Submitted on: 8/16/2014

Collecting bits and pieces – the development of methods for handling e-legal deposit of on-line news material at The National Library of Sweden

Pär Nilsson
Newspapers, Radio and Television, National Library of Sweden, Stockholm, Sweden.
par.nilsson@kb.se

Abstract:

The aim of this paper is to describe the efforts and experiences of The National Library of Sweden in developing systems and policies to collect, store, preserve online news material published in digital form. It will give a background to the new legislation, describe the experiences so far in the implementation of the new law and discuss possible future developments.

The National Library has collected printed newspapers since the first legal deposit law was introduced in 1661. The collection contains about 122 million pages and continues to grow with about 2 million pages per year. The legal deposit law was modified in 1979 to include audio-visual material and some changes were also made for the newspapers, but since then this legislation was left unchanged despite the development of new channels for publishing newspaper content from the 1990s and onwards.

A legal deposit law for digital materials had however been proposed and discussed in government reports since the 1990s and in July 2012 such a law was introduced. For an interim period (April 1st, 2013 - December 31st, 2014) a limited number of institutions, national agencies and media corporations will deliver content to the National Library, which since 2010 also contains the collections of the Swedish National Archive of Recorded Sound and Moving Images.

Items subject to delivery are e.g. materials from websites, pdf-documents and TV and radio podcasts. The materials should have some connection to Sweden, e.g. directed to a mainly Swedish audience, produced mostly in Swedish, authored or performed by a Swedish writer, performer, etc. The new law will come into effect in its entirety in 2015. The materials collected from e.g. newspaper websites through this law will not be complete and websites will only be delivered as separate articles.

Keywords: E-legal deposit, web harvesting, newspapers, news media

Background on legal deposit in Sweden

The first legal deposit legislation in Sweden was introduced in 1661. It was part of a series of reforms of the political system in the country and its main focus was on control and not on building a collection of printed publications in Sweden. This is obvious in the introduction of the text, where it is stated that “it is deemed to be useful and necessary that Their Royal Majesties may have knowledge about what books and other writings are printed and brought to light in the realm and the provinces”.

But the fact that two copies were to be delivered, one to the National Archives and one to the Royal Library, gives the impression that there was indeed also some thought given to the idea of preserving what was published for future generations. It should also be noted that not only books were included, but also all other published writings such as newspapers, magazines and ephemera.

The law was not an immediate success and it was amended in 1674 and 1707 to motivate the printers under the threat of fines to send in more of their production and also documents stating what had been printed. The number of recipients of copies was also increased and from 1707 it included the universities of Uppsala, Lund, Åbo and Dorpat, although Uppsala and Lund had already started receiving copies by special provisions in 1692 and 1698.

The first freedom of the press legislation was introduced in Sweden in 1766. This was amended in 1809 and was made more liberal. In 1812 a new version of the legislation was introduced, where a system of registered publishers (responsible for the content) of periodical publications was put into practice.

This system was used by government authorities to stop e.g. a newspaper from being published for political reasons, but it was also famously circumvented by the still existing Aftonbladet, which made a slight change to the title and introduced one new registered publisher after another in order to continue publication. The link between the freedom of the press legislation and legal deposit was strong, since legal deposit was an integrated part and one of the copies was to be delivered was sent to the national archives for control of its content.

Amendments to the legal deposit legislation in the 20th century

The freedom of the press legislation from 1812 remained largely unchanged until 1949, but in the new laws introduced that year, legal deposit was finally separated and a more modern law for this was created and remained largely intact for 30 years. The next revision of the legislation in 1978 introduced microfilming of newspapers and, even more importantly, added legal deposit for sound and moving images.

Between 1993 and 2004 further changes were made to keep up with the technological development and to include e.g. electronic documents in fixed form, but it was not until 2012 that a new law on e-legal deposit material (SFS 2012:492) was introduced, making it possible for the national library to start collecting electronic material distributed by different kinds of networks, such as the Internet. This new legislation has its own history and it took almost fifteen years to go from the first reports to the new e-legal deposit law.

Government reports on legal deposit and electronically published online material

The possibility to include electronically published online material in the legal deposit legislation has undergone investigation a number of times. The government e-legal deposit report of 1998 (SOU 1998:111: "E-plikt – att säkra det elektroniska kulturarvet") stated that the aims and objectives of e-legal deposit was to preserve and provide access to the Swedish cultural heritage, i.e. manifestations of Swedish life, Swedish society and Swedish culture for posterity, and that the material collected as far as possible should be available to all. According to the report, there were large amounts of published electronic material that fell outside the legal deposit law and that it was essential to collect and maintain it in a systematic way.

The electronic online information to be collected should, according to the report, be "widely available in this country and related to Swedish conditions". As to what would count as "widely available" the report judged that such information that the user must supply a password or pay a fee to gain access to may be considered publicly available, provided that, in principle, anyone is able to access the information. This means that web published material behind paywalls from e.g. newspapers would be included.

However, information published on e.g. a company's internal web site should not count as widely available. Nor should electronic originals of printed publications be included in the e-legal deposit law because they have never been widely available in electronic form. The National Library should however act for the collection of such material through voluntary agreements. Furthermore, more computer programs and computer games should be collected than had previously been done.

The electronic online information intended to be spread widely should according to the 1998 report be collected as completely as possible, as is done the printed and audio-visual material, and top priority should be given to publications produced by professional publishers and producers. Regarding other categories of information, such as private web pages, information from local associations and the like, it should be enough to save a selection, collected four times a year. Furthermore, a sample of the collections in digital format (databases) not delivered through e-legal deposit should be collected once per year in the form and with the content they have at the time. The preferred method for collecting most of the material should be web harvesting, which was already in use in the Kulturarw^3 project at the library.

To enable the collection and archiving of computer-based materials the report proposed that the Act on Copyright in Literary and Artistic Works (SFS 1960:729) should be complemented by provisions which would make it possible to make copies of works, including computer programs, available through online connection. The conclusion of the report regarding the possibility of providing access to the material with respect to the Copyright Act and the Personal Data Act (SFS 1998:204) was that access to the computer-based material copied must be restricted to researchers.

The report and the subsequent referral treatment showed, according to the government, that a large part of the referral opinion was sympathetic to the widest possible preservation of the Swedish digital heritage (Government bill 2000/01:3). The government made it clear that it was important that for posterity and for research purposes to preserve and provide access to manifestations of Swedish life, Swedish society and Swedish culture, but that it was not possible to preserve everything. The digital cultural heritage must therefore be secured by selection, but a wide selection without quality criteria was needed to create a rich collection for researchers. In the material to be collected there would, according to the government, be a large number of computer programs and in order for the report's proposals to be implemented modifications to the EU Directive on the legal protection of computer programs were required.

In 2003 the question of e-legal deposit was discussed in a broader government report about the work and future of the National Library (SOU 2003:129: "KB – ett nav i kunskapssamhället"). The report concluded that digitally published material must be covered by the legal deposit system because the cultural heritage and the right to transparency and information will otherwise be eroded when new methods and technologies for publishing are introduced. It recommended that the existing legal deposit legislation should also include "remotely transmitted digital materials", defined as "such materials that are made available to the public via remote transmission over a network". The concept of what constitutes "Swedish conditions", as a basis for legal deposit, was clarified.

For legal deposit it would also be required that the material had some permanent character, i.e. material whose content was not intended to change over time. Furthermore the report suggested that the producer or provider of content on a web page should be bound to deliver e-legal deposit material if he or she either already has a publication license (i.e. a certificate of no legal impediment to publication) for their printed material in accordance with the freedom of expression legislation or could seek and be granted such an publication license. E-legal deposit would then be mandatory for newspapers, municipalities, authorities, etc.

According to the report digital material would be deemed to have been made available even if it was protected by password, provided that the only condition for access is a user ID. The producer or provider would then deliver the material covered by legal deposit without having to disclose any passwords. As the producer or provider was supposed to actively deliver the material and not do so by requisition from the library, the problem that the library does not have actual knowledge of the material being distributed was avoided.

The government's reaction to the report (Government bill 2004/05:80) was that there was a need to expand legal deposit to include online digital material, but that the definition of material to be covered by this new legislation, the copyright aspects of the material and the potential impact on the integrity of sensitive data needed investigation. The proposal of the 2003 report did therefore not lead to changes in the legislation.

Web harvesting – the Kulturarw^3 project

The proposals on e-legal deposit in the 1998 and 2003 reports did not lead to changes in the law. However the National Library continued the collection of Swedish material published on the Internet by using the automatic web harvesting that the project Kulturarw^3 initiated in 1997. Some legal support for this project was introduced when the government in 2002 issued a regulation (SFS 2002:287) concerning the processing of personal data in the National Library's digital cultural heritage projects.

The starting point of Kulturarw^3 project was that all Swedish web pages were to be saved. The main reason for an almost complete collection instead of a careful selection was the lessons learned from the collection of printed material, namely that it cannot be known what material will be regarded as valuable and therefore be in demand in the future. Making a sophisticated and manual selection among millions of web pages would require significant staff effort. In many respects the information collected in this manner is more akin to a large ephemera collection than to a catalogued collection of books and periodicals. The library does not make any attempt to convert the material collected into a uniform formats and everything is saved in the format it had at the time of collection.

Creating a complete collection of web pages that are updated very frequently is of course impossible. The library's strategy was instead to take snapshots of the whole Swedish web couple of times per year. An exception to this major sweep was and is a selection of about 140 newspapers on the Internet which since June 2002 have been harvested on a daily basis. The harvesting involves as little human intervention as possible. One consequence of this is of course that updates that occur between the sweeps are not collected. Nor are the updates that occur during the day on the newspaper web sites collected. What is collected is a static snapshot of the web and not the changing and complex collection of information that the web is today.

The web pages are harvested using automated software, a so-called crawler. The program works by retrieving one or more start pages and scanning them for links. New links are then added to the queue of material to be downloaded. Then the downloaded the new pages are searched to find new links, etc. Each page is collected only once per sweep.

When it comes to defining what should be considered as Swedish website material the library uses the top-level domain of the web page address. In the beginning there were almost only websites with the top domain se. However, many Swedish companies, organizations and private users registered their domain under international top-level domains like com, org or net. On a lighter note, it has also been popular to use the top domain nu, assigned to the island state of Niue, since “nu” means “now” in Swedish. The library has cooperated with the organization responsible for nu top level domain to get information about which of the nu-addresses are used for Swedish web sites. As for web sites using other top-level domain names, they are harvested if the server is located in Sweden, as determined by the IP address.

All Swedish web pages discovered by the crawler are downloaded and stored on magnetic tape. There are two archives, one for the larger sweeps and one for newspapers. The magnetic tape is stored in two copies in different locations. The first collection in 1997 was about 3.4 million pages and 161 GB of data. Ten years later, in 2007, the corresponding sweep harvested approximately 135 million pages and 11 TB of data. Today there are over 1.7 billion items in the archive corresponding to approximately 72 TB of data

The regulation (SFS 2002:287) issued by the government concerning the processing of personal data in National Library's digital cultural heritage projects regulates accessibility in relation to the Personal Data Act (SFS 1998:204). The regulation allows personal data to be processed and stored to address the need for research and information, but the archive is only available on at the library on computers without connection to other networks or the Internet, since the material may not be disseminated electronically. Furthermore, the material in the archive is only retrievable by URL and no full text search is yet available.

The quality of the harvesting varies. In some cases images and style sheets are missing. Each web site is harvested only to a certain depth and with a limitation as to how many objects (text files, images etc.) may be retrieved. More complex web sites that require a lot of user activity and interaction are of course more difficult or impossible to harvest. Since only a few sweeps are completed each year for the majority of Swedish web sites changes in content are poorly reflected in the archive. Only the aforementioned 140 newspaper web sites are harvested every day, which certainly gives a better picture of these sites, but not a complete picture since they are updated almost every minute.

The web harvesting performed by the National Library since 1997 has doubtlessly preserved a lot of online material that otherwise would have disappeared. Web harvesting was also the preferred method according to the 1998 government report, but as will be seen in the description of the present e-legal deposit legislation, web harvesting is not considered to be the only method to be in the preservation of online material.

Proposed e-legal deposit legislation

In February 2009 the government initiated a new investigation concerning e-legal deposit legislation and in November the same year the memorandum on legal deposit for electronic documents (Ds 2009:61: "Leveransplikt för elektroniska dokument") was published. This memorandum contained a proposed new legislation which picked up where the 2003 report had left off.

The starting point was that the web pages and similar dynamic material should not be included, but only unchanging electronic documents or more precisely "a defined unit of electronic materials with text, sound or image that has a predetermined content intended to be presented at each use". The electronic documents should therefore be "permanent and complete" and documents where the intention is to continuously enter or change information (e.g. blogs) were therefore not to be covered by the obligation to deliver.

It would further be required that the material was "related to Swedish conditions", i.e. aimed at people who understand the Swedish language, included works by a Swedish author or a performance by Swedish artist or was otherwise mainly targeted at the general public in Sweden.

In three cases the electronic documents were to be exempted from e-legal deposit. Firstly, documents that have only insignificant content were excluded, e.g. small decor elements on a webpage. Secondly, commercial advertisements were not to be included, with the exception of self-advertising in e.g. newspapers. Thirdly, electronic documents that have the same or the same content as materials already delivered by legal deposit.

For government and municipal agencies only "actual publications", such as memoranda, reports, investigations, guides and similar publications, were to be included.

The material should be delivered by the publisher "within three months of making it available". It should be sent to the National Library on a data carrier (e.g. CD or USB stick) in the logical format that is made available via the network. The material should also be accompanied by information regarding where and when it was made available, the electronic document's logical format and connections with other documents, as well as codes etc. necessary for taking part of the contents.

The proposed new legislation was intended to be separate from and subsidiary to the existing legal deposit legislation

In the results of the referral process the majority of respondents did on the whole accept the proposal for new legislation on e-legal deposit for online electronic materials. However, two prominent critics of the proposal were Swedish Media Publishers' Association (TU) and Swedish Magazine Publishers Association. In the words of TU “the proposal provides a technically complicated and costly task imposed upon those under obligation to deliver” and “lacks a thorough analysis of the implications of the proposed legislation for the suppliers”. The proposal was “not practical and above all economically indefensible”.

This reaction was understandable since the proposed legislation was (and still is) closely connected and subsidiary to the already existing legal deposit legislation. This means e.g. that the newspapers should deliver each and every news article, image, video etc. not already delivered under the legal deposit legislation covering printed and audio-visual material. This would mean that e.g. the newspaper publishers would have to keep track of what material is published in print and what material is unique for the web sites. The critics explained that it is increasingly common for an article produced for print to be changed when it is published on the Internet – whether it be that the title or preamble is changed or that the article is rewritten, updated, shortened, added to a slide show, etc. To pick out what is to be sent or not on a daily basis was therefore impossible. These complaints did not in any way change the plans for an e-legal deposit law, but are instead handled in a pragmatic way by the National Library in the implementation of the legislation.

The e-legal deposit legislation for electronically published material

Finally, in March 2012 the Ministry of Education and Research had drafted the government bill on e-legal deposit and it was submitted to the Swedish parliament and accepted June 13 2012. The new legislation (SFS 2012:492) became effective July 1 2012. The legislation closely follows the ideas in the proposal from 2009, but instead of “document” the term “material” was preferred since “document” was considered to be too closely tied to the existing legal deposit legislation and printed documents.

Three groups of publishers are covered by the law:

- Publishers that have constitutional protection (e.g. newspaper publishers or TV and radio companies)
- Government and municipal agencies
- Companies which professionally produce electronic documents, e.g. e-books, e-music and e-movies

Electronic documents produced or provided by private individuals are not generally to be included, e.g. private blogs, but it is possible for the library to include this kind of material by agreement with the publisher.

The new law is implemented in two steps. From July 1 2012 to December 31 2014 only a limited number of publishers are included. These are the ten largest (printed) newspapers, the ten largest (printed) magazines and journals, a number of radio and TV companies, and a number of government agencies. The intention is to try out the legislation and the systems and methods to be used.

The second step in the implementation starts January 1 2015 and involves identification of and information to all publishers covered by the law. It will also introduce the group of publishers that is possibly most difficult to define, namely those who professionally produce electronic documents, e.g. e-books, e-music and e-movies. Some of these will be quite easy to target, but since the limit between "professional publishers" is more blurred in online publishing than in print, fixed-media music/video/etc. and broadcasting it will probably take some effort to make reasonable decisions.

The material to be delivered must meet the prerequisites of the legislation. Some examples of material are news articles, columns, opinion pieces, finished blog entries, reviews, advertisements related to the supplier's own business, brochures, guides, guides, web video, podcasts, and images. The material should according to the law be published only online, but since most suppliers of e-legal deposit material will not be able to differentiate and deliver only "web unique" content this is handled in a pragmatic way by the National Library and publishers are allowed to deliver material even if it has also already appeared in print.

Types of material not covered by the legislation are entire web pages, software or files that contain code, program code that builds databases, live broadcasts, continuously updated material (e.g. wiki sites, blogs, chats), content on intranets, privately published pictures, music, videos or blogs, calendars, schedules, seminar invitations, general commercial advertising.

The material should also be "related to Swedish conditions", by which is meant that it is aimed at people who understand the Swedish language, includes works by a Swedish author or a performance by Swedish artist or is otherwise mainly targeted at the general public in Sweden.

Systems, methods and organization for handling e-legal deposit

For handling e-legal deposit and other types of digital material the National Library has developed its own system (Mimer). Development was slow in the beginning and it was not until the library started using the system for archiving digitized newspapers that some of the worst obstacles were overcome. Mimer follows the OAIS reference model and is also integrated with other systems like LIBRIS, the joint catalogue of the Swedish academic and research libraries, and the library's system for handling the audiovisual material. Fedora Commons is used as a repository to store metadata about the files and keep a structural representation of the data and a combination of an HSM system and the cloud storage platform EMC Atmos is used for storage.

Even though the e-legal deposit law states that the material should primarily be delivered on a physical carrier this clearly is not the preferred method, neither of the National Library nor of a majority of publishers. It can perhaps be viewed as the lowest common denominator, providing even the smallest publisher with a realistic delivery method. So far two other methods have been used for the limited number of publishers now covered by the law.

FTP has been used for some material and will perhaps mostly be used for larger files and especially for audio-visual material, setting up different accounts for different publishers and even multiple accounts for some publishers. The publisher can receive a receipt when the files sent have been successfully processed and archived by the library.

However, for a majority of publishers delivery with the help of RSS will perhaps be a more convenient method, especially for frequently updated web sites like those newspapers and radio and TV, making it possible for the library to regularly download new items without any further work for the publisher. The supplier sets up a custom RSS service following a certain format (schema), which is a combination of Dublin Core and Yahoo's Media RSS, using specifications provided by the National Library, which then is responsible for retrieving the data flow. The retrieval is based on the publication date and time of the items in the RSS feed and new material will be fetched roughly every hour. This means that most of the updates from news web sites will be preserved, but depending on the frequency of updates for a file with the same address, some updates may occur between the downloads and thus will not be archived.

A third method is under development. This will be a web ingest form for uploading material through a web browser and it will probably be best suited for smaller publishers who deliver new material monthly or bi-monthly.

So far the contacts with the publishers have been handled by email and face to face meetings, but in order to handle the large increase of the number of suppliers of e-legal deposit material in 2015 when the legislation comes into full effect, the library is also developing a platform for guiding all potential suppliers to the right method of delivery depending on the size and nature of the material. This platform will also give information about what material is to be included and provide automated processes for the registration and connection of each supplier, once the library has checked that they are indeed a supplier of e-legal deposit according to the legislation and that they meet the technical requirements. Since the National Library is a state agency there is also a need to keep track of the contacts with the publishers both for the Swedish rules about public access to official records and in case a publisher does not deliver e-legal deposit material and thereby risks imposition of a conditional fine.

In order to make it possible to monitor what is delivered the Mimer system also has a user interface (Oden) for the library staff, where it is possible to see when and how much each publisher has delivered or how much has been fetched by the library and the status of the material, i.e. if it was actually archived or if there is a need to investigate possible problems. There are also some possibilities for actually viewing the material through the Oden interface.

This interface will be developed further and will include more sophisticated report tools based on e.g. how much each publisher is expected to deliver and the possibility to trigger alarms if the expected amount of material changes significantly. From 2015 and onwards there will be thousands of publishers of different types and it will of course not be possible to monitor the stream of material manually. Instead automatic methods based on rules and statistics are necessary, but the experiences the library has from having a lot of traditional audio-visual material from TV and radio stations delivered as files instead of on tape indicate that some degree of manual control is necessary. This includes the need for making certain that the files actually contain the expected material and also that all types of material from e.g. a newspaper is delivered. To make it possible to check the content another kind of interface is needed which would be more of a presentation system for the material and perhaps the first step towards an interface for researchers and users.

The administration of e-legal deposit involves many parts of the library and finding the right organization for this has taken some time. In the beginning much of the work was done in a new and quite small separate e-legal deposit division with technical support from different divisions in the IT department, but after a re-organization of the National Library the e-legal deposit work has been more integrated in different divisions under the two collections departments, Digital Collections and Physical Collections. Development of the different systems and technical IT support are now handled by the new and more integrated Information Systems Department. Legal support is available through the Corporate Services Department. This way of organizing the e-legal deposit work is both more integrated with the rest of the library work and less vulnerable than having one small and separate division, but there may be need for further changes when the legislation comes into full effect in 2015.

E-legal deposit metadata

To make it possible to search the e-legal deposit material in the future, it is necessary that certain mandatory metadata accompanies the delivered files. This applies regardless of delivery method. The information that must be provided is:

- Where and when the files are first made available
- The format in which the files are first presented
- Codes to open password protected files
- The relationship of the material with other material delivered by e-legal deposit, such as the relative order of the files in an article
- The relationship between the delivered files and analogue material delivered by legal deposit

This is of course a very limited set of metadata and there have been discussions about extracting further metadata from the material itself when it is delivered, but so far this has not been implemented in the system used for ingest. No indexing of the material has been done so far.

Future development and conclusion

The National Library is expected to report back to the government about the implementation of the e-legal deposit legislation and there are already certain problems that need to be addressed. One problem is the prescribed method of delivery, namely on physical carrier. A publisher that wishes to deliver via the Internet is obliged to seek permission to do so. Since network delivery will be the preferred method of both the library and the publishers it is desirable that the law supports this as the default method.

When the law comes into full effect in 2015 the National Library will gain a broader experience as to which publishers will actually be covered by the law and this experience might be used to create a better definition in the legislation of the rather vague “enterprises professionally producing electronic materials”.

The present legislation does not offer the National Library any legal support for making the e-legal deposit material available and this will need to be addressed in a not too distant future. As often in the library world it is only when you actually try to make the collected material available that some of the needs in the collection process become apparent.

And since what the library now is able to collect with the help of the e-legal deposit law is to a large extent the bits and pieces that make up web sites, without context or structure, it is really a necessity to tie together the traditional web harvesting process with the archive of the more complete content. It is by no means an easy task, but it could be very rewarding and give a reasonable picture of what is published on the web.

The new law is in many respects a good start and makes it possible for the National Library to start preserving also the electronically published part of the Swedish cultural heritage for future research and studies.