La conservation du document numérique dans le domaine de l'édition électronique. Pourquoi ? Quoi ? Comment ? L'exemple du centre de ressources numériques TELMA

Gautier Poupeau

La conservation du document numérique dans le domaine de l'édition électronique
Pourquoi ? Quoi ? Comment ?

L'exemple du centre de ressources numériques TELMA

I- Comment en est-on arrivé à la question de la conservation du document numérique ?

En guise d'introduction, il me semble important de légitimer ma présence aujourd'hui, en vous expliquant ce qui a amené l'École des chartes à s'intéresser de façon concrète à la conservation du document numérique, alors qu'une sensibilisation à cette problématique était dispensée dans les cours d'archivistique donnés à l'École.

L'Ecole nationale des chartes mène depuis quatre ans une politique de publication électronique qui a abouti à la création de deux sites Web :

- **THELEME**, Techniques pour l'historien en ligne, études, manuels, exercices. un manuel en ligne de sciences auxiliaires de l'histoire constitué de bibliographies des différents enseignements de l'établissement, d'introductions de cours ou d'outils comme un dictionnaire des abréviations et de dossiers documentaires qui abordent toutes les techniques amenant à la synthèse historique à partir d'exemples concrets mis au point par les élèves avec l'aide de l'équipe pédagogique de l'établissement.

- **ELEC**, Éditions en ligne de l'École des chartes, rassemble des éditions de sources, des projets de numérisation, des bases de données, des instruments de référence et des recueils d'articles en accès libre et gratuit.

Cette expérience a représenté un moyen de mieux faire connaître et partager les compétences de l'École des chartes en matière de traitement et d'exploitation des sources historiques, en particulier l'édition critique de sources. L'édition électronique sur le Web représentait un vecteur de diffusion sans précédent et souple et limitait les moyens logistiques par rapport à la publication sur le support papier qui n'était pas abandonnée pour autant. Ainsi, le résultat d'entreprises de longue haleine comme l'édition de sources ou la publication d'instruments de référence pouvait être mis à disposition sur le Web au fur et à mesure de leur réalisation, sans attendre leur mise au point finale illusoire dans le cas des instruments de référence.

Notre réflexion et notre démarche ont été influencées par trois points fondamentaux qui nous ont

accompagnés tout au long de la réalisation et l'élaboration de nos différents projets :

- **L'utilisation de logiciels libres**. La question du coût économique a été évidemment un facteur dans ce choix, il ne faut pas le nier. Mais, au-delà de cet argument pragmatique, l'utilisation de logiciels libres permet aussi de s'astreindre de la tutelle d'un éditeur de logiciels et des choix qu'il est amené à faire. Nous nous retrouvions plus libre d'agir à notre guise. De plus, les logiciels libres sont construits bien souvent sur des formats ouverts et libres et non sur des formats propriétaires dont l'évolution et les spécifications dépendent de l'éditeur.
- **L'accessibilité au sens large**. Le concept d'accessibilité renferme en fait plusieurs aspects. Tout d'abord, nous avons toujours privilégié des interfaces graphiques qui permettent une appropriation en douceur de ce nouveau média par les chercheurs en sciences humaines et plus particulièrement en histoire. Ainsi, nous avons plutôt cherché à adapter les modèles de publications existants, plutôt qu'en inventer de nouveaux qui auraient pu être rejetés par les chercheurs. De plus, nous avons eu le souci constant de proposer des URL visibles, simples à taper et avec un minimum de significations et de ne mettre en place aucune barrière technique et/ou économique pour accéder aux ressources que nous mettions à disposition. Enfin, nous avons appliqué les règles d'accessibilité des sites Web selon les recommandations du W3C reprise ensuite par l'ADAE en France. Même au niveau des pages Webs, nous avons mis en place une séparation stricte de la mise en forme et du contenu.
- **Utilisation de standards ouverts et libres**. J'y ai déjà fait allusion dans les deux points précédents. Nous nous sommes efforcés de n'utiliser que des formats ouverts et libres. Cela constitue le point fondamental de notre démarche. Les technologies du W3C représentent bien-sûr le cœur des formats que nous utilisons : XML, XSL, Xquery, HTML, CSS. Mais, nous avons adopté la même attitude quant aux choix des grammaires XML utilisés : TEI dans la plupart des cas, mais aussi METS et MODS, pour n'en citer que deux.

L'expérience acquise par l'École des chartes dans la mise en ligne de ressources issues de la recherche explique en partie le fait qu'elle a été retenue en compagnie de l'IRHT dans le cadre de la réponse à un appel d'offres lancé par la direction de l'information scientifique du CNRS pour la constitution de centre de ressources numériques. En juillet 2005, la direction de l'information scientifique nouvellement créée dans le cadre de la réforme du CNRS engagé par Bernard Laroutourou et dirigée par Laurent Romary a lancé un appel d'offres visant à l'identification de pôle de compétences existant dans le domaine de la mise à disposition sur le Web des données de la recherche. Il est alors apparu à la DIS que cet appel devait en priorité concerner les sciences humaines et sociales pour lesquelles ce besoin était le plus important. Dans la vision de la DIS, ces centres ont vocation à fédérer et à accompagner les initiatives dans le domaine de la création, de la gestion et de la diffusion de l'information scientifique (données ou corpus) et des outils dans un souci de mise à disposition gratuite sur le Web et de pérennisation. En identifiant et aidant ces pôles, le CNRS

encourage ce type d'action dans un souci d'harmonisation des standards et des moyens pour éviter la dispersion, cause de nombreuses pertes de données. A travers cette initiative de fédération, le CNRS cherche aussi à valoriser au même titre qu'une publication la mise en place souvent longue et fastidieuse de base de données aidant à la recherche. Plusieurs domaines plus précis dans le département SHS avaient été identifiés : les corpus oraux, les corpus textuelles, les sources historiques non imprimées ou encore les sources spatiales numériques pour n'en citer que quelques uns.

Dans ce cadre, l'École des chartes et l'IRHT (Institut de recherche et d'histoire des textes), unité propre du CNRS dont les domaines d'activité recoupent en partie les activités de l'École et qui possède aussi une expertise dans la mise en place de bases de données de repérage des sources ont présenté un projet commun dans le domaine des sources historiques non imprimées. Ce centre dont, vous l'avez compris, le projet a été retenu s'est donné diverses missions :

- Mettre en place une plate-forme technique permettant d'accueillir, gérer, conserver et diffuser les données accueillis par le centre, entre autres les données des deux institutions porteuses ;
- Devenir un relais d'information et de soutien technique pour la communauté des chercheurs en histoire dans le domaine du traitement numérique de l'information scientifique et technique par des actions de formation, par exemple ;
- Devenir un intermédiaire entre la communauté des chercheurs et les institutions de conservation, en particulier leurs tutelles pour mener des actions de numérisation partagées.
- Assurer une veille technologique et représenter la communauté des chercheurs en histoire dans les organismes de normalisation comme le TEI consortium.

Cette liste n'est pas sans rappeler les missions que doit remplir une archive suivant le modèle OAIS. En fait, il est vite apparu que ce modèle allait pouvoir nous aider à mettre en place ce centre, car plus qu'un modèle pour gérer la conservation à long terme des documents numériques, l'OAIS peut constituer un très bon guide voire un *vade-mecum* pour la gestion quotidienne de l'information numérique. D'ailleurs, dans la lettre de missions adressée par le CNRS et qui a vu la création du centre, six des sept missions reprennent les entités définies dans le modèle OAIS. Finalement, comme beaucoup d'autres et comme le rappelle le modèle OAIS, nous n'avons fait que constater que « la conservation de l'information sous forme numérique est beaucoup plus complexe que la conservation de l'information sur supports papier ou film. Ceci n’est pas seulement un problème pour les Archives traditionnelles, mais également pour de nombreux organismes qui, jusque-là, n’avaient jamais eu conscience d'assurer une fonction d'archivage ».

II- Que doit-on conserver ?

Avant de commencer à implémenter un système de conservation du document numérique

conforme au modèle OAIS, il est essentiel de se demander à quoi il va servir et ce qui va être conservé vu l'investissement humain et économique que représente un tel déploiement. Pour cela, je vous propose de dresser une rapide liste des caractéristiques de l'édition scientifique sur le Web. Commençons par l'essentiel. Dans ce contexte, le support de l'ouvrage n'est disponible qu'à un seul endroit : sur le serveur. Il faut aussi ajouter qu'il n'existe pas de dépôt légal par l'éditeur pour les éditions électroniques sur le Web, mais nous reviendrons plus loin sur ce problème. De ce fait, le support unique des informations se situe chez l'éditeur, il a la responsabilité d'en garantir l'accès, alors que dans le cas du papier le support de l'information se décline en autant d'exemplaires existant. Dans le cas de l'édition scientifique en sciences humaines, les données mises en ligne sont vouées à être utilisées par les chercheurs dans 50 ans voire plus, car il n'est pas rare d'utiliser des éditions critiques de sources réalisées au XIX^{e} siècle pour effectuer nos recherches. Il faut se rappeler que les vingt dernières années ont vu l'élaboration de nombreux corpus ou bases de données dans les laboratoires de recherche français qui sont aujourd'hui complètement inutilisables à cause de leurs formats ou du logiciel qui a été utilisé... Il est essentiel si l'on veut qu'un rapport de confiance s'établisse entre les chercheurs et les données numériques qu'ils puissent s'y référer à long terme. S'y référer signifie aussi qu'ils puissent citer la ressource. Or, si la ressource est citée dans une bibliographie, elle doit pouvoir être retrouvée et accessible. Ainsi, la mission de l'éditeur électronique dans le milieu scientifique est de pouvoir garantir l'accès à long terme aux ressources qu'il produit. Évidemment, garantir l'accès signifie assurer la préservation des fichiers qui permettent la génération de la ressource.

Existe-t-il alors une différence entre le but poursuivi par un éditeur et par un bibliothécaire ou un archiviste ? Je dirais que oui. Alors que pour les seconds, l'intérêt principal concerne le document ou les fichiers, elle n'est qu'une conséquence pour les éditeurs pour qui l'intérêt principal est de préserver l'accès. Même si elle paraît infime, cette différence a des conséquences sur la gestion de l'archive et le cycle de vie des fichiers qui vont permettre la gestion de la ressource.

Pour bien comprendre les enjeux, je vous propose une analogie avec l'édition traditionnelle sur le support papier. Prenons le manuel sur le Moyen Âge paru aux Presses universitaires de France et écrit par Claude Gauvard, professeur à l'université de Paris 1. Il existe quatre éditions de cet ouvrage parues en 1996, 1997, 1999 et 2004. Chacune de ces éditions est conservée à la BnF au titre du dépôt légal. Entre l'édition de 1997 et la dernière en 2004, la couverture n'est plus la même, car l'éditeur a choisi de refondre sa collection « Premier cycle » au sein de la collection « Quadrige Manuels », les chartes graphiques ont été harmonisées. Or, dans son catalogue, l'éditeur ne propose que la dernière édition de sa ressource. Dans ce cas-là, les PUF continuent à vendre son ouvrage, mais uniquement la dernière édition comme nous l'avons vu, mais si demain cet ouvrage est en rupture de stocks et que l'éditeur ne souhaite pas le rééditer, l'ouvrage sera toujours accessible dans les bibliothèques et une référence précise à une édition de cet ouvrage trouvée dans une bibliographie quelconque sera toujours accessible.

Passons cet exemple dans le contexte de l'édition électronique et essayons d'en tirer les conséquences pour la gestion de l'archive. Une ressource électronique peut évoluer à trois niveaux : son contenu, son aspect graphique, le format utilisé. Il faut garder à l'esprit que, contrairement à l'éditeur papier, les notions d'exemplaires et de rupture de stocks n'existent pas dans le paradigme électronique, car, comme nous l'avons vu, le support n'existe qu'à un seul endroit : sur le serveur de l'éditeur.

Si on change le contenu de la ressource, c'est à dire concrètement si on effectue une modification sur le fichier XML, dans le cadre du papier, on sort une nouvelle édition, mais dans le cadre de l'électronique, ces changements peuvent arriver à n'importe quel moment, nous n'avons pas besoin d'attendre une nouvelle édition, c'est bien justement un de ses avantages. Première question : doit-on conserver les deux versions du fichier XML ?

Si on change la charte graphique du site pour la faire évoluer par exemple ou pour des raisons de cohérence avec le reste du site dont la charte change aussi, des modifications sont apportées. Deuxième question : dois-je donner un moyen de visualiser les deux versions : la nouvelle et l'ancienne ? Est-ce que cela rentre dans le cadre de mes missions d'éditeur électronique ? Les historiens du livre ont montré l'importance de la mise en page et de la mise en forme pour étudier les réceptions du contenu de l'ouvrage. Cette question n'est donc pas anodine.

On peut aussi être amené à changer les formats des fichiers pour des raisons de cohérence éditoriale et technique. Par exemple, actuellement l'ensemble des éditions critiques proposées sur le site ELEC sont encodées au format P4 de la TEI. Or, le consortium TEI qui gère ce format est en train de mettre au point une nouvelle version dite P5. Il s'avère que la compatibilité descendante n'est pas assurée ce qui signifie qu'un fichier en TEI P5 n'est pas directement interopérable avec un fichier en TEI P4. Pour des raisons diverses nous pouvons être amené à utiliser un élément qui n'est proposé que par P5 pour un nouveau projet. Dans ce cas, pour limiter les exceptions dans notre système d'interrogation et de diffusion et dans le souci de conserver une cohérence éditoriale, nous allons être obligé de migrer nos fichiers, un peu à l'image des PUF qui harmonise la charte graphique de l'ensemble des ouvrages d'une collection. Dans le cas d'une archive OAIS dans une institution de conservation, le fichier n'est migré qu'à partir du moment où il devient illisible, ce qui n'aurait pas été le cas dans mon exemple, un fichier XML étant toujours consultable actuellement, les besoins de la communauté d'utilisateurs cible ne se situent pas au même niveau. Mais, cela pose une troisième question : dois-je conserver les fichiers dans les différents formats sachant qu'ils sont toujours lisibles ? Un historien des technologies du Web dans 50 ans pourrait avoir besoin de comprendre où se situait l'évolution entre l'utilisation de P4 et P5 dans notre corpus.

Ces trois questions appellent des réponses différentes. Pour la première concernant les différentes versions du contenu, *a priori* tout comme un éditeur papier ne met à disposition que la dernière édition de ces ouvrages, nous ne gérons pas les différentes versions et la ressource accessible en ligne

n'est générée qu'à partir de la dernière version mise en ligne. Cela pose évidemment un problème d'intégrité et donc des questions de validation scientifique. Imaginons qu'une erreur ait été commise dans l'édition ou qu'une interprétation soit remise en cause par une autre publication scientifique. Suite à cela, le fichier est modifié. Par honnêteté scientifique, nous devons être en mesure de prévenir le lecteur de cette modification. Pour autant, cela impose de mettre en place un système de *versionning* ce qui complique l'application et donc le coût de la mise en place de l'archive. La réponse à cette question n'est donc pas tranchée et nous devons absolument intégrer les chercheurs dans notre réflexion à ce niveau. Mais, un choix devra être fait et qui peut être lourd de conséquences. En tant qu'éditeur, il me semble que cela compliquerait considérablement la tâche et que le rapport entre coût économique et intérêt scientifique doit être pris en compte. Une solution intermédiaire serait de conserver les différentes versions dans le système, mais que ne soit accessible sur le Web que la dernière version avec la mention de la date. Les précédentes versions pourraient être accessibles sur demandes dans une version dite dégradée.

En revanche, la deuxième question appelle une réponse directe. Il ne fait pas partie du rôle de l'éditeur électronique de garder la mémoire des différentes chartes graphiques, car elle ne constitue pas d'impératifs scientifiques. Ce cas relève clairement du rôle de conservation des supports, dévolu aux bibliothèques. Nous sommes toujours en accord avec les responsabilités obligatoires qui échoient à une archive OAIS qui doit s'assurer que les informations à conserver sont compréhensibles par la communauté des utilisateurs cible. Par ailleurs, cette tâche sera assurée par le dépôt légal du Web. Il n'est pas pertinent dans le cas du contenu, car les collectes ne sont qu'annuelles et sélectives et il sera impossible de s'y fier à 100%, mais dans le cas de la charte graphique les échantillons indexés seront suffisant pour se faire une idée précise de son évolution.

La troisième question pose en quelque sorte la même que la précédente : la constitution de la mémoire du « support » qui n'a ici pas tout à fait le même sens. Pour autant, dans ce cas, le dépôt légal du Web ne constitue pas une solution, car il n'indexera pas le fichier source de nos éditions, simplement la ressource qui est générée. De ce fait, la question reste en suspens, mais pour l'éditeur, la réponse est claire : cela ne fait pas partie de sa mission. Qui pourrait accueillir ces fichiers ? L'IMEC (Institut pour la mémoire de l'édition contemporaine) est-elle en mesure de le faire ? Ces fichiers relèvent-t-ils des archives produites par l'établissement versées aux Archives nationales au titre de notre production courante ? Je profite de la présence aujourd'hui d'archivistes pour poser la question.

Les contours et les buts de l'archive OAIS étant maintenant dressés, passons au vif du sujet en vous proposant l'état de notre réflexion sur la mise en place concrète de l'archive OAIS.

III- Comment allons-nous mettre en place notre archive OAIS ?

La question de la conservation du document numérique a évolué significativement ces trois dernières années. Des réflexions, des normes et les premières solutions sont apparues. Pour autant, la question reste très complexe à aborder par la profusion des problématiques soulevées. On pourrait dresser un inventaire à la Prévert : XFDU, MPEG21, SCORM, DIDL, PREMIS, METS, répertoire de formats, PRONOM, OAIS, LOCKS, ADORE, OpenURL, ARK, handle et bien d'autres. Quand on essaye de prendre le problème par un bout, la pelote se déroule rapidement et il faut bien se rendre à l'évidence : il faut aborder le problème dans son intégralité. Et pour paraphraser Géronte dans les *Fourberies de Scapin*, on se demande bien ce qu'on est allé faire dans cette galère...

Pour régler rapidement ce problème, on peut tout à fait utiliser les solutions intégrées existantes. Nous en avons étudié trois : LOCKSS dont nous venons d'entendre une présentation, ADORE et Dspace. Chacune présente des particularités intéressantes. LOCKSS dont les éditeurs sont la principale cible rassure dans un premier temps en rejetant la question du stockage sur des tiers archiveurs. Mais, il ne résoud pas tous les problèmes et, surtout, il faudrait que ce système soit adopté en France ce qui n'est pas le cas à ma connaissance actuellement. Quant à ADORE, la réflexion menée par Herbert van Sompel et son équipe est intéressante, mais impose de toute façon d'avoir réfléchi au préalable à la composition de ces métadonnées de préservation et à mettre en place un système d'identifiants pérennes. De plus, le déploiement et la mise en production d'ADORE est assez complexe ; ainsi, l'architecture de ce système reposant sur l'interrogation d'un entrepôt OAI, il nous a semblé que les temps de réponse pouvait être long et donc pénaliser l'accès à la ressource pour les utilisateurs. Enfin, Dspace convient assez bien à des entrepôts d'articles, ce qui ne correspond pas parfaitement à la granularité des informations que nous mettons en ligne. Bref, la solution miracle n'existe pas pour nos besoins et il nous fallait donc repartir du modèle OAIS et reprendre les choses calmement et par étapes.

A- La stratégie de déploiement

Vu les moyens mis à notre disposition par nos institutions et le CNRS pour mettre en place cette archive, il nous faut adopter une stratégie de déploiement sur plusieurs années. Mais cela ne nous empêche pas d'avoir une réflexion préalable pour penser l'ensemble de l'architecture et des fonctionnalités et les rôles qu'elle doit tenir.

En tant qu'éditeur, la diffusion des données représente notre intérêt principal. A ce titre, l'entité « Accès » constitue notre principale mission. Par conséquent, nous consacrons actuellement nos efforts sur la réflexion et le développement de la constitution de paquets de diffusion (DIP) et des outils permettant aux utilisateurs d'accéder aux données mis à disposition dans notre archive.

Possédant déjà un certain nombre de corpus en ligne, ce travail consiste essentiellement à repenser le système actuel de diffusion pour le rendre plus robuste et performant et à faire en sorte que cette partie puisse s'intégrer sans problèmes dans le système final.

De façon concomitante, nous nous attachons à définir les métadonnées de préservation qui vont accompagner chacun de nos corpus. A l'image du système ADORE pour lequel il faut au préalable disposer d'un format de métadonnées gérant les objets complexes comme METS, XFDU ou DIDL, mettre au point les métadonnées de préservation et les appliquer constituent une bonne part de réussite de la mise en place d'une archive OAIS. Cet intérêt précoce pour les métadonnées a pour conséquence de les placer au cœur de notre système permettant à la fois la constitution des paquets de diffusion en fonction de la requête des utilisateurs et la mise en place à terme d'une base de données qui gérera concrètement l'archive et les paquets d'archivage (AIP).

N'ayant pas dans l'immédiat de plans de migration conséquents à prévoir et disposant de métadonnées très détaillées, nous pouvons nous permettre de mettre en place ce système, qui correspond aux entités « Gestion des données », « Administration » dans un second voire un troisième temps. Il faut aussi avouer que ces entités constituent certainement les parties les plus difficiles à déployer dans le modèle OAIS.

Évidemment, de façon continue, une importante activité de veille et de mise en place d'une documentation est menée pour assurer les choix technologiques et éventuellement prévoir un plan de migration.

B- La constitution des trois types de paquets

1- Le paquet de versement

Nous sommes en mesure de maîtriser les formats des fichiers conservés dans l'archive, ce qui est un avantage voire même une obligation au regard des moyens à notre disposition. En effet, il est plus simple et économique pour nous de faire un investissement en amont, en faisant migrer des données qui ne seraient pas conformes aux règles établies pour l'archive, plutôt que d'accepter des formats hétérogènes dont nous ne pouvons maîtriser parfaitement l'évolution et d'en assurer la conservation à long terme. Ainsi, la composition d'un paquet de versement (SIP) est assez drastique.

Pour les données, nous n'accepterons que des fichiers au format XML. En revanche, nous n'obligerons pas les producteurs à utiliser une grammaire particulière, même si nous ferons des recommandations en fonction du type d'information. Ainsi, s'il s'agit d'une édition de sources et plus généralement de textes, nous privilégierons la TEI, de même que pour une base de données bibliographiques MODS sera le format pivot proposé. Si un producteur dispose d'une base de données relationnelles, selon les règles que je viens d'expliquer, elle sera migré vers XML. En fonction de l'accord signé entre l'archive et les producteurs, nous pourrons faire migrer ces données vers un XML

« plat » ou vers une grammaire particulière. Mais, pour limiter le nombre de grammaires maintenues, la première solution ne sera envisagée qu'en dernier recours.

Le traitement des fichiers XML qui permettra d'assurer une compréhension immédiate par les utilisateurs de ces données sera assuré par trois formats : XSL, XSL-FO et Xquery qui présentent l'avantage d'être des formats ouverts et libres mis au point au W3C et pour les deux premiers d'utiliser la syntaxe XML avec tous les avantages que cela présente. Une feuille de style au format CSS complète les moyens de traiter les données pour mettre en page les fichiers HTML issus du traitement XML.

Deux utilisations des images peuvent être distinguées. Les images issues d'une numérisation devront être conservés selon deux formats, comme c'est maintenant l'habitude : en TIFF et en JPG. Les images utilisés par la charte graphique devront être au format JPG ou PNG, éventuellement au format GIF, car le logiciel Internet Explorer de Microsoft ne supporte pas actuellement toutes les fonctionnalités du format PNG.

Ces différents formats recouvrent 95% de nos besoins. Au cas par cas, nous pourrons édicter des règles pour les formats sons ou vidéos.

La définition d'un paquet de versement est assez simple à déterminer, puisqu'il correspondra en gros à un corpus ou à un projet, c'est à dire l'équivalent actuellement d'un numéro dans notre collection ELEC. Évidemment, ce paquet sera complété par un fichier rassemblant l'ensemble des métadonnées de préservation. Il existe plusieurs formats pour gérer les objets complexes, ici notre corpus : METS, XFDU, SCORM et DIDL. Notre choix s'est rapidement orienté vers METS, format maintenu par la *library of congres*, car nous disposions déjà d'une expertise et d'une expérience avec ce format l'utilisant dans le cadre d'un projet de numérisation. De plus, METS répondait à nos besoins. Ce fichier METS intègre des métadonnées au format Dublin Core et ONIX pour les métadonnées dites descriptives et PREMIS pour les métadonnées administratives.

Les producteurs pourront déposer leurs paquets de versement sur un support physique, sur un serveur FTP sécurisé ou le mettre à disposition dans un entrepôt de type OAI dont les métadonnées Dublin Core seront complétées par la norme OLAC mis au point par la communauté des linguistes. Une fois récupérée, la conformité du SIP avec nos procédures sera vérifiée, les métadonnées complétées et nous procéderons éventuellement à des migrations selon les modalités définies dans l'accord signé entre le producteur et l'archive.

2- Le paquet d'archivage

Le stockage représente un des problèmes importants auquel nous devons faire face. D'après nos calculs, nous arriverons à un total dépassant le téra-octets dans les cinq prochaines années. Cela peut paraître peu au regard d'autres institutions. Pour autant, il faut quand même disposer du personnel et du matériel pour faire face. De plus, même si cela paraît peu, nous préférons virtualiser ce stockage,

c'est à dire héberger les données sur une machine et l'applicatif sur une autre machine. Cela impose d'attribuer un identifiant pérenne à chaque fichier. Cet architecture permet de pallier à l'éventualité fort possible que les fichiers changent de lieu physique de stockage. En effet, dans un premier temps, nous prévoyons de stocker les données en interne, mais espérons pouvoir rapidement trouver une solution d'externalisation comme l'IN2P3, nous sommes d'ailleurs ouverts à toutes les propositions. Tous les fichiers seront donc stockés en externe, sauf le fichier de métadonnées qui sera stocké sur la même machine que l'applicatif et indexée dans une base de données XML.

3- Le paquet de diffusion

En fonction de la nature de la requête, le paquet de diffusion sera composé d'un fichier XML contenant les données, d'un ou plusieurs fichiers XSL ou XSL-FO pour traiter le fichier XML et des images adéquats pour la charte graphique voire le cas échéant l'image numérisée. Dans le cas où l'utilisateur fait une requête sur la base de données XML, le paquet de diffusion contiendra un fichier XML généré dynamiquement correspondant à la réponse, la feuille de style XSL pour mettre en forme la réponse et les images. Dans les deux cas, le paquet de diffusion comprendra une feuille de style au format CSS pour mettre en forme le fichier HTML généré. Au final, l'utilisateur recevra un fichier HTML, texte ou PDF généré automatiquement. Il est à noter qu'à aucun moment ces fichiers ne seront archivés et conservés. Cela relève de la mission du dépôt légal du Web et, surtout, rend indépendant les paquets de diffusion de la plate-forme utilisé et des évolutions du format HTML par exemple. Ainsi, dans l'hypothèse d'une adoption par les concepteurs de navigateurs Web de XHTML 2 actuellement mis au point au W3C, il faudra alors modifier ou ajouter des feuilles de style XSL *ad-hoc* ce qui limite le nombre de fichiers à conserver.

C- Comment répondre aux requêtes des utilisateurs ? L'entité « Accès »

L'application au cœur de l'entité « accès » doit être capable de répondre à deux type de requêtes. Le premier type de requête correspond à la demande d'un utilisateur pour afficher tout ou partie d'un corpus dans les formats texte, PDF ou HTML dans la limite des feuilles de style définies par les producteurs de l'information. Deux arguments obligatoires et deux arguments optionnels sont requis :

- Le nom du corpus
- Le nom de l'ouvrage dans le cas où le corpus est composé de plusieurs ouvrages, c'est à dire concrètement de plusieurs fichiers XML. Dans le cas contraire, le nom du corpus est confondu avec le nom de l'ouvrage.
- Le nom de l'unité structurelle, c'est à dire le nom de la division de l'information que l'utilisateur veut afficher et qui correspond concrètement à une feuille de style XSL ou XSL-

FO.

- L'identifiant de l'unité structurelle dans le cas où l'ouvrage contient plusieurs unités structurelles du même type.

Par exemple, dans le cas des cartulaires numérisés d'Ile-de-France, si un utilisateur veut visualiser l'acte 105 du tome premier du Cartulaire des Vaux-de-Cernay, le nom du corpus sera : cartulaires, le nom de l'ouvrage: vauxcernay1, le nom de l'unité structurelle : acte et l'identifiant de l'unité structurelle : 105 ce qui donne concrètement dans l'URL : http://elec.enc.sorbonne.fr/cartulaires/vauxcernay1/acte105/.

Trois formats seront proposées en sortie à l'utilisateur : HTML pour l'affichage dans le navigateur, PDF pour l'impression et le texte brut pour que les chercheurs puissent récupérer le fichier afin d'y appliquer des programmes de statistiques textuelles. Mais, cette liste est certainement amené à changer avec l'évolution des technologies.

Le second type de requête correspond à l'interrogation de la base de données XML qui indexe l'ensemble des fichiers XML conservés dans l'archive. L'interrogation pourra être effectuée directement sur le site Web de TELMA grâce à des formulaires ou par l'intermédiaire d'un Web-service. Deux types d'interrogation seront proposés. Il sera ainsi possible d'interroger un ou plusieurs corpus en texte intégral ou selon des critères spécifiques à chaque corpus qui pourront le cas échéant être identiques entre plusieurs corpus.

Il faut aussi noter qu'à terme l'application devra être capable de gérer des niveaux de droits d'accès différents aux informations ce qui imposera la mise en place d'un système de DRM. Actuellement, ce n'est pas le cas, car l'ensemble des informations sont consultables librement et gratuitement, mais il se pourrait que certaines informations soient régies par des licences qui n'autorisent pas ce genre de diffusion.

L'application gérant ce système s'appuiera sur le serveur Tomcat et le framework Cocoon, deux logiciels libres mis au point et maintenus par la fondation Apache, à l'origine du serveur éponyme. Le logiciel eXist sera utilisé pour gérer la base de données XML. Il présente l'avantage de s'appuyer aussi sur le framework cocoon et constitue à l'heure actuelle une des deux solutions libres sérieuses dans le domaine. Les fichiers XML étant indépendant de la plate-forme utilisée, cette application peut être changé à tout moment sans perdre les données conservés. Il n'est donc pas utile de la pérenniser.

D- Gérer l'archive et les données sur le long terme

L'application gérant l'archive ce qui correspond à tout ou partie des entités « Gestion des données » et « Administration » s'appuiera intégralement sur la base de données XML dans laquelle sera indexée les fichiers de métadonnées décrivant chaque corpus mis en ligne par le centre. Il est donc essentiel de disposer d'un fichier METS intégrant l'ensemble des métadonnées indispensables au bon fonctionnement de cette application et donc de l'ensemble de l'archive. En plaçant le fichier METS au

centre du fonctionnement de l'ensemble de l'archive, étant à la fois requis pour construire les paquets de diffusion et constituant la mémoire de l'archive, cela permet de limiter la redondance des informations et la multiplication des informations. Pour autant, cette solution ne peut être envisagé que dans le cas où la constitution de ce fichier se fait au moment de l'élaboration du paquet d'archivage et que les informations ne sont pas fournies par ailleurs par des applications tiers existantes. Elle offre une souplesse et une simplicité d'usage et de maintenance qui n'est pas négligeable dans le cas d'une archive relativement restreinte.

L'application devra effectuer deux tâches principales. En ce qui concerne l'interrogation et la mise à jour de la base de données, elle doit être capable de répondre à des requêtes permettant de dresser la liste de corpus existants, la liste des fichiers pour chaque corpus avec l'indication de leurs formats, de retracer le cycle de vie des fichiers et les éventuelles modifications qu'ils ont subis, bref, de donner facilement une idée précise des différents fichiers contenues dans l'archive. Ainsi, l'application pourra prendre en charge la mise à jour éventuelle des informations, la notificationr aux administrateurs de l'archive et la production des différents rapports indispensables au bon fonctionnement de l'archive.

Par ailleurs, l'application devra intégrer un système de migration des fichiers. Dans le cas de fichier utilisant la syntaxe XML dont il faudrait simplement faire migrer la grammaire utilisée, la migration est assez simple. Après avoir effectué une requête sur la base de données pour rassembler tous les fichiers utilisant une grammaire précise, l'application devra être en mesure d'y appliquer une feuille de style XSL produite en amont par les administrateurs de l'archive, puis d'archiver les nouveaux fichiers obtenus et de le notifier dans le fichier de métadonnées. Dans le cas de fichiers binaires, dans l'avenir proche, il devra être possible d'insérer dans l'application des modules capable de migrer ces fichiers d'un format à un autre. L'architecture actuelle se basant pratiquement exclusivement sur les technologies Java, cette manipulation ne devrait pas poser de problèmes insurmontables. Pour autant, il ne faut pas se leurrer, la migration de ces fichiers représenteront des difficultés de taille dont la résolution ne saurait être trouvée aussi facilement.