de fusion-acquisition d'entreprises par exemple). Ainsi *BioText Search Engine* fournit à l'usager la possibilité de rechercher des informations présentes dans les tableaux, graphiques et illustrations contenus dans les articles biologiques. Dans le domaine de la recherche d'information également, ce processus de déconstruction du document est à l'œuvre avec les systèmes de recherche en question/réponse qui visent à retrouver la réponse exacte à une question précise dans une collection de documents et non l'ensemble des documents comportant les mots-clés mentionnés dans la requête. Ce genre de systèmes, en recontextualisant immédiatement la réponse fournie par rapport à la question initiale après l'avoir extraite d'un document, illustre également le mouvement de déconstruction/reconstruction mentionnée ci-dessus.

Nous évoquerons une dernière occurrence de ce phénomène de fragmentation qui se joue cette fois au niveau de la (multi)localisation des unités documentaires sur le réseau. L'effacement de l'unité-collection est spatialement renforcé par la multiplicité de la localisation des documents tels que les pré-publications et les publications. Sites personnels de chercheurs, archives institutionnelles (comme PubMed ou HAL Inserm), archives ouvertes (@rchiveSIC), bibliothèques numériques commerciales (Springer Link, Science Direct d'Elsevier, Proquest…) sont autant de lieux de dépôt obéissant à des stratégies différentes, tant de la part des déposants, de leur tutelle institutionnelle que des organismes à l'origine de ces lieux. Du point de vue qui nous intéresse ici, la situation est qu'une même ressource informationnelle peut être accessible à différents URLs mais dans des conditions techniques et légales différentes. De plus, certains documents disparaissent après quelques mois (le cas d'un *pre-print* accepté par une revue par exemple) alors que d'autres apparaissent en ligne après expiration de la barrière mobile. Ce mouvement d'apparition/disparition diffère selon les lieux et le respect des contraintes légales, mais n'est pas simultané et introduit une dimension temporelle à la question de la fragmentation. La fragmentation de l'unité documentaire s'effectue alors selon un double axe, spatial et temporel.

Enfin, pour conclure cette section, nous évoquerons l'exemple cité par Xavier Sense (Sense, 2008) qui illustre le double mouvement de déterritorialisation/reterritorialisation qui peut être à l'œuvre sur internet. Se référant au projet artistique de Reynald Drouhin,

Rhizomes[2], il souligne précisément la capacité d'internet à « distribuer l'ensemble des données discrètes d'un document sur plusieurs machines, et non simplement celle de les déposer sur un seul et même serveur » puis de les « recomposer sur une seule et même page web, hébergée quant à elle sur un serveur ».

Après avoir illustré la phase de déconstruction de l'unité documentaire, nous présentons maintenant différentes approches qui reconstruisent une « unité » mais selon des principes différents.

3. Une approche (re)structurante : XML

La première approche s'appuie sur une (re)structuration des fragments documentaires en vue de favoriser l'accès à l'information, soit par recherche, soit par navigation. Habituellement, la structure des documents n'est prise en compte ni au niveau de la requête pour exprimer des proximités structurelles, ni au niveau de la réponse qui pourrait alors ne retourner que les parties des documents pertinents vis-à-vis de la requête. Les systèmes de recherche d'information structurée proposent de nouvelles possibilités d'accès portant à la fois sur la structure interne du document, sur sa structure externe (effet d'appartenance à une collection) et son contenu.

Le langage de structuration XML permet de décrire les documents numériques de façon aussi fine que nécessaire en explicitant les liens hiérarchiques et de références entre ses différentes parties. De plus, associés à XML, des langages normalisés sont maintenant disponibles pour manipuler les documents ainsi décrits. Les évaluations effectuées dans le cadre de plusieurs projets de bibliothèques numériques ont montré que très peu d'articles et de thèses numérisés sont lus en entiers. La consultation se fait plutôt sur des parties du document comme le résumé, l'introduction, le sommaire… En raison des contraintes de temps (téléchargements), du dispositif de visualisation utilisé (l'écran de l'ordinateur par exemple), de lieu, de coût, l'usager ne lit pas le document numérique intégralement, mais essaye de formuler une représentation globale du contenu textuel en lisant quelques parties du document afin de décider de la pertinence des informations contenues dans le texte. Plutôt que de donner accès aux documents dans leur entier, l'approche consiste à donner à l'utilisateur la possibilité de les consulter à travers des modes de visualisation (des « vues ») ne retenant que les aspects pertinents pour un usage précis. On peut ainsi définir le système de vues comme un outil de filtrage de l'information s'appuyant simultanément sur la structure interne du document et sur les besoins informationnels des usagers. Le système de vue est un mode de consultation qui permet à l'usager de visualiser des facettes différentes du même document ou du sous-ensemble de la base documentaire. C'est ce que nous avons développé dans le cadre du projet CodeX (Chaudiron, Role, Ihadjadene, 2000) pour lequel un ensemble de 5 vues statiques a été défini, permettant la consultation de différents agrégats de fragments documentaires :

- Visualisation du titre, de l'introduction et de la conclusion ;
- Visualisation du titre et du résumé d'auteur ;
- Visualisation du titre, du résumé d'auteur et de la bibliographie ;
- Visualisation des titres de sections et/ou de sous-sections (reconstitution d'un équivalent de la table des matières) ;
- Visualisation de l'index auteur offrant une navigation intra/inter documentaire.

[2] http://www.incident.net/works/rhizomes

Dans la continuité de ces travaux, nous travaillons actuellement sur l'intégration d'un outil de recherche s'appuyant simultanément sur la structure des thèses et sur une indexation automatique des différents fragments. Le choix de la pondération des termes est fondé conjointement sur le calcul des fréquences des termes dans un document, l'importance locale du terme dans un fragment, ainsi que sur les fréquences inverses qui permettent de rendre compte de l'importance globale du terme respectivement dans la collection de documents et dans la collection des fragments. Le but est alors d'utiliser cette structure et ces pondérations afin de renvoyer à l'usager des éléments de granularité appropriés à ses besoins et classés par ordre de pertinence.

Cette fragmentation en de multiples unités peut aussi s'accompagner d'une personnalisation de la recherche reposant sur d'autres critères : domaine d'intérêt, temps, espace, etc. Certaines applications s'appuient sur la description d'un texte selon différents niveaux d'analyse correspondant à des usages différents. C'est ce que Noureddine Chatti et Sylvie Calabretto (Chatti et Calabretto, 2007) appellent la multi-structuration, en ce sens que plusieurs structurations différentes peuvent être associées à un même document, telles que la structure logique, physique, sémantique, temporelle ou spatiale.

En ce qui concerne l'accès par navigation, la généralisation des accès au niveau académique par l'intermédiaire de portails scientifiques de type *Science Direct* d'Elsevier ou *Pubmed* offre aux chercheurs un « univers informationnel » structuré par les hyperliens. Sur des plates-formes comme *Crossref*, l'article ou la référence bibliographique est enrichi d'un ensemble de liens offrant à l'usager la possibilité de constituer sa « propre » bibliothèque numérique au gré de ses recherches et de sa navigation. À ces liens inférés automatiquement à partir des banques de données, sont parfois ajoutés d'autres services permettant par exemple une navigation conceptuelle et participative à travers les *tags* des usagers, des *blogs*, des *wikis*…

Un autre exemple de la reconstruction de l'unité documentaire est issu du monde des bibliothèques dans lequel il existe un ensemble de travaux qui visent à modéliser les pratiques de catalogage et d'indexation en refondant l'unité documentaire de base qu'est la notice bibliographique pour se référer à la notion d'œuvre intellectuelle ou artistique. Dans cette approche, la notion d'œuvre permet de lier par exemple un roman et ses traductions ou ses adaptations, ce que les catalogues en ligne sont actuellement incapables de faire. Le modèle FRBR (*Functional Requirements for Bibliographic Records*) proposé par un groupe d'experts de l'IFLA **(**Madison, 2006) offre un cadre conceptuel pour la mise en œuvre d'un catalogue enrichie de ces relations en proposant quatre niveaux :

- *œuvre* : une création intellectuelle ou artistique (par exemple, *Germinal* de Zola) ;
- *expression* : une réalisation de cette création intellectuelle (par exemple, la traduction anglaise de *Germinal* par Roger Pearson) ;
- *manifestation* : la matérialisation d'une expression (par exemple, *Germinal* de Zola, traduit par Roger Pearson et publié chez Penguin Books en 2004) ;
- *item* : un exemplaire isolé d'une manifestation (par exemple, l'exemplaire de *Germinal* de Zola, traduit par Roger Pearson et publié chez Penguin Books en 2004, qui se trouve à la bibliothèque municipale de Perpignan).

En plus de ces entités, le modèle FRBR introduit d'autres types de liens qui relient la ressource décrite aux autorités impliquées dans le processus de création et de production ainsi qu'aux concepts qu'elle contient, constituant ainsi une base de connaissance pour l'usager que l'augmentation des points d'accès que les actuels catalogues en ligne ont tendance à favoriser.

4. Les approches collaboratives

Les folksonomies, terme né de la contraction de l'anglais « *folks* » (les gens) et « *taxonomy* » (classification hiérarchique), sont des étiquettes apposées par les internautes eux-mêmes sur des ressources documentaires en vertu de leur double statut d'éditeurs et de consommateurs d'information. Les folksonomies attestent de la possibilité de construire un processus d'interaction collective sans se connaître grâce à l'utilisation d'artefacts spécifiques. Dans le processus de reconstruction de l'unité documentaire, le principe de l'étiquetage collaboratif nous semble particulièrement intéressant. Par ce jeu communautaire émergent des agrégats de ressources dont l'unité n'est plus interne (comme dans le cas de la classification ou de l'indexation automatique) ni médiée par des professionnels (comme dans le cas de l'indexation contrôlée) mais externe et impulsée par les pratiques informationnelles des internautes.

Un exemple intéressant dans le monde des bibliothèques est celui de *LibraryThing*[3] qui permet aux usagers de mettre en ligne leurs propres bibliothèques et de la partager. Comme avec *Del.icio.us* ou *FlickR*, les internautes utilisent des étiquettes pour décrire leurs livres, enregistrer leurs critiques et les partager. L'indexation collaborative des ouvrages est aussi dynamique (la description d'une ressource donnée peut changer au cours du temps). Cette description « documentaire » se fait d'une façon libre sans se référer à des normes ou à l'idée de collection pré-établie. L'internaute identifie simplement les mots-clés qui lui semblent les plus pertinents pour un *item* particulier. Contrairement aux langages documentaires qui sont des outils de médiation entre des collections et des usagers et au processus d'indexation qui consiste à identifier dans un document certains éléments significatifs qui serviront de clés pour retrouver le document au sein d'une collection, les pratiques de *tagging* se fondent sur des items isolés hors de toute collection.

Toutefois, dans cette approche hors de toute médiation documentaire au sens professionnel du terme, se recrée une unité documentaire, presque à l'insu des internautes. L'utilisation des nuages de *tags*, la fréquence d'apparition des titres, la popularité des livres sont autant d'éléments qui rétablissent l'idée de la collection (personnelle ou collective). En particulier, le regroupement des *tags* en *nuage de mots clés* offre une visualisation des mots clés les plus fréquemment utilisés, dans un annuaire de sites syndiqués.

5. L'« écriture » des données brutes de la science

Enfin, le dernier exemple de reconstruction de l'unité documentaire concerne un domaine particulier de l'IST, celui des données brutes de la science. L'enjeu est double : il s'agit d'une part de conserver de manière pérenne les résultats des observations scientifiques, les données initiales sur lesquelles se fondent les constructions scientifiques et qui aboutissent aux diverses formes de publications des résultats et d'autre part d'élaborer une « écriture » permettant de communiquer les résultats, adaptée aux technologies numériques et à internet. Avec cette écriture, il s'agit par exemple d'associer une publication scientifique aux manipulations expérimentales qui en sont à l'origine afin, par exemple, d'assurer la reproductibilité de la démarche.

---

[3] http://www.librarything.com. (les membres de Librarything ont catalogué à ce jour plus de 21 millions de livres).

Dans le domaine des SHS, des recherches ont été engagées depuis le milieu des années 1980 sur la modélisation *a priori* des textes scientifiques pour proposer une écriture multimédia adaptée à internet qui tente à la fois d'assurer la manipulation aisée des constructions scientifiques et l'accès à l'ensemble des données qui les fondent. Une telle approche, proposée par Jean Claude Gardin (Gardin et Roux, 2004 et Roux, 2004), suggère ainsi de restituer l'architecture de nos constructions scientifiques sous forme de schématisations à travers un formalisme particulier, le format SCD (*Scientific Constructs & Data*). Ce format réorganise l'écriture des textes scientifiques selon une analyse logiciste qui distingue quatre niveaux de consultation de la publication, identifiant clairement les données, les propositions initiales, la démarche et le raisonnement et les résultats. Le projet ARKEOTEK[4] est un bon exemple de mise en œuvre de ce mode de publication dont l'objectif est la constitution de bases de connaissances.

Pour notre propos, l'intérêt de travaux de Jean Claude Gardin se situe dans la démarche de communication des données brutes de la science qui s'inscrit dans une préoccupation très actuelle, celle de l'*e-science*, concernant non seulement la conservation pérenne des données recueillies mais aussi de leur articulation avec la publication des résultats. On assiste ici également à un processus de redéfinition de l'unité documentaire dans le domaine de l'IST qui concernant plusieurs disciplines, l'astronomie avec l'*International Virtual Observatory Alliance* ou la bioinformatique avec le projet *Uni Prot Knowledge Base* qui contient près de cinq millions de structures de protéines associées à des outils informatiques permettant de manipuler les séquences. En France, on peut mentionner le projet Adonis qui vise à offrir aux chercheurs un « espace de navigation unifié » pour explorer les documents et les données numériques des sciences humaines et sociales.

Les travaux actuels sur les cyber-infrastructures questionnent la démarche scientifique et les moyens permettant de valider le processus de collecte, d'analyse, de traitement et de restitution des données. Le terme de cyber-infrastructure décrit un ou des environnements unifiés de recherche dans lesquels les outils informatiques sont à la disposition des chercheurs dans le cadre d'un réseau interopérable. Au-delà du document numérique (article, monographie…) ou de la donnée brute, il s'agit dorénavant de proposer un accès aux protocoles d'expériences, aux collections de données (corpus de textes, de sons, d'images et vidéos), et aux logiciels utilisés dans les expérimentations scientifiques. Ces travaux posent non seulement le problème de la place des infomédiaires et de leurs pratiques dans les cyber-infrastructures mais aussi, et c'est ce qui nous intéresse ici, la question de l'unité documentaire. Se construit ainsi un continuum dans le « construit » scientifique où les données observées, les protocoles, les hypothèses, les inférences, les résultats, etc. s'articulent dans une structure en réseau. Pour le bibliothécaire ou le documentaliste, l'enjeu ne sera plus simplement de gérer la publication des résultats de la recherche mais de gérer les données de recherche elles-mêmes, notamment en les annotant sémantiquement et en veillant aux choix de métadonnées intéroperables.

Cette mise à disposition des données brutes et des outils de manipulation interroge également les pratiques scientifiques des chercheurs des différentes disciplines et la place du chercheur au sein de cet environnement. Pour les disciplines expérimentales par exemple, la disponibilité de ces données brutes est une condition de la reproduction et de la réplication des résultats expérimentaux, nécessaire à de toute connaissance scientifique. Pour une discipline comme la nôtre, les SIC, cela peut également induire une modification de nos

[4] http://arkeotek.org/

pratiques de chercheurs, une réflexion sur nos modes de « pensée », d'argumentation mais aussi une évolution des modes d'évaluation « scientifique ».

6. Conclusion

Dans cette communication, nous avons essayé de montrer que l'économie du numérique induit simultanément un double mouvement de fragmentation et de recomposition de l'unité documentaire. Alors que l'unité documentaire a été relativement stable depuis au moins un siècle, le phénomène de fragmentation affecte le document à de multiples niveaux. Inversement, nous avons également essayé de montrer que nous sommes dans une phase de mutation qui inclut des phénomènes de reconstruction de (nouvelles) formes de collection. Les innovations techniques ainsi que les nouveaux modèles économiques ont abouti à suggérer des logiques d'accès renouvelées aux ressources scientifiques et à induire par conséquent de nouveaux modes de lecture. À travers la syndication de contenu par les fils RSS, l'achat d'un article, d'un chapitre d'ouvrage ou l'abonnement à un « pack » informationnel (d'ailleurs plus ou moins imposé par les éditeurs), la question de l'unité documentaire se renouvelle. Avec l'émergence des infrastructures numériques, d'autres enjeux et d'autres potentialités émergent qui remettent en cause l'idée traditionnelle de la collection documentaire et par conséquent les outils utilisés, les normes, les pratiques, mais qui suggère une nouvelle appréhension de la notion d'unité.

Le point commun de ces différents phénomènes de déconstruction/reconstruction est une recomposition de l'idée documentaire. Le mécanisme de construction des bibliothèques personnelles et donc de reconstruction d'une collection individuelle, déjà à l'œuvre depuis plusieurs années, est amplifié. En conclusion, on peut avancer l'idée qu'il n'y a pas de disparition de l'unité documentaire mais que celle-ci se recrée désormais du côté de l'usager. Les dispositifs de reconstruction de l'unité documentaire que nous avons montrés visent en effet à permettre de personnaliser l'accès aux collections et aux œuvres en s'appropriant les différentes ressources numériques disponibles. C'est donc l'usager qui recompose l'unité informationnelle, la reconstruit constamment dans un processus interactif de recherche et de navigation.

7. Références

Chatti N., Calabretto S., 2007, « Les documents multistructurés. Encodage et interrogation », *Document numérique*, vol. 10, n° 1, pp. 39-62.

Chaudiron S., Role F., Ihadjadene M., 2000, « CodeX : un système pour la définition de vues multiples guidées par les usages », *CIDE 2000*, 4-6 juillet 2000, Université de Lyon III, pp. 71-81.

Gardin J.-C., Roux V., 2004, « The Arkeotek project : a european network of knowledge bases in the archaeology of techniques », *Archeologia e Calcolatori*, n°15, pp. 25-40.

Lainé-Cruzel S.,Guinet E., 2000, « Fragmentation et enrichissement de textes scientifiques sous forme électronique », *Document numérique*, , vol 4, n°1 et 2, pp.59-84.

Madison O., 2006, « Utilizing the FRBR Framework in Designing User-Focused Digital Content and Access Systems », *Library Resources & Technical Services*, vol. 50, n° 1, pp. 10-15.

Pédauque R. T., 2007, *La redocumentarisation du monde*, Éd. Cépadues, 212 p.

Régimbeau G., Couzinet V. (dir.), *Actes du colloque Organisation des connaissances et société des savoirs : concepts, usages, acteurs*, Chapitre français de l'ISKO et Université de Toulouse III, 7 et 8 juin 2007, Toulouse, 2007.

Roux, V., « Faciliter la consultation de textes scientifiques. Nouvelles pratiques éditoriales », *Hermès*, n°39, 2004, pp. 151-159.

Sense X., « De l'espace du document numérique à l'espace d'internet : projet d'une pensée (im)possible », *Etudes de Communication*, n°30, 2007, pp. 99-114.

Smiraglia R. P., 2007, « The "works" phenomenon and best selling books », *Cataloging & Classification Quarterly*, vol. 44, n° 3 et 4, pp. 179-195.