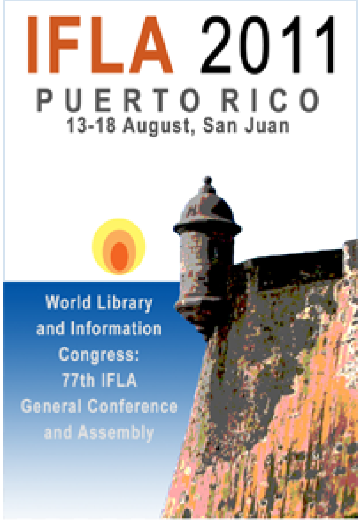
http://conference.ifla.org/ifla77 Date submitted: June 28, 2011


IFLA 2011
PUERTO RICO
13-18 August, San Juan
World Library and Information Congress: 77th IFLA General Conference and Assembly

Managing Legal Deposit for Online Publications in Germany

Renate Gömpel
&

Dr. Lars G. Svensson
Deutsche Nationalbibliothek
Frankfurt, Germany

Persistent Identifier: urn:nbn:de:101-2011061609

Meeting: **193 — *e-Legal deposit: from legislation to implementation; from ingest to access* — Bibliography Section with IFLA-CDNL Alliance for Digital Strategies Programme (ICADS), Information Technology, National Libraries and Knowledge Management**

Abstract:

A new law regarding the German National Library (DNBG) came into force in 2006. The most extensive modification compared to the preceding law was to the e-legal deposit, the collecting, cataloguing, indexing and archiving of resources in a non-physical format.

The German National Library has been implementing a fully-automated workflow for the acquisition, cataloguing and archiving of all kinds of online material since then. Using this workflow, we can handle e-books, online theses, e-journals and digital newspapers and we are working on archiving websites.

The creation of the workflow included dealing with metadata formats of different sources e.g. publishers, academic communities, individuals, but also with different data formats which must be archived. It also meant attending to copyright law.

The workflow facilitates the delivery of online publications as far as possible. For the time being, we have three submission interfaces:
- via web form (used for small numbers of publications)
- via OAI PMH (used e.g. for publishers and universities)
- via ftp or WebDAV to a hot folder (used mostly by commercial publishers).

In all three cases, metadata are supplied by the creator or the publisher and are transferred to the catalogue without intellectual intervention. As soon as the resource is transferred to the repository and archived, the title can be seen in the catalogue and the publication can be read in the reading room. All newly-submitted online publications are recorded in the German National Bibliography, in an extra series known as the O series.

Long time preservation is guaranteed but not all kinds of data formats can be stored so far. The German National Library is working to enlarge the range of processable metadata and data formats to meet the challenges of the electronic era. To ensure that our preserved digital resources can be reliably cited and accessed, we assign persistent identifiers to all online publications entering the library through legal deposit. The use of persistent identification not only ensures that the resource will be identifiable and accessible even if it changes location or is migrated, but is also important in the larger context of national bibliographies and for aggregators such as Europeana or the German Digital Library, which collect data from several sources.

The paper will present the German National Library's workflow for resources in a non-physical format and the use of persistent identifiers for long-term accessibility.

Introduction

It is the mandate of the German National Library (Deutsche Nationalbibliothek, DNB) to collect, catalogue, index and archive all German and German-language media works from 1913 on, to preserve them long term, and to provide the general public with access to them. Since the Law about the German National Library (DNBG) came into force on June 22 2006, this task has been extended to the collection of media works in "immaterial form", or to use the generally-established term, online publications.

The policy is not only named in the law (DNBG), but elaborated on in the legal deposit regulation. In this statutory statement, the how and when of the requirements for the deposit of media works (including online publications), and the extent and restrictions of the deposit regulation, are named.

After the legal deposit regulation became law on October 17 2008, a new version of the collection guidelines, with individual examples, was published as a working tool, thus refining the collection brief, which contains statements about online publications in a separate section.

What are online publications?

All representations available in public networks in the form of written text, images and sound, belong to the collection brief. Examples are e-books, e-journals, digital representations, music files and websites. The question of which electronic products are included in the collection is answered in the revised version of the Legal Deposit Regulation and the collection guidelines (see also http://files.d-nb.de/pdf/sammelrichtlinien.pdf (available in German only).

Which technical avenues for submission are available?

As is the case with print publications, any party releasing an online publication, including commercial but also private persons or organisations, is legally obliged to submit its publications to the DNB. Not all parties are familiar with the transmission of files and metadata yet. It was therefore necessary to find a simple way to enable submission without too much effort on the one hand, while making it possible for the DNB to automatically process the bulk of the vast amount of electronic publications on the other hand.

Development of an automated workflow – implementation in stages

As a result of the new legal mandate, the DNB began to develop a workflow with the goal of providing automated submission features for online publications, their display in the catalogue, and the archiving of the files.

With that aim, a step-by-step procedure, which offers the advantage of allowing testing and improvement of the technology with smaller quantities of data, in real operation, was chosen.

As a first step, the registration of depositors of online publications was planned for all procedures and interfaces. The recorded data (such as name, address and email address), is visible only to the submitter and to the DNB employees who work in the area of online publications. The registration process takes into consideration cases where service providers, who distribute electronic media for publishers, are depositing for a third party.

Submission via web forms

Web forms were the first option used for submission of different types of online publications. The new forms are based on the form originally used for monographs (e-books), and another used for online dissertations before 2006, on a voluntary basis. The new forms permit a simple transmission of further online publications: the form for monographs can be used for e-books, online dissertations and music. However, there are very few common mandatory fields (e.g. title, publication date, address of the publication), and depending on the type of publication, further fields may be added (e.g. information about the dissertation in the case of online dissertations or specific identifiers such as ISMN for music). A further form allows the transmission of titles of electronic periodicals and a third the transmission of periodicals. The forms are connected: since a title is only registered once, when the submitter signs in the next time, he/she obtains a list of his/her periodicals, and by clicking, most fields for the submission of a periodical are automatically filled out.

The web forms are ideal for smaller quantities of publications since the submission of the metadata takes place manually. Additionally, there is a maximum allowance of 50 MB for the upload of individual pieces and of 500 MB for submission via URL.

Provision for collection via interface

As a next step, an interface through which the publications are made available by the depositor and picked up by the DNB, was set up. A HTTP-based harvesting protocol, developed by the Open Archive Initiative (OAI-PMH, Open Archive Initiative, Protocol for Metadata Harvesting), is used for this service. When using OAI-PMH, a client or harvester requests data by sending HTTP GET-requests to a server or repository. Using this, metadata is picked up from the server of the depositor by the DNB through a fully automated process, which requires no manual intervention on either side. In a next step, a transfer URL is included in the metadata with which the publication is also retrieved automatically. The metadata is brought into the DNB catalogue without the transfer URL and the files are brought into the repository. Further information on setting up an OAI interface is available under http://www.d-nb.de/netzpub/ablief/pdf/automatisierte_ablieferung.pdf (available in German only). After a test period, this process now runs automatically on both sides and is suitable for larger amounts of files.

Deposit via Hotfolder

An additional interface has been in operation since April 2011. Hotfolders are suitable for the transfer of larger amounts of data and are sent by the depositor to this monitored folder. The folder is called "hot" because each step of the process that takes place is monitored by another process. After registration for an account by depositors, the publications are held in a Zip-Container along with the metadata. Available methods for transmission of the container are FTP (File Transfer Protocol) or WebDAV. Via an automated procedure, the metadata is integrated in the catalogue and the files are archived in the repository. The Hotfolder requires the depositor to actively provide the publications and the data, however, the interface was requested by publishers because of their familiarity with its data transfer options (such as FTP).

In all three cases, the metadata or records and the electronic object are collected.

Use of data formats

Online publications are collected in the data format in which they were issued. It is important to note that a transferable online publication must form an independent logical unit that can be separated from its environment. It should not be dependent on a server connection in the background, in which considerable parts of the content have to be requested dynamically and ad hoc from a data storage system at the moment of interaction with the user.

Currently, in addition to PDF (PDF/A and all other types of PDF), the format EPUB and documents in HTML can be automatically archived.

All data formats must be transferred without encoding in order to guarantee long-term usability of the documents with a minimum of effort.

Use of metadata formats

It is the depositor who enters the metadata in the case of use of a web form. There is no standard to adhere to, only the requirements outlined in the form are relevant. The metadata from the fields which have been filled out is mapped and transferred to the appropriate fields of the internal catalogue format.

In the case of the other two methods of deposit, the metadata for the automatic workflow must be submitted according to a defined and agreed upon standard. Presently, the acceptable formats are ONIX for Books, MARC-XML or XMetaDissPlus; further formats will be added eventually.

For the purposes of simplifying the deposit process, the minimal requirements for the metadata elements were defined in a set of core metadata elements. Details are available at http://www.d-nb.de/netzpub/ablief/pdf/metadaten_kernset_definitionen.pdf (available in German only).

How are online publications handled?

Automated descriptive and subject cataloguing

On the whole, metadata is integrated into the catalogue as it is. As the metadata is loaded, links to authority files are not made and only a few mandatory fields are monitored. However, the contents of the fields do not get checked and because of the vast number of records cannot be manually processed. As part of a project at the DNB, scenarios are being developed that improve the contents of the records through automated descriptive and subject cataloguing:

- names of persons provided (e.g. author, editor, translator) are checked against the name authority files and linked with authority records retroactively.
- the new records for online publications are verified for the existence of parallel print editions and, in the case that they exist, linked to them – at this point information from the intellectually catalogued print editions is carried over into the online publication records.
- online publications will automatically be indexed with Dewey classes and subject headings, especially in those cases where no parallel print version exists.

As part of another project, search engine technology is used to improve catalogue search and retrieval processes especially for online publications.

Access to metadata in the catalogue and to publications in the reading rooms

The metadata for the submitted and archived online publications is freely available for viewing in the catalogue of the German National Library, whereas the rights management for the digital objects is more complex. When submitting an object, the depositor can indicate which distribution rights he grants the DNB. The rights range from end-user access only in the library's reading-room, access over the internet for registered users so that they can access the publication from home, to worldwide unlimited access for any user. Commercial publishers generally offer paid access to the publication over their web-site and restrict the use to on site access in the library.. In those cases when a publication can be accessed only in the library's reading-room, end-users cannot make digital copies of the objects or send them per email, due to copyright restrictions. For information about the legal foundations see also http://bundesrecht.juris.de/dnbg/index.html (available in German only) and http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001L0029:EN:HTML.

The German National Bibliography – the O series

Immediately when an object has been deposited, the metadata is available in the library catalogue. Further, all incoming online publications are advertised in what is known as the O series (online publications) of the German National Bibliography. Once a month a list of the incoming records is compiled and this file can be downloaded free of charge. The records are available in the MAB and MARC 21 data exchange formats. Unlike the other series of the National Bibliography, this file is not offered in PDF form.

Citability and archiving

All digital publications deposited with the German National Library are not only indexed in the National Bibliography, but also assigned a so-called Persistent Identifier (PI). A PI is an identifier that is unique and is used worldwide for the identification of addressable objects such as documents, images, sound recordings, animations or metadata description. When a PI is used, an index known as a resolver is activated between the name and the address of the digital object. The task of the resolver is to ensure the assignment of a link between the name and the address. In the case of a change of address, they are updated in the resolver so that the latter always points to the correct address. Thus the references remain stable and the expended effort is manageable. The separation of the identification of the objects and their location by means of a distinct character string is a fundamental principle of PIs.

As part of the Epicur project (2002-2005), the German National Library introduced a PI infrastructure, which uses Uniform resource Names (URN) from the namespace NBN (National Bibliography Number) as identifiers – a namespace which was developed specifically for the identification of bibliographic resources. In order to allow a decentralised administration, urn:nbn is geographically structured; the German National Library is responsible for Germany and therefore of the sub-namespace urn:nbn:de.

For the conversion of the PIs into access addresses, the German National Library operates a resolver under http://nbn-resolving.org/. The resolver not only handles names from the German namespace urn:nbn:de, but also those from the relevant Swiss and Austrian namespaces (urn:nbn:ch or urn:nbn:at). Additionally, it allows the forwarding of queries about persistent identifiers which it does not administer itself. The resolver forwards queries to urn:nbns from the Czech Republic, Finland, Hungary, the Netherlands, Norway and Sweden to the relevant national resolvers, and queries about persistent identifiers from the DOI schemes (Digital Object Identifier), Handle, and Ark (Archival Resource Key) to the various service providers. The resolver was developed as part of EuropeanaConnect and is also part of the infrastructure of Europeana (http://europeana.eu).

The resolver of the German National Library does not only hold persistent identifiers for digital legal deposit. Any issuer of digital content can depose persistent links to objects. In May 2011, over 400 institutions were using this feature and had registered close to 5 million URNs. Counting approximately 3.500 queries a day, the access numbers are still relatively moderate, and with increasing use of persistent identifiers, these numbers are bound to grow.

The German National Library archives all deposited digital objects in a repository. This digital object repository is responsible for providing access to working copies of the publications; this can be a copy of the original publication or – particularly in the case of digitised items – a derivative such as a JPEG copy of a TIFF image. When an online publication is deposited in the repository, an automated process will create a data package for digital long term preservation and submit it to a separate archive system responsible for the long term preservation of the object.

The combination of digital preservation and persistent identifiers offers many advantages. With the help of persistent identifiers, researchers can make references to digital objects, knowing that they will remain stable long term and other researchers can call up the cited resource to verify information without further ado. Publishers and other issuers of documents can request a persistent identifier from the DNB for purposes of listing it in documents before publication, so that the persistent identifier will also be available in e.g. print versions of the document. The use of the resolver also allows publishers or data curators to store equivalent objects in different locations and then tell the resolver which address has the highest priority: E. g. a publisher can use the same URN to identify a publication available both in the publisher’s online shop and in the library’s repository. When resolving the URN, the address to the object in the online shop will have the higher priority so that the user will be directed to the publisher’s (commercial) offer. Should the publisher for any reason go out of business or simply not maintain the link, the reference to the library’s digital archive will still be available.

Long-term digital preservation is a task which is to be treated seriously, whence the importance of ensuring that the data is protected as far as possible from changes and influences from outside. The digital archive currently implemented by the German National Library is based on the software DIAS, a system developed in the KOPAL project in a co-operation between the DNB, the Niedersächsische Staats- und Universitätsbibliothek Göttingen and IBM Germany. Already at this stage, it is obvious that it will be necessary to replace classic long-term archives like DIAS with a next-generation, internationally-linked archiving infrastructure in which documents are deposited redundantly in several, interconnected long-term archives. The DNB is part of the EU project SHAMAN (Sustaining Heritage Access through Multivalent Archiving), which aims to develop technical and conceptual foundations for a new generation of linked systems and to link the long-term archives with the help of GRID technologies, in order to make the complex and resource-costly tasks of digital preservation possible. In order to meet future needs for a flexible and scalable IT infrastructure, the DNB is currently upgrading its data center to ensure that we can easily extend the storage for the digital repository and the long-term archive, as well as plug in further server nodes to provide sufficient processing power.

Long-term accessibility does not only mean preserving documents in an unchanged form, but works under the obligation that they will also be usable and readable in the future. Whereas current digital archives, such as DIAS, store only the bit string, the German National Library is well aware that the challenge is to ensure that the object is usable with future operating systems and other software generations needs to be met. Within the EU project KEEP

(Keeping Emulation Environments Portable), DNB and other partners are looking at how a relatively exact representation of static and dynamic objects can be achieved. This includes all types of objects: text, sound, images, multimedia documents, websites, databases, and video games, amongst others. The goal of the project is guaranteed long-term accessibility by means of the development of flexible access and storage tools. When those tools are in place and preservation actions on objects in the digital long term archive are performed, separate processes will ensure that the migrated object is synchronised back to the digital repository in order to provide end-user access to the most current version.

Challenges, problems and opportunities

The rapidly changing technical conditions clearly belong to the challenges involved in the collection of online publications. Not only the form of publication constantly changes due to the use of newer technologies, but also the options available for the transmission of electronic data and its use continually advance. In the not so distant past, the ability to store, change, and process texts on a computer was considered technical progress. Variability of forms of presentation soon followed, along with the inclusion of multimedia content and the ability to produce works for which ultimately there is no recognizable final version comparable to traditional forms of publication. Since the collection policy of the German National Library is not only to collect, but also to provide access and in particular to archive the collection, we need comprehensive solutions in order to overcome limitations set by formats.

Practical problems already exist with current, well-established processes, in which texts are stored in an internal system and not brought together as a work until the moment of interaction with the user. We are all familiar with reference works such as Wikipedia or also information databases, the content of which cannot be used in a useful manner without the software in the background.

Even new reading formats pose new challenges for the library – often the formats, in which the files are stored, can only be opened with the appropriate reader and have encoding features, which allow neither general accessibility nor archiving.

Besides the challenges arising from new techniques we have to deal with metadata issues. In the course of the last few years many discussions with publishers have taken place. Through these it has become apparent that because libraries are accustomed to the exchange of metadata and the application of standards in that process, shared use of metadata is a matter of course for them. Publishers and all enterprises that sell products have a different view on metadata: to them, data is seen mainly as a means to drive business, whereas libraries – particularly those having a mandate to record a country's publications – have other requirements. Those different views on metadata can make harmonisation difficult and a smooth data exchange can only be expected between participants having the same expectations and use commonly agreed standards building on the same understanding of metadata, e.g. between libraries or also with national book trade organisations where metadata creation is considered a high profile activity.

Next steps

In addition to the monographic online publications, electronic periodicals and newspapers are collected. Since May 2010, in a DNB project, 300 daily newspapers are collected and recorded in the cataloguing system as well as archived in the repository on an ongoing basis with the help of a service provider. DNB is now focussing on the collection of e-journals. An idiosyncrasy of e-journals compared to e-books, is that having no metadata standards, e-journals use individual formats for metadata. The transfer of available metadata is thereby difficult to nearly impossible. Consequently, alternative methods to catalogue them need to be found.

In the last years, printed books and periodicals have been digitised on a large scale by German libraries. Provided that they are publicly available, they fall under the legal deposit act for the library. To improve the archiving process, rather than collecting the file formats which have been made available to users, the German National Library aims, where possible, to obtain the master files, which allow saving without loss of information. The German National Library is in discussion with libraries and public institutions which have experience with digitisation in order to approach the task cooperatively.

We do not yet collect websites. However, as part of a project, the German National Library is working on implementing a workflow for the collection, archiving and indexing of websites as well.

The spectrum of data formats for the digital objects must be widened so that the segment of online publications that we would like to preserve long term can be expanded. With that aim, the handling of material that is protected by encoding mechanisms needs to be tested for purposes of long-term preservation and accessibility.

Our aim is to provide additional ways for the delivery of online resources in order to make it possible for more publishers to become depositors through at least one of the interfaces. The appropriate mappings are in preparation and will necessarily be updated continuously.