First a trial balloon, now an established workflow: collecting electronic theses at the German National Library

Uta Ackermann
German National Library
Adickesallee 1, 60322 Frankfurt am Main, Germany,
u.ackermann@dnb.de

ABSTRACT

Since 1998 the German National Library (DNB) has been collecting electronic dissertations and postdoctoral theses. At the beginning of 2011 the total number of collected online dissertations reached the 100.000 mark, which makes it the largest national collection of online dissertations in Europe.

This success could only be achieved by close co-operation with German universities, their libraries, institutional repositories and library service centres. This co-operation was conducted as a string of projects funded by the Deutsche Forschungsgemeinschaft and is known as 'DissOnline', a name, which has almost become its own brand.

Since 2006, as a result of changes in the law regarding the German National Library, the DNB has faced a much greater task than before: To collect and archive not only all German physical media but all German online publications. Faced with this challenge, the DNB was able to utilise the experience with ETDs gained through the DissOnline project.

Keywords (Required)

national library, co-operative project, large national collection of online dissertations, online publications in general, DissOnline, metadata, ingest, URN,

INTRODUCTION

When the German National Library (DNB) started collecting online dissertations centrally in 1998, there were 600 of them in Germany; last year we reached 100.000. After a steady rise the percentage of online dissertations and post doctoral theses in all types of media is now around 40% of all dissertations.

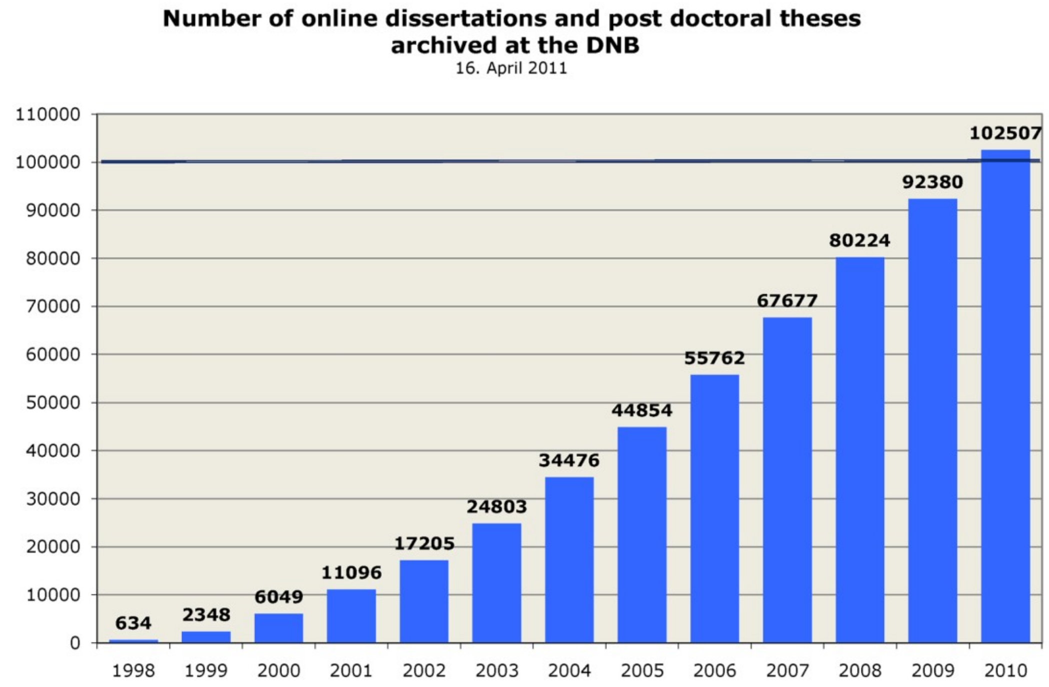

Figure 1. Number of online dissertation and post doctoral theses archived at the DNB

The aim of this talk is to outline how the collecting and storing of online dissertations have developed in Germany since the late 90s. During this process many difficulties had to be resolved: Firstly, users were reluctant to adopt the new technology. Secondly, there were legal issues to be resolved. Another area was the technology itself: the file formats to be used, author support, cross-referencing and data protection. In addition, there was the problem of how to deal with metadata and arguably the project's greatest challenge, the long-term preservation of ETDs (electronic theses and dissertations).

It will be shown that this dynamic process would not have been possible without extensive cooperation among many different institutions.

The German National Library is interesting in this context for many different reasons. Not only has it the largest national collection of online dissertations in Europe but there were factors unique to Germany that made this development possible. One factor is that every PhD candidate is obliged to publish his or her dissertation. Only after publication of the thesis the candidate has a legal right to use the title 'Dr'. This is meant to guarantee access and enable readers to quote from these sources. Another factor is that the DNB has been collecting all German dissertations right from the foundation in 1912. Since then it has been its task to collect, catalogue, index, archive and make accessible all publications from within Germany and written in the German language. Initially, this was done on a voluntary basis. Later a 'legal deposit right' was established, which means that since 1913 all German dissertations are available through the German National Library. This made it an ideal focus for installing the new digital library.

THE DEVELOPMENT OF ONLINE DISSERTATIONS IN GERMANY

The founding of DissOnline

The advantages of online publications for both authors and users are obvious: the authors can publish quickly and cheaply, the users gain easy access to the latest scientific research. However, implementation was a lengthy process that required the cooperation of many stakeholders.

It all began with an initiative of the learned societies. They wanted to work together to develop and use digital information and communication technologies for their members, scientific authors and readers (Diepold, 2000). One branch of this project dealt with ‘dissertations online’. Its main aim was to achieve easy access to research publications worldwide as well as a reduction of cost for the authors.

This initiative led to a proposal to the German Research Foundation (DFG - Deutsche Forschungsgemeinschaft) to fund an interdisciplinary project to present dissertations online on the internet, involving five German universities (Berlin, Duisburg, Erlangen, Karlsruhe and Oldenburg) and five academic fields, i.e. chemistry, education, information technology, mathematics and physics (Diepold 2000).

The electronic publishing of dissertations was made possible in 1997 when the KMK, ‘The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany’ allowed the possibility of electronic publishing in their ‘Principles of Publishing of Dissertations’.

Soon more partners for this project were found: university libraries, IT centres and eventually the German National Library. With its pre-established task of collecting all printed German dissertations and theses it was ideally placed to start collecting online dissertations as soon as 1998. During the first few years, the submission of online dissertations was entirely voluntary.

The German Research Foundation (DFG) supported and funded a string of projects on the topic of ETDs, which is now known as ‘DissOnline’, a name that has almost become its own brand.

Problems and their solutions

The topic of electronic publishing led to a widespread discussion among authors, PhD candidates, universities, academics, publishers, libraries and IT centres. There were many legal and technical issues that needed to be resolved, methods of submission had to be developed and questions regarding standards and security for both file formats and metadata required an answer.

Acceptance

Initially there was great scepticism towards the introduction of electronic publishing in academic circles. Fortunately, the good work done by the learned societies and the cooperation among different universities helped to overcome the general reluctance to adopt the new technology. Statistics show, however, that authors in the humanities are much less prepared to publish online than those in sciences (STM – Science, Technology and Medicine). Particularly conservative in their attitude are members of the legal profession as well as those working in the fields of business studies and economics. In these disciplines it seems that the printed word is still indispensable for the foundation of a professional reputation.(Wollschläger. 2003)

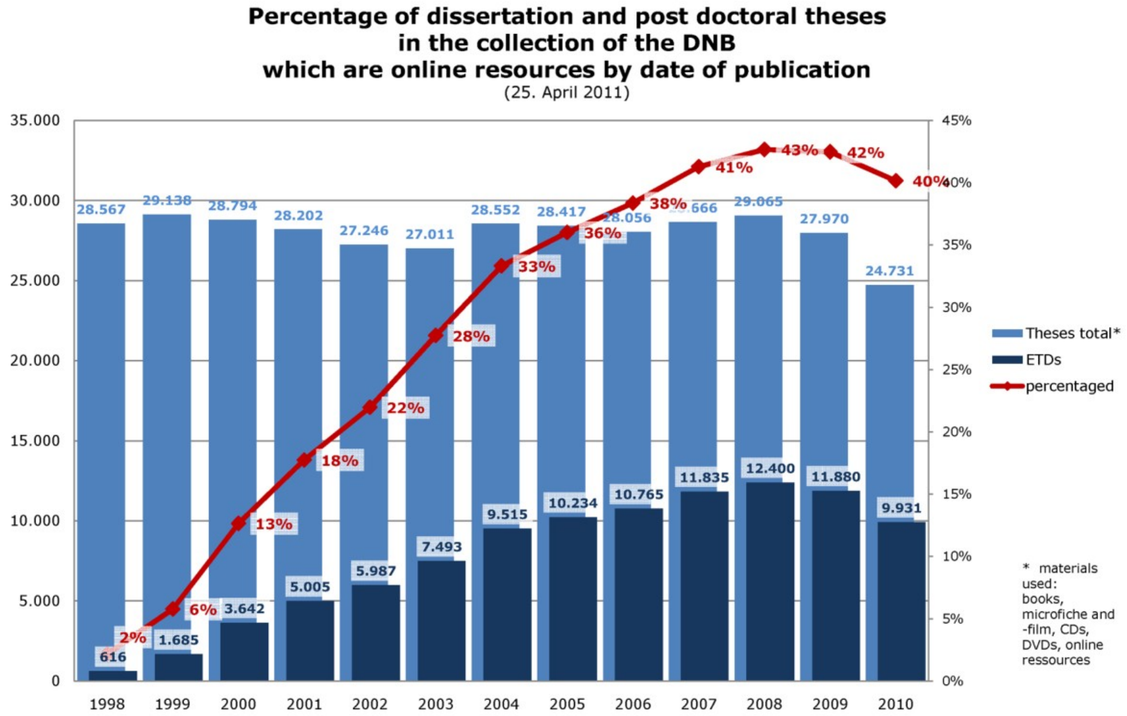

Figure 2. Percentage of dissertation and post doctoral theses in the collection of the DNB which are online resources by date of publication

Creating a legal frame work

One of the most pressing tasks at the beginning was the creation of a legal framework for online dissertations. The two main issues were standardization and copy-right.

In Germany, it is the universities or even the faculties that are responsible for the regulations regarding dissertations. These rules had to be adapted and approved by all the relevant institutions. Some regulations were adopted by all universities, but when asked in 2003, almost a third of German universities said that it was still their faculties that were responsible for online dissertations. (Wollschläger. 2003) The DissOnline projects were trying to achieve a standardization of dissertation rules across all universities and their legal implementation.

The question of copy-right was also discussed. The stakeholders developed recommendations for contracts between author and publisher, which keep the option of an online publication on the university server open. For the universities they drew up consent forms in which the copy-right of authors and libraries or IT centres was clarified and secured.

Technical issues

DissOnline profited much from the cooperation with DINI, the German Initiative for Network Information. DINI is committed to improve the information and communication services in higher education institutions and learned societies, and to provide the necessary information infrastructures regionally and nationally. Its members are service institutions for the world of academia, like libraries, IT centres, media centres, as well as different associations, academic institutions and organisations. More importantly, with its ‘DINI Certificate for Document and Publication Services’ the DINI sets a clearly defined standard in this field (information about the Certificate under http://www.dini.de/dini-zertifikat/english/).

The findings of this cooperative commitment and the resulting recommendations are published on a website (www.dissonline.de). In addition, regular workshops provided an opportunity to present and discuss new developments. Both website and workshops run under the ‘brand’ name DissOnline.

Formats

Much discussed were the formats in which dissertations should be submitted for publication. Initially, Word or PostScript was used, then XML was the preferred option for a while. This, however, turned out to be labour intensive. For a few years now, PDF, being both proprietary but disclosed has become the established format for publication.

DissOnline Tutor

Another important topic for DissOnline was the support of the authors and provision of a forum where their questions could be answered. Thus 'DissOnline Tutor' was developed in a special project. DissOnline Tutor provides authors of online dissertations and theses with word processing tools, such as Microsoft Word, Open- and StarOffice, and the document mark up language and document preparation system LaTex. In addition, authors can access interactive learning and teaching modules as well as materials. The aim was mainly to improve the technical quality of online university publications which are supposed to be archived for long term preservation.

DissOnline Portal

Eventually, the 'DissOnline Portal' was developed. This portal makes the collected German dissertations centrally accessible for users. Apart from searching metadata, a full-text search option is also available.

Global Access

To improve global access to European research theses, the DNB passes on the metadata of the ETDs from its collection to DART, the European E-theses Portal. The DNB participates in this partnership of research libraries and library consortia that are working together to achieve this aim.

Compound Objects

One problem that is still largely unsolved is the topic of compound objects. These are ETDs that are combined with research data or multimedia or dynamic elements. So far, the opportunities for using these have not been used to the full. Today, only a few years after their appearance, Compound Objects present an often insurmountable challenge to the provider as far as immediate access and long term preservation is concerned.

FROM COLLECTING ONLINE DISSERTATIONS TO COLLECTING ALL ONLINE PUBLICATIONS AT THE DNB

Since 2006, as a result of changes in the law regarding the German National Library (available only in German http://bundesrecht.juris.de/dnbg/index.html), the DNB has faced a much greater task than before: To collect, record, archive and make accessible not only all publications traditionally printed from within Germany and written in the German language but all resources in a non-physical format, that is all German online publications.

Faced with this formidable challenge, the DNB was able to utilise the extensive experiences gained through the Dissonline projects.

Transfer of Online Publications to the DNB

Since the early days different methods of submission have been tried.

Email

Submission via email was used for a while but as it made a fully automated process impossible, it was abandoned last year.

Web forms

Some of the first ETDs were also submitted using a web form which has been improved several times since. Experience showed that the web form should be kept as simple as possible. That way the chance of receiving correct metadata was higher even if it meant fewer information was sent. Another advantage of its simplified format is that since the extension of the deposit law in 2006 the form is mainly used by self publishers and small publishers who might not have the necessary know-how for dealing with bibliographic metadata. This makes a short, easy-to-use way of submitting the metadata and the object absolutely necessary. This problem does not arise so much in university publications where the trained personnel of the university libraries fill in the fields of the forms.

The new web 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). Another form allows the transmission of titles of electronic periodicals and a third the transmission of periodicals itself. 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.(Gömpel and Svensson 2011)

Initially, the submission of the web form was followed by a semi-manual treatment when metadata were transferred to the catalogue system using a small tool and during this process were checked intellectually. Now the metadata and resources are transferred in a fully automated manner but without any intellectual checks.

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. (Gömpel and Svensson 2011)

OAI-PMH

Since 2005 another method of ingest has been in use: the OAI-PMH method. This is an HTTP based harvesting protocol developed by the Open Archive Initiative. This interface is very popular with universities and is used by the DNB not only to harvest metadata but also to store them its repository. During this process, a so-called transfer-URL, which is a mandatory metadata tag, is used to save the object onto the repository of the DNB. Similar to the web form, manual interference was initially necessary in order to transfer the metadata into the library system and to download the object via the transfer-URL. Nowadays this process is also fully automated. Metadata are harvested up from the server of the depositor and transferred to the catalogue. In a second step, the object is located through the transfer-URL and stored in the repository. This process runs automatically on both sides and is suitable for larger numbers of files.

For commercial publishers it is, of course, especially important that the transfer-URL, through which the object can be accessed directly and at no cost, does not fall into the wrong hands. The DNB therefore guarantees that this link is used exclusively for the transfer of the object and is not passed on to unauthorised persons. Overall, the OAI-PMH is still unpopular as a method for submissions of publications in the commercial sector on a large scale. People still associate it with Open Access and are concerned about security. These concerns could not be dissipated even by technical facts. For that reason, the following ingest workflow was developed, especially with producers of e-books and e-journals in mind.

Deposit via Hotfolder

An additional interface has been in operation since April 2011. Hotfolders are suitable for the transfer of larger amounts of files which are sent by a 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. Via an automated procedure, the metadata is integrated in the catalogue and the objects 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 active procedures, 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. (Gömpel and Svensson. 2011)

URNs as Persistent Identifiers

One of the greatest challenges in the context of digital media is their long term preservation. Naturally this question was discussed at DissOnline. Especially in the world of research it is of greatest importance that published works are easily found and quoted from. A link that ends on page 404 (HTTP code ‘not found’) undermines the reliability of an academic paper. To eliminate this danger, it was decided to use ‘persistent identifiers’. These are unchanging, reliable markers that stay connected with the objects even if they are saved in different or changing locations or even in different or changing formats and systems. The task of the persistent identifier is to ensure long term access to the object.

There are different systems of persistent identification, for example Handle, DOI or PURL. Within the framework of the CENL (the Conference of Europeans National Librarians) some European national libraries chose URN (Uniform Resource Name) as their persistent identifier system. This has its own namespace with the IETF (Internet Engineering Task Force), in order to enable cataloguing of digital publications in bibliographic resources and registration (RFC 3188, using National Bibliography Numbers as URNs).

Therefore in 2001 the DNB started to build an infrastructure of an URN Service consisting of resolver, database and transfer surfaces.

So how does an URN work? In a database the URN as an identifier is matched with at least one address or reference called URL. Through the URN an address of the object can be reached using a resolver.

The URN has a hierarchical structure. Every URN is marked by the name 'urn' and consists of a prefix and a suffix. The prefix contains the NID (Namespace Identifier), in our case :nbn: for National Bibliography Numbers, and the SNID (Subnamespace Identifier). By dividing the URN into subnamespaces the internationally used hierarchical structure can be continued on a national level. A central element of this is the country code. An URN that starts with urn:nbn:de shows that it is a German publication that can be opened using the URN resolver at the DNB. Beyond the country code the URN structure can be further defined. This opens the possibility to give institutions their own subnamespace within which they can independently give out URNs for their own publications.

Giving the partners in DissOnline, usually university libraries, their own subnamespace meant that the successful cooperation could continue in the area of Persistent Identifiers.

Now the university libraries give out an URN for every online dissertation that is published by them. These URNs are registered with the URN service of the DNB. After the submission of the object itself the DNB archive link in the URN database is added, in addition to the address on the original server.

This concept, tested on ETDs, was extended by the DNB to other online publications. All online publications archived in the DNB get an URN from the namespace 'urn:nbn:de'. The depositors can give out URNs themselves using their own subnamespace. If that does not happen, an URN is given to every publication within the ingest workflow, so each digital object that is collected and archived by the DNB possesses an URN as a persistent identifier.

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 July 2011, over 400 institutions were using this feature and had registered over 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.

Metadata

Even in the early days of ETDs metadata were a key issue. DissOnline projects developed their own metadata format for online dissertations and theses. The first format MetaDiss was still embedded into the html basis. To enable transfer via OAI-PMH this was changed to the XML based XMetaDiss in 2005.

Like its predecessor MetaDiss, XMetaDiss is a sophisticated format that has about 50 elements and enables the DNB to store detailed information, for example about the people involved in a dissertation and their involvement in universities. Over the years it has been shown, however, that the possibilities of this metadata format are very rarely used to full capacity. The DNB has learned from this experience. For the submission of online publications a very simple core set of fields for metadata was developed. Especially when the web form is used, metadata are limited to this core set.

XMetaDiss also contains a section for technical metadata but these are not used any more by the DNB for legal deposits. Instead it was easier to obtain this information during the automated process. So technical metadata that are needed for long term preservation are gathered during routine transactions.

Last year the XML based metadata format was – again by close co-operation with our partners – extended to XMetaDissPlus (2010). Now it is no longer limited to dissertations and postdoctoral theses but also includes a great variety of publications found in university depositories, for example books, articles, journals, different theses, pictures etc.

It was down to the close connection and cooperation with the DissOnline partners, the German university libraries and IT centres, that the transfer of objects with metadata already attached has been realised in such a comfortable, now even fully automated manner. How close this connection is can be demonstrated by the fact that XMetaDiss was already an integral part of the repository software OPUS, which is the most popular in Germany. One could even say that the world of repositories in Germany was transformed by the developments coming out of DissOnline. The group of OPUS repositories also initiated the opening up of the format for all types of academic papers. In order to expand this development, the cooperation with DINI was sought.

A focus group of DINI (the AG Elektronisches Publizieren der DINI e.V.) developed a 'Common Vocabulary of Types of Publications and Documents'(2010). On the basis of that, the reference descriptions and examples for XMetaDissPlus were developed in close cooperation between the BSZ (library service centre Baden-Wuerttemberg) and the DNB.

In the DNB the introduction of XMetaDissPlus is closely linked to the completely automatized ingest of university theses. The new interface, now using XMetaDissPlus, can transfer metadata as well as objects into the catalogues or the repository.

XMetaDissPlus is downward compatible with XMetaDiss, which means that it is down to the universities whether they limit themselves to ETDs or pass on further university publications to the DNB.

Beyond XMetaDissPlus there are other metadata schemes which could be used to ingest dissertations and other online resources like e-books. Right now the DNB accepts also ONIX 2.1 and MARCXML metadata

SUMMARY AND OUTLOOK

Academic publishing in online form is firmly established in Germany today. The close, long lasting link between the stakeholders through DissOnline, now especially through DINI and the Open Access movement, which is very active in Germany, facilitates innovations and their general implementation. DissOnline did make important contributions and pave the way for new developments.

In 2006, the law demanded a new perspective from the DNB, and it rose to the challenge. Looking at the sheer volume of publications, electronic publishing and its possibility of automated processes is here to stay, and much more work is to be done.

The DNB continues to strive for the same success and highest standards in the collection of all its online publications that it has already achieved with online dissertations.

REFERENCES

1. Diepold, P. 2000. Dissertationen Online : The ETD Project of the German Learned Societies, in *Liber Quarterly* 10 (2000), No. 1, 31-40. Available http://liber.library.uu.nl/publish/articles/000349/article.pdf
2. Wollschläger, T. 2003. Die aktuelle Abgabepraxis von Online-Hochschulschriften an den deutschen Hochschulen, in *Bibliotheksdienst* 37 (2003), H. 11, 1422-1437. Available only in German http://bibliotheksdienst.zlb.de/2003/03_11_05.pdf
3. Gömpel, R., Svensson, L. G. 2011. Managing Legal Deposit for Online Publications in Germany, World Library and Information Congress: 77th IFLA General Conference and Assembly, 13-18 August 2011, San Juan, Puerto Rico. Available http://conference.ifla.org/sites/default/files/files/papers/ifla77/193-goempel-en.pdf
4. RFC 3188, using National Bibliography Numbers as URNs. 2001. Available http://www.ietf.org/rfc/rfc3188.txt
5. XMetaDissPlus. 2010. XMetaDissPlus - Format des Metadatensatzes der Deutschen Nationalbibliothek für Online-Hochschulschriften inklusive Angaben zum Autor (XMetaPers). Available only in German http://www.d-nb.de/standards/pdf/ref_xmetadissplus_v2-0.pdf
6. Common Vocabulary of Types of Publications and Documents. 2010. Gemeinsames Vokabular für Publikations- und Dokumenttypen. Available only in German http://edoc.hu-berlin.de/series/dini-schriften/12/PDF/12.pdf