Submitted on: 17.06.2015

Ensuring metadata quality of e-legal deposit in an ever-changing environment

Stina Degerstedt
Metadata and Systems Support, Department of Information Systems, National Library of Sweden, Stockholm, Sweden.
E-mail address: stina.degerstedt@kb.se

Joakim Philipson
Metadata and Systems Support, Department of Information Systems, National Library of Sweden, Stockholm, Sweden.
E-mail address: Joakim.philipson@kb.se

Abstract:

The new Swedish Law on legal deposit of electronic documents, fully effective from January 1 this year, poses an exceptional challenge for the National Library of Sweden (NLS) by the sheer volume of documents of a wide variety of media types expected to be delivered by thousands of publishers (suppliers), such as government agencies, online news media, publishing houses etc. This naturally requires a high level of automation in the data processing, from ingest to validation, transformation, enrichment and storage, while at the same time attaining best possible metadata quality. To meet the challenges thus encountered the NLS has developed new electronic systems and workflows which will be roughly explained in this paper. We also try to touch on some of the issues that appear on the road regarding metadata quality, library catalogs and the ever-changing environment.

Keywords: E-deposit, Sweden, Metadata, Automation, Workflows.

The Swedish e-deposit law in brief

Since January 1 this year the Swedish Legal Deposit Act for analog material finally got a fully effective complement – The Legal Deposit Act for Electronic Materials. The stipulations of this act are binding for individual suppliers – agents, producers, publishers, distributors, federal and municipal authorities under governmental auspices, and government agencies (although the latter are subject to slightly different regulations).

Subject to legal deposit are electronic materials that have been made available to the public in Sweden by transfer via network, single files, completed and of permanent nature, and intended to be presented in the same way at each viewing. The content of the files may be of any nature or combination of text, sound and image. Some examples are news articles consisting of text and image, sound bites and video clips, entire programs from web television and play services, music and film files, images, podcast, brochures, reports and e-books, etc. Examples of material not subject to legal deposit are entire websites or databases, computer programs and other software, live streams, materials updated continually (for example, wikis), privately published images, music, films and blogs, content of the internal networks of companies, online games, etc.

According to the law each electronic item shall be delivered on a data carrier in its original form. Each item shall be accompanied by information about the place and time of publication, about the material's format, passwords necessary to access these materials, as well as information on how this particular material is related to other material subject to Legal Deposit. More information in English can be found in the following 2 brochures:

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

It should be mentioned that harvesting of Swedish webs sites has been carried out by the NLS since 1997, but the new law makes it possible to cover the gaps that arise when harvesting only skims the surface a couple of times a year.

The challenges and opportunities to begin with

The NLS has worked actively in the preparations for the law and for it to be introduced. Now we can seriously work toward becoming the nation's memory even in the digital field. Nevertheless, it poses an exceptional challenge just by the sheer volume of documents, and the wide variety of media types expected to be delivered by thousands of suppliers.

The fact that each electronic item shall be delivered on a physical carrier, requirements for metadata are very small and choosing file format is not possible, meant that we saw before us a growing mountain of USB sticks to take care of, filled with unidentifiable data files and inadequate metadata. However, the legislator was well aware of the fact that most suppliers of e-legal deposit would prefer more convenient methods of delivery than USB-sticks. So, according to an accompanying decree, we are accorded the right to decide about other possible methods of delivery, and with that also come the possibility to require more metadata. A great deal of effort is put on finding suitable delivery methods, as well as metadata standards that are both simple enough to use and widely adopted by suppliers as container formats, at the same time being expressive enough to carry all the metadata needed for the deliveries.

A high level of automation is required to handle all this. NLS has built a new technical platform with its central point in the digital preservation system called "Mimer". The system handles everything from ingest to validation, transformation and enrichment of metadata, storage as well as creating metadata records in LIBRIS, the national union catalog.

In addition to this technical development, a new profession has emerged - "e-administrators". These maintain contact with the suppliers, supported by IT staff, metadata experts, legal experts and others at NLS. All staff working with e-deposit, whatever position in the workflow, including catalogers, needs to acquire more knowledge about how the publishing business on the web works today and how it develops.

Delivery methods and metadata standards

Presently, the NLS provides four different methods of online delivery, i.e. RSS feeds, FTP, OAI-PMH and, upload via a web form. The decision by a prospective supplier to use either of these methods for e-legal deposit is considered to imply an agreement with the NLS to follow also a certain metadata standard specification, which may be more extensive than the limited requirements for metadata by the law.

RSS Feeds

RSS feed is implemented both as a method of delivery and as the preferred metadata format for news feeds, however it may be used for any type of publication. RSS feeds are harvested at regular intervals from the web sites of the news providers (online newspapers, radio- and TV-stations etc.), or any other supplier, validated against our adapted xml-schema and "split up" into single items before further processing (see more under Mimer, normalization of metadata and AIPs).

RSS 2.0 in particular is designed to be a very simple and easy to use standard, with very few mandatory elements or attributes. To overcome some of the limitations we added certain elements from MediaRSS and Dublin Core (dcterms) as mandatory in our implementation of the RSS specification for e-legal deposit. All in all there are seven unconditionally mandatory elements and further three that are mandatory if applicable, for example title, identifier, internet address (url), publishing date, publisher, accessibility at the time of publishing and file format. Among the optional metadata elements in our specification, providing information that we thus cannot automatically count on are creator, contributor, keywords, subject headings, etc. These have been considered to be less important for news feeds and part of a price to pay for keeping it as simple as possible, for the mutual benefit of publishers and the NLS. The full specification for RSS feeds can be found at the following address: http://www.kb.se/namespace/digark/deliveryspecification/deposit/rssfeeds/

FTP or OAI-PMH

FTP is the preferable delivery method for suppliers with large files (for example media broadcasting companies) and to deliver via OAI-PMH is an already well established practice among Swedish universities and colleges. So far we have only one metadata specification that works with these methods - the "Common Specification for deposit of single electronic publications (FGS-PUBL)". This FGS-PUBL is one of many forthcoming specifications for information packaging, developed in cooperation with the Swedish National Archive and other public Swedish archives.

Default metadata standard in all FGS specifications is METS, Metadata Encoding and Transmission Standard. METS is a container format housing bibliographic, administrative as well as preservation and structural metadata. In the FGS-PUBL specification MODS

(Metadata Online Description Schema) is added for the bibliographic part of METS (dmdSec).

With FTP and OAI-PMH deliveries we have a good chance to get much better metadata even though the minimum default requirements are approximately the same as for RSS feeds. Since METS and MODS are more expressive metadata formats the FGS-PUBL specification is more suitable for those suppliers who want their publishing appear in the national catalog and therefore voluntarily submit more qualified descriptions than in the RSS case. The full specification for FGS-PUBL can be found at:

http://www.kb.se/namespace/digark/deliveryspecification/deposit/fgs-publ/

Upload via a web form

A web form has been posted on the NLS website for "small" publishers who might give out a few titles a year. For each publication, the supplier manually fills in a metadata form and uploads the file or files. A delivery package in METS and MODS according to the FGS PUBL specification is created in the background, something the supplier need not care about.

Mimer, normalization of metadata and AIPs

Mimer is an electronic archive for ingest and storage of e-legal deposit and other digital collections at the National Library of Sweden (NLS). The architecture of Mimer follows the OAIS, the Open Archival Information System standard. The processes of Mimer include ingest of data files and metadata, validation of deliveries against schemas, checking for version of an already ingested item, normalization and enrichment of ingested metadata (both bibliographic and administrative metadata). The result is the creation of an AIP, an Archival Information Package, with a metadata record better aimed to serve purposes of future access and preservation. The original metadata from the supplier is always stored together with the AIP in the archive for reference.

Normalization is the transformation of the original supplied metadata to the canonical archival metadata format common to all AIPs, irrespective of delivery method or original metadata format. It also involves enrichment of metadata from external sources, including the addition of administrative and technical metadata for preservation. There are essentially four different data sources for this transformation: (i) The original supplied metadata ; (ii) The "Supplier registry" which holds information about each and all suppliers of e-legal deposit and the publishers that they are serving as delivering "carriers" for; (iii) A so-called "channel record" in the national catalog LIBRIS which holds some bibliographic information that is (expected to be) common to all items delivered on a particular channel (a URL or a FTP account), for example genre, language, host url; (iv) File data gleaned from the actual data files belonging to an item, such as MIME-type, format name, format key and format version; elements that are vital for preservation purposes. Further, information about file sizes is added and fixity checks using MD5 check sums.

As archival format in Mimer we have chosen the METS-MODS standards, the same as for the delivering format FGS-PUBL described above. Technical information about each data file together with metadata about actions (events) performed during the archiving process is stored in PREMIS, the metadata standard for preservation purposes.

What we have achieved here is a stable system that regardless of delivery method or material types archives everything the same way and that can be developed in pace with external accelerating change.

Catalog records, deduplication and version control

From the normalized metadata in the AIP a record is also created in the national library union catalog LIBRIS. Today this is done by an entirely automated transformation from internal AIP metadata format in Mimer via MARCXML to MARC21.

Before a record is created Mimer performs a search in order to avoid creating a duplicate. If a record of the resource searched for already exists only a holding is added to the preexisting bibliographic record. A similar, and more common case, is when we get updates, i.e. new versions of earlier published items already ingested to Mimer. This is of course a regular phenomenon in online publishing, as news stories grow with the events unfolding sometimes hour by hour, minute by minute. Naturally, we do not want to create a new library catalog record with every update of an item. In both cases, duplicates and new versions, we are dependent on the supplier/publisher to use unique and fairly persistent identifiers for every resource. Two examples of catalog records in LIBRIS automatically created from e-deposit ingest:

1. From an RSS feed, http://libris.kb.se/bib/17370222
2. From FTP and the FGS-PUBL specification (MODS and METS metadata), http://libris.kb.se/bib/17564144

For some government agencies, cultural heritage institutions and university libraries, for example, there is the prospect of replacing earlier resource demanding in-house manual library cataloging with automated, machine generated catalog records as a byproduct of e-legal deposit. But this does not come without a substantial initial effort on the supplier side in setting up and managing their system for delivery. For those suppliers and publishers willing to make that extra effort, a fairly high quality of library catalog records thus produced may be attainable for delivery by means of FTP and OAI-PMH as we can see in example 2 above.

One side effect of the new legislation is the substantial risk of completely inundating the library union catalog LIBRIS with records for every news item ever published, to the point where other document records will be more or less submerged in a sea of news bites. Every day Mimer receives in the order of 6000 e-legal deposit packages. Assuming as a rather cautious hypothesis that only half of these result in new library catalog records (due to deduplication, version control, etc.), this would still mean about 3000 new catalog records daily produced solely by e-legal deposit. That is one reason why it was decided to simply suppress all web articles and newspaper issue records created by Mimer from display in the web search interface of LIBRIS.

Future developments or decisions

Our model for taking care of e-legal deposit is continuously being developed. In the near future we consider offering other methods of delivery and metadata formats as well, such as Atom, DDEX, RDFa, ONIX.

To further improve the quality of our library catalog more methods for metadata enrichment need to be introduced. There will also be a need for tools to facilitate manual

post-processing such as matching incoming metadata with already existing authority data in the library catalog (for example names of people, institutions and places).

One big change yet to come is the reshaping of the LIBRIS system into something entirely new. This new system, based on linked data and discarding the MARC format, will eventually require new transformation schemes as well as it will involve other possibilities to showcase data about e-deposit, which will benefit the end users.

Some bibliographical issues

Difference in bibliographic level - what is the lowest minimum metadata level accepted? Some of the metadata delivered may look strange to the eye of a cataloger, but is it qualified enough to be useful for the end user?

We receive an enormous amount of material which is usually not described in the catalog, e.g. web articles, journal issues, movie clips, etc. How much of it should be exposed in the online catalog, and in what way?

The same question applies to national bibliographies- what should you (or not) be able to find in a national bibliography? All kinds of resources published in a country, or only text material?

These are just a few questions that have arisen concerning automated workflows, metadata quality and traditional bibliographic activities. There are perhaps no obvious answers to these questions, or the questions that will come. One thing is certain, however, electronic publishing and e-deposit legislation is bringing fundamentally change into our bibliographic world and the NLS is obliged to at least try to keep up with those changes.