DRM and digital preservation: A use case at the German National Library

Stefan Hein
Deutsche Nationalbibliothek
Adickesallee 1
60322 Frankfurt am Main, Germany
+49 69 1525 1722
s.hein@dnb.de

Tobias Steinke
Deutsche Nationalbibliothek
Adickesallee 1
60322 Frankfurt am Main, Germany
+49 69 1525 1762
t.steinke@dnb.de

ABSTRACT
Digital Rights Management (DRM) is in use for many digital publications. Digital libraries with a mandate to collect and preserve publications have to deal with technical challenges for preservation of DRM restricted objects. In the European project APARSEN a systematical classification of DRM methods and its risks for digital preservation was introduced. The German National Library handles the different types of DRM protections within the ingest workflow of the archival system by analysis and case-by-case distinction.

General Terms
Management, Measurement, Standardization, Verification.

Keywords
Digital rights management, Digital preservation, Deutsche Nationalbibliothek, APARSEN, Ingest Level

1. INTRODUCTION
Digital Rights Management (DRM) of digital publications like e-books, multimedia disks and audio files could come in many different ways. Restrictions for access and usage are often implemented by technical measures. Typical restrictions are prevention of creating copies or usage in not allowed software environments.

Libraries collect, archive and give access to digital publications. The challenge of digital preservation is dealing with obsolescence of hardware and software. Especially for national libraries with a mandate to preserve their collected objects for an unlimited time, dedicated preservation strategies and actions are needed. File format migration and emulation of old systems environments are common ways to handle the task.

The technical measures of DRM could be a problem for preservation actions. File format migration means converting and copying files, emulation means using an object in another technical environment. Both strategies might be in opposite to the intended restrictions of DRM. There are also other potential problems like dependencies on online sources for verification.

The German National Library, Deutsche Nationalbibliothek (DNB), collects many different types of digital publications within legal deposit legislation. As a partner of the European project APARSEN[1] DNB worked on a systematical approach to classify the challenges that DRM could be for digital preservation.

2. DRM: A CHALLENGE FOR DIGITAL PRESERVATION
Through the integration of proprietary rights control mechanisms as an integral component of digital objects, a new problem has arisen regarding long-term preservation (LTP). The main cause of this problem has been that restrictions of access and usage could hinder the preservation of the object. If access to the content is already blocked, the problems involved in executing LTP measures are clearly apparent. Preservation measures without access to the actual content are not viable. Technical or other types of metadata (e.g. bibliographic) can only – if at all – be extracted to a limited extent from protected files. According to OAIS, however, these data need to be incorporated in the data management and are essential for meaningful preservation planning and the execution of preservation actions ([1]). The encrypted content could also conceal malware (viruses, trojans) which could enter the archive and remain undiscovered by virus scanners.

2.1 Scale for Long-Term Preservation Risk
In order to evaluate the risk of different DRM technologies, APARSEN defined the following scale (Long-Term-Preservation Risk (LPTR)):

Table 1. Long-Term-Preservation Risk (LTPR)

LTPR	Characterization
no risk	No risk for future LTP measures
medium	Possible to use at present (at time of publication) in up-to-date hardware and software environment, current LTP measures restricted, no external dependencies, medium risk for future LTP measures
high	Use and LTP measures already (currently) restricted, high risk for implementation of LTP measures in the future as result of external dependencies

In summary, the higher the LTPR value, the greater the risk in archiving and maintaining the usability of the object concerned.

[1] http://www.alliancepermanentaccess.org/index.php/aparsen/

This appraisal contains a prediction component, meaning that 100% guarantees cannot be offered.

2.2 Classification and Assessment

In the APARSEN "Report on DRM preservation" ([2]) four DRM variants are identified and assessed:

Data carrier copy protection, LTPR = medium: Data carrier migration is a key LTP measure, meaning that the prevention of all activities aimed at separating the data stream from the carrier should be regarded as risky. The data carrier copy protection prevents in principle copying. If the data stream cannot be separated from the data carrier, this carries a high risk for future LTP measures because the necessary players and/or software may no longer be available. Usage is, however, possible at present with common player devices. Depending on the kind of data carrier protection, data carrier migration might be possible with current equipment, albeit with restrictions e.g. a loss of quality in case of a digital-analogue conversion of audio material.

Lightweight DRM, LTPR = no risk: Lightweight DRM (LWDRM) refers to all mechanisms which do not of themselves restrict access to digital objects or their use, but which serve the detection and tracking of legal infringements [3]. This is mostly achieved through the use of marking techniques such as digital watermarks. Digital watermarks may be applied to the digital object in a way which is invisible to the user but which allows the content providers to detect their works e.g. on illegal file-sharing sites. Lightweight DRM involves no restrictions on access or usage. The marking of digital objects therefore poses no risk for use or LTP measures.

Encryption-based password protection, LTPR = medium: This variant focuses on DRM mechanisms which require no connections to external components (such as authentication servers) during use and which basically manage the access and usage possibilities of objects. The term "access" here signifies the opening of a file object using pre-defined player and display software - even though the act of opening could itself be interpreted as the most basic form of use. Use is therefore always conditional upon having access to the object. An example of this is Adobe's PDF format. It contains functions which render access and usage and it is manageable in a variety of forms (like print, edit document, copy content, extract pages). This kind of limitation of use is one of the most common DRM variants that libraries such as the German National Library face, primarily in the context of online publications (e.g. e-books) and dissertations. The access to the data stream and the use of the content is predicated upon knowing the password. The password must be saved separately and linked to the actual content. The user must be given the password when access is granted. If only limited usage rights, such as text extraction, are granted yet the content can still be displayed, it can no longer be predicted with any certainty whether the conversion tool will require precisely this feature in the future. The execution of current and future LTP measures therefore carries risks.

DRM Systems, LTPR = high: This DRM category focuses not only on selected aspects already presented above, but also attempts, by means of a system of diverse components and technologies such as the digital watermarks and encryption methods already examined, to cover all the core DRM areas. The architecture of a DRM system (figure 1) is outlined by Bill Rosenblatt ([4]) and consists of the three linked components of content server, licence server and client. The different DRM components can be geographically distributed and communicate via the Internet. This results in a range of dependencies which can affect everything from generation and content through to use. The client, e.g. the media player or the document reader, therefore no longer functions independently as a gateway to the actual content. It is apparent that precisely this interaction between the different components markedly increases the complexity of DRM systems in comparison to the DRM variants already presented.

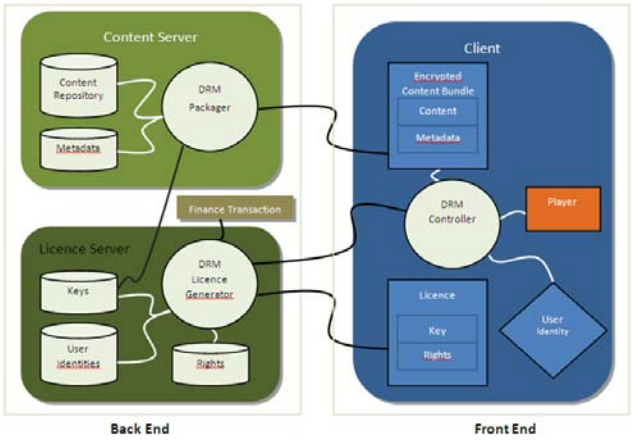

Figure 1, Architecture of a DRM System (adapted from [4])

Given that access to and use of the content is restricted similar to the "encryption-based password protection" variant, objects protected by DRM systems also carry the same risks. A further problem factor is the existence of an external license server, and connection to it is a precondition for encryption. Even today, use may be impaired or prevented entirely in the event of the content provider going out of business, network problems etc.

3. DRM AT THE GERMAN NATIONAL LIBRARY

DNB takes care that all digital publications can be utilized in accordance with legal regulations. Depending on the rights that the content producer grants DNB during the submission process, some publications can be provided in-house only, while others are remotely accessible. DNB receives DRM protected material but does not produce material that is DRM protected. In general publications which are published by the DNB are DRM free. Also DNB advises its deliverers to abstain from the use of DRM mechanism for the delivery to the DNB. In the past DRM mechanism of digital objects were only detected manually. However, no statistical recordings of DRM mechanisms detected were implemented. But it can be assumed that the proportion of DRM protected material has been increasing in parallel to the further development of DRM techniques and format capabilities.

3.1 Data Types

The following data types are occasionally submitted with integrated DRM measures to the DNB:

- Doctoral theses and teaching theses of German universities
- DNB digitized print media
- e-books
- e-journals

- e-papers

The use of DRM techniques and tools depends on the file format and its capabilities, the data type and the publisher. The following techniques were detected so far:

- PDF document restrictions (password protection and print, copy restrictions)
- Adobe's LifeCycle Management (mostly publishers)
- encrypted ZIP container

3.2 Approach

DNB considers DRM measures as a potential risk to fulfill its legal obligation. Since the end of 2012, DNB uses tools to detect DRM measure of digital objects during the ingest process. Before that time the detection was manually done by random sampling.

In accordance with the decision to preserve unaltered originals and to abstain from normalization measures at the time of ingest, the DNB tries to collect the unprotected version of the digital object whenever it is possible.

The approach for online publication contains the decision to refuse "DRM suspicious" material after detection and give the publisher or the delivering institution the possibility to remove the protection for a second delivery. "DRM suspicious" means the existence of DRM techniques which were assessed as medium or high (LTPR).

The Ingest Level concept that is in use at the German National Library leads to provisional rejection of all objects with any kind of DRM ([5]). An Ingest Level is an assigned risk of preservation. This is based on five criteria: file integrity (FI), file format identification (ID), technical restrictions (TR), format specific metadata (MD) and file format validity (V). These criteria are automatically checked within the ingest workflow and an Ingest Level of 0 to 4 is assigned (table 2). Any kind of DRM restriction means level 0 or 1 and a provisional rejection.

Table 2, Ingest Level and criteria

	FI	ID	TR	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

It is, however, not always an option to reject DRM protected objects, respectively, to request DRM free versions, especially when the producer cannot be identified anymore. Furthermore, not every content provider is immediately willing to provide its objects without DRM to the preservation institution.

In these cases, it can only be attempted to create awareness for the problem on the side of the producer / content provider. In the case of the DNB, the legal mandate can be used as an argument. Also the guarantee that the rights will be protected via an institutional access management, so that no disadvantages result from DRM free objects for the content provider, can assist the argumentation. This approach, however, implies additional effort, namely in the implementation of such an access management.

For the automatic detection DNB uses the support of open-source tools. In the case of encrypted ZIP containers the regular unpack routine would report the protection measure. For some time now the automatic generation of technical metadata using metadata tools has been a recognized and established component of the ingest process. The DNB has long been using the File Information Tool Set (FITS)[2] 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 and JHOVE2)[3] tool, the Digital Record Object Identification (DROID)[4] tool and the NLNZ Metadata Extractor. Use of a tool set widens file format support and reduces the risk of errors in the identification and validation of the file format. Some of the above tools (e.g. JHOVE) also permit the recognition of document restrictions such as password-protected PDF files.

As a wrapper for FITS we use a self-developed tool called didigo (diagnose digital objects). FITS is called from didigo for every file and the FITS output of the different analysis tools is used to calculate the Ingest Level. The Ingest Level is compared to the expected value for the files and actions are initiated accordingly.

One result of the automated ingest routine, the provisional rejection of DRM protected objects and the request for re-submission of unprotected material is that the number of ingested DRM protected PDFs in the DNB collection has been very low since the end of 2012: Only 146 PDF documents out of a total of 1,630,600 PDF documents that were ingested between December 2012 and March 2014 are DRM protected (figure 2).

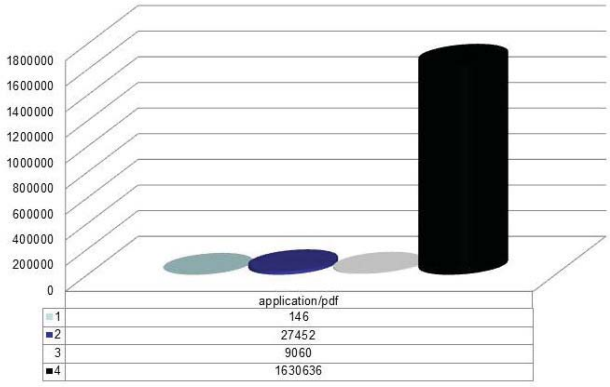

Figure 2, Number of PDF files per Ingest Level

According to its legal mandate the DNB takes preservation actions like migration on archived publications. Where DRM mechanisms inhibit preservation actions, an agreement between the German Publishers and Booksellers Association, the national association of the phonographic industry and the DNB, allows the DNB to remove DRM mechanisms for archival purposes. In particular this is important for post processing the stock of already archived objects, which have unrecognized DRM mechanisms.

As mentioned before, DRM was only detected manually in sample checks between 1998 and 2012. DNB has accepted quite a

[2] http://fitstool.org/

[3] http://jhove.sourceforge.net/

[4] http://www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm

number of DRM protected objects into the archive. This was, however, not documented, and therefore, no statistical figures are available.

During 2014, DNB will re-ingest all "old" objects in the collection with the new, automated ingest workflow that makes use of several metadata tools. With this, DNB will at least, or in a first step, be able to identify the DRM protected objects in its collection for further treatment.

Based on the statistical findings and the DRM analysis, it is possible to plan countermeasures. This will probably become a project in its own right. Where possible, DNB will try to get in touch with the publishers and request re-submission or will try to remove the DRM protection.

3.3 Limits

One limit of the approach of refusing "DRM suspicious" material lays in the limited capabilities of the used metadata tools. So the tools have to be up to date to support new formats and format versions. Unfortunately FITS is not able to determine all variants of PDF restrictions. But if that would be possible another question would arise: Which restrictions are real risks for long-term preservation activities? If the user is not allowed to print the document, it might not be a risk for a conversion in the context of format migration actions. In cases of format transformations a further question still arises as to whether and how such usage restrictions should be preserved.

The alternative approach of removing DRM mechanisms implies many problems in itself. Removing technical mechanisms needs corresponding tools and might change the authenticity of the object. In general it is not easy to acquire a software tool that violates the current legislative. If there aren't any tools or they are not allowed to use the last approach for encrypted documents could be trying every combination of possible password characters. That approach is known as a brute-force attack and is very expensive, because it needs a lot of hardware resources like processor time. For long password lengths it takes a very long time to crack the password, in the worst case the cracking attempts are nearly infinite.

4. CONCLUSION

Technical measures of DRM can be classified in four categories. The most critical category for digital preservation is related to external dependencies like online verification. Local encryption and hardware protection might be a serious threat for preservation actions as well, but there could be ways to maintain access by specific solutions or agreements.

The German National Library uses file analysis tools within the ingest workflow to recognize and categories possible threats for digital preservation. If a protection with high or medium risk is detected the publishers are requested to re-submit the files without protection. Older collected objects with protections could be a problem. DNB has an agreement with the right holders that allows the removal of technical protection measures for archiving proposes, but this was not yet done. In future projects the existing collections will be checked and protected files will be changed if it is possible and feasible.

In general the increase and change of file formats, their implementations and the DRM techniques that they contain are some of the biggest challenges. Therefore it is necessary to keep the used analyzer tools and reading platforms up to date. Furthermore new technologies like tablet PCs and portable e-book readers with new embedded techniques to protect digital rights have to be considered.

It is important to detect DRM measure as early as possible – then there is a good chance to contact the author or publisher for a DRM-free version. The more time has passed, the smaller the chance to get in contact with the rights holder. That increases the risk to have to deal with a restricted version of a publication for preservation and access.

5. REFERENCES

[1] CCSDS, "Reference Model for an Open Archival Information System (OAIS)," June 2012. [Online]. Available: public.ccsds.org/publications/archive/650x0m2.pdf. [Accessed 26 11 2013].

[2] K. Kaur, S. Hein, S. Schrimpf, M. Ras and M. Holzmayer, "Report in DRM preservation", 2014. [Online]. Available: http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=D31.1+Report+on+DRM+preservation. [Accessed 06 03 2014].

[3] R. Grimm and C. Neubauer, "LWDRM - An Alternative Rights Management System," 2004. [Online]. Available: http://waste.informatik.hu-berlin.de/Grassmuck/drm/Folien-Grimm-Neubauer-eng.pdf. [Accessed 25 11 2013].

[4] B. Rosenblatt, "Enterprise Digital Rights Management," 14 July 2005. [Online]. Available: http://www.giantstepsmts.com/Authentica-RMS%20Whitepaper.pdf, pg. 5. [Accessed 06 03 2014].

[5] Schmitt, K., & Hein, S., "Risk Management for Digital Long-Term Preservation Services", from IPRES 2013 : proceedings / of the 10th International Conference on Preservation of Digital Objects: http://purl.pt/24107/1/iPres2013_PDF/Risk%20Management%20for%20Digital%20Long-Term%20Preservation%20Services.pdf. [Accessed 06 03 2014].