Submitted on: 16/08/2015

Garantir la qualité des métadonnées du dépôt légal numérique dans un environnement en perpétuelle évolution

French translation of the original paper: "Ensuring metadata quality of e-legal deposit in an ever-changing environment".

Translated by: Emilie Liard, S.C.D. de l'Université de Poitiers, Poitiers, France

Le texte de ce document a été traduit de l'anglais, des différences avec l'original peuvent exister. Cette traduction vous est uniquement fournie à titre d'information

Stina Degerstedt
Support informatique et métadonnées, Département des systèmes d'information, Bibliothèque Nationale de Suède, Stockholm, Suède.
Adresse courriel: stina.degerstedt@kb.se

Joakim Philipson
Support informatique et métadonnées, Département des systèmes d'information, Bibliothèque Nationale de Suède, Stockholm, Suède.
Adresse courriel: joakim.philipson@kb.se

Résumé:

La récente loi suédoise sur le dépôt légal des documents numériques, entrée en vigueur au 1er janvier cette année, représente un défi exceptionnel pour la Bibliothèque Nationale de Suède (Kungliga biblioteket ou KB), ne serait-ce que par le volume de documents sur des supports très variés que l'on s'attend à recevoir de milliers d'éditeurs (fournisseurs), tels les organismes gouvernementaux, la presse en ligne, les maisons d'édition, etc. Ceci nécessite évidemment un niveau élevé d'automatisation dans le traitement des données, depuis leur intégration jusqu'à leur validation, en passant par leur conversion, enrichissement et stockage, tout en obtenant la meilleure qualité possible de métadonnées. Pour se préparer à ces défis, la KB a développé de nouveaux systèmes et flux de travaux électroniques qui seront exposés grosso modo dans cette contribution. Nous tenterons également ici d'évoquer quelques unes des questions soulevées en chemin concernant la qualité des métadonnées, les catalogues de bibliothèque et l'environnement en constante évolution.

Mots-clés: dépôt légal des documents numériques, dépôt légal numérique, Suède, métadonnées, automatisation, flux de travaux, workflow.

En résumé, la loi suédoise sur le dépôt légal des documents numériques

Depuis le 1er janvier cette année, la Loi Suédoise sur le Dépôt Légal des documents analogiques s'est enfin munie d'un complément applicable avec la Loi sur le Dépôt Légal des Documents Numériques. Les clauses de cette loi sont contraignantes pour les fournisseurs individuels – agents, producteurs, éditeurs, distributeurs, autorités fédérales et municipales sous égide du gouvernement, organismes gouvernementaux (bien que ces derniers soient soumis à des règles légèrement différentes).

Sont soumis au dépôt légal les documents numériques accessibles au public par transfert sur le réseau, fichiers individuels, complets et pérennes, conçus pour un affichage identique à chaque ouverture. Le contenu des fichiers peut être de tout type et combiner texte, son, et image. On peut citer comme exemple les articles de presse constitués d'image et de texte, extraits sonores et clips vidéos, émissions entières de télévision et services en ligne, fichiers audio et vidéos, images, podcasts, brochures, rapports et livres numériques, etc. Des exemples de documents non soumis au dépôt légal incluent les sites Web ou bases de données complets, programmes informatiques et autres logiciels, flux, ressources continuellement mises à jour (par exemple les wikis), ressources publiées à titre privé (images, music, films, blogs), contenus des intranets d'entreprises, jeux en ligne, etc.

Conformément à la loi, tout document numérique doit être fourni dans sa forme originelle sur un support de données. Chacun doit s'accompagner d'informations sur le lieu et la date d'édition, le format du document, les identifiants nécessaires à son accès, ainsi que des informations sur les relations entre le document déposé et d'autres documents soumis au dépôt légal. On trouvera des renseignements complémentaires en anglais dans les deux plaquettes suivantes :

> Legal deposits of electronic materials in Sweden. For individual suppliers.
> Legal deposits of electronic materials in Sweden. For government agencies.

Précisons que le moissonnage des sites Web suédois est assuré par la KB depuis 1997, mais la nouvelle loi permet de combler des lacunes inévitables dès lors que le moissonnage ne s'effectue qu'en surface et deux fois par an.

Des défis et opportunités pour commencer

La KB s'est beaucoup impliquée dans le travail préparatoire à la loi et dans sa mise en œuvre. Nous pouvons désormais nous consacrer à devenir la mémoire de la nation jusque dans le champs numérique. Néanmoins, cela représente un défi exceptionnel de par le volume même de documents, et les nombreux types de médias que l'on s'attend à recevoir de milliers de fournisseurs.

Du fait que chaque document électronique doit être fourni sur support physique, que les contraintes sur les métadonnées sont minimales et que le choix du format de fichier ne peut être imposé, nous nous attendions à devoir traiter une montagne grandissante de clés USB, emplies de fichiers de données non-identifiables et de métadonnées inexploitables. Cependant, le législateur était bien conscient que la plupart des fournisseurs du dépôt légal

numérique préféreraient des moyens de dépôt plus pratiques que la clé USB. Par conséquent, un amendement par décret nous autorise à convenir d'autres modalités de dépôt, et nous donne par ailleurs la possibilité d'exiger des métadonnées supplémentaires. L'effort porte en grande partie sur la recherche de modalités de dépôt adéquates, ainsi que sur des normes de métadonnées qui soient à la fois simples à appliquer et adoptées largement par les fournisseurs comme formats de contenu, tout en étant suffisamment élaborées pour véhiculer toutes les métadonnées nécessaires aux dépôts.

Un fort degré d'automatisation s'avère indispensable pour gérer tout cela. La KB a conçu une nouvelle plate-forme technique dont le cœur se situe dans le système de conservation numérique appelé « Mimer ». Ce système gère tout depuis l'intégration jusqu'à la validation des données, en passant par leur conversion, enrichissement et stockage, ainsi que la création de notices de métadonnées dans LIBRIS, le catalogue collectif national.

Parallèlement à ces développements techniques, un nouveau métier a émergé : les « administrateurs numériques ». Ils entretiennent le contact avec les fournisseurs, assistés des personnels informaticiens, experts métadonnées, experts juridiques et autres à la KB. Tous les personnels travaillant sur le dépôt légal numérique, quelle que soit leur position dans le flux de travaux, et notamment les catalogueurs, doivent acquérir de plus amples connaissances sur le fonctionnement actuel de l'industrie de l'édition numérique et sur son développement.

Modes de dépôts et normes de métadonnées

Actuellement, la KB propose quatre solutions de dépôt en ligne : les flux RSS, les protocoles FTP et OAI-PMH, et le chargement via un formulaire en ligne. Le choix par un fournisseur potentiel d'utiliser l'un ou l'autre des modes de dépôt légal numérique l'engage implicitement auprès de la KB à respecter les spécifications associées à une certaine norme de métadonnées, qui peuvent être plus nombreuses que les seuls éléments exigés par la loi.

Flux RSS

Les flux RSS sont implémentés à la fois comme moyen de dépôt et format de métadonnées privilégié pour les flux d'actualité, mais peuvent être utilisés pour n'importe quel type de publication. Les flux RSS sont moissonnés à intervalle régulier sur les sites Web des fournisseurs d'actualité (journaux en ligne, stations radio et chaînes de télévision, etc.), ou tout autre fournisseur, confronté pour validation à notre schéma XML adapté, et « éclaté » en éléments isolés avant traitement subséquent (voir le paragraphe « Mimer, normalisation des métadonnées et AIPs »).

RSS 2.0 en particulier est conçu comme une norme très simple et facile d'utilisation, avec très peu d'éléments et attributs obligatoires. Pour pallier certaines de ses limites, nous lui avons adjoint quelques éléments obligatoires de MediaRSS et Dublin Core (dcterms) dans notre implémentation des spécifications RSS pour le dépôt légal numérique. Au final, il y a sept éléments incontournables et 3 éléments qui ne sont obligatoires que si le cas s'y prête, par exemple titre, identifiant, adresse internet (url), date d'édition, éditeur, modalités d'accès au moment de la publication, et format du fichier. Dans nos spécifications, parmi les éléments de métadonnées optionnels susceptibles d'apporter des informations complémentaires, mais sur lesquels on ne peut donc compter systématiquement, on trouve créateur, contributeur, mots clés, autorités matières, etc. Nous avons considéré ces derniers comme moins importants pour des flux d'actualité, et comme faisant partie du prix à payer pour que le processus reste simple, pour le bénéfice conjoint des éditeurs et de la KB. Les spécifications

complètes pour les flux RSS peuvent être consultés à l'adresse suivante : http://www.kb.se/namespace/digark/deliveryspecification/deposit/rssfeeds/

FTP ou OAI-PMH

Le FTP est le moyen privilégié pour les éditeurs qui ont des gros fichiers à déposer (par exemple les sociétés de médias audiovisuels), et le dépôt via OAI-PMH est une pratique déjà bien répandue au sein des universités et instituts suédois. Pour le moment nous n'avons qu'un type de spécifications de métadonnées associé à ces méthodes : les « Spécifications communes pour le dépôt des publications électroniques isolées ». Ces FGS-PUBL sont l'une des nombreuses spécifications disponibles pour la transmission d'information, développées conjointement par les Archives Nationales Suédoises et d'autres archives publiques.

Dans toute spécification FGS, le format de métadonnées par défaut est le METS, Metadata Encoding and Transmission Standard. METS est un format qui permet d'exprimer des métadonnées bibliographiques, de gestion, ainsi que des métadonnées structurelles et de conservation. Dans les spécifications FGS-PUBL, du MODS (Metadata Online Description Schema) est ajouté à la partie bibliographique du METS (dmdSec).

Avec des dépôts FTP et OAI-PMH, nous devrions recevoir de bien meilleures métadonnées, même si par défaut les exigences minimales sont sensiblement identiques à celle des flux RSS. Dans la mesure où METS et MODS sont des formats de métadonnées plus expressifs, les spécifications FGS-PUBL sont plus adaptées pour les éditeurs qui souhaitent que leur production apparaissent dans le catalogue national, et qui par conséquent soumettent des descriptions plus précises que dans le cas du RSS. Les spécifications complètes pour FGS-PUBL sont disponibles à l'adresse suivante : http://www.kb.se/namespace/digark/deliveryspecification/deposit/fgs-publ/

Chargement depuis un formulaire en ligne

Un formulaire en ligne est disponible sur le site Web de la KB pour les « petits » éditeurs qui ne publieraient que quelques titres par an. Pour chaque édition, le fournisseur complète à la main les métadonnées dans un formulaire et télécharge le ou les fichier(s). Les données sont produites dans le système en METS et MODS conformément aux spécifications FGS PUBL, ce qui reste transparent pour le fournisseur.

Mimer, normalisation des métadonnées et AIPs

Mimer est une archive en ligne destinée à la réception et au stockage du dépôt légal numérique et autres collections numériques de la Bibliothèque Nationale de Suède (KB). L'architecture de Mimer est conforme à l'OAIS, norme pour l'Open Archival Information System (Système ouvert d'archivage d'information). Les opérations gérées par Mimer incluent l'intégration de fichiers de données et métadonnées, la validation des dépôts en fonction des schémas, la vérification de la présence d'anciennes versions d'un fichier déjà intégré, la normalisation et l'enrichissement des métadonnées réceptionnées (à la fois les métadonnées bibliographiques et administratives). En résulte la création d'un AIP, Archival Information Package (objet d'archivage), avec une notice de métadonnées plus adaptée à l'accès et à la conservation futurs. Les métadonnées d'origine livrées par le fournisseur sont toujours conservées avec l'AIP dans la base d'archivage, pour référence ultérieure.

La normalisation consiste à convertir les métadonnées d'origine vers le format canonique de métadonnées d'archive, commun à tous les AIPs, sans tenir compte du moyen de dépôt utilisé ou du format d'origine des métadonnées. Cela inclut par ailleurs l'enrichissement des métadonnées depuis des sources externes, notamment l'ajout de métadonnées de gestion et métadonnées techniques pour la conservation. On compte essentiellement quatre sources de données pour cette opération : 1) les métadonnées d'origine ; 2) le « registre des fournisseurs » qui contient des informations sur chaque fournisseur du dépôt légal numérique, ainsi que les éditeurs auxquels ils servent d'intermédiaire ; 3) une sorte de « notice de canal » dans le catalogue national LIBRIS qui contient des informations bibliographiques (censées être) communes à tous les documents déposés via un canal particulier (une URL ou un compte FTP), par exemple le genre, la langue, l'URL de l'hébergeur ; 4) des données sur le fichier glanées parmi les fichiers de données directement liées à un document, par exemple le type MIME, le nom du format, la clé du format, et la version du format, éléments indispensables dans le cadre de la conservation. De plus, des informations sont ajoutées sur la taille des fichiers et leur empreinte numérique est vérifiée grâce à l'algorithme MD5.

Comme format d'archivage dans Mimer, nous avons choisi les normes METS-MODS, comme pour le format de dépôt FGS-PUBL décrit précédemment. Des informations techniques sur chaque fichier de données, ainsi que les métadonnées concernant les opérations (événements) effectuées au cours du processus d'archivage sont conservées en PREMIS, la norme de métadonnées dédiée à la conservation.

Ce que nous avons obtenu ici est un système stable qui, indépendamment des modalités de dépôt ou du type de document, archive tout de la même manière, et qui peut être développé au rythme de l'accélération des changements extérieurs.

Notices bibliographiques, dédoublonnage et contrôle de version

A partir des métadonnées normalisées dans l'AIP, une notice est également créée dans le catalogue collectif national LIBRIS. Cette opération est actuellement entièrement automatisée, par conversion du format de métadonnées interne AIP dans Mimer via MARCXML vers MARC21.

Avant qu'une notice ne soit créée, Mimer effectue une recherche pour éviter de générer un doublon. Si une notice existe déjà pour la ressource concernée, seul un nouvel exemplaire est ajouté à la notice bibliographique existante. Un cas similaire, et plus fréquent, se présente lorsque nous recevons des mises à jour, c'est à dire des nouvelles versions d'un document paru précédemment et déjà intégré dans Mimer. Il s'agit bien entendu d'un phénomène très courant dans l'édition en ligne, dans la mesure où l'actualité s'enrichit des événements en cours, parfois heure par heure, minute par minute. Naturellement, nous ne voulons pas d'une nouvelle notice de catalogue de bibliothèque à chaque mise à jour d'un document. Dans les deux cas, doublons et mises à jour, nous dépendons de la bonne volonté du fournisseur/éditeur à utiliser un identifiant unique et raisonnablement pérenne pour chaque ressource. Deux exemples de notices dans LIBRIS créées automatiquement à partir du dépôt numérique :

1. A partir d'un flux RSS http://libris.kb.se/bib/17370222
2. A partir d'un FTP et des spécifications FGS-PUBL (métadonnées METS et MODS) http://libris.kb.se/bib/17564144

Pour certains organismes gouvernementaux, institutions de gestion du patrimoine culturel et bibliothèques universitaires, par exemple, naît la perspective de voir remplacés des documents exigeant précédemment un catalogage manuel interne par des notices de catalogue produites automatiquement par le système comme produit dérivé du dépôt légal numérique. Mais cela n'est envisageable qu'avec au départ un effort substantiel du côté du fournisseur pour mettre en place et maintenir leur système dans le cadre du dépôt. Grâce aux fournisseurs et éditeurs prêts à faire cet effort supplémentaire, on peut obtenir des notices bibliographiques de qualité assez élevée dans le cadre du dépôt par FTP et OAI-PMH, comme nous pouvons le voir dans l'exemple 2 ci-dessus.

Un des effets secondaires de la nouvelle loi est le risque non-négligeable d'inonder complètement le catalogue collectif national LIBRIS de notices pour chaque nouveau document publié, au point que toute autre notice de document se verra plus ou moins noyée dans une mer de brèves. Chaque jour, Mimer reçoit environ 6000 paquets par le dépôt légal numérique. Partant de l'hypothèse plutôt prudente que seulement la moitié de ceux-ci donneront lieu à de nouvelles notices dans le catalogue de bibliothèque (en raison du dédoublonnage, du contrôle de versions, etc.), cela représente toujours environ 3000 nouvelles notices produites quotidiennement du fait du seul dépôt légal numérique. C'est l'une des raisons pour lesquelles il a été décidé de simplement supprimer les notices d'articles en ligne et journaux créées par Mimer lors de l'affichage des résultats de recherche dans l'interface web de LIBRIS.

Développements futurs et orientations choisies

Notre modèle de gestion du dépôt légal des documents numériques est constamment en développement. Dans un avenir proche, nous envisageons de proposer d'autres modes de dépôt et d'autres formats de métadonnées tels Atom, DDEX, RDFa, ONIX.

Pour améliorer encore la qualité de notre catalogue de bibliothèque, d'autres méthodes d'enrichissement des métadonnées doivent être intégrées. Des outils seront également nécessaires pour faciliter le traitement manuel postérieur, comme l'appariement des métadonnées entrantes avec les données d'autorité présentes dans le catalogue de bibliothèque (par exemple les noms de personne, collectivités et lieux).

L'un des gros changements à venir est le remodelage du système de LIBRIS en quelque chose de tout à fait nouveau. Ce nouveau système, basé sur les données liées (linked data) et bannissant les formats MARC, nécessitera à terme de développer des nouveaux schémas de conversion, de même qu'il ouvrira des possibilités pour mettre en valeur les données du dépôt numérique, au bénéfice de l'usager final.

Quelques questions bibliographiques

La différence de niveau bibliographique – quel est le niveau minimum de métadonnées acceptable ? Une partie des métadonnées déposées peut paraître étrange à l’œil du catalogueur, mais sont-elles d'une qualité suffisante pour être utiles à l'usager final ?

Nous recevons des quantités impressionnantes de documents qui ne sont pas habituellement décrits dans le catalogue, comme les articles en ligne, les journaux, des séquences vidéo, etc. Quelle proportion de ces documents devrait être exposée dans le catalogue, et de quelle manière ?

La même question se pose pour les bibliographies nationales : que devriez-vous trouver (ou non) dans une bibliographie nationale ? Tous les types de ressources produites dans un pays, ou seulement les documents textuels ?

Ce n'est qu'une partie des questions qui sont survenues concernant les flux de travaux automatisés, la qualité des métadonnées et l'activité bibliographique traditionnelle. Il n'y a peut-être aucune réponse évidente à ces questions, ni à celles qui se poseront à l'avenir. Une chose est néanmoins certaine, c'est que l'édition numérique et la loi sur le dépôt légal des documents numériques apportent des changements radicaux dans notre monde bibliographique, et la KB se doit d'au moins tenter de suivre ces changements.