2015 IFLA International News Media Conference

"Transformation of the online news media: implications for preservation and access"

Sub-theme: Issues of e-legal deposit: online news media

PAPER

Online news media in Mimer:

the challenge of metadata quality in e-legal deposit in Sweden /

Joakim Philipson (joakim.philipson@kb.se), Stina Degerstedt (stina.degerstedt@kb.se)

Introduction

Mimer is an electronic archive for ingest and storage of e-legal deposit and other digital collections at the National Library of Sweden (NLS), handling a wide variety of media types. The architecture of Mimer follows the OAIS, the Open Archival Information System standard. A considerable part of the e-legal deposit material handled by Mimer comes from online news media.

At the National Library of Sweden we have until now implemented RSS both as the method of delivery and, in combination with MediaRSS and Dublin Core (dcterms), as the preferred metadata format for news feeds. RSS feeds are harvested at regular intervals from the web sites of the news providers (online newspapers, radio- and TV-stations etc.), validated against our adapted xml-schema, "split up" into single items, and together with the associated media files subsequently "repackaged" into a SIP, a Submission Information Package.
Each SIP is then processed separately. This processing includes new validation, checking for version of an already ingested item, normalization and enrichment of both bibliographic (MODS) and administrative metadata (PREMIS). 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. Original RSS metadata from the SIP is always stored together with the AIP in the archive for reference. From the normalized metadata in the AIP a record is also created in the national library union catalog LIBRIS.

Particular to online news media is that they often involve multiple media types, e.g. text, images, video clips, audio files, while the original metadata provided is rather poor. This poses particular challenges for the possible enrichment of metadata in the data processing and also for the handling of files of different media types in the same metadata record. But these are not the only stumbling blocks encountered in the struggle to ensure highest possible metadata quality for e-legal deposit online news media.

Stumbling blocks

1. Negative expectations
2. The Law - what can we legally ask for?
3. Positive expectations – too much, too soon!
4. The Publishers
5. Format / metadata standard constraints
6. Deduplication and version control
7. Inundation

1. Negative expectations

When the new law on e-legal deposit was still in preparation, among those most critical of the proposed law was the Swedish Media Publishers' Association ("Tidningsutgivarna", TU). Their argument was that the proposition was too technically complicated and too resource demanding, and thereby not feasible either economically or practically.[1] They suggested as an alternative method for the NLS to collect the content of Swedish online news media by means of a web crawler or robot harvesting web pages. Harvesting the Swedish web by means of a robot was something that was in fact being implemented by the NLS already since 1997. But the legislator did not see this as sufficient. One argument was that a good part of the electronic material intended to be covered by the new law is protected by passwords or by other means, in such a way that would prevent access for a web robot. Another argument, found in the committee report preceding the law, is that the possible metadata gathered by a web robot would be insufficient for secure preservation and future retrieval. Whatever the possible strengths and weaknesses of the arguments pro and con different methods of effecting e-legal deposit, the resulting final bill submitted and passed by the Swedish parliament on June 20, 2012 was fairly modest in this respect. The legislator was sufficiently cautious not to demand too much of the prospective providers of e-legal deposit, among whom, to be sure, were also online news media publishers. This cautiousness on the part of the legislator affected in particular the requirements on metadata following the actual deposits.

[1] Regeringens proposition 2011/12:121 Leveransplikt för elektroniskt material. http://www.regeringen.se/content/1/c6/18/94/43/0abb8a0c.pdf, p. 14

2. The Law – what can we legally ask for?

The NLS' interpretation of the Swedish Law (2012:492) on e-legal deposit[2] presumes, at least implicitly, the default method for submission of electronic material to be a physical carrier, such as a USB-stick. This is a consequence of the requirement that anyone who is a prospective supplier subject to this law must be able to deliver electronic material "on a data carrier in the logical format in which it was made available on a network" (§10), assuming that there might be some prospective suppliers / publishers lacking the technical means for automated, electronic transfer of files and metadata. Confident that deposit by means of physical carriers will in itself be cumbersome and not the most effective way of handling deliveries for most suppliers, by contrast, the requirements on metadata following the electronic material deposited are kept to a bare minimum of information about net address (URL), access condition and time of publishing. In case a deposited electronic resource is associated with or related to another resource that is subject to legal deposit (electronic or analog), there is also a requirement of information about this relationship. Further, there is a legal demand for information about files being part of a deposited electronic resource: filenames/-identifiers, file formats and, if applicable, encryption keys or passwords needed for access. That's about it.

However, since the legislator was probably well aware of the fact that most suppliers of e-legal deposit, as well as the NLS itself, would prefer more convenient methods of delivery than USB-sticks, according to an accompanying decree[3], NLS is accorded the right to decide about other possible methods of delivery of e-legal deposit material, viz. electronic file transfer over a network. Presently, the NLS provides for four different methods of online delivery, i.e. upload via a web form, FTP, OAI-PMH and, thus, RSS. 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.

On the other hand, it is in the self-interest of the NLS not to impose too strong requirements for metadata on prospective suppliers choosing electronic network delivery methods such as these, in order not to deter them to the point of resorting to the default physical carrier instead. Thus, to minimize this risk, our different metadata format specifications have to strike a balance of interests between the requirements of the law, the metadata standard itself, a sufficiently rich metadata quality to support preservation and future access, and imposing a fair enough work load on the part of the supplier. Striking this balance is something that cannot be fixed once and for all; it is an ever ongoing process. With this in mind we will soon take a look at our just recently updated RSS specification.

[2] Lag (2012:492) om pliktexemplar av elektroniskt material, http://www.riksdagen.se/sv/Dokument-Lagar/Lagar/Svenskforfattningssamling/sfs_sfs-2012-492/

[3] Förordning (2012:866) om pliktexemplar av elektroniskt material, http://www.riksdagen.se/sv/Dokument-Lagar/Lagar/Svenskforfattningssamling/Svensk-forfattningssamling-201_sfs-2012-866/

But first, we will just briefly stumble upon some of the other hurdles on our way to ensure a good enough metadata quality for our electronic archive, Mimer.

3. Positive expectations: too much, too soon, without much effort

While some of the prospective suppliers, among them notably online news media publishers, expected the worst from the new law on e-legal deposit, others, - not only suppliers, but also libraries and end users, may have expected too much, too soon. 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. Such expectations are not wholly unwarranted. In fact, for some suppliers of e-legal deposit who have opted for delivery via FTP, using the METS and MODS metadata standards for their SIPs in accordance with our FGS-PUBL and MODS-specifications, [4] Mimer already produces machine generated records in the national library union catalog LIBRIS, that are comparable in quality to corresponding manually cataloged records. 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 (of which the latter will also use the METS-MODS metadata standards already for SIPs). However, deposit effected through RSS, as is the case for the bulk of online news media publishers, will almost always fall short of such high quality of library catalog records. This is partly due to limitations in the metadata standard itself, as we will see later. Nevertheless, in the normalization process, by means of metadata enrichment from external sources, we strive to overcome at least some of these limitations.

Other stakeholders, rather on the end user side, might expect eventually to get free, ubiquitous access to copyrighted documents, to which access is otherwise restricted by payment or password control. Anyone holding such hopes is bound to be disillusioned, at least in the foreseeable future. There are some very substantial legal, financial and technical issues to deal with before we will get anywhere near free access, regardless of user location, to documents stored and preserved in Mimer. The NLS, so far, has only just begun to deal with the technical issues, perhaps the easiest part to solve, by planning for the access to DIPs, the dissemination information packages. These issues involve the means to give access, the prospective GUI, search options etc. Here we are also bound by an earlier management strategic decision on LIBRIS, the national library union catalog, to serve as "bibliographic metadata master" for most services provided by the NLS.

[4] http://www.kb.se/namespace/digark/deliveryspecification/deposit/fgs-publ/FGS-PUBL_eng.pdf
http://www.kb.se/namespace/digark/deliveryspecification/deposit/fgs-publ/mods/MODS_enligt_FGS-PUBL.pdf

4. The Publishers

The news media publishers are active in a very volatile and vulnerable market. It is no secret that they are facing huge challenges right now, struggling to find new feasible business models in the transition from analog to digital production and distribution. This also seems to bring about an accelerating rate of mergers, ownership changes and discontinuances. For NLS and Mimer, this means that keeping track of online news media publishers becomes a real challenge. In order to get reliable, unique identifiers of both publishers and documents (items), we need publisher-IDs that are persistent, unique and distinct from the supplier-IDs, which are often shared by a number of online news media belonging to the same publishing house or media syndicate. At times, online news media and newspapers seem to change affiliation and ownership overnight. There is also the problem that the legally responsible agent registered for e-legal deposit is often the supplier or distributor, rather than the publisher in a bibliographic sense (amounting to an entry in MARC 260 #b in a library catalog record). In particular, in our RSS specification, there is a mandatory element <dcterms:publisher> with as required value a unique publisher-ID consisting of a base-URL "http://id.kb.se/organisations/SE" + [Swedish official organization no.]. But since this organization number is often shared by several online news media (e.g. newspapers) belonging to the same publishing house, we demand that in these cases a - suffix identifying e.g. a particular online newspaper within a consortium is added. These presumably unique publisher-IDs are subsequently used to create likewise globally unique identifiers of *resources* (documents). This is particularly important in those cases where the original metadata that we receive with the SIP has only local identifiers of resources. For all we know, local identifiers might consist only of an object number in a local database with as little as one digit. Thus, to ensure that resources (documents) get at least one potentially globally unique identifier we also need to get unique publisher-IDs from which to construct them.

5. Format / metadata standard constraints

All metadata standards and formats have their limitations. RSS 2.0 in particular is designed to be a very simple and easy to use standard, with very few mandatory elements or attributes, even without a namespace of its own, in order to ensure compatibility with previous versions.[5] To overcome some of the limitations that this entails when it comes to richness of metadata, so-called "modules" such as MediaRSS have been introduced. Simply put, this means that an RSS feed may contain elements and attributes not described in the general RSS 2.0 specification, "only if those elements and attributes are defined in a namespace" (*ibid.*). The NLS and Mimer takes advantage of this possibility to "mix and match" by adding 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

[5] RSS 2.0 Specification http://www.rssboard.org/rss-specification

mandatory elements and further three that are mandatory if applicable. For most elements their status as mandatory is derived from the metadata requirements of the law which - as we have seen - are rather limited in scope. This concerns elements for identifier (*guid*), internet address of resource and constituent files (*link, media:content/@url*), publishing date (*pubDate*), publisher (*dcterms:publisher*), accessibility at the time of publishing (*dcterms:accessRights*), file format of resource and constituent files (*dcterms:format, media:content/@type*). Further, the specification provides for the option of delivering what amounts to constituent files separately by other means, e.g. via FTP. This is something that is in demand e.g. by one online media publisher with large video files. Thus, if this option is used, then the corresponding metadata element (*dcterms:references*) is mandatory. The only mandatory element not derived from the law in any sense, but rather as a byproduct of the RSS format itself and for practical reasons is the title element. According to the RSS 2.0 general specification, "all elements of an item are optional, however at least one of title or description must be present." We simply found it more convenient to make *title* mandatory, rather than having either the title or a description, as prescribed by the RSS 2.0 general specification, never knowing beforehand which one we would get. This seems to be alright also with the suppliers, for which providing a title of each item appears to be the rule, whereas a description or abstract is not always present.

Among the optional metadata elements in our specification, providing information that we thus cannot automatically count on having for our AIPs and library catalog records, are elements for license, statement of responsibility *(creator, contributor*), keywords, categories (subject headings). 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.

6. Deduplication and version control

Another challenge typical for online news media is the handling of new versions of the same items, as news feeds seem to be updated at an ever faster pace. Efforts of deduplication and version control run into a particular problem here as a result of a format limitation briefly touched upon above. The fact that the main identifier used for items according to the general RSS 2.0 specification, *guid,* can contain almost anything, a string or a URI, since its data type is not specified. This makes it very difficult to use *guid* for deduplication and version control searches. Only recently the NLS is trying to come to terms with this limitation by means of the introduction of another, optional and repeatable *dcterms:identifier* element, with an *xsi:type* attribute enabling the specification of identifier type (such as, e.g. *doi, hdl,* etc.) and, thus, making for effective deduplication.

Deduplication here means an effort not to create a duplicate record in the national library union catalog, if such a record of a particular item already exists. In that case, Mimer will only add its holdings 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. So, what if we actually never receive persistent and unique identifiers, with specified identifier types, for these items? To be sure, we are still dependent on the supplier / publisher at least to use *the same* identifier value, the same *guid*, for every update of one and the same item. We have seen some unfortunate cases of version updates, where the *guid* supplied was simply a copy of the URL in the *link* element, something that is by itself completely permissible according to our specification. However, in these particular case the *link* URL was misspelled in the first instance, and then subsequently corrected in the updated version. When this correction was copied to the *guid,* thus making the second instance different from the first, there was no way to escape a duplicate record in Mimer.

7. Inundation and concealment – when the levee breaks

Despite all the efforts made to achieve highest possible metadata quality, struggling to overcome some of the stumbling blocks above, we cannot do very much about the forces of nature. The naked truth is: there are simply too many online news items being published out there. Even if we were to be 100% successful in our deduplication and version control, we still run a 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, delivery errors etc.), this would still mean about 3000 new catalog records daily produced solely by e-legal deposit. (On top of that, besides e-legal deposit, the NLS together with the MKC, the Media Conversion Center, are running a continuous digitizing of analog newspapers, producing on average 118 issues daily, amounting to about the same number of new catalog records.) Evidently, there is the possibility that all these records already create too much “noise” in the catalog, making e.g. ordinary title searches impractical with unwieldy hit-lists. 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. This is done by means of a special use of certain metadata tags added in our xsl-transformation to LIBRIS. However, this has already proven to be insufficient, since the suppressed records are still pouring in to the Voyager database and cataloging client at a rate that the system has ever greater difficulties to bear, markedly slowing down the processing times. Therefore, since LIBRIS is anyway in the process of shifting from MARC and Voyager to an in-house built system and cataloging client based on linked data (JSON-LD), as a first step **all** e-legal deposit records will soon be redirected to a special isolated section of the new LIBRIS XL database. From there, only those records that would anyway not be suppressed from display in the web interface under the old regime will also, during a transition period, make their way into the Voyager database. This means that most online news items that we harvest from RSS feeds will no more, at least not during this transition period, have any visible representation in the national library union catalog. So why then should we still care about the quality of metadata representing these items? Well, for the same reasons that we should care for the future, for our children and grand-children. Even if we cannot be absolutely sure that they will ever appear in public, on stage or on the screen.

Processing

One thing should be made very clear: the library catalog records in LIBRIS that are produced in the processing of e-legal deposit in Mimer are not the only representations of the resources collected and, consequently, not the only use of metadata received, reproduced and enriched in the process. The even more important use of metadata in this respect is naturally for the electronic archive itself, for the AIPs in Mimer. Without sufficiently rich metadata in the AIPs, we will never be able to fulfil the overriding purposes of preservation and giving access in the future – near or distant. For the very same purposes, the process of *normalization*, the transformation of SIPs into AIPs is absolutely essential. Now, in the case of RSS, there are some important processing steps preceding normalization. Notably, there is the feed-reader, for fetching the feeds and their referenced files (in *link* and *media:content*). There is validation of compliance with our RSS specification against our xml schema. And there is the split-up of feeds into single items and (re)packaging of these together with the associated files into proper SIPs.

However, we will not go in to detail of these preceding steps here. Suffice it to say that the validation is done against an XML-schema (1.1) adapted for our "mix-and match" RSS-implementation, which can be downloaded from our website by suppliers wishing to test and validate if their feeds are fit for e-legal deposit. Just recently we have put out an online validation service, so that suppliers no longer need to download the xml-schema(s) themselves to test their feeds.

In the following we will concentrate on the normalization process.

Normalization

Normalization is the transformation, with XSLT as the main vehicle, of the original supplied metadata in the SIP, 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 in the SIP itself, ii) the supplier registry, iii) a so-called "channel record" in LIBRIS

(Voyager), which in the case of RSS is associated to all feeds published on a certain URL, and iv) file data (such as file size, file formats etc.) gleaned from the actual resource files associated with each item of a feed.

i) The supplied original metadata that we get from the SIP was treated as one of the stumbling blocks encountered above, the inherent constraints of the metadata format itself. Suffice it to say here that it is one of the main tasks of the normalization process to overcome at least some of these limitations. As archival format in Mimer we have chosen the METS-MODS standards, for several reasons. First, both standards are produced and administered by the US Library of Congress, apparently a warrant in itself for their future maintenance, widespread use and resilience. Secondly, METS, the Metadata Encoding and Transmission Standard, is evidently one of the main standards for implementing the OAIS reference model for digital preservation in the library and archival community. It serves particularly well for cooperation and exchange with other national libraries and was already in use by our cooperation parner the Swedish National Archive. METS is a container format housing bibliographic (descriptive), administrative as well as preservation and structural metadata. For the descriptive part of METS (dmdSec), MODS was chosen as the most expressive bibliographic metadata format tailored specifically to library needs and fairly easily transformed into current library cataloging formats (still MARC21, but soon to be replaced by linked data formats).

ii) The supplier registry holds information about each and all suppliers of e-legal deposit and the publishers that they are serving as delivering "carriers" for, as well as the "channels" used for delivery (e.g. url:s of different RSS-feeds). Apart from providing information about suppliers and publishers, such as publisher's real name, the supplier registry also serves as the access point to the "channel record", by supplying its LIBRIS record number.

iii) The "channel record" holds bibliographic information that is (expected to be) common to all items in every RSS-feed published on a particular channel (url). As we have seen already the RSS metadata format is rather limited in scope, even in the adapted version we use that is supplemented with Dublin Core and MediaRSS. This means that the possibility of enrichment of metadata with information from the "channel records" is particularly relevant for RSS-feeds. Admittedly, the importance of "channel records" has somewhat diminished since we started processing in test mode back in 2013, as a result of more values and parameters being controlled directly in normalization and transformation. Nevertheless, we are still dependent to a varying degree on the channel records for metadata elements such as language(s), digital origin, genres and host publications.

iv) Finally, the file information gleaned from the actual data files belonging to an item includes metadata on format, such as MIME-type, format name, key and version; elements that are vital for preservation purposes. Further, file information metadata includes file sizes and fixity checks using MD5 check sums. These are all part of the METS amdSec, the administrative metadata section used for preservation metadata.

Preservation

The Swedish law on e-legal deposit accords the NLS no right to specify preferred file formats to be supplied. Mimer must accept and be able to ingest all types of electronic resources in the file formats that they have been published online. Only if a resource is published online in multiple different file formats can we specify which is to be the preferred format.

The NLS has not yet commenced any preservation activities on the material contained in Mimer. In order to do any preservation planning and apply the right measures at the right time (e.g. migration), we need to know which data formats we have in store. Therefore, as much information as possible about the data files are saved in the archival packages. The information we receive from suppliers vary widely depending on type of material, e.g. whether it is e-deposit born digital web articles, or older newspapers digitized in-house. At a minimum, it is mandatory for the supplier to inform us of MIME types. But since this information is not sufficient Mimer always performs a format validation using DROID from which information about the format name and format key in the format register PRONOM is downloaded. To further ensure the material's authenticity all the identifiers of the supplier are saved plus checksums if any. Mimer always adds its own checksums.

We use the metadata standard PREMIS as the obvious choice since, as far as we know, there are no other options for preservation metadata. PREMIS stores information about which actions (events) Mimer has performed on each data file in the validation and normalization processes, the results of these actions and the applications software (agents) that performed the events. In this way at least we have prepared ourselves for future preservation planning and preservation activities.

Giving access

So far, the steps taken in Mimer to give access to ingested e-legal deposit material are only preliminary and preparatory. Primarily, they involve currently the transformation of AIP metadata to MARC catalog records in LIBRIS, the national library union catalog, by means of XSLT. LIBRIS has been designated as “bibliographic metadata master” of most NLS information systems and it will be the natural access point for e-legal deposit material in the future, although the LIBRIS system itself is in the process of intense development into something entirely new. This new system being born, based entirely on linked data and discarding the MARC format, will eventually require new transformation schemes for e-legal deposit metadata. Mimer is prepared to take on that challenge, and we believe it may benefit the end users in the sense that we will be able to display also information and links that are presently hidden from view, such as provenance metadata. Even more important, we expect that a new LIBRIS system with a new web catalog interface will ultimately allow us to display and give access also to all the online news media articles that are now suppressed from view simply due to their overwhelming numbers.

The future – near and far

Our model for taking care of online news media through 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. But the biggest change yet to come involves giving access by means of a reshaped web catalog, where MARC will no longer be the premium format. However, this presumes not only a change of format and the development of a new and more user friendly interface. It requires also a (re)solution of fundamental legal and financial issues that may not be so easy to resolve. How near, or far, this future awaits us, is a question that only time will tell.