Arguments bibliothéconomiques

La chaîne de traitement des documents numériques

Caractéristiques et mise en œuvre

Christian Lupovici

Directeur de la bibliothèque
de l'université de Marne-la-Vallée
lupovici@univ-mlv.fr

Le document numérique est désormais incontournable et il va gagner rapidement en importance dans les procédures de traitement documentaire des bibliothèques.

C'est une évidence pour ce qui concerne la littérature scientifique : les revues qui ne sont pas encore exclusivement numériques cesseront bientôt de paraître sous leur forme traditionnelle. Le secteur de la recherche n'est pas le seul à utiliser de façon massive des documents électroniques.

Du document « objet » au document « sujet »

Outre la masse des documents accessibles sur l'Internet, les éditeurs de livres électroniques, quelle que soit la forme physique que prend leur transmission à l'utilisateur, ont commencé la diffusion de leur fonds numérique auprès des bibliothèques, non seulement en Amérique du Nord, où l'offre électronique des bibliothèques se compte en milliers ou dizaines de milliers de références, mais également en Europe.

Pour donner un exemple français, les Instituts d'études politiques de Grenoble et de Lyon, ainsi que l'université de Marne-la-Vallée vont étudier, dans le cadre d'un projet commun de « campus numérique » avec Vivendi Universal Publishing et les Éditions De Boeck, les conditions

Christian Lupovici *est directeur du service commun de la documentation de l'université de Marne-la-Vallée. Analyste-programmeur, titulaire du DSB, il est l'actuel président de l'ADBU. Il a publié plusieurs articles concernant ses recherches sur le document électronique.*

de la mise en ligne et de l'utilisation de manuels universitaires numériques. Il faut ajouter à cette offre commerciale l'offre institutionnelle de documents produits, voire édités par les organismes d'enseignement et de recherche.

De la gestion de l'accès à la gestion du document

La documentation acquise par les bibliothèques n'est plus exclusivement – et depuis longtemps déjà – une documentation papier, ces dernières cataloguent couramment des documents audiovisuels sur différents supports. La chaîne de traitement des documents numériques ne diffère en rien de la chaîne de traitement des documents papier, pas plus que de celle des cassettes ou des CD en tant qu'objets catalogués. Nous sommes toujours en présence de documents physiques qui suivent une procédure de traitement séquentielle, de l'acquisition à la mise en rayon en passant par le catalogage. Ce traitement consiste à la fois en une description (externe), une indexation d'accès et une indication de localisation. Les données administratives qui permettent la gestion du document (les règles de communication, la confidentialité, les mouvements de prêt…) sont réparties entre différents systèmes qui vont du module de prêt au règlement intérieur de la bibliothèque.

En revanche, le document numérique, au sens utilisé dans cet article, est un document qui n'est lié à aucun support physique, lequel n'est qu'un véhicule transitoire : c'est un document « structuré » (plus ou moins) qui contient les éléments essentiels de son identification, de ses accès (liés à l'identification mais aussi à sa structure) et de son administration. Jusqu'à présent, le document numérique n'a pas été traité par les bibliothèques en tant que tel.

Si l'on considère les revues électroniques, leur traitement échappe aux bibliothèques tant qu'elles sont accessibles auprès de l'éditeur. C'est lui qui règle les questions de création, de chargement, de gestion, d'architecture du système, de mise en ligne, d'interface, de sécurité, d'archivage…

Si l'on considère les monographies électroniques, elles ont été traitées soit en tant que support, comme des objets audiovisuels classiques et cataloguées comme tel, soit en tant que bases de données de contenu, comme les bases bibliographiques *(cf.* FRANTEX du CNRS) et considérées selon la même approche que les revues électroniques aujourd'hui. Il n'y a pas eu d'appropriation de la gestion des documents par la bibliothèque.

La bibliothèque et le document numérique

Les bibliothèques sont désormais confrontées à la gestion de documents numériques pour de multiples raisons. D'abord, parce qu'elles font partie d'une institution qui crée elle-même des documents de ce type. C'est particulièrement le cas des universités, qui produisent des cours, des travaux de recherche et des travaux d'étudiants qui doivent tous être valorisés, « publiés », mis en accès public et éventuellement conservés sur le long terme. Ensuite, parce qu'elles ont pu numériser des documents qu'elles doivent désormais gérer et préserver. C'est particulièrement le cas des bibliothèques qui conservent des fonds patrimoniaux.

Enfin, toute bibliothèque est amenée à se poser la question de son rôle dans la chaîne éditoriale, parce que son institution édite de nombreux documents qui vont de la communication institutionnelle à la publication scientifique (universitaire ou de société savante). Cette implication de la bibliothèque n'est pas exclusive, mais elle est nécessaire et légitime car elle permet, aux différents stades du cycle de vie des documents, de prendre totalement en compte leurs aspects documentaires. Ces derniers concernent les métadonnées de description (analogue au catalogage descriptif), les données de structure (la hiérarchie des paragraphes, les liens internes et externes, les notes, la bibliographie, l'appel aux fichiers images, vidéo et son…), les données administratives (gestion administrative et juridique, comme les autorisations d'accès ou la confidentialité, éventuellement le prix…) et les données de préservation (pour permettre la migration future vers d'autres plates-formes de lecture).

La compétence sur les formats de documents et sur les métadonnées relève de la responsabilité des professionnels de la documentation. C'est pourquoi le travail d'édition électronique ne saurait se passer des bibliothécaires et des documentalistes.

Les thèses comme exemple de document numérique structuré

La circulaire du ministère de l'Éducation nationale (n° 2000-149) du 21 septembre 2000 sur les thèses annonçait comme un objectif à atteindre le dépôt électronique des thèses.

Ce type de document est en effet produit en format électronique natif et il suffit de lui appliquer quelques règles de structure pour le rendre utilisable en tant que document structuré pour la recherche documentaire, la publication sur le web et pour sa gestion administrative en bibliothèque. Pour la première fois en France, étaient définis des formats de production, un format de conservation, des métadonnées descriptives et administratives et leur formalisation. La mise en application des recommandations de cette circulaire va permettre à une vingtaine d'établissements de tester des organisations de chaîne de traitement des documents numériques, principalement à partir de l'expérience des universités de Lyon 2 et de Marne-la-Vallée.

Le principe repose sur la définition d'un modèle qui identifie tous les éléments de données et qui décrit et structure l'ensemble du document selon une formalisation normalisée. Ainsi, les auteurs qui respecteront ce modèle (en utilisant leur traitement de texte habituel) auront l'opportunité de bénéficier d'un outil d'aide à la rédaction. De plus, ce document, en passant dans une chaîne de traitement informatique, sera reformaté en XML (eXtensible Markup Language) selon un modèle de description normalisé (DTD : Description de Type de Document) qui lui assurera une indépendance par rapport à toute plateforme de lecture. Ceci est un élément important pour la garantie de pérennité et de réutilisation du document sur le long terme.

De plus, la structure XML, si elle peut être exploitée par un logiciel de recherche documentaire, offre des possibilités très étendues de navigation et d'accès.

Analyse des traitements à opérer

L'organisation de la chaîne de traitement peut commencer par l'identification d'une entité spécifique, séparée de l'organisation traditionnelle dans la bibliothèque. L'intérêt de cette méthode est de ne pas perturber l'organisation en place et de mieux identifier les problèmes qu'engendre la nouvelle organisation. De ce point de vue, le service « Édition électronique » (ex. SENTIER) créé à l'université de Lyon 2 pour le traitement des thèses en est un bon exemple*.

Néanmoins, il est sans doute pernicieux, voire dangereux, de maintenir une organisation séparée chaque fois que l'on est confronté à une nouvelle technique, à un nouveau type de document. Il faut très rapidement étudier la façon d'intégrer cette nouvelle entité dans l'organisation générale du traitement des documents.

* Voir la page : http://sophia.univ-lyon2.fr/index_EdE.html

Les traitements du document numérique

Le document numérique se distingue du document papier par le fait qu'il nécessite l'intégration des différentes phases de traitement. Dans le cas de l'appropriation de documents sur l'Internet ou de documents produits dans l'institution locale, les phases d'acquisition, de « catalogage » et d'indexation sont successivement effectuées par la même personne. Il y a donc intégration des procédures d'acquisition et de catalogage. Dans le cas de documents achetés à un éditeur, les phases d'acquisition et d'indexation (catalogage et indexation) doivent être automatiques et échapper au contrôle visuel humain. Par ailleurs, de nouvelles fonctions s'ajoutent aux fonctions traditionnelles.

« L'acquisition » du document et les questions juridiques

Lors de l'acquisition du document numérique, des questions d'ordre juridique sur le statut du document et de ses composants vis-à-vis du droit d'auteur et du droit commercial se posent. Elles ne se posaient pas pour le document papier et assez peu pour le document audiovisuel, parce que l'on en effectuait l'achat dans un environnement juridique bien éprouvé. Dans la prise en charge de documents issus du web ou créés localement, la vérification des droits attachés au document s'impose. De plus, il faut assurer l'intégrité de la procédure de transfert du document vers la base de données de gestion et d'interrogation à travers les traitements intermédiaires. Le document ne doit pas pouvoir être modifié ni subir d'amputation ou d'altération (de codage de caractères par exemple). Des protocoles de traitement et des processus de vérification (automatiques) doivent donc être mis en place.

La vérification de la structure du document : le « stylage »

Contrairement aux documents manufacturés (articles de revue, livres électroniques commerciaux) qui doivent entrer directement dans le système d'information après la mise au point du protocole de chargement, les documents créés localement (d'origine numérique, ou numérisés) doivent faire l'objet d'une prise en charge préalable de la bibliothèque pour ajouter des métadonnées et vérifier la structure et sa conformité au modèle prescrit. Cette phase est très coûteuse en temps de personnel pour des documents peu ou pas structurés selon le « style » de référence. Au contraire, cette phase peut être assez légère si l'auteur s'est parfaitement conformé au modèle de référence. Aussi la formation des auteurs est-elle d'une importance capitale, au moment où ils vont se mettre à composer leur document.

L'ajout de métadonnées administratives et descriptives

Pour tous les documents créés localement et, dans certains cas, pour ceux acquis à l'extérieur, qu'il s'agisse de documents structurés ou d'une collection d'images, l'intervention humaine est nécessaire. Elle permet une plus grande finesse de structure (pour donner un accès plus précis dans un document composé d'images numériques). Elle permet également d'ajouter des données concernant la gestion administrative, le statut juridique, la gestion de la préservation, voire des données d'autorité et d'indexation spécifique pour améliorer les possibilités d'accès à l'information.

La vérification de format et le recyclage des erreurs

L'entrée dans le système d'information constitue l'étape ultime de vérification de structure des documents. Il s'agit du contrôle primordial

dans le cas de documents commerciaux. Il peut, en effet, arriver que des documents ne « passent » pas ce contrôle et qu'ils soient rejetés en erreur. Qu'en faire ? Une procédure doit être établie pour chaque type d'erreur, en fonction de l'erreur elle-même et de l'accord que l'on peut avoir conclu avec le producteur (commercial ou local), pour conduire l'intervention nécessaire. Cette procédure indique comment les erreurs doivent être recyclées : renvoyées au producteur à l'unité ou par lot, ou bien traitées localement avec rapport au fournisseur.

Analyse de l'organisation de la chaîne de traitement

L'organisation du traitement des documents numériques doit prendre en compte trois séries de facteurs : les phases et la nature des traitements ; les types de documents et leur origine ; les types de compétences nécessaires.

Organisation de la collecte des documents

Organisation technique

Les documents numériques de type monographie sont de plus en plus « lourds » en terme de poids informatique, que ce soient des cours, des rapports de recherche, des mémoires ou thèses. Même les articles de périodique peuvent prendre une place importante pour peu qu'ils comprennent des images en couleur.

La tendance des documents à augmenter en volume est un phénomène à considérer très sérieusement dans la réflexion sur la chaîne de traitement et sur l'architecture du système d'information. Les documents, en effet, comprendront non seulement plus d'images en couleur, mais aussi de la vidéo, du son et du logiciel intégré. Déjà, la sauvegarde d'une thèse en sciences humaines ne peut plus être faite sur quelques disquettes, mais sur cédérom et bientôt sur DVD. Elle ne peut donc pas être transmise en « document attaché » par messagerie électronique. Il est nécessaire de créer un serveur FTP où les auteurs ou producteurs viendront déposer leurs documents et où le bibliothécaire le prendra pour le traiter. La procédure doit cependant être sécurisée, pour éviter toute interférence indésirable entre le dépôt par le producteur et « l'acquisition » par le bibliothécaire. La gravure d'un cédérom ou d'un DVD doit être possible pour assurer le dépôt ou l'archivage de la version déposée (comme témoin de recours).

Organisation humaine

L'interface avec le fournisseur, qu'il soit un fournisseur commercial ou un auteur local, demande une grande attention et une compétence étendue sur les aspects d'organisation scientifique, technique et juridique. C'est en effet lors du premier contact que se concrétisent les procédures qui vont suivre : la mise en conformité avec la structure demandée, la garantie d'intégrité du document, les garanties de respect de la réglementation sur le droit d'auteur et les droits collatéraux. On doit être capable d'indiquer à l'interlocuteur ce qui va advenir de son document d'un point de vue technique, administratif et juridique, ainsi que les délais de traitement.

La vérification du modèle de structure

Dans le cas des thèses, la version soumise au jury sera celle que le doctorant aura déposée avant vérification du stylage. Néanmoins, on peut prévoir qu'avec la généralisation de l'utilisation des modèles normalisés, l'auteur pourra présenter au jury la version déjà vérifiée, et, pourquoi pas, la version issue du fichier XML.

L'avantage de la situation actuelle est d'éviter toute discordance entre ce qu'a voulu l'auteur et ce qui est imprimé pour le jury. L'inconvénient, c'est que la conformité du document avec les intentions de l'auteur en bout de chaîne reste à faire approuver. Inversement, l'avantage de l'édition de la version soumise au jury à partir de la version (re)stylée, voire de la version XML, c'est d'avoir une assurance de parfaite conformité. De plus, dans le cas de documents non imprimables (logiciels, multimédia…), c'est la seule procédure possible. L'inconvénient, c'est la relative lourdeur du travail, qu'il faut effectuer dans un temps limité, et le fait que la mise en forme doive être validée par l'auteur à ce moment-là.

La vérification de la conformité de la structure avec le modèle est une phase intéressante, parce qu'elle fait apparaître avec clarté la structure intellectuelle du document, et donc la pensée de l'auteur. Il existe des cas où la difficulté à trouver la logique de la structure oblige le bibliothécaire à recourir à l'auteur pour résoudre le(s) problème(s). Cette phase est également fastidieuse, parce qu'elle implique une vérification formelle et répétitive, le parcours mécanique du document sans s'arrêter sur le contenu. Elle s'effectue dans des conditions analogues au catalogage (en particulier la dérivation de notice) avec un recours à la dactylographie pour les corrections de forme. Aucune initiative n'est permise sur le fond, et très peu sur la présentation : le travail effectué ressemble à celui d'un imprimeur.

La version correctement mise en page et vérifiée est une version qui réclame l'approbation de l'auteur pour être considérée comme la matrice de la version de référence.

L'organisation du traitement

La vérification de structure et l'ajout des métadonnées sont des opérations qui peuvent s'apparenter au catalogage. Elles demandent les mêmes compétences, mais sont effectuées dans une philosophie bien différente. Il s'agit ici de compléter le document lui-même avec tous les

éléments qui assureront son autonomie en termes de gestion technique et administrative, d'accès et de préservation.

Les opérateurs travaillent sur des documents complets au lieu de ne travailler que sur la notice descriptive. Les postes de travail doivent comporter les logiciels utilisés par les créateurs de document (textes, images, vidéo, son…) pour permettre ce travail. Les opérateurs doivent pouvoir graver un cédérom ou un DVD de sécurité au moment où, récupérant le document du serveur FTP, ils ont vérifié que le document était complet. Ils vont alors commencer à travailler sur la vérification de structure et l'ajout des métadonnées.

L'organisation de la gestion des cycles d'édition et de correction

Entre l'auteur et le bibliothécaire, le document va décrire un ou plusieurs cycles de vérification, pour que chacun s'assure que la version numérique est conforme à sa représentation (généralement l'impression papier) et que celle-ci est conforme aux intentions de l'auteur.

Dans le cas contraire, et particulièrement dans le cas des corrections demandées par un jury (thèse) ou un éditeur (article ou livre), le bibliothécaire doit corriger ou intégrer les corrections de l'auteur pour se conformer au désir de l'éditeur. Le bibliothécaire doit aussi tenir compte des contraintes et nécessités techniques pour que le document soit techniquement le plus correct possible et permettre ainsi la pérennité de son utilisation.

Dans l'avenir, tous les documents stockés en XML devront avoir une feuille de style d'impression associée (XSL ou XML Style Sheet). Celle-ci devra faire l'objet d'une validation par l'auteur, non seulement pour la conformité de la mise en page, mais aussi pour la conformité de la représentation des caractères spéciaux. Ces opérations n'ont rien à voir avec le travail d'éditeur scientifique *(editor),* mais elles sont très proches du rôle de l'éditeur commercial *(publisher),* hormis pour ce qui concerne la diffusion et la vente. Dans le cas d'une diffusion électronique gratuite sur le web, nous sommes très proches d'accomplir la totalité du cycle de publication du document. C'est pourquoi l'université de Marne-la-Vallée considère qu'elle est véritablement éditeur des documents ainsi produits, auxquels elle adjoint une mention de *copyright.*

L'organisation de la mise à jour du système d'information

Les types et les volumes de documents que les bibliothèques auront bientôt à gérer ne correspondent plus à une solution GED (gestion électronique des documents) ni à une gestion en répertoires sur le web, même si le démarrage de cette gestion peut commencer ainsi. La nécessité d'un système d'information propre à la gestion des documents électroniques se fait très vite sentir. Le problème, c'est qu'il n'existe actuellement aucun système de gestion de documents comparable aux systèmes commerciaux de gestion de bibliothèque. C'est pourquoi l'université de Marne-la-Vallée a été chargée par le ministère de l'Éducation nationale de créer un tel système pilote avec pour mandat qu'il puisse être reproductible dans les autres universités.

Ce projet, dénommé PELLEAS, organise la gestion des documents numériques de tout type, de leur entrée dans le système d'information jusqu'au contrôle d'accès aux documents et à l'habilitation des utilisateurs. Le bibliothécaire commande l'ordre d'entrée des documents dans le système de base de documents. Le système vérifie que les contrôles automatiques de cohérence et d'intégrité se sont bien déroulés. Dans le cas contraire, le bibliothécaire enclenche le processus de recyclage des documents rejetés.

Une fois les documents entrés dans le système PELLEAS, les métadonnées sont extraites automatiquement des documents pour générer les index. C'est là que s'achève le rôle de la production. Le relais est pris par l'administrateur du système d'information.

Dans le cas d'un système d'information de type PELLEAS, l'édition à la demande des documents XML s'effectue en utilisant une feuille de style d'impression. Cette feuille de style peut être paramétrée pour tenir compte des intentions précises de l'auteur et devra, dans la plupart des cas, faire l'objet d'une validation de sa part. Elle restera attachée au document correspondant pour être réutilisée à chaque impression.

Élargissement et intégration des compétences

Le traitement des documents numériques implique un élargissement des compétences : les tâches techniques sont plus diversifiées et plus intégrées que celles qu'exige le traitement catalographique.

Les qualités et les compétences utiles au traitement des documents numériques correspondent sur beaucoup de points à celles des bibliothécaires adjoints spécialisés (BAS) dans les bibliothèques du secteur public. En effet, l'apprentissage de la structure des notices bibliographiques et de leur codage est une bonne introduction à la structure du document et à son codage. Plus encore, les métadonnées sont, par analogie, des données de catalogage dont l'objectif et le champ d'application sont modifiés, et dont la formalisation est différente, mais qui demandent de la part de l'opérateur les mêmes qualités de précision et de clarté des concepts.

Il semble donc naturel que l'organisation d'une chaîne de traitement des documents numériques s'organise au sein du service du catalogage. Il est logique que l'on emploie les mêmes personnes pour traiter indifféremment les documents numériques et les documents traditionnels. Ainsi, le personnel formé aux différentes techniques peut se consacrer à

l'une ou l'autre des chaînes de traitement selon les variations de la charge de travail. Il faut noter que seules les bibliothèques disposent d'un personnel bien formé et en nombre assez important pour pouvoir optimiser ces charges de travail, extrêmement variables, toute l'année.

La compétence et la formation

Il est urgent d'introduire, à tous les niveaux de formation, un apprentissage au traitement des documents numériques. Les personnels des catégories A et B de la filière des bibliothèques doivent savoir traiter les formats de fichier, en particulier les formats de la famille SGML, comme le XML et le HTML, et les DTD du monde documentaire et éditorial (EAD, TEI, ISO 12083…), de la même manière qu'ils ont appris les formats catalographiques ou les formats du papier et du livre.

Il est également important de les former aux modèles conceptuels de données et à leurs descriptions, comme ils ont appris les règles de catalogage et la construction des vedettes. Le profil de « réviseur » de documents numériques va bientôt faire son apparition dans les propositions de postes. Il faut que la formation initiale, la formation continue et les descriptifs de métiers en tiennent compte.

La collaboration

La mise en œuvre d'une chaîne de traitement du document numérique nécessite une bonne collaboration des personnels de bibliothèque avec différents acteurs de l'établissement :

Avec les informaticiens

La collaboration avec les informaticiens est une composante essentielle pour la mise en œuvre technique de la chaîne, depuis l'installation du serveur FTP pour l'acquisition des documents envoyés par les auteurs jusqu'à la mise en place du système d'information, en passant par les configurations logicielles (quelquefois relativement exotiques) et l'équipement des postes de traitement des documents avec leurs périphériques (scanners, graveurs de disques et imprimantes).

Avec les auteurs

La réussite de la chaîne dépend aussi de l'investissement personnel des auteurs dans les procédures mises en place. Cela passe d'abord par la compréhension et le respect du modèle de structuration du document, puis par la participation à la pose de liens et à l'indexation spécifique, ainsi éventuellement qu'à la détermination des métadonnées de gestion et de préservation.

Avec les éditeurs commerciaux

Que ce soit dans le sens de l'acquisition de documents commerciaux, ou pour fournir des documents d'auteur à un éditeur commercial, la collaboration est nécessaire pour mettre au point les bonnes interfaces et les protocoles de traitement de l'information qui passe de l'un à l'autre.

L'impact sur l'organigramme du Service commun de documentation

L'organisation d'une chaîne de traitement des documents numériques est à la portée de toute bibliothèque, même de taille modeste. Si l'activité se révélait insuffisante sur le seul aspect du traitement des documents numériques, il suffirait que les personnes affectées aux tâches de traitement aient une autre activité. Néanmoins, la technicité du traitement des documents numériques nécessite une activité soutenue (au moins un mi-temps), afin d'entretenir les compétences.

La refonte des services techniques

Il est tentant d'imaginer que l'organisation du travail de traitement des documents devra se fonder sur la chaîne numérique, et de considérer que la chaîne papier n'est qu'un cas particulier de l'organisation. C'est en tout cas à méditer.

Les chaînes de traitement qui se mettent actuellement en place dans les bibliothèques en sont souvent encore à un stade expérimental. Il faudra attendre de disposer de leurs retours d'expérience pour analyser les changements intervenus dans la structure même de l'organisation de l'institution. Il est en tout cas prévisible qu'une restructuration sera nécessaire, et il faut y préparer nos institutions. À l'université de Marne-la-Vallée, la fusion du service des acquisitions avec le service du catalogage en un département technique semble être une évolution évidente. Dans cette nouvelle entité, les opérateurs seront organisés en « ligne de produit » plutôt que selon la façon taylorienne actuelle.

Vers des presses universitaires numériques

Dans les universités qui n'ont pas de presses universitaires, le service commun de la documentation a l'occasion de mettre en avant l'intérêt d'une activité de publication, en particulier sous forme électronique. Elle peut faire valoir sa légitimité pour mettre en place cette activité au sein du service commun, sachant que la commercialisation d'une telle activité n'est plus de son ressort.

Dans les universités qui ont déjà des presses universitaires, le service commun peut – et doit – jouer un rôle en amont de la publication. Il doit également pousser les responsables des presses universitaires à évoluer dans leurs techniques de préparation et de publication des documents pour qu'ils adoptent des formats adaptés à la publication électronique, selon les normes des grands éditeurs scientifiques. L'édition papier est alors considérée comme un sous-produit du document électronique.

Novembre 2001