Submitted on: 5/30/2014

How to handle the masses – automated workflows as a solution for the collection and preservation of e-books in the German National Library

Cornelia Diebel
Information Technology, Deutsche Nationalbibliothek, Frankfurt a. M., Germany
c.diebel@dnb.de

Abstract:

The number of e-books and online resources on the German book market as well as in other countries is increasing rapidly. In order to fulfil its legal mandate, the German National Library has decided to develop automated workflows for submission and processing of online publications, for their display in the catalogue and for the digital preservation of the files. In practise, no librarian or specialized employee performs any tasks related to the handling of an individual online publication. Instead, the effort is invested in motivating and supporting the publishers in fulfilling their duty to deliver the online resources on the one hand, and the preferred use of automated interfaces on the other hand. In addition, quality management monitors the automated processes, because automated processes need permanent improvement.

The paper focuses on the technical background and describes the four conditions necessary for automated workflows. Basically, the German National Library has established cooperations with publishers as a starting point for the collecting itself, which requires standardized metadata formats and interfaces for the ingest workflow processing. Another aspect is the technical solution to verify the quality and integrity of the digital objects. This part is important to assure the availability and usability for library users and the transfer into our digital preservation system.

Keywords: e-book, ingest, automatic workflow, long term preservation.

1. Introduction

The mandate of the German National Library (Deutsche Nationalbibliothek, DNB) is to collect, catalogue, index and archive all German and German-language media works published since 1913, to preserve them permanently and to make them accessible to the general public. Since the revised Law regarding the German National Library (DNBG)[1] came into effect on 22 June 2006, this task has been expanded to the collection of online publications.

However, the year 2006 is not considered the starting point for the collection of online publications, even before 2006, the DNB has gained experience in this respect. On a voluntary basis, electronic university dissertations and publications of some academic publishers were collected since the late 1990s. After the adoption of the new law and the related expansion of the legal collection mandate it was clear that former procedures and processing routes were no longer sufficient to cope with real mass business and that there was a pressing need to find alternatives.

The number of e-books and other online resources has grown enormously - internationally as well as in Germany[2]. In order to cope with the increased amount of electronic media, especially in view of the fact that the number of print publications does not decrease, the DNB decided to develop automatic procedures for the transmission, the processing, the transfer into the long-term archive and finally the provision of digital resources.
This in practice means that in the processing chain for online resources usually no intellectual or manual activities are necessary and involved. So the effort can be invested in new and further development of existing processes and of course in the motivation of the depositors to deliver electronic legal deposit to the DNB.

At the end of April 2014, the German National Library's collection of born-digital online resources has significantly exceeded the one million mark and currently consists of the following objects:

total stock (on 30.04.2014)		1.205.466
thereof	monographs	717.257
	audio books	950
	current periodicals	2.652
	periodical parts	483.920
	websites (title)	957
	websites (time sections)	3.267

Figure 1: Total stock of online resources

2. Organizational Aspects

The number of depositors in Germany who are legally obliged to deliver online publications to the DNB is large. The obligation applies to all commercial and non-commercial publishers,

[1] Gesetz über die Deutsche Nationalbibliothek
[2] Börsenverein des Deutschen Buchhandels (2013)

to cooperate bodies and entities established in Germany, e.g. universities, public and private institutions and organizations, as well as to self-publishing authors.

Due to the high number of depositors, practice has proven to cooperate with aggregators and service providers as much as possible. Whereas in the academic field most of the university libraries are faced with the task to deliver all electronic university publications, some public institutions are responsible for the legal deposit of other authorities. For example, as a consequence of Germany's federal structure, the German Federal Statistics Office (Statistisches Bundesamt) plays the role of an aggregator in representation of the Statistical Regional Authorities (Statistische Landesämter).

In the German commercial publishing business there are a lot of publishing service providers, who inter alia handle the delivery of e-books to e-book selling platforms. Some of these service providers have been successfully convinced to integrate the legal deposit obligation to the German National Library in their service portfolio. In this context it is helpful to use delivery procedures similar to common extradition practices.
Of course, there are publishers who handle the delivery process to the DNB on a stand-alone-basis, independently of service providers. But cooperation with aggregators and service providers facilitates the task for the DNB enormously because communication can be focused and concentrated. In both ways, direct contact between publisher and the DNB in advance is absolutely necessary.
If the option "delivery via service provider" is taken, publishers as well as self-publishing authors have to pass through an initial testing phase at the DNB. This means that small test sets are checked intellectually for content and technical suitability. Depending on the outcome, responsible contact persons report back errors and/or possible improvements to the publisher or service provider. After the "go-live process", quality checks are made only as random samples.

To cope with the large number of technical and organizational challenges, the DNB has decided to expand the collection of online resources gradually in a step by step approach. One important step is the collection of e-books. Other building blocks like the systematic collection of online university dissertations continued of course but have been adapted to the newly developed procedures. Further dedicated steps will be the collection of object types like e-journals, e-papers, download audio books and websites.

This paper deals with the solutions which the DNB has developed for the collection of e-books. Other object types are collected via automatic methods as well, but require different solutions in detail.

3. The Technical Workflow

The technical workflow consists of several steps which are shown in the following:

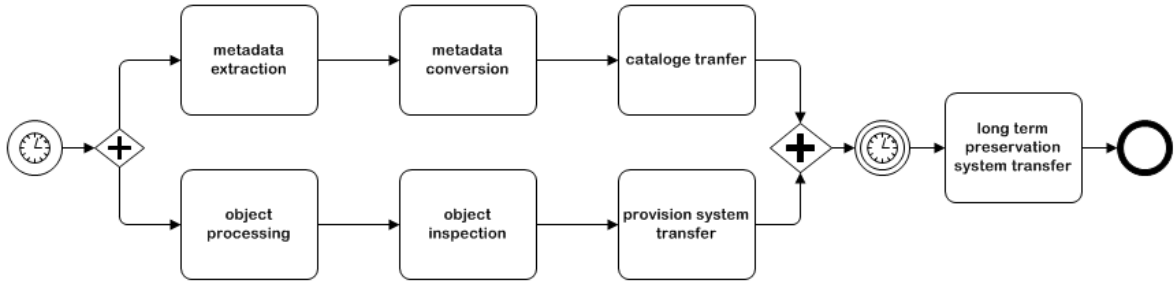


Figure 2: Detailed overview of the technical workflow

The DNB provides a form based ingest option and, relevant in terms of quantities, machine-based interfaces as well. Each depositor who delivers via these mass deposit interfaces corresponds to an individual ingest process definition built by the tools mentioned above. They are specifically configured and can be synchronised between daily, weekly, monthly and other frequencies for processing depending on the delivery quantity. Deliveries via web forms are controlled by the delivery itself. At the runtime of a task, the processing of metadata and object is running in parallel.
At present there are over 650 individual ingest tasks implemented which are distributed on different days and different times with the help of self-developed scheduling tools.

Metadata are extracted, checked for validity, converted into the internal metadata format of the DNB and transferred into the DNB's catalogue system. The object is subjected to a detailed inspection (see section 4) and transferred into the deployment archive. Further routines of the ingest process are for example, checking for duplicates and assigning a persistent identifier. If all actions are successful, the object is available to the users immediately. Regardless to this, the transfer of the object into the long-term archive system is controlled by an independent scheduling module.

If errors occur during the delivery processes they will be tracked, managed and processed by a connected ticket system, which serves in this case as a workflow engine. The different errors are categorized and they are finally handled by a special group of library staff, who take care of the completeness and validity of the collection.

3.1. Requirements for an automated process

To establish an automated workflow, four conditions have to be fulfilled (for the time being):

1. metadata and the electronic object are collected together
2. to agree on the metadata format
3. to agree on the accepted file formats
4. to agree on the interfaces for data exchange

<u>Metadata and electronic objects are collected together</u>
At the moment, the first condition is a crucial one for simply technical reasons. Moreover, it minimises the risk of getting metadata of lower quality that causes problems to save the

relation between object and metadata. The consequence is that the DNB does not collect online resources from those who cannot fulfill this requirement. It has been shown that almost all of the depositors (not only the commercial) are able and willing to deliver objects and metadata together.
The understanding that the salability and retrievability of the products depends on the quality of the metadata has experienced large spread in the book industry in recent years.

However, this condition should be modified in the future and DNB should extend its capacities: The goal should be the possibility to collect online resources without metadata and then create metadata from the objects themselves.

Metadata formats

As long as metadata and electronic objects are collected together it is necessary to agree on the metadata format which is delivered to the DNB. Currently, the DNB has several options for metadata delivery.
In the case of using a mass delivery interface, the metadata for the automatic workflow have to be submitted according to a defined and agreed standard. Presently, the accepted metadata formats for monographic online publications are ONIX 2.1 for Books[3], MARCXML[4] and XMetaDissPlus[5]; further formats will be added eventually. To simplify the deposit process, the minimal requirements for the metadata elements have been defined in a set of core metadata elements[6]. The DNB chose metadata standards which are widely used in the library community as well as in the book sellers' community.

ONIX is an international XML-based standard for book metadata, providing a consistent way for publishers, retailers and their supply chain partners to communicate extensive information about their products. It is widely used throughout the book and e-book supply chain in North America, Europe and in the publisher's community in Germany, so ONIX metadata are available for most of the classic e-book titles.

MARCXML is the well-known international bibliographic metadata format derived from MARC 21 which is also used as a metadata standard by some international publishers.

XMetaDissPlus is a special German development, with a background in the German university library and IT center community that agreed to develop an easy method for accessing research publications. Therefore a special metadata format has been developed since 1998 which is now widely used by the scientific community. The format faced different stages of development and at the moment it is not only suitable for online dissertations or electronic theses but also for all kinds of different materials like e-books or e-journal articles.

On the whole, metadata are integrated into the catalogue as they are. When the metadata are uploaded, no links to authority files are made and only a few mandatory fields are monitored. However, the contents of the fields are not checked and cannot be manually processed because of the vast number of records.

Additionally to the automatic processes in the collection of e-books, the DNB has invested in the automatic cataloguing in recent years. Since no efforts are put into the intellectual

[3] Editeur
[4] Library of Congress
[5] Deutsche Nationalbibliothek (2012b)
[6] Deutsche Nationalbibliothek (2012a)

cataloguing, various processes to improve the metadata supplied have been established. Based on the textual analysis of the objects, automatic DDC-based subject categories[7] and GND-tags[8] are assigned and transferred into the catalogue entry, on the one hand. On the other, automatic links to personal names are made in the GND authority file and online resources are linked with related print issues. Those enrichment processes improve the usability of the objects: the access points are thus increased and syndicated editions are merged.

File formats

There is a wide variety of file formats available among online resources, but the DNB cannot collect objects in all possible file formats at the moment. The DNB prefers objects that follow open standards and needs to know what formats are used. In this way, digital long-term preservation as well as accessibility of the objects in the reading rooms of the German National Library can be ensured which covers two of the main goals of the DNB.
Online publications are collected in the file format in which they were issued. There exists no workflow which is used to convert objects initially into one unified presentation or preservation file format. It is important to note that a transferable online publication must form an independent logical unit that can be separated from its environment.
Currently, in addition to PDF objects (PDF/A and all other types of PDF), the now strongly used e-book format EPUB as well as different image and audio formats can be automatically processed, provided in the reading rooms and archived. More formats will follow, for example the Amazon e-book format.

All resources must be supplied without file protection in order to guarantee long-term usability of the documents with a minimum of effort. With that aim, the handling of material that is protected by encoding mechanisms is automatically tested in terms of long-term preservation and accessibility (see section 4).

Interfaces

The use of different interfaces is the main technical challenge to be overcome. If producers and the DNB do not use the same interfaces, it is not possible to establish an automatic workflow; if the DNB does not establish a workflow it is not able to collect the respective online resources.
Currently, the DNB offers three different types of interfaces. Before using one of the submission methods it is necessary to register with the DNB and to receive authorization.

In terms of using accepted interfaces for deposit delivery, the DNB offers now different options to those who do not want to establish a technical infrastructure themselves even if this implies investing more effects in intellectual work, and to those who prefer delivering via so called mass interfaces.

[7] Deutsche Nationalbibliothek (2014a)
[8] Deutsche Nationalbibliothek (2014b)

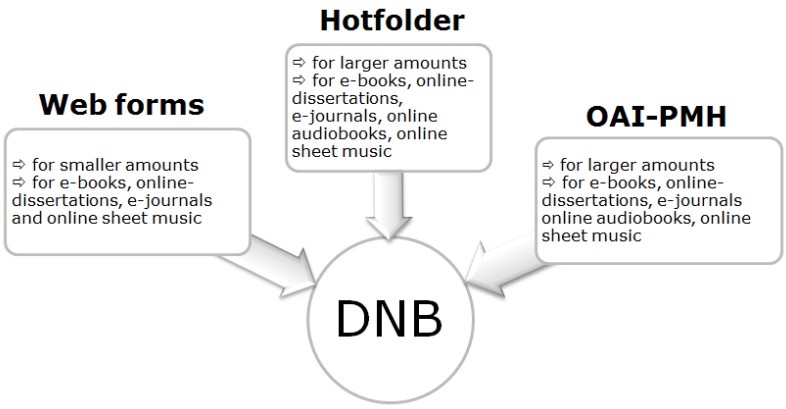


Figure 3: Overview of interfaces

For the latter option there exist two variants, however, it is necessary to fulfill the conditions for delivery as described herein. Once fulfilled, the processes usually run without further technical support in the background, except error handling.

Web forms

Among the submission methods, web forms were the first option used for submission of different types of online publications. The DNB started with e-books, went on with online dissertations and now accept the submission of online sheet music and e-journals via web forms as well.
The web form contains very few common mandatory fields, for example title, publication date, address of publication. Depending on the type of publication, further fields can be added (e.g. information about the dissertation in the case of online dissertations or specific identifiers like ISMN in the case for music).
The web form variant is ideal for submitting smaller quantities of publications, because the creation and submission of the metadata takes place manually.
Web forms are used by a large number of small depositors. The advantage is that it is not necessary to establish technical solutions within the DNB. However, the number of publications delivered via this interface is only around 5%. This figure shows the importance of those interfaces that were designed for mass delivery.

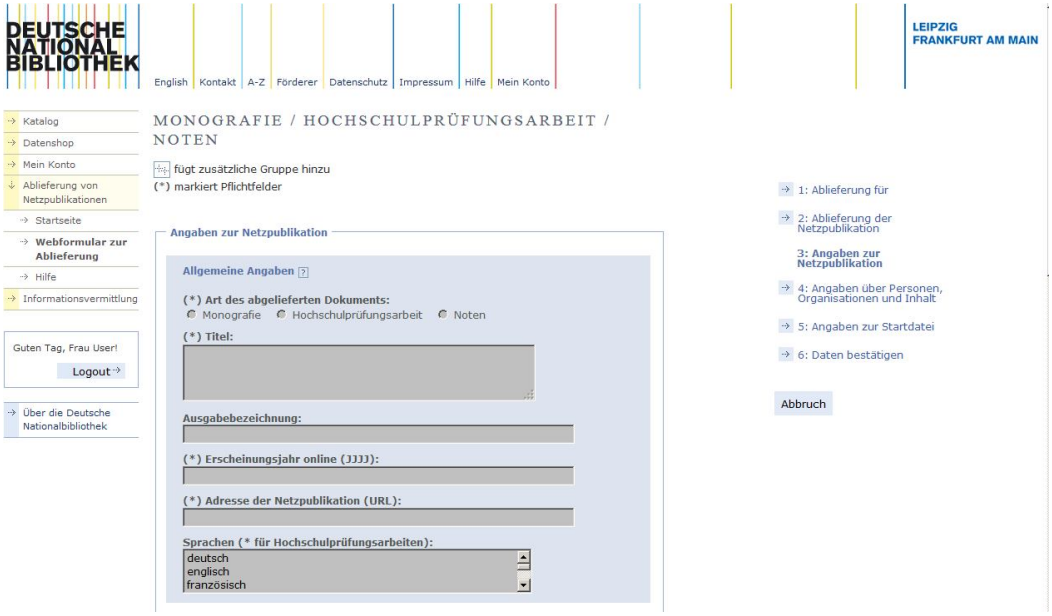


Figure 4: screenshot of the DNB web form

Open Archive Initiative - Protocol for Metadata Harvesting (OAI-PMH)

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. The Protocol for Metadata Harvesting developed by the Open Archive Initiative (OAI-PMH)9 is used for this service.
Using this option for submission, a client or a harvester requests data by sending HTTP GET-requests to a server or a repository. Metadata are picked up from the server of the depositor by the DNB in a fully automated process which requires no manual intervention on either side. In the next step, a so called transfer URL which is included in the metadata is used to retrieve the publication automatically. The metadata are transferred into the DNB catalogue without the transfer URL and the object files are transferred into the repository. This process is suitable for larger amounts of files and runs automatically by an action, which is controlled in the DNB. This interface is mainly used by scientific institutions from the university sector and there, especially in combination with the metadata format XMetaDissPlus.

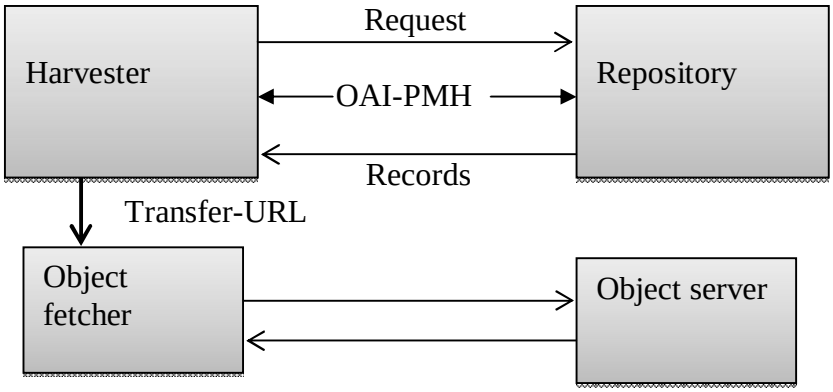


Figure 5: OAI Harvesting Procedure

Hotfolder

Since April 2011, an additional interface is in operation, which the DNB calls "Hotfolder". Like the OAI harvesting, the submission procedure via Hotfolder is also suitable for transferring larger amounts of data that are sent by the depositor to a monitored folder at a DNB server. Each step of the process that takes place is monitored by another process.
After depositors have registered for an account and have been assigned a storage location, the online resources that are held in a structured zip-container along with the metadata, will be transferred to the storage location. Transmission methods for the container are Secure File Transfer Protocol (sFTP) or WebDAV. The metadata are integrated in the catalogue and the files are archived in the repository via an automated procedure. The Hotfolder requires the depositor to actively provide the publications and the data. Nevertheless, the interface was requested by publishers because of their familiarity with its data transfer options (such as FTP).
Even the delivery to e-book selling platforms described above, often follows this automatic submission procedure. Currently, the Hotfolder is the most widely used interface. At the moment most of the depositors choose this variant and most of the deliveries via Hotfolder are based on ONIX 2.1.

[9] Open Archives Initiative

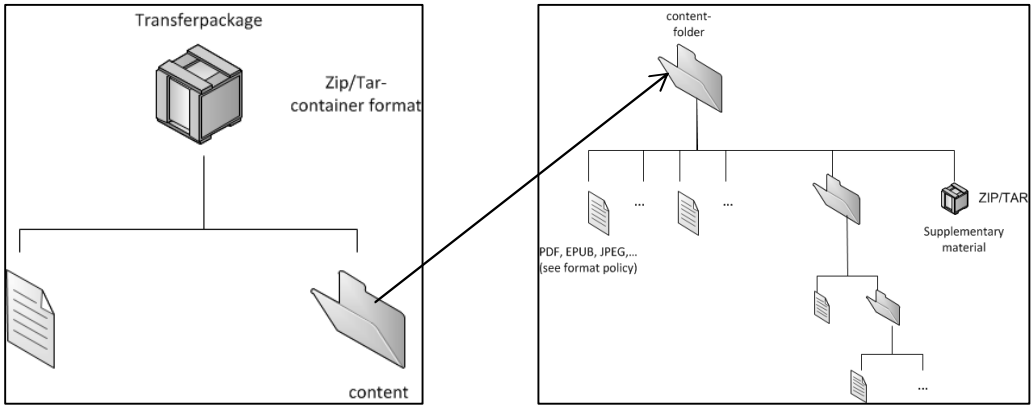


Figure 6: Structured zip container with metadata and objects

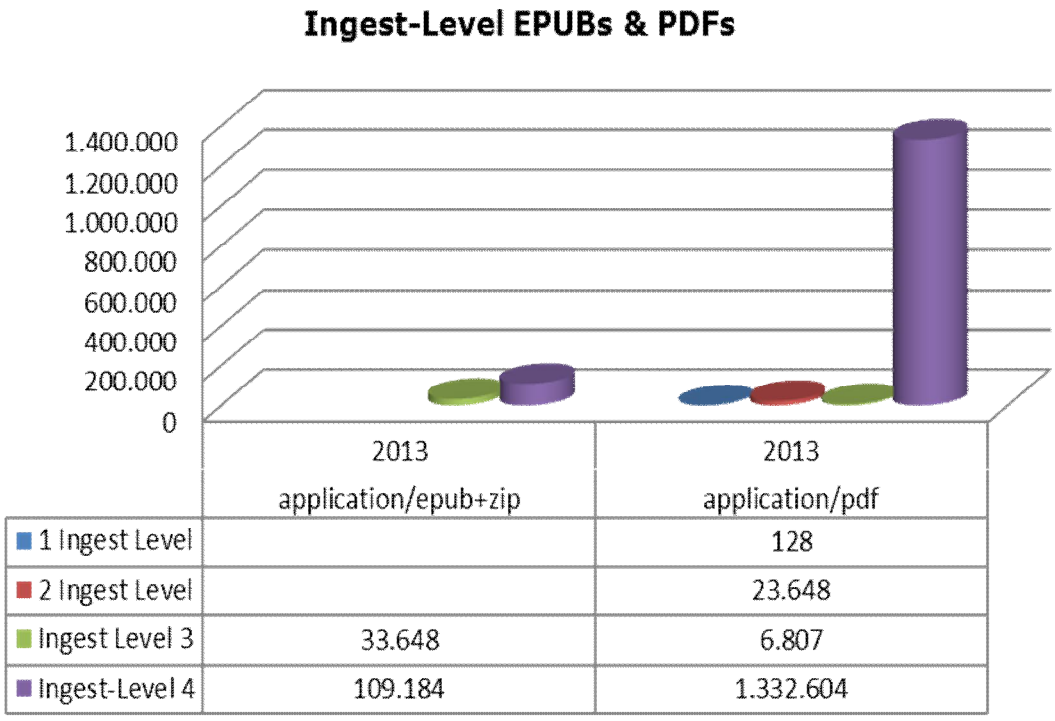
4. Object Inspection

For all digital objects which are delivered through the mass deposit interfaces OAI-PMH or Hotfolder, a fully automatic quality inspection during the ingest process takes place within the DNB.
The inspection of the object aims on the one hand on safeguarding the authenticity of all delivered digital objects, on the other hand on analyzing, whether technical restrictions exist which hinder or even prevent the task of long-term preservation, use or provision of digital objects.
In this context, workflows have been developed which recognize technical restrictions as well as define and initiate, if necessary, countermeasures.

Along with tasks like checking for duplicates or allocation of persistent identifiers, which take place in the ingest process but are not described in this paper, the object inspection comprises basically the steps checksum testing, generation of technical metadata and the awarding of ingest levels for each object. At this, ingest levels represent the result of a multistage testing procedure in which the probable long-term preservation and usability of an object is determined on the basis of technical criteria. According to the DNB, the higher the detected ingest level, the higher the probability of the long-term usability.

Fundamental to the preservation of the objects is the proven receipt of the bistream over an indefinite period of time. A method to determine this is the checksum test. DNB is therefore allocating ingest Level 0 to an object after successful application. A successful identification of the accepted file format leads to the assignment of ingest Level 1. To award the next higher ingest level (ingest level 2), limiting mechanisms, which restrict or even prevent the use and functionality of the object, may not be found in the analysis. In the case of PDF documents this would be for example passwords, copy or printing restrictions, which would prevent the award of ingest level 2.
If sufficient additional format-specific technical metadata for long-term preservation measures could be extracted, ingest level 3 is assigned. For this, an example would be the determination of the color space or the exact size of an image.
Digital objects, which additionally could be tested positive in terms of the validity of an accepted file format, for example EPUB oder PDF, achieve currently the highest and therefore "best" level, ingest level 4. The validation of the form is a key requirement for the provision and in terms of long-term preservation as well. Modern presentation systems are quite capable of representing even file formats, which are not completely valid, apparently

error-free. However, this cannot be assumed automatically for the future provision or long-term conservation measures. So to be prepared best, the demand of the validity of the form is mandatory.



	2013 application/epub+zip	2013 application/pdf
1 Ingest Level		128
2 Ingest Level		23.648
Ingest Level 3	33.648	6.807
Ingest-Level 4	109.184	1.332.604

Figure 7: Ingest level of all EPUBs and PDFs in 2013

It can be seen that only EPUBs with ingest-level 3 or 4 have been ingested in the collection of the DNB; 23.5% of all EPUBs were not completely valid in format. Although for this group there is an increased risk in terms of long-term usability, the DNB decided to ingest those EPUBs. The situation in case of PDF is slightly different, but there it appears that most of the objects reach ingest level 4, but some just reach ingest level 2. Countermeasures, like the technical correction of the files or a controlled initial transformation would be a possible conservation measure. For EPUBs, PDFs or other digital objects, which are not free of restriction (DRM protection), corrective measures are impossible or only with great effort. According to the DNB, in this case, the risk of being not able to secure long-term usability is too high. For this reason, objects which do not reach ingest level 2 (restriction-free) are rejected. If an error in the inspection process occurs, the library staff will contact the depositor and ask for a version that is free of restriction. Normally these errors are isolated cases, because we aim to prevent them right from the beginning before the go live with each depositor.
Therefore, the object inspection, well established in the business processes of the DNB, includes a risk management, which already starts at the ingest process of the objects10.

5. Provision

The use of online resources is one of the central concerns of the DNB. In this context, compliance with the legal framework is a basic requirement. This means to respect the usage restrictions which are made by the publishers. This is limited to the provision of online resources only in the reading rooms of the DNB, if this constraint is not explicitly repealed. Moreover, the possibilities of further processing like object download are suppressed. Open Access objects are freely offered. For use, it is also important that the appropriate viewer and player are available. So for each new file format a solution for deployment must be found. Currently, to give access to the objects, there are different and free available viewers in use. Given the fact that long-term archiving is now well established the use of online resources can be ensured for the future.

All online resources archived by the German National Library, get an URN from the namespace "urn:nbn:de". Compared to URLs, URNs offer a unique and permanent identification of digital objects, independent of their storage location. This is an important factor in maintaining long term preservation. If the storage location of a publication changes, for example if it is moved from one server to another, the access address (URL), which is combined with an URN, can be adjusted. With this, the URN keeps its validity and still refers to the corresponding publication. This makes the citation of an URN a reliable method to ensure a long term reference for digital objects.

Automatic cataloguing processes (see section 3.1) aim to provide additional access points and therefore the findability and usability of the online resources is improved.

6. Summary and Outlook

The German National Library made grand investments to develop the automatic processes and routines. The result is visible through the high amount of online resources which are currently ingested and useable in the library. This encourages the DNB to follow this way.

[10] For more information about the ingest-level concept, refer to Schmitt, Hein 2013.

This paper focused on e-books, but it must be noted that every object type has its own requirements with regard to ingest workflows and process handling. So the decision to take a step by step approach seems to be a reasonable roadmap to include different online resources one after another and enlarges the collection of online resources in the German National Library.
The DNB has to face new challenges, for example to ingest e-books which are published as apps or apps which are dynamic applications. Therefore the German National Library has to provide new solutions and of course has to improve its existing workflows.

References

Börsenverein des Deutschen Buchhandels: Von der Perspektive zur Relevanz - Das E-Book in Deutschland 2012. Frankfurt 2013. Available at http://www.boersenverein.de/sixcms/media.php/976/E-Book-Studie_2013_PRESSEMAPPE.pdf, cited 15 May 2014.

Deutsche Nationalbibliothek: DDC Deutsch - German Dewey Decimal Classification. Last updated 05 May 2014. Available at http://www.dnb.de/EN/Wir/Projekte/Abgeschlossen/ddcDeutsch.html, cited 15 May 2014.

Deutsche Nationalbibliothek: Integrated Authority File (GND). Last updated 03 February 2014. Available at http://www.dnb.de/EN/gnd, cited 15 May 2014.

Deutsche Nationalbibliothek: Metadata Core Sets for Automatic Delivery. Last updated 18.12.2012. Available at http://www.dnb.de/EN/Netzpublikationen/Ablieferung/MetadatenKernset/metadatenkernset_node.html, cited 15 May 2014.

Deutsche Nationalbibliothek: XMetaDissPlus - Format des Metadatensatzes derDeutschen Nationalbibliothek für Online-Hochschulschriften inklusive Angaben zum Autor (XMetaPers). Leipzig, Frankfurt 2012. urn:nbn:de:101-2012022107. Available at http://d-nb.info/1020009535/34, cited 15 May 2014.

Editeur: ONIX for Books, Previous Releases. Available at http://www.editeur.org/15/Previous-Releases/, cited 15 May 2104.

Gesetz über die Deutsche Nationalbibliothek. Available at http://www.gesetze-im-internet.de/dnbg/index.html, cited 15 May 2014. [For a non-official translation see: Draft Law regarding the Deutsche Nationalbibliothek (DNBG). Available at http://www.dnb.de/SharedDocs/Downloads/EN/DNB/wir/dnbg.pdf?__blob=publicationFile, cited 15 May 2014].

Library of Congress: MARCXML : MARC21 XML Schema. Available at http://www.loc.gov/standards/marcxml, cited 15 May 2014.

Open Archives Initiative: Protocol for Metadata Harvesting : Protocol Version 2.0 of 2002-06-14, Document Version 2008-12-07T20:42:00Z. Available at http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm, cited 15 May 2014.

Schmitt, K., Hein, S.: Risk Management for Digital Long-Term Preservation Services IPRES 2013, p. 314 – 317. In: Proceedings of the 10th International Conference on Preservation of Digital Objects. 3.-5. September Lisbon – Portugal. ed.Jose Borbinha, Michael Nelson, Steve Knight.