World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

August 4, 2005

Code Number: 060-F
Meeting: 150 SI - ICABS (IFLA/CDNL Alliance for Bibliographic Standards)

NORMES DE METADONNEES RELATIVES A LA PRÉSERVATION DES RESSOURCES NUMERIQUES : CE DONT NOUS DISPOSONS ET CE QUI NOUS MANQUE

Sally H. McCallum
Library of Congress
Washington DC, USA

Traduction: *Rachel Cervera*
(documentaliste, France)
rcervera@chez.com

Résumé:

Les métadonnées permettant la mise en place de procédés automatisés de préservation, seront un élément clé d'une préservation réussie des ressources numériques. La quantité de documents numériques empêchera tout traitement humain, et le fait que ces ressources soient électroniques rend logique les traitements des activités de préservation par ordinateur. Durant ces dix dernières années, un certain nombre d'expériences de systèmes de dépôts numériques ont approché le problème de manières variées, développant et utilisant différents modèles de d onnées, et qui, de façon générale, ont amélioré notre compréhension. Cet article rapporte une initiative récente, PREMIS, qui s'appuie sur les concepts et les expériences actuels. Cette initiative mérite d'être évaluée avec soin pour s'assurer que les métadonnées identifiées puissent être utilisées de façon générale et devenir

un socle pour des métadonnées plus détaillées. Et quelle quantité sera encore nécessaire pour les activités de préservation ? Seront aussi discutées les initiatives pour des métadonnées techniques supplémentaires et des enregistrements de formats de documents.

Métadonnées fondamentales pour la préservation : PREMIS

Les origines

Le projet Stratégies pour l'implémentation de métadonnées relatives à la préservation (Preservation Metadata Implementation Strategies, PREMIS) est né des expériences de cette dernière décennie (1). Un important travail a été effectué sur les systèmes de dépôts au sein de la communauté des bibliothèques et particulièrement parmi les institutions constitutives de l'ICABS et de leurs collaborateurs. Inévitablement, ce travail impliquait la conception de modèles de données formelles ou informelles et l'identification des éléments de données pour la fonction de préservation, même si ses buts dépassaient souvent celle-ci, se focalisant sur les questions d'accès et de distribution. Parmi ces projets se trouvent le projet Networked European Deposit Library (NEDLIB) mené par la Bibliothèque Nationale des Pays-Bas et la Bibliothèque Nationale de France, le projet CURL Exemplars in Digital Libraries (CEDARS) du Royaume Uni, le projet Pandora de la Bibliothèque Nationale d'Australie, et diverses initiatives institutionnelles telles que celles entreprises par l'OCLC, l'expérience de la National Digital Library de la Bibliothèque du Congrès, etc.

Il est intéressant de remarquer que tous ces projets font référence à un moment ou à un autre à leur relation avec le modèle de référence du Système ouvert d'archivage d'information (Open Archival Information System, OAIS) (2), qui fut tout d'abord mis au point pour des données spatiales puis devint une norme ISO (ISO 14721). Le modèle OAIS eut un effet unificateur sur les recherches de ces dix dernières années, au moins pour fournir un langage évolué adapté à des discussions. Les paquets d'information de l'archivage, de la soumission et de la propagation (respectivement AIP, SIP, DIP) sont communément admis comme des éléments conceptuels de base dans l'implémentation d'entrepôts numériques. Ces paquets d'information sont composés de quatre parties en rapport avec l'Objet information traité : l'information contenue, l'empaquetage de l'information, la description de l'information et, ce qui nous intéresse, la préservation de l'information. En 2002, un projet sponsorisé par l'OCLC et RLG fit un excellent travail en rassemblant dans une seule structure les modèles et métadonnées mentionnés dans les projets ci-dessus et en les intégrant dans les concepts généraux du modèle de référence OAIS (3). Par conséquent, la première tâche du groupe de travail PREMIS était de reprendre ces résultats et de les traduire en un ensemble d'éléments de données implémentables par l'intermédiaire d'un dictionnaire de données.

Les objectifs

Le projet PREMIS était un groupe de travail multi-annuel dont l'intention était de travailler avec des institutions ayant une importante implantation mondiale. Des représentants d'Australie, de Nouvelle Zélande, des Etats-Unis, de Grande-Bretagne, des Pays-Bas et d'Allemagne y ont contribué de diverses façons, quelques-uns se levant tôt afin de participer aux conférences téléphoniques hebdomadaires. Le travail planifié pour une année nécessita deux ans de travail mais il en résulte un ensemble hautement perfectionné d'éléments qui peuvent servir de base à des implémentations.

L'effort poursuivait plusieurs objectifs connexes, tous pratiques et destinés à donner aux concepts une base d'implémentation. Ces objectifs originaux incluaient l'identification d'un ensemble « fondamental » de métadonnées et le développement d'un dictionnaire de

données pour ces métadonnées, les deux sont maintenant réalisés avec succès. L'expérimentation du dictionnaire de données sera le meilleur moyen d'articuler des stratégies alternatives de mise en oeuvre, le troisième objectif. Les objectifs finaux, tests pilote du dictionnaire de données et programmes de coopération fondés sur les éléments fondamentaux, doivent poursuivre le travail actuel.

L'enquête

Le travail commença par une enquête sur un certain nombre d'implémentations de projets numériques d'entrepôts afin d'identifier les pratiques actuelles et les tendances des projets numériques. 48 réponses provenant de 13 pays sont revenues, un bon taux pour un sujet en cours de développement. Les conclusions générales de l'enquête (4), qui ont servi à renseigner sur le travail ~~sur~~ le dictionnaire de données qui avançait en parallèle et suivait les progrès de l'enquête, peuvent être résumées ainsi :

- On remarque un usage largement répandu du modèle de référence OAIS pour la structure et comme point de départ de la conception du système de dépôts.
- Stocker des métadonnées de façon redondante dans les systèmes de dépôts est une pratique courante : dans une base de données en version X ML ou une base de données relationnelle pour une recherche rapide, une communication flexible et à l'aide du contenu même de l'objet pour une auto-définition et la préservation à long terme.
- On note une extension de l'utilisation de la norme de métadonnées pour l'encodage et la transmission (Metadata Encoding and transmission standard, modèle METS) pour l'encodage d'un éventail plus large de métadonnées nécessaires aux objets numériques, dont les métadonnées de préservation ; avec des MIX (Métadonnées pour les Images en XML) intégrées dans des METS pour les métadonnées techniques de l'image.
- La tendance actuelle est de garder l'original et aussi de stocker plusieurs versions normalisées et/ou migrées du contenu de l'objet, chacune avec les métadonnées qui lui sont rattachées.
- L'utilisation de multiples stratégies, à l'intérieur d'une même institution, est courante dans un domaine expérimental, en cours de développement, comme celui-ci.

De plus, l'enquête montra que plusieurs différences étaient faites pour les métadonnées se rapportant à différents types d'objets (les flux de bits, fichiers, collections, objets logiques, etc.) et que l'information indiquant les liens entre les objets était souvent enregistrée. Alors qu'une enquête instrumentale dans un domaine d'activités en développement comme celui-là n'est pas définitive, les résultats sont à la fois intéressants et utiles pour le travail sur le dictionnaire de données.

Le dictionnaire de données

Partant du précédent projet de structure (et indirectement de plusieurs projets majeurs de la dernière décennie) et de l'information apportée par l'enquête sur les entrepôts numériques, le dictionnaire de données des éléments fondamentaux fut alors développé par le groupe de travail PREMIS (5). Plusieurs décisions prises lors des premières étapes du projet sont importantes car elles ont des répercussions pratiques.

Les éléments de données *fondamentaux* étaient interprétés par le groupe de travail comme signifiant « les choses que la plupart des entrepôts opérationnels à vocation de préservation ont probablement besoin de savoir afin de tolérer une préservation numérique. »(6) Intentionnellement, le groupe ne traita pas de quelques aspects bien connus de la préservation, tels les métadonnées techniques détaillées des différents médias. Seules les métadonnées techniques généralement applicables à tous les formats de fichiers étaient retenues par le groupe de travail PREMIS.

Une autre considération importante prise en compte par le groupe de travail était que les métadonnées mentionnées devaient pouvoir, autant que possible, être fournies et utilisées automatiquement. Cela mena à une préférence pour des valeurs extraites de listes autorisées plutôt que des descriptions textuelles. Cela est également lié à l'intention du groupe de travail de permettre une implémentation indépendante du dictionnaire des données. Comme l'enquête l'a montré, des entrepôts sont déjà opérationnels et pour ceux qui sont en phase d'élaboration, l'environnement système dans lesquels ils évolueront peut avoir des caractéristiques spécifiques. Les éléments fondamentaux de PREMIS qui doivent être disponibles pour l'entrepôt ne sont pas forcément et explicitement stockés dans celui-ci. Les éléments pourraient être stockés dans des systèmes auxiliaires, pourraient être implicites dans les procédures utilisées par l'entrepôt, ou alors stockés à l'intérieur d'une base de données locale ou d'un format. Le point important est que les données fondamentales puissent être disponibles pour être converties dans une autre norme en cas d'échange. Ou que les données soient d'une manière prévisible disponibles pour une quelconque application informatique que l'entrepôt pourrait choisir et dont on attend que les données fondamentales PREMIS soient accessibles. Les systèmes n'ont pas besoin d'être ré-implémentés ou spécialement élaborés afin de maintenir les fondamentaux de PREMIS dans un certain format normalisé. Cela amena le groupe de travail à définir des « unités sémantiques » dans le dictionnaire de données plutôt que des « éléments de métadonnées ».

Modèle des données

Alors que l'article est trop court pour une description détaillée du modèle des données, il faut cependant noter quelques caractéristiques importantes. (Le modèle est bien expliqué dans sa totalité dans le rapport de PREMIS, voir la référence (5)).

Le modèle est *simple*. Il n'existe que 5 sortes d'entités : *Objets, Manifestations, Agents, Droits et l'Entité Intellectuelle* elle-même. Ce qui est au coeur de l'information et qui était inclus dans le dictionnaire des données fut attentivement examiné. Ainsi, par exemple, les métadonnées descriptives décrivant l'Entité Intellectuelle, qui peut être un livre, une carte, un site web, etc. sont laissées aux nombreuses normes telles MARC, MODS (metadata objet description standard, norme des métadonnées de description de l'objet), et DC (Dublin Core) qui existent déjà. De même, les données détaillées sur les Agents sont laissées aux formats MARC, MADS (metadata authority description standard, norme pour les métadonnées de description de l'auteur), vCard et autres normes. Les données concernant les Droits sont limitées à celles appartenant aux autorisations pour les activités de préservation, puisque les droits associés à l'accès et à la distribution de l'Objet ne sont pas au cœur des activités de préservation. Les métadonnées techniques détaillées ainsi que celles pour les médias et la documentation du matériel informatique sont exclues mais leurs spécificités sont laissées à l'initiative d'experts des formats.

Figure 1 : le modèle de base des données PREMIS

Les unités sémantiques des Objets, le concept central du modèle, peuvent être spécifiées à trois niveaux, fournissant une flexibilité à inclure de l'information à des niveaux adéquats pour le matériel et pour le fonctionnement d'un système de dépôts. Ces niveaux sont le flux de bits, qui est un composant du niveau supérieur, le *fichier* (ou le *flux de fichiers*). Un ensemble de fichiers nécessaires pour un rendu complet d'une Entité Intellectuelle constitue le plus haut niveau, la *représentation*.

L'entité Manifestation, qui renseigne les actions relatives à l'Objet, est une partie importante du modèle. Une grande variété d'actions affecte la préservation du matériau numérique dont la modification de l'Objet, les contrôles de validité et d'intégrité effectués, même les requêtes pour une propagation ou des rapports. Les Manifestations sont aussi fréquemment reliées aux liens, puisqu'une Manifestation dérivée produit un autre Objet et l'enregistrement du lien entre les Objets est généralement important à des fins de préservation. Le dictionnaire des données fournit plusieurs unités sémantiques de liens relatives à l'enregistrement d'informations sur les documents dérivés et les liens structurels, de dépendances, et autres liens.

Un aspect essentiel du modèle de données est ce que le groupe de travail appelle le principe 1 :1. Les nouveaux Objets créés à partir d'Objets existants (copies, versions, transformations, etc.) sont traités comme de nouveaux Objets et liés à l' « ancien » Objet par Manifestation et les informations relatives aux liens. Une des conclusions de l'enquête était que les entrepôts gardent fréquemment plusieurs copies d'un Objet, et que, à des fins de préservation, il est important que les données de chaque Objet soient complètes. Par conséquent, l'information sur les liens fournit la liaison sans diminuer ni compliquer l'enregistrement complet de l'information à des fins de préservation des documents dérivés. Alors qu'en interne un entrepôt peut construire des arbres de données afin de réduire la redondance des données, lors d'un échange, le système de dépôts doit pouvoir être capable de ressortir un Objet indépendant avec les métadonnées de préservation complètes.

Prochaine étape : la phase de test

L'envergure de PREMIS a été soigneusement étudiée. Une collaboration internationale a produit un dictionnaire des données de métadonnées ayant le potentiel de permettre un échange normalisé des informations de préservation des archives électroniques à l'aide de matériel numérique. Il n'impose pas d'architecture particulière pour les entrepôts mais fournit des conseils pour les métadonnées fondamentales à des fins de préservation. Bien que global quant à la participation, le projet PREMIS fut sponsorisé par l'OCLC et le RLG, la Bibliothèque du Congrès a pris la responsabilité du site web officiel pour la prochaine phase (7). Tous les documents du projet et son actualité peuvent être obtenus par l'intermédiaire de ce site.

Les objectifs finaux du projet, un banc d'essai du dictionnaire des données et une coopération construite autour de la question des métadonnées, peuvent dorénavant être planifiés. Récemment, un schéma XML a été élaboré pour les unités sémantiques identifiées dans le dictionnaire des données (8). Il a besoin d'être utilisé et évalué par de nouveaux projets et pour l'échange. Cependant, l'espoir est que des implémentations existantes d'entrepôts ou des projets en cours avec des architectures particulières participeront également au banc d'essai, en analysant leurs métadonnées, implicites et explicites, en les comparant aux unités sémantiques du dictionnaire des données. Pendant cette période, le dictionnaire des données et le schéma XML resteront stables mais sujets à des révisions de maintenance au fur et à mesure de l'expérience acquise par le banc d'essai.

Les autres parties du puzzle

Comme indiqué ci-dessus, d'autres parties nécessaires aux métadonnées de préservation des médias numériques existent mais ne furent pas définies par le groupe de travail B de PREMIS, par exemple les métadonnées de l'extension des droits et les métadonnées techniques détaillées, y compris l'information du format numérique.

Les métadonnées relatives aux droits

Les données concernant les droits sont étroitement définies dans le cadre PREMIS, et l'on pourrait argumenter que certaines informations relatives à l'accès et à la propagation sont importantes pour les fins de préservation. D'ailleurs, plusieurs initiatives s'intéressent au langage de l'expression des droits et aux questions de normes de messagerie en relation avec l'accès et la propagation. Le travail Indecs de l'Union européenne, les efforts d'ONIX de groupes d'éditeurs, et l'Initiative de la gestion des droits électroniques (the Electronic Rights Management Initiative, ERMI) de la Digital Library Federation (DLF) sont quelques-unes des recherches majeures.

Les métadonnées techniques

L'enquête du PREMIS a trouvé que beaucoup de systèmes de dépôts utilisaient des METS afin d'établir des lots de leurs métadonnées d'objets numériques et qu'il existait une variété de sortes et de quantités de métadonnées techniques maintenues, selon ce que l'entrepôt permettrait de collecter automatiquement. L'unique domaine où le travail sur les normes a significativement progressé est celui des métadonnées pour les ressources d'images. Un dictionnaire des données normalisées était mis au point dans le cadre de NISO avec une période d'essai en 2002. (9) D'ailleurs, les MIX, schéma d'extension des METS fondé sur le dictionnaire des données de NISO, sont déjà largement utilisées. (10) L'intérêt rapide suscité par cette norme et ce schéma indique que les systèmes de dépôts sont très intéressés par les normes et des conseils pour une information technique détaillée. Pour disposer de métadonnées techniques détaillées, la communauté des bibliothèques a besoin de collaborer à toutes les normes d'industries émergentes, ou au moins d'en prendre soigneusement note, étant donné que ce niveau de métadonnées nécessite d'être dérivable des objets B au-delà du niveau d'information PREMIS. Le site web des METS présente plusieurs schémas de métadonnées techniques développés localement pour différents types de matériel qui peuvent sans doute servir de point de départ à des efforts plus larges afin de développer des normes comparables à celles des données pour l'image. (11)

Enregistrements des formats

Une seconde suite à donner, potentiellement utile aux métadonnées de préservation, est de faciliter l'accès aux spécificités du format des données électroniques. Ce renseignement peut parfois être trouvé sur les sites web de sociétés responsables de divers formats de données, si un tel site existe, mais ce n'est pas une méthode efficace pour obtenir de l'information. Dans la perspective de préservation, la connaissance des formats de données aide à la validation des objets numériques insérés ou des contrôles d'intégrité, elle aide à évaluer le risque associé aux différents formats numériques, et elle précise les passerelles appropriées de migration des objets numériques. Une compréhension du format de fichiers peut aussi aider à déterminer les métadonnées qui pourraient être extraites de l'objet numérique, aidant ainsi à répandre PREMIS et les bases de données de métadonnées techniques détaillées.

Deux projets importants ont développé des répertoires collaboratifs sans cesse mis à jour, mais il n'est pas encore clair s'ils peuvent ou non être soutenus. Un projet est nommé PRONOM, des Archives Nationales du Royaume-Uni (12). Au départ, cet enregistrement était un outil élaboré au niveau local car les Archives Nationales avaient besoin de conseils pour la migration de documents de façon à remédier à l'obsolescence des applications informatiques. En 2004, il devint accessible sur le web, en 2005, une nouvelle version hautement améliorée est sortie. S'occupant surtout d'archives publiques, cet enregistrement a été particulièrement travaillé pour les formats d'applications informatiques orientées vers le texte.

Un second projet a atteint le stade de la preuve par l'application des concepts. C'est le Global Digital Format Registry (GDFR) qui débuta suite à une réunion sponsorisée par la DLF (13). Dès l'édition de ce modèle d'enregistrement par une équipe d'Havard, l'Université de Pennsylvanie développa un service prototype des formats nommé Format Registry Demonstration (FRED) (14). Par l'intermédiaire de FRED, les développeurs de systèmes de dépôts peuvent évaluer les possibles utilités de ce service, ses offres de service, la question de la maintenance, etc.

Ce domaine n'est pas prestigieux mais semble être important pour la préservation à travers tous les médias B. Un enregistrement commun serait efficace pour la communauté.

Conclusion

Pas à pas, se bâtissant à partir des modèles conceptuels passés et des expériences d'implémentations, des guides et des normes de métadonnées émergent et supportent les activités de préservation des systèmes de dépôts.
Les constructeurs de ces systèmes ne sont désormais plus obligés de partir d'une "feuille blanche". Aujourd'hui, l'ordre du jour des futurs développements est : le test des éléments fondamentaux de PREMIS, une attention portée sur les nécessités techniques détaillées, et la collaboration pour un format d'enregistrement de données.

Références

(1) Site web officiel PREMIS : <http://www.loc.gov/standards/premis>
(2) *Reference Model for an Open Archival Information System (OAIS)*. Washington, DC : Consultative Committee for Space Data Systems, 2002. (<http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>)
(3) *A Metadata Framework to Support the Preservation of Digital Objects*. Dublin, Ohio : OCLC Online Computer Library Center, 2002. (<http://www.oclc.org/research/projects/pmwg/pm_framework.pdf>)
(4) *Implementing Preservation Repositories for Digital Materials : Current Practice and Emerging Trends in the Cultural Heritage Community*. Dublin, Ohio : OCLC Online Computer Library Center, 2004. (<http://www.oclc.org/research/projects/pmwg/surveyreport.pdf>)
(5) Data dictionary for Preservation Metadata : Final Report of the PREMIS Working Group, May 2005. (<http://www.oclc.org/research/projects/pmwg/premis-final.pdf>)
(6) Ibid., p. ix.
(7) Site web officiel PREMIS : www.loc.gov/pre/
(8) Les schémas PREMIS peuvent être obtenus sur le site : <http://www.loc.gov/standards.premis/schemas.html>
(9) Data dictionary B Technical Metadata for Digital Still Images, NISO Z39.87-2002/AIIM 20-2002. (<http://www.niso.org/standards/resources/z39_87_trial_use.pdf>)
(10) Les MIX peuvent être obtenues sur le site : <http://www.loc.gov/mix>
(11) Voir <http://www.loc.gov/mets>
(12) Pour plus d'informations : <http://www.nationalarchives.gov.uk/pronom/>
(13) Pour plus d'informations : <http://hul.harvard.edu/gdfr/>
(14) Pour plus d'informations : <http://tom.library.upenn.edu/fred/>