L'édition électronique change tout et rien. Dépasser les promesses de l'édition électronique

Gautier Poupeau

L'édition électronique change tout et rien

Dépasser les promesses de l'édition électronique

Gautier Poupeau, gpoupeau@enc.sorbonne.fr, Ecole nationale des chartes.

L'édition électronique[1] a fait l'objet de beaucoup d'attention depuis plusieurs années[2]. Les débats sur le bien-fondé[3] et les promesses[4] ont maintenant laissé place aux expériences[5]. Pour autant, les particularités et le mode de fonctionnement de ce nouveau support sont loin d'être complètement maîtrisés par les concepteurs et par les utilisateurs. L'histoire des médias montre qu'à chaque apparition d'un nouveau support, les hommes ont cherché à reproduire le média existant. Ce constat est aussi vrai pour Internet et le Web, ce qu'Olivier Guyotjeannin a résumé par cette phrase : « L'édition électronique change tout et rien »[6]. Cette phrase visait la présentation et les modes d'accès au Cartulaire blanc[7]. Pourtant, il résume aujourd'hui notre sentiment à l'égard de l'édition électronique.

L'introduction de l'informatique dans la discipline historique a provoqué une première révolution dans le traitement des données, ce qui faisait dire à Emmanuel Le Roy Ladurie : « l'historien de demain sera programmeur ou il ne sera plus »[8]. Trente ans plus tard, le constat est moins catégorique, surtout grâce à l'avènement de l'informatique personnelle dans les années 1980 et 1990. Aujourd'hui, Internet et surtout une de ses composantes, le Web[9], sont en train de provoquer une deuxième révolution, ou plutôt évolution – soyons prudents – dans

[1] Dans le cadre de cet article, nous concevons l'édition électronique comme le fait d'éditer en ligne des documents consultables par l'intermédiaire d'un navigateur Web.

[2] *Le médiéviste et l'ordinateur* aborde cette problématique depuis 1994, année de l'apparition du premier navigateur Web. Depuis cette date, tous les numéros de la revue contiennent au moins un article abordant la mise en ligne de textes ou d'outils sur Internet.

[3] L' édition électronique a souvent été accusée à tort de risque de supplanter détruire le support papier, cf. en particulier le débat autour de l'article de Robert Darnton, « The new age of the book », *New York Review of books*, 18 mars 1999, http://www.nybooks.com/articles/546 et la polémique qui en a découlé : Daniel Garcia, « 'L'affaire' Darnton, les sciences humaines tuées Net ? », *Livres hebdo*, n°331, 2 avril 1999 et Daniel Garcia, « Les thèses provocantes de Robert Darnton », *Livres hebdo*, n°331, 2 avril 1999.

[4] Voir, par exemple, l'article de Marin Dacos, « Le numérique au secours du papier. L'avenir de l'information scientifique des historiens à l'heure des réseaux », *Cahiers d'histoire*, n°1, 1999, http://ch.revues.org/document48.html, consulté le 25 mai 2004.

[5] Pour prouver la multiplication des expériences, il suffit de voir l'augmentation des différents annuaires de liens spécialisés en histoire.

[6] Guyotjeannin Olivier et Poupeau Gautier, « Le projet d'édition électronique du Cartulaire blanc de Saint-Denis et projets électroniques de l'Ecole nationale des chartes », dans le *Médiéviste et l'ordinateur*, t. 42, printemps 2003, la Diplomatique, [**en ligne**], http://lemo.irht.cnrs.fr/42/mo42_12.htm, consulté le 9 mai 2004.

[7] Olivier Guyotjeannin (dir.), *Le Cartulaire blanc de l'abbaye de Saint-Denis*, [**en ligne**], http://www.enc.sorbonne.fr/cartulaireblanc/.

[8] Emmanuel Le Roy Ladurie, *Le territoire de l'historien*, 1973, p. 74.

[9] Nous rappelons que le Web (ou World Wide Web) n'est qu'une des applications d'Internet avec le protocole http.

l'accès aux données de la recherche[10] et donc dans l'utilisation et l'exploitation de ces données[11]. Pour autant, les bases de la discipline historique n'ont pas changé ; les mécanismes de production et de diffusion des résultats de la recherche sont toujours les mêmes. Les revues restent primordiales dans l'actualité de la recherche, les monographies et les thèses ont gardé leur importance dans la reconnaissance du chercheur et les colloques restent le moyen privilégié de débats concernant un sujet précis. Il est important de ne pas perdre de vue ces mécanismes et de les prendre en compte, pour que l'édition électronique acquière une légitimité auprès des chercheurs.

Cette prise en compte ne doit pas faire oublier les spécificités du support numérique, et du Web en particulier. Comme l'édition papier, la mise en ligne de documents sur le Web impose la connaissance de techniques, de langages et de codes. Cette connaissance garantit à la fois la cohérence de l'information, son accès, son exploitation et sa conservation, quatre principes anciens et valables pour tous les médias. C'est pourquoi il nous a paru intéressant dans le cadre de cet article de s'attacher à ces principes en essayant de montrer pour chacun d'entre eux les techniques à notre disposition aujourd'hui et l'état de la réflexion.

[10] Voir en particulier l'article de Steven Harnard, « Lecture et écriture scientifique "dans le ciel" : Une anomalie post - gutenbergienne et comment la résoudre », *Colloque virtuel text-e.org, Ecrans et réseaux, vers une transformation du rapport à l'écrit*, http://www.text-e.org/conf/index.cfm?fa=texte&ConfText_ID=7, consulté le 12 mai 2004.

[11] Par exemple, le DEEDS poject mené par Michael Gervers, http://www.utoronto.ca/deeds/research/research.html et l'article, Michael Gervers, The DEEDS Project, « Towards the dating and analysis of english private charters of the twelfth and thirteenth centuries », dans *le Médiéviste et l'ordinateur*, t. 41, L'apport cognitif, [**en ligne**], http://lemo.irht.cnrs.fr/41/mo41_07.htm, consulté le 9 mai 2004.

I- L'information historique et l'édition électronique

Différents types de documents et de publications sont manipulés par l'historien. Les sources primaires, les éditions de sources, les outils de recherche (dictionnaires, outils bibliographiques, inventaires et catalogues…), les articles (périodiques, actes de colloques, mélanges…) et les monographies sont les principaux. Chacun tient une place précise voire institutionnalisée dans le cadre de la recherche. Il n'est donc pas étonnant que les projets d'édition électronique soient, dans la grande majorité des cas, des essais d'adaptation de ces différentes publications. L'adoption de ce nouveau support est ainsi facilitée par la conservation des repères des chercheurs. Chaque type de publication possède ses caractéristiques et ses règles en termes d'écriture, de structuration logique et de présentation. Les technologies utilisées pour mettre en ligne ces différents types de publication devront respecter ces différentes caractéristiques et ainsi garantir la cohérence de l'information.

A- La source brute

La source primaire est le principal matériau dans la recherche de l'historien. Documents d'archives, manuscrits, incunables, livres imprimés ou encore documents figurés et films, elle prend des formes diverses. Cette diversité a des conséquences sur sa mise à disposition sur le Web. Pour autant, un souci majeur et commun se pose pour toutes les sources : garantir l'intégrité originale de la source, afin d'assurer sa véracité et donc son examen par les chercheurs. La numérisation en mode texte ou image, que nous concevons comme la mise à disposition d'un fac-similé le plus proche possible de l'original[12], semble la technique la plus appropriée dans cette optique. Elle diffère de l'édition électronique dans le fait que les sites dont la vocation est la numérisation proposent une valeur ajoutée non pas en termes de contenu éditorial par rapport à l'original, mais plutôt en termes d'accessibilité et de disponibilité. C'est pourquoi la numérisation s'intègre parfaitement dans les missions des institutions de conservation. Elle permet, de plus, de garantir la conservation de l'original, les chercheurs pouvant travailler sur ce fac-similé, et de résoudre le difficile problème ces institutions dont le rôle est à la fois de conserver tout en permettant un accès aux documents.

[12] La numérisation provoque forcément une perte d'informations par rapport à l'original. Par exemple, dans le cas d'une numérisation en mode image, une image numérisée ne pourra rendre toutes les teintes et tous les infimes détails de l'original.

La numérisation en mode image est souvent décriée, car elle ne permet pas a-priori la recherche à l'intérieur des documents[13]. Pourtant, son utilisation correspond à deux besoins : la mise à disposition d'une masse importante de documents dans un souci de valorisation du patrimoine[14] et la possibilité d'offrir au public le fac-similé le plus proche de la forme originale du document. Cette dernière application trouve toute son utilité dans le cadre d'études codicologiques, iconographiques ou paléographiques, par exemple, pour lesquelles le mode image est indispensable[15]. Dans ce cadre, la numérisation en mode image permet certaines manipulations des images impossibles sur l'original. Le zoom permet, par exemple, de mettre en lumière des détails impossibles à voir à l'œil nu.

La numérisation ne doit pas se faire sans quelques règles. Les formats d'images sont très nombreux et certains semblent plus pérennes que d'autres ou répondent à des besoins différents. Ainsi, dans un souci de conservation, il est toujours intéressant d'enregistrer les documents sous deux formats : un dédié à la conservation et qui présentera le moins de perte de qualité par rapport à l'original, comme TIFF, et un dédié à la diffusion beaucoup plus léger en poids et, par conséquent, de moins bonne qualité, comme JPEG[16]. On pourra éventuellement proposer en complément des formats moins pérennes mais qui peuvent rendre ponctuellement des services intéressants comme DjVU[17] ou PDF[18] qui permettent la manipulation de l'image (déplacement, zoom…) grâce à l'installation d'un greffon[19].

La numérisation fait perdre la structure physique du document numérisé, puisque nous disposons dans la plupart des cas de fichiers disparates correspondants à une page du document. Toutefois, la plupart des projets de numérisation recréent artificiellement la structure physique

[13] Il existe néammoins des sites qui proposent des services d'OCR à la volée, qui permettent l'interrogation en texte intégral du mode image.

[14] A cet égard, Gallica, la bibliothèque numérique de la BnF est un exemple. Le but de cette bibliothèque numérique est de composer une bibliothèque encyclopédique de langue française et de permettre l'accès simplifié à un grand nombre de documents. Or, le coût aurait été 10 fois supérieurs s'il avait fallu numériser en mode texte. Pour plus de renseignements, vous pouvez vous reporter à la FAQ de Gallica : http://gallica.bnf.fr/faq.htm, consulté le 12 mai 2004.

[15] cf. le numéro du médiéviste et l'ordinateur consacré à cette question : *Le médiéviste et l'ordinateur*, La numérisation des manuscrits médiévaux, t. 40, Automne 2000, [**en ligne**], http://lemo.irht.cnrs.fr/40/mo40-toc.htm, consulté le 12 mai 2004.

[16] Pour plus de renseignements techniques, Ministère de la Culture, *Numérisation du patrimoine culturel : informations techniques*, octobre 2000, [**en ligne**], http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_04.htm, consulté le 12 mai 2004 ou IRHT, *Cours de numérisation d'images, de l'acquisition à la numérisation*, [**en ligne**], http://www.irht.cnrs.fr/formation/cours/intro.htm, consulté le 17 mais 2004.

[17] DjVU pour déjàvu, pour plus d'informations concernant ce format, cf. Nicholas Brousseau et Gautier Poupeau, « La numérisation et la mise en ligne des diplômes de Charles le Chauve à la lumière de l'expérience du projet 'Kaiserurkunden in Abbildung ' mené à la Bayerische Staatsbibliothek à Munich » dans *Le médiéviste et l'ordinateur*, n°42, printemps 2003, éd. IRHT, Paris, France, 2003, [**En ligne**], http://lemo.irht.cnrs.fr/42/mo42_02.htm, consulté le 12 mai 2004.

[18] PDF pour Portable Document Format, format semi-propriétaire créé par la société américaine Adobe.

[19] Greffon ou plugin : petit logiciel à télécharger qui permet d'ajouter une fonctionnalité au navigateur Web.

du document par l'interface de navigation. Un standard, tel que METS[20] permet grâce à un fichier XML de reconstituer logiquement la structure physique du ou des documents. Dans ce cas, ce n'est pas l'interface mais une construction logique et organisée qui permet de reconstuire la stucture du document numérisé et d'en garantir l'intégrité physique.

B- L'édition critique de sources

L'édition de sources est « un secteur à part entière du travail de l'historien »[21], pourtant elle occupe une place particulère dans ses productions. Peu pratiquée, elle demande la connaissance de nombreuses techniques dont l'enseignement est lacunaire à l'université. De plus, peu d'éditeurs prennent le risque économique de publier des éditions scientifiques de sources, les acheteurs étant pratiquement exclusivement des bibliothèques, des dépôts d'archives ou des instituts de recherche. Les tirages sont souvent inférieurs à 500 exemplaires, ce qui rend l'ouvrage confidentiel dans bien des cas. Enfin, les éditions de sources sont souvent des entreprises de plusieurs années et la sortie de l'ouvrage sur le support papier doit attendre la mise au point définitive de l'édition, ce qui peut être assez long. Pourtant, l'édition de sources s'avère essentielle dans le travail de recherche, puisqu'elle évite la consultation de l'original, permet l'économie, non négligeable dans bien des cas, de la transcription de la source primaire et la mise à disposition d'un texte établi, tout en donnant les premières clefs d'interprétation délivrées par l'éditeur scientifique.

Depuis plusieurs années, l'existence des cédéroms rassemblant des corpus de sources[22] a permis de prendre conscience de la richesse du support numérique pour l'interrogation et l'exploitation des éditions de sources[23]. L'historien ne fait pas une lecture exhaustive d'une édition de sources, mais il y cherche la portion de texte qui s'insère dans son corpus de sources ou qui va répondre à une question de sa problématique. C'est pourquoi les accès rapides à l'information permis par le support numérique, en particulier la possibilité de recherche en texte intégral, donnent de nouvelles perspectives à l'édition de sources. De plus, le support numérique règlera définitivement le débat entre les tenants de l'édition interprétative et de

[20] Library of congress, *METS, Metadata Encoding and Transmission Standard*, http://www.loc.gov/standards/mets/, consulté le 12 mai 2004.

[21] Ecole nationale des chartes, Françoise Vielliard et Olivier Guyotjeannin (coord.), *Conseils pour l'édition des textes médiévaux*. 1. Fascicule I : conseils généraux, éd. du CTHS, Paris, 2001, p. 18.

[22] Il existe une quantité de cédéroms de ce type. A titre d'exemple, on peut citer le CETEDOC, *library of christian latin texts*, éd. Brepols, 1996 ou *Patrologia latina*, éd. Chadwyck-Healey. Vous trouverez une liste de cédéroms intéressants les médiévistes sur le site de Ménestrel, http://www.ccr.jussieu.fr/urfist/menestrel/medcdrom.htm, consulté le 1er juin 2004.

[23] Cf. à ce propos, les réflexions d'Amiri Bassir, « Réflexions sur les enjeux et perspectives des recherches lexicales et sémantiques assistées », *Le médiéviste et l'ordinateur*, n°41, hiver 2002, *L'apport cognitif*, éd. IRHT, Paris, 2002, [en ligne], http://lemo.irht.cnrs.fr/41/mo41_02.htm, consulté le 24 mai.

l'édition imitative qui, en fin de compte, ont chacun leurs rôles dans la recherche historique, car, comme le rappellent Olivier Guyotjeannin et Françoise Vielliard, « les potentialités ouvertes par la mise à disposition de bases textuelles informatisées et de corpus numérisés permettront très vite de résoudre le dilemme, en juxtaposant commodément plusieurs moyens de prendre connaissance des textes médiévaux : aspect physique des manuscrits et actes originaux, éditions de travail, éditions accessibles à la lecture courante… »[24]. Enfin, l'édition électronique permet de mettre à disposition au fur et à mesure de sa réalisation le texte édité, puisque les corrections et les ajouts sont possibles sur ce support. Toutes ces raisons font du support numérique et de l'édition électronique sur le Web des vecteurs de diffusion à privilégier aujourd'hui pour l'édition de sources historiques.

Une fois ce constat établi, des questions restent en suspens. En particulier, comment profiter pleinement des avantages du support numérique tout en assurant des critères scientifiques et en garantissant les règles d'édition scientifique établies ? Et comment assurer pleinement la conservation à très long terme d'un tel ouvrage, la vocation d'une édition de sources, si elle est bien faite, étant d'être encore utilisée dans cent ans et plus ? Autant le dire tout de suite, une réponse définitive à la dernière question n'existe pas, mais nous y reviendrons plus loin. Quant à la première question et à la lumière des premières expériences menées à l'Ecole nationale des chartes, les technologies XML semblent les plus appropriées.

Le but du XML[25] est de caractériser par des balises le rôle intellectuel tenu par des mots, des groupes de mots, des phrases ou des portions de textes à l'intérieur de l'information. Ce balisage permet ensuite de mettre en valeur les différents types d'informations ou de faire resortir un élément balisé dans le cadre d'une recherche précise ou d'un index. A la différence d'une base de données relationnelle, le XML permet de hiérarchiser plus précisément l'information et de gérer des granularités hétérogènes. Il impose l'utilisation d'une grammaire qui définit le nom des balises et leurs règles d'agencement, appelée DTD ou XML schéma[26]. Il est possible de développer un schéma qui va permettre le balisage de tous les types d'informations de la source éditée. Mais le défaut d'un tel schéma est souvent de n'être adapté qu'à un type de sources ce qui limite les possibilités d'interrogation. De plus, l'élaboration d'un schéma et sa maintenance se révèlent très fastidieux. C'est pourquoi il est plus profitable de se tourner vers un schéma existant comme la TEI[27]. L'avantage d'utiliser la TEI est de disposer

[24] Ecole nationale des chartes, Françoise Vieillard et Olivier Guyotjeannin (coord.), *Conseils pour l'édition des textes médiévaux*. 1. Fascicule I : conseils généraux, éd. du CTHS, Paris, 2001, p. 14.

[25] XML pour eXtensible Markup Language : http://www.w3.org/XML/. Une bonne entrée en matière sur le XML est l'ouvrage de Eric Lease Morgan, *XML pour les bibliothécaires : un manuel et un atelier*, traduit en français par Nicolas Morin, [**en ligne**], http://morinn.free.fr/xml/, consulté le 16 mai 2004.

[26] La syntaxe utilisée pour écrire une DTD (Document Type Definition) a été remplacée par celle des schémas XML, http://www.w3.org/XML/Schema.

[27] TEI pour Text Encoding Initiative, http://www.tei-c.org, pour plus de renseignements sur l'utilisation de la TEI pour l'édition de sources, cf dans ce même numéro du *médiéviste et l'ordinateur*, l'article consacré à cette

d'un schéma adapté à l'édition en sciences humaines et sociales, de proposer des balises correspondant à plus de 90% des besoins d'un chercheur et d'un éditeur scientifique pour l'édition de sources, d'obtenir l'aide d'une communauté bien établie et d'utiliser un standard maintenu et mis à jour régulièrement[28]. Cependant, le temps nécessaire pour s'approprier et utiliser parfaitement cette DTD, puis pour encoder un fichier XML en TEI en font un outil complexe.

Les technologies XML sont promises à un brillant avenir ; pour autant leur jeunesse explique l'immaturité des logiciels et implique la manipulation de différents langages informatiques pour une implémentation optimale.

C- Les outils de recherche

« Usuels », « Outils de recherche », ces termes désignent les ouvrages dont les chercheurs se servent couramment pour les aider et les diriger au cours de leurs recherches. Dans la plupart des cas, il s'agit d'ouvrages de références, de dictionnaires de tous types et sur tous les sujets, de bibliographies rétrospectives, des inventaires d'archives ou des catalogues de bibliothèques par exemple. Leur carastéristique réside dans leur accessibilité et leur facilité d'utilisation, car l'intérêt est d'y trouver rapidement l'information recherchée.

Dans la plupart des cas, il s'agit d'ouvrages collectifs, rassemblant tous les spécialistes d'un domaine par exemple dans le cas des dictionnaires. Le niveau d'exigence scientifique du contenu dépend de la problématique de l'ouvrage et du public visé. Leur mise au point est donc souvent longue et fastidieuse, mais ce sont surtout leurs mises à jour qui posent le plus de problèmes. En effet, il arrive bien souvent qu'une notice dans un dictionnaire soit dépassée au moment de la sortie de l'ouvrage ou peu de temps après et l'auteur doit attendre une éventuelle réédition pour pouvoir corriger l'erreur et actualiser son ouvrage. La souplesse éditoriale de l'édition électronique permet à l'auteur d'introduire des modifications dans son dictionnaire, par exemple de nouvelles références bibliographiques, à n'importe quel moment et ainsi le mettre à jour dès que la recherche a fait des avancées concernant une question.

L'intérêt réside aussi dans la conception de ces outils de recherche. Les bases de données qui constituent le cœur de ces ouvrages dans la plupart des cas peuvent être mises en ligne dès la genèse du projet. Tous les participants à l'élaboration de l'outil peuvent alors intervenir

question.

[28] La nouvelle version de la TEI, dite P5, est actuellement en préparation au sein du consortium TEI, http://www.tei-c.org/P5/, consulté le 25 mai 2004.

directement en ligne sur la base de données pour y ajouter des notices, les modifier voire les supprimer. Cette possibilité représente un gain de temps considérable et surtout offre une réactivité que le support papier ne permet pas. Pour faciliter ces tâches, des technologies tels que PHP[29] permettent la mise en place d'interfaces accessibles avec un navigateur Web, grâce à des formulaires Web, pour effectuer les tâches de maintenance de la base de données dans une zone sécurisée accessible avec des mots de passe. Le site Web se met alors à jour dynamiquement, puisque le contenu affiché proviend de la base de données. A cet égard, le projet « sermones.net : le thésaurus des sermons de Jacques de Voragine »[30], mené par l'UMR 5648 est un exemple[31].

Enfin, l'apprentissage par les historiens des technologies telles que PHP ou de l'utilisation de SGBD tels que MySQL peut permettre la mise en ligne des bases de données mises au point dans le cadre d'une recherche précise, mais dont le contenu pourrait intéresser d'autres chercheurs. Par exemple, les bases de données prosopographiques constituent des dictionnaires biographiques intéressants[32]. La mise à disposition de ces données peut non seulement venir appuyer l'argumentation développée dans le cadre du travail de recherche, mais aussi fournir une mine d'informations pour d'autres chercheurs.

D- Articles, monographies, essais

Une fois l'étude et le dépouillement des sources effectués, il reste l'essentiel du travail de l'historien : mettre en place son argumentation en fonction d'une problématique et en s'appuyant sur les sources. Comme le rappelle Philippe Carrard, « " mettre en texte " est une étape majeure dans l'entreprise historique »[33]. Ces textes peuvent prendre plusieurs formes : articles, monographies ou essais. Mais, ils possèdent des caractéristiques communes qui permettent de les identifier comme des analyses scientifiques : « l'histoire savante se signale, en effet, par des signes extérieurs beaucoup plus évidents, et notamment la présence d'un apparat

[29] PHP pour Hypertext Preprocessor, http://www.php.net/, est un langage de script côté serveur qui présente en particulier l'avantage d'interfacer des SGBD tels que MySQL, http://www.mysql.com. Ces deux applications sont en open source. Leur équivalent chez Microsoft est appelé ASP avec les SGBD Access ou SQLserver.

[30] *Sermones.net : le thésaurus des sermons de Jacques de Voragine*, édition électronique d'un corpus de sermons latins médiévaux, [**en ligne**], http://www.sermones.net/, consulté le 25 mai 2004.

[31] Pour plus de renseignements sur ce projet, cf l'article de Marjorie Burghart dans ce même numéro du médiéviste et l'ordinateur.

[32] Cf les réflexions à ce sujet de Jean-Philippe Genêt, « les bases de données à distance : une expérience *», le Médiéviste et l'ordinateur*, n°41, hiver 2002, l'apport cognitif, [**en ligne**], http://lemo.irht.cnrs.fr/41/mo41_10.htm, consulté le 25 mai 2004.

[33] Philippe Carrard, *Poétique de la nouvelle histoire : le discours historique français de Braudel à Chartier*, éd. Payot, Lausanne, 1998, p. 87.

critique, de notes en bas de page »[34]. Ainsi, il semble que les différences entre ces types de publication ne se situent pas au niveau du contenu lui-même, mais au niveau de leur caractéristique physique, en particulier la longueur du texte et la problématique choisie, celle d'un article étant plus restreinte dans la plupart des cas.

Les textes historiques se composent donc de deux éléments essentiels : l'argumentation développée par l'historien et les « marques d'historicité »[35], c'est à dire les références données par l'historien pour appuyer son argumentation, sous la forme des notes infrapaginales ou de documents annexes. L'édition électronique doit aussi mettre l'accent sur ces deux éléments avec une attention particulière pour la structuration globale car, comme le rappelle Alain Prost, « Généralement, elle [l'argumentation] en commande le plan, et c'est pourquoi il n'est pas injuste de juger les livres d'histoire à leur plan »[36]. Le plan doit donc être facilement identifiable et mis en lumière par l'organisation logique de l'information.

Contrairement aux éditions de sources, ces textes comportent peu d'informations assez significatives pour nécessiter un balisage fin, aussi leur codage en XML peut-il se contenter d'une grammaire simple, comme XHTML[37] ou TEI Lite[38]. Cette simplicité, ajoutée à la place prépondérante de ces types de publication dans la recherche, explique l'intérêt précoce dont ils ont fait l'objet et l'existence d'outils destinés à faciliter leur mise en ligne. Chacun présente des spécificités et est souvent axé vers tel ou tel type de publications. Parmi ces outils, qui ont l'avantage d'être des logiciels libres téléchargeables librement et gratuitement sur le Web, nous pouvons citer trois projets français :

- Cyberdocs[39] mis au point dans le cadre du projet Cyberthèses est une plate-forme de traitement et de diffusion des thèses;
- Lodel[40], logiciel d'édition électronique dédié à la mise en ligne d'articles, en particulier pour les revues;
- Hyper Article en Ligne, pour mettre en place une archive ouverte, dont le développement est assuré par le Centre pour la communication scientifique directe du CNRS[41].

[34] Alain Prost, *Douze leçons sur l'histoire*, éd. du Seuil, Paris, 1996, p. 263.
[35] Pomian Krysztof, « Histoire et fiction », *Le Débat*, n°54, mars-avril 1989, p. 114-137 repris par Alain Prost, *Douze leçons sur l'histoire*, éd. du Seuil, Paris, 1996, p. 263.
[36] Alain Prost, *op. cit.*, p. 256.
[37] XHTML est la version la plus récente de HTML respectant les règles du XML. W3C, *XHTML*, [**en ligne**], http://www.w3.org/MarkUp/, consulté le 17 mai 2004.
[38] TEI Lite est une version simplifiée de la TEI comportant les balises principales de cette DTD. Lou Burnard et C. M. Sperberg-McQuenn, *La TEI simplifiée : une introduction au codage des textes électroniques en vue de leur échange*, traduit en français pas François Rôle, [**en ligne**], http://www.tei-c.org.uk/Lite/teiu5_fr.htm, consulté le 31 mai 2004.
[39] Cyberdocs, http://sourcesup.cru.fr/cybertheses/, consulté le 31 mai 2004.
[40] Lodel, http://www.lodel.org, consulté le 31 mai 2004.
[41] Centre pour la communication scientifique directe, CCSD-CNRS, http://www.ccsd.cnrs.fr, consulté le 31 mai 2004.

II- Permettre l'exploitation de l'information

L'organisation logique de l'information ne suffit pas pour permettre son exploitation, même si elle va induire, en grande partie, sa présentation physique[42]. Il est impensable de proposer aux chercheurs-internautes de lire directement le document à la source du fichier en XML. C'est pourquoi il est important de réfléchir à la structuration physique de l'information, c'est à dire dans le cadre de l'édition électronique à l'interface graphique du site Web. En ce sens, le travail de l'éditeur électronique ressemble à celui de l'éditeur traditionnel dans la phase d'élaboration de l'ouvrage.

A- Naviguer dans l'information[43]

L'hypertexte et la lecture à l'écran sont les deux caractéristiques qui différencient l'édition électronique du support papier. L'hypertexte se caractérise par le fragment : c'est à dire qu'il se constitue de pages-écran et de liens, qui relient les pages-écrans entre elles[44]. A l'inverse, une publication sur le support papier conçue de façon linéaire a pour unité textuelle matérielle la page et c'est la reliure de l'ouvrage qui permet d'organiser les différentes pages entre elles. L'avantage de l'hypertexte par rapport à une publication papier est de multiplier les points d'accès directs à l'information. Par exemple, l'utilisation d'un index sur le support papier impose l'action de feuilletage pour retrouver une référence, ce qui est moins direct qu'un simple clic qui renvoie directement sur une page-écran.

Il existe deux types de liens : d'une part les liens « tabulaires », c'est à dire toutes les informations sur le texte qui ne sont pas contenues à l'intérieur de celui-ci (la table des matières, les différents types d'index par exemple), et d'autre part le lien « textuel », c'est à dire qu'un mot du texte sert de lien vers une autre page-écran dont le sujet renvoie à ce mot. Pour

[42] Les anglo-saxons emploient pour désigner ce travail le terme d'architecture de l'information, pour plus de renseignements voir le site de l'institut Asilomar pour l'architecture de l'information dont une partie est traduite en français, http://aifia.org/fr/, consulté le 27 mai 2004.

[43] Cette partie et la suivante reprennent de façon synthétique les conclusions auxquels nous sommes parvenus dans le cadre d'un DEA de sciences de l'information : Gautier Poupeau, *L'information historique à l'épreuve de l'édition électronique, un exemple : la monographie*, DEA de sciences de l'information sous la direction de Ghislaine Chartron, école doctorale EDIIS, Lyon, juin 2003.

[44] Pour plus de renseignements sur le principe de l'hypertexte et les conséquences de son utilisation sur la lecture et l'écriture, voir BALPE, Jean-Pierre, *Hyperdocuments, hypertextes, hypermédias*, Paris, Eyrolles, 1990, LAUFER, Roger, et SCAVETTA, Daniel, *Texte, hypertexte, hypermédia*, éd. des PUF, Paris, 1992, (*Que-sais-je ?)* et VANDENDORPE, Christian, *Du papyrus à l'hypertexte : essai sur les mutations du texte et de la lecture*, Paris, éd. La découverte, 1999, (*cahiers libres*).

profiter de tous les avantages donnés par l'hypertexte, il faut pouvoir utiliser ces deux types de liens, sans pour autant en multiplier le nombre ce qui pourrait désorienter le lecteur.

Pour optimiser la navigation en utilisant au mieux le principe de l'hypertexte et les avantages du support numérique, on identifie avant la conception les différentes utilisations que le lecteur peut faire de l'information, pour que l'interface et la navigation à l'intérieur de l'ouvrage prennent en compte ces différents usages. Dans le cadre de l'information scientifique, trois utilisations systématiques peuvent être identifiés :

- Une lecture linéaire et complète de l'information ;
- La « lecture-zapping » ; l'utilisateur ne sait pas s'il va trouver une référence intéressante dans le texte proposé, il va « naviguer » dans l'ouvrage sans but précis.
- le lecteur recherche une référence précise dans le texte sans être certain de sa présence. Il va donc lancer une recherche en texte intégral.

Pour la lecture linéaire, l'idéal est une interface qui permet une navigation aisée d'une page écran à l'autre avec des flèches de navigation par exemple. Pour éviter la désorientation à l'intérieur du site Web, il est important d'aider le lecteur à se situer, en indiquant, par exemple, le titre courant de la page-écran.

Pour une « lecture-zapping », on propose aux lecteurs différents moyens de connaître le contenu du document afin d'y accéder rapidement. La tabularité du texte peut nous aider à mettre en place de telles interfaces. Comme nous l'avons dit, grâce à l'hypertexte, le principe de feuilletage, est simplifié et accéléré. Ainsi, les index permettent à partir de mots-clefs d'accéder à des thèmes, des personnes ou des lieux intéressant le lecteur. Un sommaire détaillé affichant le résumé de chaque partie voire de chaque page-écran donne au lecteur le moyen de prendre connaissance du contenu des parties et ainsi d'accéder à l'information qui l'intéresse. Une table des matières présente sur toutes les pages-écrans permet une navigation rapide à l'intérieur du site Web. Dans le cadre d'une monographie ou d'une édition de textes, une page listant tous les documents annexes avec un renvoi vers les pages y faisant référence peut aussi créer un parcours de lecture original que le papier ne permet pas. Grâce à la tabularité du texte, c'est le lecteur qui crée son propre parcours de lecture en fonction de ses centres d'intérêts et non plus l'auteur qui impose sa propre vision du texte. Ces interfaces permettent de mettre en valeur toute la richesse de l'information.

Enfin, pour le lecteur en quête d'une référence précise, les modules de recherche permettent de lancer une recherche en texte intégral, une recherche sur les différents index ou une recherche croisée plus précise prenant en compte différents types d'informations. Cette utilisation donne toute sa raison d'être à l'édition électronique et à la sémantisation de l'information.

B- Visualiser l'information

La lecture sur écran n'est pas encore aisée et constitue le principal défaut actuel de l'édition électronique. Les usagers se plaignent souvent de la fatigue visuelle engendrée par l'écran. Il est donc important de faciliter le processus de lecture, en pensant à ces difficultés.

Lorsque le texte est affiché sur toute la largeur de l'écran, le lecteur se fatigue trop vite, le nombre de signes étant trop important. Pour une lecture fluide, une ligne de texte comporte entre 90 et 100 caractères espaces inclus au maximum. Dans la partie de la page réservée au texte, la barre de navigation verticale à gauche de l'écran engendre une rupture dans la lecture, car l'œil, au moment du changement de ligne, est attiré par cette barre de navigation. C'est pourquoi il est préférable de placer le texte au milieu de la page encadré par deux bandes verticales vierges de textes plus ou moins grandes en fonction de la résolution de l'écran. Tandis que dans l'édition papier, l'habitude était d'occuper le plus de place possible sur le support pour diminuer les coûts de fabrication, dans l'édition électronique, passée une première impression de perte de repères, le lecteur se rend vite compte du confort de lecture induit par l'importance du blanc sur la page. La barre de navigation essentielle pour le passage entre les différentes parties de la publication trouve sa place à l'horizontal en haut de la page. De cette façon, le lecteur qui opte pour la lecture linéaire pourra le faire avec un confort visuel optimal.

Le passage au support électronique inclut une réflexion sur la typographie du texte[45]. Ainsi, les polices de caractère dites serif utilisées habituellement pour le papier sont moins adéquates pour l'écran qu'une police sans serif, telle qu'Arial dans une taille standard. Sa simplicité et les formes de ses lettres permettent une visualisation aisée sur un écran. Il est intéressant de noter qu'habituellement, ce genre de police n'est utilisé que pour les titres ou les descriptions brèves, car peu adapté à la lecture de longs textes. L'écart entre les lignes est équivalent au standard du support papier, soit entre une ligne et une ligne et demi.

Enfin, le support électronique impose un minimum de graphisme. Il ne faut pas oublier que, dans l'environnement du Web, ce n'est pas la textualité qui est essentielle au premier abord mais l'image. Cependant le graphisme doit se mettre au service de l'information et non l'inverse. Sans alourdir graphiquement la page pour attirer des internautes, ce qui risquerait de

[45] L'exemple de l'étude mené par l'ISDN sur les ebooks est de ce point de vue significatif, puisqu'un typographe professionnel a participé à l'expérience, *Contrats de lecture, Rapport sur une expérimentation de prêts de livres électroniques en bibliothèques : dimensions socio-économiques et psycho-cognitives*, ISDN, 2002. [**en ligne**], http://isdn.enssib.fr/archives/axe2/contratslecture/Rapport_CLLe.pdf, consulté le 25 mai 2004.

détourner l'attention du lecteur, on peut utiliser le graphisme pour aider à la mise en valeur du texte. Si nous prenons l'exemple de l'édition des édits de pacification sur le site de l'Ecole des chartes[46], la page-écran est entourée d'un cadre noir, avant tout conçu pour des raisons esthétiques, mais qui aide le lecteur dans le processus de lecture en reproduisant l'environnement visuel d'une page d'un livre. De la même manière, l'association d'un fond d'écran clair et d'une police de caractère foncée est plus reposante pour les yeux.

Au final, le lecteur a face à lui une page-écran qui ressemble à une page de livre contenant le même nombre de signes par ligne, le même écart entre les lignes et même des cadres matérialisant la page ce qui n'a rien d'étonnant, car le livre tel qu'il est conçu aujourd'hui est le résultat de nombreux tâtonnements qui avaient tous pour but le confort optimum de lecture et la recherche permanente d'une lecture fluide et aisée. Il n'est pas utile de réinventer ce que nos prédécesseurs ont déjà fait, il suffit de l'adapter[47].

[46] Bernard Barbiche (dir.), *L'édit de Nantes et ses antécédents (1562-1598)*, éd. Ecole nationale des chartes, Paris, 2003, coll. ELEC, n°5, [**en ligne**], http://elec.enc.sorbonne.fr/editsdepacification/, consulté le 25 mai 2004.

[47] Cf à ce propos les analyses, riches d'enseignements, d'Henri-Jean Martin dans un livre d'entretien avec Jean-Marc Chatelain et Christian Jacob, en particulier le dernier chapitre intitulé « la fabrique du lisible ». Henri-Jean Martin, *Les métamorphoses du livre, entretiens avec Jean-Marc Chatelain et Christian Jacob*, éd. Albin Michel, Paris, 2004.

III- Accéder à l'information

Une fois l'information mise à disposition sur le Web, il faut pouvoir y accéder. Or, comment la retrouver parmi les 4 milliards de pages Web indexées par un moteur de recherche tel que Google[48] ? Le principe du réseau est de pouvoir partager l'information. Comment utiliser les possibilités offertes par les principes du réseau dans le cadre de la diffusion de l'information scientifique ? Accéder à l'information signifie aussi garantir son accessibilité à long terme et donc la pérenniser : la question de la conservation doit être posée. Enfin, il faut pouvoir garantir l'accès à l'information au plus grand nombre quels que soient les systèmes d'exploitations, les navigateurs et les capacités et/ou déficiences éventuelles.

A- Le référencement et les métadonnées

Exister aujourd'hui sur le Web, c'est être indexé dans les moteurs de recherche[49]. Ils garantissent la visibilité du site au milieu des milliards de pages accessibles sur le Web et donc leur légitimité d'existence auprès des organes décisionnels. Pour autant, leur fonctionnement est opaque : quelles sont les critères utilisés pour indexer les pages ? Comment ces moteurs calculent-ils la pertinence d'une réponse par rapport à une autre ? Autant de questions dont les réponses sont laissées au bon vouloir des entreprises mettant en place ces moteurs[50].

Malgré tout, un certain nombre de réponses peuvent être apportées grâce à l'utilisation quotidienne et l'étude de ces moteurs. La première d'entre elles est la primauté de l'indexation en texte intégral, c'est à dire que les moteurs de recherche scannent l'ensemble du contenu d'une page Web. Or, cette indexation est faite automatiquement par des programmes informatiques, appelés robots ou crawlers, qui ne font pas la différence entre la partie du code concernant le graphisme et la partie concernant le contenu à proprement parler. Par conséquent,

[48] Google, http://www.google.fr, est devenu aujourd'hui le moteur de recherche de référence pour trouver une information sur le Web. Son succès provient essentiellement du nombre de pages indexées, de la pertinence des résultats grâce à son algorithme, le page rank, et de la sobriété de présentation et d'utilisation. Il faut tout de même avertir les utilisateurs de la position dominante acquise aujourd'hui par ce moteur de recherche. Pour plus de renseignements sur les moteurs de recherche et Google en particulier, voir le débat virtuel sur ce sujet sur le site de la BPI, http://debatvirtuel.bpi.fr/moteurs, consulté le 25 mai 2004.

[49] La première réaction d'un utilisateur, lorsqu'il cherche une référence sur le Web, est de se précipiter sur Google.

[50] Pour voir les problèmes posés par ces questions, Marin Dacos, « Google, une enquête », *les enjeux culturels des moteurs de recherche*, débat virtuel de la BPI, 2004, [**en ligne**], http://debatvirtuel.bpi.fr/moteurs/papers/2, consulté le 25 mai 2004.

il est essentiel de séparer la forme du fond. C'est pourquoi le code source en HTML de la page Web gagne à respecter les standards existants[51], être parfaitement sémantisé et ne contenir aucune des informations graphiques qui seront rassemblées au sein de la feuille de style CSS[52]. Les pages Web parfaitement sémantisées sont mieux indexées par les moteurs de recherche, car elles ne comporteront pas d'informations inutiles pour la recherche[53], le bruit des réponses étant déjà suffisamment important.

Un autre moyen de garantir une indexation efficace par les moteurs de recherche est l'utilisation des métadonnées. Les métadonnées rassemblent des informations sur les données, c'est pourquoi elles sont souvent définies par la périphrase « des données sur les données », c'est à dire « des balises ou des jalons qui permettent de circonscrire l'information sous toutes ses formes »[54]. Dans le cadre d'une page Web, les metadonnées sont des marqueurs spéciaux situés dans l'en-tête du document HTML, donc invisibles pour l'internaute, qui aident les moteurs de recherche à indexer les sites Web.

Les professionnels de l'information ont souvent comparé les métadonnées au catalogage. Il est vrai qu'actuellement l'utilisation et la mise en place des métadonnées y ressemblent. Mais, à notre avis, cette comparaison est erronée, même provoque des erreurs d'interprétation sur leurs buts. Une des principales différences avec le catalogage vient du fait que ce sont les producteurs de l'information qui précisent les métadonnées d'une page, donc en vue de sa diffusion. Il ne s'agit pas d'un processus a posteriori en vue de la conservation comme pour le catalogage. Le but n'est pas de cataloguer l'information à l'intérieur d'une entité géographique définie : la bibliothèque, mais de donner le moyen à des programmes informatiques d'indexer plus précisément l'information sans notion d'espace géographique. Enfin, les métadonnées sont, dans le cas d'une page HTML, incluses dans le document, à la différence d'une notice de bibliothèque qui en est séparée physiquement. Ce dernier argument pourrait à terme faire toute la différence. En effet, une technologie émergente appelée Web sémantique[55] est fondée sur l'utilisation des métadonnées pour naviguer à l'intérieur du Web.

---

[51] Les standards utilisés sur le Web sont émis par le W3C, organisme rassemblant des universitaires, des chercheurs et des industriels basé au MIT, http://www.w3.org.

[52] Nous reviendrons plus loin sur un autre intérêt de la séparation de la forme du fond, mais pour plus d'informations sur l'utilisation des feuilles de style CSS, cascading StyleSheet, vous pouvez reporter à l'article de Thierry Buquet dans ce même numéro du médiéviste et l'ordinateur.

[53] Pour plus de renseignements sur l'utilisation des standards Web et leur impact sur le référencement, Denis Bourdeau et Tristan Nitot, « Pourquoi les standards du W3C ? », *OpenWeb Group*, mars 2003, [**en ligne**], http://openweb.eu.org/articles/pourquoi_standards/, consulté le 25 mai 2004.

[54] James M. Turner et Véronique Moal, « Que sont les métadonnées », dans le site *Métrometa*, [**en ligne**], http://mapageweb.umontreal.ca/turner/meta/francais/metadonnees.html, consulté le 25 mai 2004. Ce site contient aussi une amusante cartographie des différents types de métadonnées existantes sous forme de plan de métro, http://mapageweb.umontreal.ca/turner/meta/francais/metrometa.html.

[55] Pour plus de renseignements sur le Web sémantique, voir le site en français, http://websemantique.org, consulté le 25 mai 2004.

Le principal frein actuel au développement de l'utilisation des métadonnées est la méconnaissance des standards existant pour les décrire. Cette méconnaissance a eu pour conséquence une utilisation disparate des métadonnées par les concepteurs de sites. C'est une des raisons pour lesquelles la plupart des moteurs de recherche n'utilisent que très peu les métadonnées pour l'indexation, privilégiant le texte intégral. A l'heure actuelle, les moteurs de recherche ne prennent en compte que le titre de la page, les mots-clefs et la description de la page. Pourtant, une norme existe depuis 1994 : le Dublin Core. Mise au point par le Dublin Core Metadata Initiative[56], organisme international, cette norme vise à promouvoir une interopérabilité entre les métadonnées et à développer un langage spécifique pour la description des ressources présentes sur le réseau. « La norme du Dublin Core comprend 15 éléments dont la sémantique a été établie par un consensus international de professionnels provenant de diverses disciplines telles que la bibliothéconomie, l'informatique, le balisage de textes, la communauté muséologique et d'autres domaines connexes. »[57]. Longtemps resté confidentiel, le Dublin core est aujourd'hui de plus en plus utilisé, surtout depuis qu'il est devenu une norme ISO[58].

Un des intérêts du Dublin Core réside dans son utilisation par le protocole OAI[59]. Mis au point dans le cadre des archives ouvertes, c'est à dire de réservoirs d'articles en accès libre sans barrière économique et juridique, le protocole OAI est basé sur l'indexation par un moteur de recherche, appelé moissonneuse ou *harvester*, de métadonnées au format Dublin Core présentes sur chaque document rassemblé au sein d'un site Web, appelé « entrepôt ». OAI est un standard ouvert et libre[60], c'est pourquoi il est aujourd'hui incontournable dans la communauté scientifique. Son intérêt réside dans la possibilité d'interroger avec une même interface différents entrepôts avec des critères précis, puisqu'ils reprennent les 15 éléments du Dublin Core. Un regret toutefois : le protocole OAI n'intègre pas à l'heure actuelle l'indexation de l'ensemble du texte, l'interrogation en texte intégral est donc a-priori impossible avec des moteurs de recherche OAI.

---

[56] Dublin Core Metadata Initiative, http://www.dublincore.org, consulté le 25 mai 2004.

[57] Diane Hillmann, *Guide d'utilisation du Dublin Core*, 2001, [**en ligne**], http://www.bibl.ulaval.ca/DublinCore/usageguide-20000716fr.htm, consulté le 25 mai 2004.

[58] ISO 15836:2003, *l'ensemble des éléments de métadonnées Dublin Core*, http://www.iso.ch/iso/fr/CatalogueDetailPage.CatalogueDetail?CSNUMBER=37629&ICS1=35&ICS2=240&ICS3=30, consulté le 25 mai 2004. Il ne s'agit que de la fiche permettant d'acquérir la norme ISO, mais l'ensemble de la recommandation est consultable gratuitement sur le site du Dublin Core metadata initiative, http://www.dublincore.org.

[59] OAI pour Open Archive Initiative est un mouvement de chercheurs internationaux dont le but est de promouvoir des standards interopérables et de définir un ensemble de protocoles techniques liés à l'interrogation des données et à leur description, http://www.openarchives.org/, consulté le 26 mai 2004. Pour plus de renseignements sur le mouvement des archives ouvertes, Ghislaine Chartron, *les archives ouvertes dans la communication scientifique*, 2003, [**en ligne**], http://www.ccr.jussieu.fr/urfist/archives-ouvertes.htm, consulté le 26 mai 2004.

[60] Cela signifie que les scripts permettant de mettre en place un entrepôt OAI et une moissoneuse OAI sont accessibles librement et gratuitement sur le site de l'Open archive initiative.

B- La syndication de contenu

Un des avantages de la mise en réseau des informations est de pouvoir les partager. Dans le cadre de l'information scientifique, ce principe peut permettre l'échange de contenu entre différents sites Web. Dans le cadre du Web, on parle de syndication de contenu. Elle permet de rendre visible sur un site Web A les dernières informations parues sur un site Web B de façon synchrone, puisque la mise à jour des informations sur le site B se fait automatiquement. Par exemple, en page d'accueil du site du département SHS du CNRS[61], sont affichées les dernières nouvelles parues sur le site Calenda.org, calendrier en sciences sociales[62]. Dès qu'une nouvelle information paraît sur Calenda, la page d'accueil du département SHS du CNRS va afficher cette nouvelle information.

La syndication de contenu est rendue possible par des fichiers XML dont il existe plusieurs formats comme RSS[63] ou ATOM[64]. La plupart des sites Web sont aujourd'hui dynamiques, le contenu est géré indépendamment de l'affichage et stocké dans une base de données ou dans des fichiers XML. A partir des informations de la base de données ou du fichier XML, plusieurs formats de sortie peuvent être générés en HTML pour l'affichage ou dans un format permettant la syndication. Le site met ensuite à disposition librement le fichier de syndication sur son site Web et tous les sites qui veulent récupérer les informations peuvent utiliser ce fichier. Des petits scripts en PHP, par exemple, permettent ensuite de parser, c'est à dire d'analyser, le contenu du fichier XML, qui se trouve physiquement sur le serveur du site-source pour le transformer et l'afficher sur son site en HTML. Actuellement, les formats de syndication de contenu se limitent à quelques informations : le titre de l'article ou de la nouvelle, le lien vers l'article, un résumé éventuel. Mais, dans le cadre du développement des *Web services*[65], on peut espérer que la syndication et l'échange d'informations entre différents sites permettront de rassembler des

[61] Département SHS du CNRS, http://www.cnrs.fr/SHS/, consulté le 26 mai 2004.

[62] Calenda.org, calendrier en sciences sociales, http://www.calenda.org, consulté le 26 mai 2004.

[63] RSS pour Really Simple Syndication est un format de fichier XML inventé par la société netscape, les spécifications de la deuxième version du format RSS ont été traduites en français, http://www.stervinou.com/projets/rss/, consulté le 26 mai 2004, mais vous trouverez beaucoup d'autres références sur ce format sur le Web.

[64] ATOM est plus récent et semble plus complet que RSS, ainsi le W3C s'intéresse de près à ce format. Spécification du format ATOM, http://www.mnot.net/drafts/draft-nottingham-atom-format-01.html, consulté le 26 mai 2004.

[65] Les Web services sont une série de normes au niveau des protocoles de communication mis au point par le W3C pour échanger entre des applications Web des données au format XML, *Web service activity*, W3C, http://www.w3.org/2002/ws/, consulté le 26 mai 2004.

articles ou des éditions de sources issus de différents sites, pour leur faire subir un traitement lexicographique par exemple.

C- Pérenniser l'information

Editer un document signifie aussi en assurer son accès à long terme et sa conservation, ce qui représente une des difficultés actuelles de l'édition électronique. Les questions sont complexes et multiples et les réponses souvent lacunaires. En attendant que des solutions concrètes soient trouvées, nous nous limiterons à des constats, des informations et quelques conseils.

Conernant la question de la conservation du document numérique, trois aspects peuvent être dégagés :

- la conservation des données ;
- la conservation des interfaces et du contexte de navigation ;
- la conservation des supports.

A l'inverse du livre, le document numérique ne se suffit pas à lui-même pour être consulté. Nous avons besoin de la médiation d'une machine adaptée à sa consultation. Or, la conservation des supports comporte deux aspects : le support en lui-même et le matériel utile pour lire le support. Ces problèmes sont bien connus, puisqu'ils sont les mêmes que pour les disques à micro-sillons, par exemple. De plus, le cédérom présenté au moment de sa création comme un support pérenne ne semble pas tenir toutes ces promesses, puisque les premiers cédéroms présentent aujourd'hui des problèmes de conservation. Bien qu'aucune solution miracle n'existe à l'heure actuelle, deux pistes se dégagent. La première consiste à stocker les données sur des disques de verre, le verre étant un matériau inaltérable. Mais, cette technique s'avère coûteuse et, surtout, ne résoud pas le problème de la machine capable de lire ce support. La seconde consiste à stocker les données sur des machines, éventuellement sur plusieurs, et à faire migrer régulièrement sur d'autres machines. Cette solution s'avère moins couteuse, puisqu'on dispose toujours d'une machine, un serveur par exemple, disposant d'espaces libres et que ce type de mémoire n'est pas cher à l'achat aujourd'hui.

Le problème du stockage ne résoud pas celui des données. Il n'existe aucun moyen de garantir que le format dans lequel sont enregistrés les données sera toujours le même dans 10 ou 20 ans. Les technologies informatiques évoluent très rapidement et personne ne peut prédire à quoi ressemblera l'informatique de demain. Le problème se pose, d'autant plus, dans le cas où

les interfaces de navigation et les données ne sont pas séparées. Il vaut donc toujours mieux utiliser les standards existants qui privilégient la séparation de la forme et du fond. L'utilisation d'un standard n'est pas un gage de sa pérennité à très long terme, mais un standard étant maintenu et normalisé, on peut espérer la mise en place d'outils de conversion qui permettront de migrer les données dans les futurs standards informatiques. Ainsi, dans le cadre du Web, l'utilisation de XML et la séparation de la forme et du fond, même lorsqu'il s'agit de page Web[66] doit être privilégiée. L'avantage de XML est d'être ouvert et libre et donc d'être indépendant des plate-formes et des logiciels utilisés. De plus, la séparation de la forme du fond permet au moins d'envisager une conservation des données qui seront indépendantes des interfaces de navigation. En ce qui concerne le dépôt légal du Web, un consortium de bibliothèques nationales dont fait partie la BnF est en train de réfléchir aux modalités de mise en place, mais cette question s'avère tout aussi compliquée à résoudre[67]. Quelques expériences existent sur le Web, comme le site Internet Archive[68].

Un des autres aspects de la pérennisation de l'information est la possibilité de citer un document électronique dans une bibliographie, en étant sûr qu'il sera toujours accessible. Le problème de la citabilité des éditions électroniques est donc essentiel pour la survie de cette forme d'édition. Mais, la citation de l'URL pose des problèmes, car l'URL est basée sur la localisation géographique de la ressource, et non sur son identification, et à l'oubli de cette problématique, pourtant centrale. Sa prise en compte permettra la citation du document électronique dans les bibliographies et donc la légitimation scientifique de cette nouvelle forme de publication. Deux problèmes sont à prendre en compte et donc à améliorer : le changement de l'URL et son masquage[69].

En effet, il n'est pas rare qu'une adresse valide au moment de la citation dans la bibliographie ne le soit plus quelques mois plus tard. Les pages peuvent être supprimées, changées de serveurs ou déplacées d'un répertoire à un autre. Toutes ces opérations ont pour conséquence de rendre caduque l'URL. Il est donc essentiel qu'elle soit fixée au moment de la première publication et qu'elle ne change pas par la suite. Il existe des systèmes permettant d'identifier une ressource autrement que par sa localisation, comme le préconise le principe des

[66] Cela passe par l'utilisation de XHTML et CSS, cf la note 44.

[67] Pour plus de renseignements, cf les pages Web consacrés à cette question sur le site Web de la BnF, http://www.bnf.fr/pages/infopro/depotleg/dli_intro.htm, consulté le 26 mai 2004. Vous pouvez aussi consulter l'article d'un des conservateurs chargés du dossier à la BnF, Julien Masanès, « Towards continuous Web archiving, first results and an agenda for the future », *D-lib Magazine*, décembre 2002, [en ligne], http://www.dlib.org/dlib/december02/masanes/12masanes.html, consulté le 26 mai 2004 et le site du consortium chargé de cette question, International internet preservation consortium, http://www.netpreserve.org, consulté le 26 mai 2004.

[68] Internet Archive, http://www.archive.org/, consulté le 26 mai 2004.

[69] Cf l'article de l'inventeur du Web sur cette question, Tim Berners-Lee, « Cool URIs don't change », trad. française, Karl Dubost, « Les URLs sympas ne changent pas », 1998, [**en ligne**], http://www.la-grange.net/w3c/Style/URI, consulté le 26 mai 2004.

URNs et des URIs[70] : les redirections[71], les résolveurs de liens[72], le protocole standard de résolution de liens[73].

Par ailleurs, l'URL étant le moyen d'identification de la page, il faut éviter de la masquer. Il arrive que l'URL de la page en cours ne soit pas visible, en particulier à cause du procédé appelé *frames,* qui crée des cadres à l'intérieur de la page HTML ou dans les fenêtres qui s'ouvrent sans la barre d'adresses, les *pop-ups*. Il est important d'éviter de masquer l'URL pour que la référence pointe directement sur la ressource souhaitée.

D- Permettre l'accès au plus grand nombre

Les questions relatives à l'accès à long terme de l'information ne doivent pas faire oublier qu'il faut aussi avant tout garantir l'accès aux plus grand nombre d'utilisateurs dès la mise en ligne des documents. Cette question regroupe en fait deux aspects : le filtrage des accès et l'accessibilité à l'information.

Il est important d'éviter de filtrer l'information par des barrières économiques ou technologiques. Dans le cadre des instituts de recherche et des universités, l'édition électronique est, actuellement, intégralement subventionnée par l'Etat[74]. Il est donc normal qu'en retour, l'accès à ces données ne soit pas conditionné par le paiement d'un droit d'entrée. De plus, il arrive parfois que la ressource soit gratuite, mais qu'il faille s'inscrire et s'identifier pour y avoir accès. De la même façon, ces deux sortes de filtrage, économiques ou technologiques, ont pour effet de casser la chaîne de l'hypertexte. En effet, si une personne référence sur un site ou dans une bibliographie la ressource en question, les personnes ne pourront la consulter qu'après s'être acquitées de la licence et/ou s'être identifiées. Il n'y a pas d'intérêt à filtrer l'information, comme le rappelle Marin Dacos, le responsable scientifique de revues.org, « Publier, c'est donner à lire, à copier, à s'inspirer. Ce n'est pas filtrer, cacher, exclure. »[75].

---

[70] URI pour Uniform ressource identifiers, http://www.w3.org/Addressing/, consulté le 26 mai 2004.

[71] Cf le projet *PURL, Persistent Uniform Ressource Locator*, online computer library center, http://purl.oclc.org/, consulté le 2 juin 2004.

[72] The digital object identifier system, DOI, http://www.doi.org/, consulté le 2 juin 2004.

[73] The proposed OpenURL Framework Standard, NISO committee AX, http://library.caltech.edu/openurl/Standard.htm, consulté le 2 juin 2004.

[74] Cf à ce sujet, Pierre Portet, « Gratuit/payant : quelques réflexions sur les modèles de diffusion de l'information historique sur l'Internet », *La lettre n°3 de Ménestrel*, mars 2004, [**en ligne**], http://www.ccr.jussieu.fr/urfist/menestrel/med-lettre3.htm et les commentaires à cette lettre, Gautier Poupeau, « La lettre de Pierre Portet lue et commentée », avril 2004, [**en ligne**], http://www.ccr.jussieu.fr/urfist/menestrel/lettre3reponse.htm, consulté le 3 juin 2004.

[75] Marin Dacos, *L'édition électronique de périodiques scientifiques, rapport remis à Ketty Schwartz, directrice de la recherche*, ministère de la recherche, 2002, [**en ligne**], http://apropos.revues.org/document26.html, consulté le 3 juin 2004.

Par ailleurs, il arrive trop souvent que des sites soient inaccessibles pour des raisons technologiques : site non optimisé pour tous les navigateurs ou consultation conditionnée à l'installation d'un greffon particulier. Pourtant, le Web a été conçu de telle façon que sa consultation est indépendante des plate-formes, des systèmes d'exploitation ou des logiciels utilisés pour naviguer. Ainsi, comme le rappelle Tim Berners-Lee, directeur du W3C et inventeur du Web dans sa définition de l'accessibilité sur le Web, il faut « mettre le Web et ses services à la disposition de tous les individus, quel que soit leur matériel ou logiciel, leur infrastructure réseau, leur langue maternelle, leur culture, leur localisation géographique, ou leurs aptitudes physiques ou mentales »[76]. Consultant leurs statistiques de fréquentation, trop de concepteurs de sites partent du principe que 80% des internautes utilisent le navigateur Internet explorer de Microsoft et optimisent leurs sites, simplement, pour ce navigateur. Or, il existe d'autres navigateurs, comme Mozilla[77] ou Safari[78] par exemple. Pour garantir l'accès le plus large possible, les concepteurs des sites doivent, autant que faire se peut, respecter les recommandations émises par le W3C que les concepteurs de navigateurs sont censés suivre[79]. De plus, pour permettre la consultation par d'anciens navigateurs qui ne supportent pas les dernières recommandations émises, des petites astuces qui permettent au contenu d'être consultable, malgré tout[80] peuvent être utilisées. Enfin, il existe aussi une série de recommandations émises par le W3C pour garantir la consultation du Web pour les handicapés physiques (visuels et moteurs) regroupée au sein de l'initiative pour l'accessibilité du Web, WAI[81]. Ces recommandations sont à destination des concepteurs de navigateurs, mais aussi des concepteurs de sites[82]. Dans le cadre des sites du service public, il est important de ne pas négliger ce point[83].

---

[76] W3C, *Web accessibility initiative*, [en ligne], http://www.w3.org/WAI/, consulté le 4 juin 2004.

[77] Le navigateur Mozilla est un logiciel Open-souce qui a servi de base à netscape 6 et 7 conçu pour Windows, Linux et Mac, pour plus de renseignements, http://www.mozilla-europe.org/fr/, consulté le 3 juin 2004.

[78] Safari est un navigateur mis au point par Apple pour le système d'exploitation MacOS X, http://www.apple.com/safari/, consulté le 3 juin 2004.

[79] La dernière version, comme les précédentes, d'Internet explorer ne suivent pas les recommandations du W3C ce qui peut poser des problèmes, mais cela ne doit pas être un argument pour refuser le respect des standards du W3C.

[80] Vous trouverez une série d'astuces sur différents sites, en particulier, Mark Pilgrim, Dive into accessibility, traduit en français par Karl Dubost, [**en ligne**], http://www.la-grange.net/accessibilite/index.html, consulté le 4 juin 2004.

[81] W3C, *Web accessibility initiative*, [**en ligne**], http://www.w3.org/WAI/, consulté le 4 juin 2004.

[82] W3C, *Directives pour l'accessibilité aux contenus Web*, mai 1999, [**en ligne**], http://www.w3.org/TR/1999/WAI-WEBCONTENT-19990505/, traduit en français, http://www.la-grange.net/w3c/wcag1/wai-pageauth.html, consulté le 4 juin 2004. Vous trouverez une bonne entrée en matière sur le site de l'OpenWeb group, http://openweb.eu.org/articles/intro_accessibilite/, consulté le 4 juin 2004.

[83] L'Agence pour le développement de l'administration électronique, sous la tutelle du premier ministre a émis des directives dans ce sens, [**en ligne**], http://www.adae.gouv.fr/article.php3?id_article=246, consulté le 4 juin 2004.

Quinze ans après sa création au sein du CERN, le centre européen de recherche nucléaire, le Web a acquis son statut de média, au même titre que le livre, la radio ou la télévision. En tant que tel, il possède ses propres logiques et son propre mode de fonctionnement, en clair sa place au sein de nos moyens de communication. Pour autant, comme cela avait été le cas au moment de l'invention de l'imprimerie, nous avons cherché avant tout à adapter les modes de communication préexistants. Cette étape s'avère nécessaire durant la phase d'appropriation du nouvel outil. L'existence de repères permet ainsi d'en faciliter la prise en main et l'adoption. C'est pourquoi, dans le cadre de la recherche historique, nous nous sommes attachés, presque naturellement, à adapter les différents types de publications et à recréer les conditions d'une lecture rappelant les cadres du livre. Même si cette étape n'est que la première, elle ne s'avère pas simple, puisqu'elle impose de comprendre les mécanismes traditionnels de diffusion de l'information scientifique et l'apprentissage des logiques et des langages spécifiques au Web.

Cette phase d'adaptation, qui, à notre avis, n'est pas encore terminée, a permis d'identifier les particularités, les avantages et les inconvénients de ce nouveau média. En particulier, les chercheurs se sont rapidement aperçu, et cela avant même la création du Web, du potentiel que représentait le support numérique en terme de rapidité d'accès et de traitement de l'information. Dans ce mouvement, Internet et le Web ont permis de multiplier les expériences, en les mettant à la libre disposition d'un public de plus en plus nombreux.

En nous appuyant sur l'édition traditionnelle et les expériences menées, nous avons pu mettre en place des procédures de travail dont la caractéristique essentielle est la séparation entre l'organisation logique de l'information et sa présentation physique sous forme d'interfaces. Cette dichotomie permet de garantir, à long terme et à un large public, un accès aux documents mis en ligne, tout en assurant la cohérence de l'information.

Pour autant, la rapidité des évolutions technologiques dans ce domaine montre que le Web est loin d'avoir atteint sa maturité. Il reste encore beaucoup de chemins à explorer, en particulier dans la mise à disposition, l'accès et le traitement de l'information scientifique. De plus, il est évident que l'utilisation du Web va induire des changements dans les méthodes de travail, voire dans la perception cognitive de l'information scientifique. Comme le rappelle Henri-Jean Martin, la technique joue un rôle dans « la modification des méthodes du travail intellectuel, et sans doute, par cet intermédiaire, de la pensée. Ainsi les systèmes de pensée d'une société sont étroitement liés aux technologies utilisés »[84]. Notre travail est alors d'étudier ces changements, pour les accompagner au mieux.

[84] Henri-Jean Martin, *Les métamorphoses du livre, entretiens avec Jean-Marc Chatelain et Christian Jacob*, éd. Albin Michel, Paris, 2004, p. 253.