Risk Management for Digital Long-Term Preservation Services

Stefan Hein
German National Library
Adickesallee 1
D-60322 Frankfurt am Main
+49-69-1525-1722
s.hein@dnb.de

Karlheinz Schmitt
German National Library
Adickesallee 1
D-60322 Frankfurt am Main
+49-69-1525-1782
k.schmitt@dnb.de

ABSTRACT
This article presents an ingest level system which has been developed as part of the Digital Preservation for Libraries (DP4lib) project. The purpose of the system and its implementation is to facilitate automatic technical quality checking of digital materials. It represents an essential part of the risk management system within the long-term preservation processes of the German National Library (DNB). Initial practical experience is reported upon, demonstrating that a significant step has been taken towards ensuring the long-term usability of digital materials.

Categories and Subject Descriptors
Standardization, Verification

General Terms
Management, Reliability, Verification.

Keywords
Digital Preservation, Risk Management, Ingest-Level, Quality Management

1. INTRODUCTION
Handling risks is part of the daily business of long-term digital preservation. In all the areas of long-term digital preservation examined here, it is always important to recognise risks at an early stage, to assess their possible effects, to develop countermeasures and to implement these as required. Such risk management in organisations must be institutionalised in order to ensure continual monitoring of potential risk sources and to minimise any impact.

But how can comprehensive risk management be achieved for long-term digital preservation and its operational processes?

Risk management in this context is often referred to in the literature, e.g. in the OAIS reference model [1], as an integral part of preservation planning. The primary purpose of risk analysis in the ExLibris Rosetta system is to warn against the threat of obsolete file formats [2]. There it is carried out by the repository manager and is based on the data currently being managed. The approach presented here, by contrast, is distinguished by proactive measures taken right from the point at which the digital publication is ingested and regards any "inferior object quality" apparent at this time, which is based on more than an analysis of the file format, as a risk for future preservation action.

The objectives of the two-year Digital Preservation for Libraries[1] (DP4lib) project launched by the DFG were to evaluate the possibility of setting up a long-term preservation service for third parties and to implement a prototypical solution. An overview of the project results can be found in the long-term digital preservation manual [3] for service providers and users. Reflecting the main results of the project, one of the main benefit was that a suitable system of a cooperative risk management was set up consisting of automatic technical quality checking of digital objects and full reporting of all long-term digital preservation activities. The purpose was to lay the foundations for a trusted repository.

One of the main sources of risks in long-term preservation lies in the digital materials to be archived. The technical quality of the digital materials, for instance, is often both unknown and substandard, meaning that preservation of their long-term usability is already questionable with our current knowledge.

To check and if necessary avoid such risks, the service users and providers must cooperate to set up a joint risk management system which can recognise risks at an early stage and avoid them if possible.

The key component of the risk management ingest level system is described in section 2. Section 3 focuses on the technical implementation. The ingest level system ensures that risks associated with the partnership on the one hand and on the wide range of file formats on the other can be automatically recognised and communicated. The initial practical experience is presented in section 4. Finally, the last section includes a summary and the outlook for the further development of this approach.

2. THE INGEST LEVEL SYSTEM
The idea behind the ingest level system is presented in this section. The ingest levels are first defined and then the organisational integration and the contribution to risk management within the DNB are examined. The DNB actively uses the ingest level system for its internal long-term preservation processes, for ingesting digital publications as well as for the planned long-term preservation service for third parties.

The idea of using different levels for controlling and checking within long-term preservation is not new. Within PREMIS, for instance, different preservation level types were introduced which are based closely on groups of significant document properties which need to be preserved [4]. As in the ingest level system, preservation of the bitstream constitutes the first level. A similarly

[1] Project homepage: http://dp4lib.langzeitarchivierung.de/

close connection between level and preservation strategies can be found in the DHEP project [5] in which a total of 4 different levels of preservation strategies were introduced. By contrast, the ingest level system concentrates exclusively on checking the technical quality of a range of file formats and provides an indication of possible risks for the long-term usability of digital documents.

2.1 Definition and criteria

Assignment to an ingest level is the result of a tiered automatic checking process for file formats which is carried out (in part) in cooperation between the DNB and the depositing partners. By assigning an ingest level to a digital publication qualitative statements can be made about certain technical aspects of a digital object. A technical quality standard can also be expressed for the publication.

The general goals of this quality check, which is to be run for each file in each ingest transaction, are safeguarding the authenticity of the digital objects received and carrying out an analysis aimed at recognising technical restrictions at an early stage which hinder or even prevent the task of long-term preservation and also use of the digital objects.

Five test criteria, each one following on from the next, have been defined for this purpose:

1.) File integrity (DI)

The files submitted by the depositors have not changed during the course of the data transfer and processing.

2.) Identification (ID)

The file formats of the digital publication's files have been clearly identified.

3.) Lack of restrictions (LR)

The file object is free of restrictions, i.e. there are no recognisable (to the DNB) technical barriers which could impede or prevent the use or long-term preservation of the publication.

4.) Extraction of format-specific technical metadata (MD)

Format-specific metadata which are required for digital preservation could be generated.

5.) Format validity (V)

The file format (specifications) of the publication is valid.

Table 1 shows how the individual criteria relate to each other.

Table 1: Ingest level and criteria

	DI	ID	LR	MD	V
Level 0	X	O	O	O	O
Level 1	X	X	O	O	O
Level 2	X	X	X	O	O
Level 3	X	X	X	X	O
Level 4	X	X	X	X	X

Following the technical test, a digital publication is assigned level 0 if the integrity (DI) of the files belonging to the publication could be checked, confirmed and logged following the successful transfer to the DNB as the result of coordinated processes between the depositing institution and the DNB. Special procedures (checksum tests) are used for this. A digital publication is then assigned ingest level 1 if the file format could be successfully identified. No restrictive mechanisms may be detected in the subsequent analysis of the digital publication which impede or prevent the use or functionality of the publication for the issue of the next ingest level (ingest level 2). In the case of PDF documents, these include e.g. password, copy or printing restrictions which would prevent the issue of this ingest level. Ingest level 3 is assigned if sufficient additional format-specific technical metadata for long-term preservation measures could be extracted. The DNB has specified a core set of technical metadata for each file format. Currently the highest, and therefore the "best", level (ingest level 4) is achieved by digital publications if the validity of the file format used could also be positively tested.

The higher the ingest level, the more criteria have been positively tested and therefore the greater the risk management probability that the deposited publication can be preserved.

This form of technical qualitative analysis allows the DNB, for the first time, to automatically recognise long-term preservation risks for digital publications and to undertake suitable countermeasures at the time of transfer. As a consequence, the question arose as to whether countermeasures should be taken as a suitable response to the identified risks - and if so, which. The DNB has drawn up a format policy for the ingest and processing of digital publications.

2.2 Format Policy

A list of the minimum and maximum ingest levels for the file formats has been drawn up for the file formats deposited at present with the DNB on the basis of the current technical analysis possibilities. Table 2 contains an extract from this list. By setting a minimum quality standard for archivable file objects it was possible to draw up a format policy which contains rules for accepting and rejecting digital publications and also provides rules for further analysis tasks.

Table 2: DNB Format-Policy.

File Format	Min. ingest level	Max. ingest level
PDF	2	4
EPUB	2	4
…	…	…

The ingest of a publication is rejected on technical grounds if an ingest level below 2 is determined for a file of the digital publication. In such cases the DNB contacts the depositor. All other publications assigned an ingest level of 2 or higher are accepted into the archive system of the DNB. If some of the publication files have only been assigned ingest level 2 or 3, this does not constitute grounds for rejecting the publication. With regard to long-term digital preservation, the DNB is responsible for preserving the individual files of the publication in a permanently usable state and for carrying out any necessary preparatory measures.

The ingest levels are henceforth to be interpreted as new minimum expectations for the assessed quality standard of the individual file formats in the import process.

3. TECHNICAL IMPLEMENTATION AT THE DNB

The following section describes the technical implementation of the approach for risk management based on the DNB's ingest process for digital publications.

As shown in Figure 1, the DNB ingest process starts with the deposit of the digital publications via mass deposit interfaces such as OAI-PMH. It also includes the import processing chain for storage in the repository and ends with a further workflow, independent of the import process (LtpBinding), for transfer to the Long-term archive (LTA). The main steps of the risk management-enhanced import process include tasks such as checking for duplicates, issuing persistent identifiers, carrying out checksum checks, generating technical metadata and conducting the ingest level comparison.

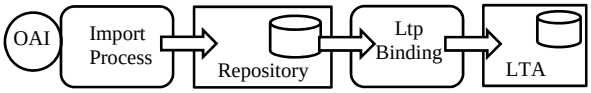

Figure 1 The Ingest Workflow.

3.1 Checksum test

The checksum test is one of the first test routines in the DNB import process; the first step involves calculating a checksum at the file level. This is then compared with that calculated and supplied by the depositor. Only if both checksums concur will the file object be assigned ingest level 0 and be forwarded for further processing. Ingest level 0 therefore constitutes the basis for all other process stages shown in Figure 2. At the DNB these are contained in a tool called *diagnose digital objects (didigo).*

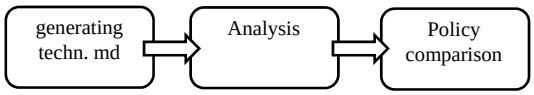

Figure 2. Diagnosis of digital objects.

3.2 Generation of technical metadata

For some time now the automatic generation of technical metadata using metadata tools has been a recognised and established component of the ingest process. The DNB has long been using the *File Information Tool Set (FITS)* as a framework for using an entire tool set. This framework provides access to a whole range of tools including the *JSTOR/Harvard Object Validation Environment (JHOVE)* tool, the *Digital Record Object Identification* (*DROID*) tool and the *NLNZ Metadata Extractor*. JHOVE cannot handle the same variety of file formats as DROID, however it does support the generation of technical metadata and also checks the formal accuracy and format validity. DROID, by comparison, merely identifies the file format and its version. Use of a tool set widens file format support and reduces the risk of errors in the identification and validation of the file format. FITS also offers significant added value in the form of easily configurable standardisation of the different tool outputs into the FITS format using XSLT. The DNB has used this function to adapt the FITS output to its own requirements, e.g. incorporating other metadata elements not included in the FITS distribution into the standardisation. However, the resulting output schema still complies with the FITS standard. This extended FITS format provides a format-specific metadata set which unifies the different technical metadata elements of a number of metadata tools and combines them structurally into a single standard [7]. A further adjustment which the DNB has made is the integration of a DNB tool to analyse files in ePub format.

3.3 Analysis

The FITS processing is followed immediately by analysis of the results. This is concluded by final calculation of the ingest level which is initially set at 0. The test criteria of restriction-free access, file format, format-specific metadata and format validity are examined - in this order - on the basis of the FITS output. Each test which is successfully passed raises the ingest level incrementally by 1, with 4 being the highest ingest level achievable by a file object. As soon as one of the above tests has been failed, the ingest level remains at its present level.

FITS yields XML objects, meaning that the technical implementation of this test can consist in querying individual XML elements using e.g. XPATH expressions. An example here is the corresponding expression for the file format test criterion:

```
/fits:identification[@status='UNKNOWN']
```

This expression checks the existence of the kind element *identification* which has the attribute *status* and the value *unknown*. The existence of such an element indicates that FITS was not able to identify the file format. This means that the test criterion for granting ingest level 1 has not been met. As noted above, the incremental increase in the ingest level stops here and the ongoing results analysis is discontinued. The file object is forwarded marked ingest level 0 to the next stage, the ingest level comparison.

3.4 Ingest level comparison

The depositor-dependent format policy is loaded for the ingest level comparison. This sets the minimum ingest level to be reached for each file format. The relation between file format and ingest level is established using the *PRONOM Unique Identifier (PUID)* issued by DROID. For example, if the definition of ingest level 2 is reached for PUID fmt/16, only file objects in PDF format version 1.2 for which

- the bitstream passes the integrity test
- the file format is identified and
- no use restrictions apply

will be ingested into the DNB repository and therefore into the preservation repository. If a publication consists of multiple files, all its elements must meet the set criteria, with the lowest value determining the overall ingest level.

4. PRACTICAL IMPLEMENTATION AND EXPERIENCE

Following on from the description of the basic idea and technical implementation of the risk management issues, the intention below is to present an overview of the experience gained to date.

The system was put into operation in December 2012 as part of the DNB operational processes for handling digital publications. The vast majority of files undergoing the risk management processes since then have been PDF and ePub objects. Figure 3 (date: 12.4.13) shows the distribution of analysed ingest levels. The visualised results show the figures for file objects submitted to the DNB which fulfil the requirements of the DNB internal format policy.

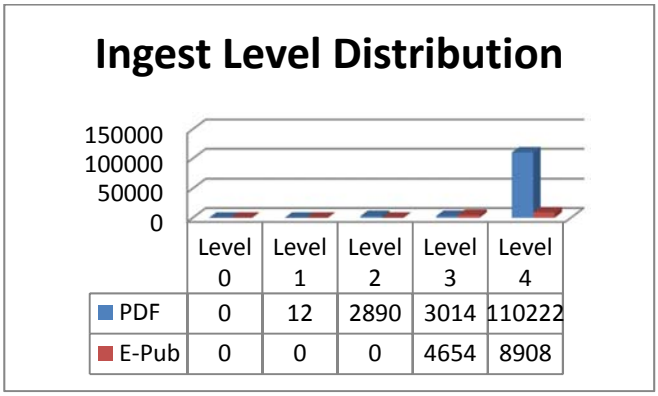

Figure 3 Ingest level distribution (PDF, ePub) in the period from 12/2012 to 04/2013

Of the total of 116,138 PDF files, the vast majority (110,222) are ingest level 4. Even though only 3,014 PDF objects had a validity problem (ingest level 3) and no technical metadata could be generated for 2890 objects (object level 2), this absolute figure is likely to rise and should not be underestimated. Several thousand problematic objects can accrue within just a few years; these need to be prioritised for preservation strategies such as format migration. With regard to the ePub format, only half of all the objects transferred to date are free of validity problems.

It should also be mentioned that the fact that a clear majority of objects are ingest level 4 does not necessarily signify that this majority automatically represents the "object quality" of the publication world. It should be borne in mind that only through the risk management measures and the resulting requests by staff for "better" versions were many objects of ingest level 3 or lower able to be raised to ingest level 4. The return of defective objects raises awareness amongst the publishers of the need to attach greater importance to the quality of their objects. In some cases this has already led to checking tools being integrated in the publishers' production processes. Despite all the automation systems, these costs associated with manual risk management activities, including e.g. necessary adaptations to the format policy, should not be neglected in any cost assessments.

4.1 Technical limits

In many cases, file objects which only achieve ingest level 2 reveal their technical limits in the validation tools used. At present, for example, some PDF variants (e.g. PDF/X) cannot be correctly processed, meaning that the resulting technical metadata deficiencies are not always due to supposedly "poor" object quality.

A clear discrepancy between theory and practice has also emerged in format validity. The differing interpretations of the HTML standards by the panoply of disparate browser providers and the resulting differences in the ways in which a website are displayed are acknowledged examples of this. Additionally, the library's ePub-Analyzer metadata tool which checks conformity of ePub files against the ePub specifications often identifies a lack of schema validity in the toc.ncx file which describes the table of contents. However, practical tests of their display and use on current devices showed that this validity problem is negligible at present. Nevertheless, from the perspective of long-term preservation it represents a significant risk factor which can be dealt with in the preservation strategy planning e.g. by means of suitable corrective measures.

A total of 12 individual ingest level 1 PDF objects are shown in Figure 3, some of which are attributable to different results obtained by the tools operating in FITS with regard to the existence of usage restrictions. In these cases, manual analysis showed that use of the objects was not restricted.

Finally, the ongoing development of file formats for electronic publications poses further demands in terms of constant updating and development of the metadata tools used. During transition periods in which tool support is still incomplete, compromise solutions, e.g. lowering of the ingest level, should be considered.

5. Summary and outlook

The present article examines the DP4lib ingest level system and its practical use in the DNB. This system introduced automatic quality checking to the DNB's long-term preservation activities as part of a comprehensive risk management system. It was shown that risks which are ubiquitous in the file formats of digital materials can be detected and classified at an early stage. The first countermeasures designed to reduce file format risks were the formulation of a format policy and the setting of a limit beyond which the task of ensuring the long-term usability of digital objects can no longer be fulfilled. Initial experience shows that the automatic quality analysis has yielded accurate findings regarding the technical quality of the library's stocks. The data can also be used as the basis of improvement processes and to reduce long-term preservation risks. The ingest level system therefore provides a practicable control instrument based on tangible limits and rules of action. It also allows depositing partners to formulate their own requirements and expectations in terms of object quality and risk analysis, thereby facilitating the creation of service agreements between DP4lib service users and providers. It should be added that this approach has also resulted in a number of terms entering the vocabulary of the specialist and IT departments of the DNB, leading to a corresponding improvement in communication.

In the future it should be established whether the five levels (and their order) in the current ingest level system and the related weighting are sufficient to address the long-term preservation risks for digital publications and the associated problems arising from the growing variety of file formats.

6. References

[1] The Consultative Committee for Space Data Systems (CCSDS), June 2012. *Reference Mode for an Open Archival Information System (OAIS), Recommended Practice.*

[2] Ex Libris Group, 2010. *Ex Libris Rosetta: A Digital Preservation System – Product Description.*

[3] Langzeitarchivierung – Ein Handlungsleitfaden für Dienstleister und Dienstnehmer. http://dp4lib.langzeitarchivierung.de/

[4] PREMIS With a Fresh Coat of Paint http://www.dlib.org/dlib/may08/lavoie/05lavoie.html

[5] Data Preservation in High Energy Physics; David South; Proceedings of plenary talk given at the 18th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2010).

[6] *File Information Tool Set;* http://code.google.com/p/fits/