The Ties that Bind - On the Impact of Losing a Consortium Member in a Cooperatively Operated Digital Preservation System

Michelle Lindlar
TIB Leibniz Information Centre for
Science and Technology
Welfengarten 1B
30168 Hannover, Germany
+49 511 762 19826
michelle.lindlar@tib.eu

ABSTRACT

Cooperatively operated digital preservation systems offer institutions of varying size the chance to actively participate in digital preservation. In current times of budget cuts they are also a valuable asset to larger memory institutions. While the benefits of cooperatively operated systems have been discussed before, the risks associated with a consortial solution have not been analyzed in detail.

TIB hosts the Goportis Digital Archive which is used by two large national subject libraries as well as by TIB itself. As the host of this comparatively small preservation network, TIB has started to analyze the particular risk which losing a consortium member poses to the overall system operation. This paper presents the current status of this work-in-progress and highlights two areas: risk factors associated with cost and risk factors associated with the content. While the paper is strictly written from the viewpoint of the consortial leader/ host of this specific network, the underlying processes shall be beneficial to other cooperatively operated digital preservation systems.

Keywords

Digital Preservation Services; Digital Preservation Networks; Consortial Systems; Risk Assessment; Exit Scenario.

1. INTRODUCTION

Digital preservation is per definition a risky business – or as Corrado and Moulaison put it: "Ultimately, digital preservation is an exercise in risk management" [1]. Much research has gone into the assessment of risks associated with digital preservation [2]: risks associated with file formats [3][4], risks associated with specific business cases and the application of risk assessment methodologies such as SPOT (Simple Property-Oriented Threat) or SWOT (Strengths, Weaknesses, Opportunities, Threats) to repositories [5],[6]. The focus of these assessments is either content-driven, i.e. focusing on problems specific to certain collections, or institutional repository driven, i.e. considering an institutional repository as a closed ecosystem.

Simultaneously, with institutions facing budget cuts, a growing number of institutions are turning to digital preservation networks, joint system implementations and preservation services such as DPN (Digital Preservation Network) or the MetaArchive Cooperative.

Despite the wide adoption of preservation networks, many supporting digital preservation actions maintain an institutional repository fixed view. Certification processes, for instance, such as the Data Seal of Approval, the nestor seal or the TRAC Audit process usually audit the participating institutions separately, even if they are participating in a single central digital preservation repository. This leads to a distinct blind spot regarding consortial management. A central question not answered by this approach is the following: what happens, if an institution leaves the consortia? While it can be assumed that the impact highly depends on the overall size of the consortia, the risks associated with an institution leaving touch on different areas and should be evaluated carefully.

Preservation networks as well as collaboratively operated systems range from *small networks* of 2-5 institutions, such as that of the National Library of New Zealand and Archives New Zealand in the National Digital Heritage Archive [7], to *mid-sized networks* of 6-20 institutions which are often found at the regional or state level, such as DA-NRW, the digital archive of North-Rhine-Westphalia in Germany[1], to *large national or international networks* with over 20 institutions, such as DPN – the Digital Preservation Network[2] – with over 60 members. More importantly, networks and collaborations differ in modi operandi regarding overall available preservation levels as well as responsibilities. In order to adequately assess the impact a leaving institution has on a consortia, a first requirement is thus a categorization of the jointly operated system.

1.1 Categorization of Cooperations

Terminology such as "digital preservation network", "digital preservation collaborations" and "digital preservations services" have been used loosely, leading to no distinct boundaries between infrastructural and service levels associated with the terms. However, to fully understand the work conducted by a participating institution versus that being taken care of by a host or service provider, infrastructural and personal responsibilities need to be defined. Unfortunately no clear categorization schema exists as of today, leading to often misleading communication about networks, collaborations and jointly operated digital preservation systems.

The cost impact analysis put forth in section 2 of this paper uses the Curation Cost Exchange (CCEx) breakdown of digital preservation activities and resources. The author proposes to use this breakdown to further categorize jointly operated digital preservation systems, preservation networks and preservation services. To achieve this, the four CCEx service/activity categories Pre-Ingest, Ingest, Archival Storage, Access[3] – are used and further divided into the resource layers "Infrastructure" and "Preservation Management". Infrastructure can be mapped to the CCEx "Cost by Resource" classification as containing purchases[4] and support/ operations staff (see Staff - Support Operations in Table 3). Similarly, Preservation

[1] https://www.danrw.de/

[2] http://www.dpn.org/

[3] See Table 2

[4] see a)i), a)ii) and a)iii) in Table 3

Management can be mapped to the CCEx "Cost by Resource" classification as containing Producer and Preservation Analyst staff (see Staff - Producer. and Staff – Preservation Analyst in Table 3). To further exemplify: "Preservation Management" includes any human task associated with the digital object (as opposed to the preservation framework) along its lifecycle. This includes tasks such as defining packaging and mapping at the pre-ingest level, conducting deposits and handling errors occurring in file format validation steps at the ingest level, preservation planning and action at the archival storage level as well as defining DIPs (dissemination information packages) and access rules at the access level. Human tasks supporting the maintenance of the digital systems, such as system and network administration is captured on the infrastructural level.

The derived criteria are listed in the first column of Table 1. In a second step, each criterion is either assigned to the host level, meaning that the hosting or leading institution/ entity is responsible, or to the participating institution level. Table 3 shows a thus completed categorization view for the Goportis Digital Archive.

Table 1: Categorization of the Goportis Digital Archive. The criteria are based on the CCEx categories.

Criteria	Reponsibility
Pre-Ingest – Infrastructure	Participating institution
Pre-Ingest–Preservation Management	Participating institution
Ingest - Infrastructure	Host
Ingest – Preservation Management	Participating institution
Archival Storage – Infrastructure	Host
Archival Storage - Preservation Management	Participating institution
Access - Infrastructure	Host
Access – Preservation Management	Participating institution

1.2 The Goportis Digital Archive

TIB hosts the cooperatively operated digital preservation system for the Goportis consortium. The consortium consists of the three German national subject libraries: TIB Leibniz Information Centre for Science and Technology, ZB MED Leibniz Information Centre for Life Sciences and ZBW Leibniz Information Centre for Economics. Furthermore, TIB is currently designing a preservation-as-a-service offer for smaller institutions. The three Goportis partners finance the digital preservation system and the human resources responsible for it from their own resources, which are firmly fixed in each cooperation partner's annual budget. The costs of jointly operating the system are currently borne equally by all three institutions. Each partner has its own digital preservation team that is firmly embedded in each institution's structure and organisational chart. TIB is the Rosetta software licensee, hosts, operates and administers the digital preservation system, and provides Goportis partners access to the system. Use and operation are regulated in cooperative agreements between TIB, ZB MED and ZBW. [5]

Reflecting on the categorization put forth in Table 1, TIB covers both roles – participation institution, as the system is used for its own holdings, as well as host. It is important to stress that this paper is only written from the viewpoint of the host role.As the Goportis consortia falls into the smallest scale of networks, it is of utmost importance to check the impact which losing an institution would have on the network.

This paper puts forth first results of TIB's analysis of risks associated with an institution leaving the consortia. The following sections highlight two key areas of risks: risks associated with the overall cost of the consortial operation of the Goportis Digital Archive and risks associated with the content belonging to the different institutions. The sections describe how the analysis was conducted and for both areas, cost and content, concrete risks are described including an impact evaluation as well as a first suggestion for mitigation strategies. While the sections 2 and 3 describe the analysis strictly from the viewpoint of TIB as the host of the consortial operation, the final conclusion and outlook section will touch on the relevance of this work to other institution and outline next steps which TIB intends to take.

2. COST RISKS

The last decade has seen a lot of research toward the cost of digital preservation [8]. While most institutions still show reluctance towards sharing cost information [9], various cost models have been put forth which allow institutions to evaluate their own cost requirements. For the evaluation of cost in the consortial context, the cost breakdown of the 4C project's CCEx (Curation Cost Exchange)[6] platform was chosen as it is based on a gap analysis of prior cost model work done in other major projects such as LIFE³ and KRDS (Keeping Research Data Safe). CCEx allows the institutions to define a cost unit, and to allocate the total cost of that unit twice: once by service/activities and once by resources (purchases and staff) [9].

The breakdown for cost by service/activities can be taken from Table 2, which indicates the relevant criteria for TIB as the hosting institution (see also Table 2).

Table 2: CCEx Service/Activity levels and corresponding responsibility level of TIB as the hosting entity of the Goportis Digital Archive

	Service/Activity	Goportis Digital Archive responsibility
1.)	Pre-Ingest	none
2.)	Ingest	Infrastructure
3.)	Archival Storage	Infrastructure
4.)	Access	Infrastructure

Within the Goportis Digital Preservation System Pre-Ingest work is strictly done within the partnering institutions' infrastructure. Data is transferred to the TIB environment for Ingest – relevant system architecture parts for the Ingest process are the network connection to the partnering institutions, allocated transfer storage as well as allocated operational storage which the digital preservation system requires for system internal ingest processes such as technical metadata generation. The archival storage is kept separate from the operational storage and keeps 2 copies plus backups. Automated processing mainly takes place during ingest and preservation action, including (re-)identification processes for file formats or the (re-)running of fixity checks. The system is currently operated as a dark archive and access only takes place for proof-of-concept purposes, for checks done by preservation

[5] See Goportis Digital Preservation Policy, available at: http://www.goportis.de/en/our-expertise/provision-of-scientific-content/digital-preservation/goportis-long-term-digital-archives-preservation-policy.html

[6] http://www.curationexchange.org/

staff or for trigger-based manual delivery of objects in case of corruption or loss of the access copy in use within external access systems. Due to this clear understanding of the resources currently used for the different activities, we can derive a rough estimate of cost percentage dedicated to the different services, as shown in Figure 1.

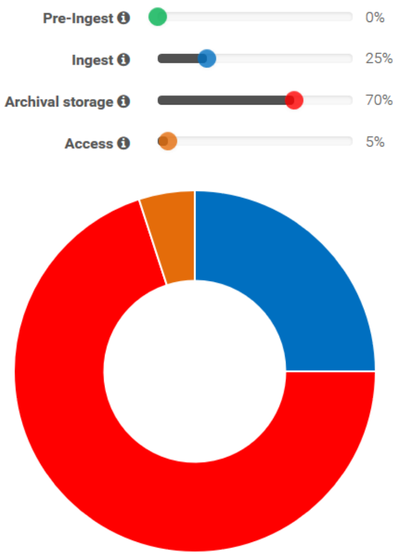

Figure 1. Estimate of cost breakdown by activity

The breakdown of cost by resources is hown in table 3. Here, the responsibility is matched to either to TIB as the host of the digital preservation system or to one or several of the participating institutions.

Table 3: CCEx Reource levels and corresponding responsibility level of TIB within the Goportis Digital Archive

Cost category	Cost	Responsibility
1.) Purchases	a) Hardware	Host
	a.) Software	Host
	b.) External or third party services	Shared by participating institutions
1.) Staff	a.) Producer	Participating institutions
	b.) IT developer	Participating institutions
	c.) Support/ operations	Host
	d.) Preservation Analyst	Participating institutions
	e.) Manager	Host
2.) Overhead	a.) Overhead	Host

While the hardware used has already been described in the analysis of "cost by service/activities", the software used is the proprietary digital preservation system "Rosetta" by Ex Libris for which the consortium shares the license cost. Further third party tools or services are currently not in use.

Within the digital preservation system the partnering institutions conduct the deposit, preservation planning and preservation action for their own content. Furthermore, each institution has full access to APIs[7] which allow the extension of the system to institutional needs. Development capacities within the institutions range between 0.25 and 1 FTEs (full-time equivalent). While developments may be used more than one institution, for example the development of a proxy mapping metadata imported from the union catalogue to the descriptive metadata, currently no dedicated consortial extension exists and the IT developer resource does not count towards the "consortial operation" cost unit. Support/operations, however, caters to all three partnering institutions. In addition to 1 FTE for system administration approx. 0.25 FTE go towards support of the partnering institutions for daily system operations including communication with the system vendor's support. Managerial work includes organizational coordination between the three institutions while overhead accounts for fixed costs such as office and server room space and electricity.

In addition to the cost unit break-down, CCEx requests a breakdown of the digital assets including an indication of type, size and volume [9]. As archival storage makes up a large cost factor, this analysis will be conducted per institution in the near future.

The break-down of the cost unit "consortial operation" by services/activities and resources allows for a good understanding of cost factors. Based on the high-level analysis, three cost risks can be determined, which are briefly discussed below: hardware/ infrastructure, software licenses and staff.

2.1 Hardware / Infrastructure

2.1.1 Risks

The estimate has shown that archival storage needs account for a large section of the overall costs. The requirements in archival storage size are naturally mandated by the archived content of the partnering institutions. In case of an institution leaving the consortium, the used storage space would be freed and would currently not be needed. The potential risk is that the infra-structure could be oversized for the existing requirements of a changing consortium constellation.

2.1.2 Impact

Impact depends on the overall size of the repository as well as the holdings and growth rates per institution. In the case of the Goportis digital preservation system the impact can currently be described as "low", as the freed storage can be easily allocated to the other two institutions without oversizing the repository or institutional storage allocation. Furthermore, TIB's infrastructure would allow free storage not used by the digital preservation system to be allocated to different services.

2.1.3 Mitigation Strategy

In addition to the CCEx recommended breakdown of digital assets in the as-is state, a prognosed growth rate per institution is collected on a yearly basis. It is advisable that the prognosis interval matches the notice period of the partnering institutions.

Furthermore, the break-down analysis of the cost-unit "consortial operation" shall be re-run once a year to check against new risks which can arise due to new requirements such as access to an institution's light archive collection.

2.2 Software Licenses

2.2.1 Risks

While a breakdown of purchase cost is currently not available, software vendor cost is always a key factor. The risk exists in form of license and support costs not tied to a specific number of institutions. In that case, an institution leaving the consortia

[7] https://developers.exlibrisgroup.com/rosetta/apis

would leave the remaining institution having to cover higher license and support costs.

2.2.2 *Impact*

Impact depends on the software license and support agreement, on the licensing and support cost as well as on the consortia size. As the Goportis consortium only consists of three institutions, the impact is defined as “high”.

2.2.3 *Mitigation Strategy*

Include scenarios for changing consortia constellations and varying consortia sizes in the vendor contract.

2.3 Staff

2.3.1 *Risks*

The majority of staff for the consortial system goes towards system administration with additional requirements for support/operation and managerial tasks. The risk exists in form of staffing requirements being oversized when an institution leaves the consortia.

2.3.2 *Impact*

Impact depends on the overall size of the consortia and the staffing requirements based on that. In the case of the Goportis digital preservation system, support/operation as well as managerial tasks are covered by various TIB digital preservation team members who also perform institutional digital preservation tasks. The system administration FTE is required regardless of the size of the consortia. Due to this, the impact on staff can be described as “low”.

2.3.3 *Mitigation Strategy*

Staffing requirements for consortial operation shall be re-evaluated on a yearly basis to check for changing risks. Spreading out support/operation and managerial tasks across different staff minimizes the risk of an oversized team structure.

3. CONTENT RISKS

An institution leaving a consortia is a concrete exit scenario. A solid exit strategy is an integral part of every digital preservation system. Certification processes such as TRAC [10], the Data Seal of Approval [11] and the nestor seal [12] require or recommend that exit strategies be in place. However, certification guidelines do not give concrete description of what exit strategies should contain. Instead, the strategy is usually considered evidence of appropriate succession and contingency plans. Commonly, the use of systems which support open standards is seen as a pre-requisite for an exit strategy [1]. However, current descriptions of exit scenarios usually pertain to the situation where an existing institutions exits from one system into another. Contingency plans covering the institution’s demise usually only focus on technical requirements for data export, such as completeness and open formats, as well as extensive representation information to allow for adequate interpretation of the digital objects. Legal aspects are highly specific to the jurisdiction of the archive and are less frequently covered in exit strategies [13][1].

As opposed to a system-wide exit scenario, a consortially operated system calls for a tiered exit scenario which clearly allows for the export and interpretation of the data pertaining to a single institution. Furthermore, two scenarios need to be considered: the institution exits because it leaves the consortia but continues to exist and the institution exits because it ceases to exist. In the latter case, the data may need to be handed over to a third-party which leads to different legal requirements and implications.

These legal implications as well as standard exit scenario requirements lead to four risks associated with the content of an institution leaving a consortium. These risks are further described in the following subsections.

3.1 Export of Institutional Data

3.1.1 *Risks*

In the case of an institution exiting a consortium the repository needs to be able to export and delete the institution’s data from the repository while leaving the data of the remaining institutions intact. The risk is that the repository is either unable to select the objects and their associated metadata per institution and/or that the exported data is incomplete or not interpretable outside of the digital preservation system.

3.1.2 *Impact*

This risk exists for any consortium, regardless of size or makeup. As the repository operator would not be able to fulfill a fundamental requirement of a trustworthy digital preservation system the impact has to be defined as “high”.

3.1.3 *Mitigation Strategy*

A consortial system shall clearly differentiate between the different institutions from the start. Ideally, different data management interfaces exist for the different institutions. Workflows shall be completely separated and the objects’ accompanying metadata shall clearly include the institution as the content owner. Additionally, separate storage locations should be set up for each institution.

3.2 Documentation of Institutional Processes

3.2.1 *Risks*

Preservation processes may include documentation which is not directly stored within the repository. Examples for this are full license agreements between a depositor and the institution. While the license text may be included in rights metadata, the signed agreement is usually stored in a rights management system or resides as a hard-copy within the institution. Another example is supporting documentation for a preservation plan.

While not directly available within the repository, this information is still essential for interpretation of the digital objects across their lifecycle. Especially in the case where an institution exits the consortium due to its demise and the digital objects are to be handed over to a new steward, either a consistent link to external information or, ideally, the entire information itself, shall be provided in a data export.

3.2.2 *Impact*

The impact is especially “high” for the archiving institution as well as for a potential third party who takes over as a steward of data in the case of the institution’s demise.

3.2.3 *Mitigation Strategy*

Consortia wide policies shall be in place to regulate the availability of complementary information for all preservation workflows. Where it is not possible to store the information in the repository, a clear description of where to find the information must be given.

3.3 Non-transferable Rights

3.3.1 *Risks*

No risk exists if an institution exits and requests an export of their objects to store in a different system or locally. However, the situation is different if an institution exists because it ceases to exit. In that case, a new steward for the institution’s objects needs to be found and the consortium leader may therefore have to pass the objects on to a third-party. The risk here resides in often non-transferable rights of digital objects [14].

3.3.2 *Impact*

The impact is particularly “high” for a future steward of information which previously belonged to an institution which ceased to exist. Unless the objects are licensed under a public license, the license will have to be re-negotiated between the creator and the data steward. This becomes particularly hard if the information provided about the creator alongside the object is only rudimentary.

3.3.3 *Mitigation Strategy*

While there is no solution for non-transferable rights, the situation can be improved by including further information about the creator. Here, particularly contact information such an email address is helpful. Also, the availability of the full original license agreement, as described in section 3.2, is beneficial.

3.4 User Names in Metadata

3.4.1 *Risks*

As part of PREMIS based preservation metadata generation, the Goportis Digital Archive gathers information about agents. These agents can be software as well as users. If a user acts as an agent, the username is captured in the metadata. If a user performs a deposit, additional information such as the full name, work address and email are captured. Full address information of the user is also included in the user's profile.

In Germany the use of personal data is protected by the BDSG (Bundesdatenschutzgesetz) law book. BDSG §20 states that public institutions – such as the three Leibniz information centres belonging to the Goportis consortia – are required to delete personal data of their employees as soon as this data is no longer required to fulfill its original purpose [15]. As in the case of non-transferable rights this becomes especially a problem when an institution exits due to its demise and the objects and their accompanying metadata are to be handed over to a third-party as the new data steward. Since the preservation metadata is an integral part of the AIP to be handed over, all user data captured within would need to be anonymized or pseudonymized.

3.4.2 *Impact*

As described above, the impact is "high" if the objects need to be handed to a third party who becomes the new data-steward.

3.4.3 *Mitigation Strategy*

An overview of where user data is captured within the metadata shall be prepared to assist in an anonymization process. It needs to be evaluated if pseudonymization is preferable, e.g. by substituting user names by a fixed set of roles. The understanding of what role triggered an event within a workflow may assist a third-party institution in better interpreting the preservation metadata as well as the lifecycle events it describes.

4. CONCLUSION AND OUTLOOK

While the analysis of the impact which an institution leaving the consortium imposes is still a work-in-progress, this paper put forth a first analysis of risks associated with the overall costing of the cooperatively operated digital archive as well as of risks associated with the content of the institution exiting.

In regards to the cost analysis, the CCEx tool proved to be extremely helpful in analyzing affected cost segments. Here, further work will be invested in two tasks: (a) gather information to allow for a better differentiation between economic and non-economic cost factors[8] and (b) a detailed analysis of the holdings per size, type and volume for each institution including effective growth over the past two years and prognosed growth for the next year

Regarding the content analysis, the results made clear that the extent of on object's description in its lifecycle – especially when the lifecycle shall foresee a transfer to a different data steward – are wider than anticipated. The two take-aways here are: (a) the Goportis digital preservation policy should be checked towards including further information regarding the availability of relevant object lifecycle information currently not stored in the repository and (b) the export of all institutionally relevant data shall be checked regularly including a strategy to anonymize or pseudonymize the user data captured in the preservation metadata.

Also, further work will go into the identification of other impact areas. The impact on "shared knowledge and efforts" is one which is currently not yet covered. For example, the Goportis Digital Archive shares networking activities and maintains a wiki to exchange results. Losing a partner would impact this form of knowledge aggregation.

The analysis in this paper was strictly conducted from the viewpoint of TIB in its role as the consortial leader and host of the Goportis Digital Archive. As such, the situation evaluated was that of TIB losing a partnering institution. Needless to be said the situation would be completely different if the institutions would lose their consortial leader and host. Despite the specific use case given here in form of a small network of three large national subject libraries, the identified risks shall apply to preservation collaboration or networks of different make-up and size. An analysis of the cost unit "consortial operation" for a different network will most likely lead to different distribution results regarding service/activities and resources as other networks may very well include pre-ingest work or share IT development resources. However, the risk breakdown of "hardware", "software" and "staff" appears to be a universal one and while the impact may of course differ, the briefly sketched mitigation strategies may be used a basis for own work. The impacts of the content and the associated risks seem to be universal regardless of preservation makeup and size. While legislation differs from country to country, the transferability of rights and the requirements to anonymize user data should still be checked.

5. REFERENCES

[1] Corrado, E., Moulaison, H.L. 2014. *Digital Preservation for Libraries, Archives, and Museums.* Rowman & Littlefield., Lanham, MD.

[2] Dappert, A. *Risk Assessment of Digital Holdings*. TIMBUS Project presentation. http://timbusproject.net/documents/presentations/9-risk-management-and-digital-preservation

[3] Graf, R., Gordea, S. 2013. A Risk Analysis of File Formats for Preservation Planning. In *Proceedings of the 10th International Conference on Preservation of Digital Objects*. IPRES2013. Lisbon, Portugal.

[4] Graf, R., Gordea, S. 2014. A Model for Format Endangerment Analysis using Fuzzy Logic. In iPRES 2014 – *Proceedings of the 11th International Conference on Preservation of Digital Objects*. iPRES 2014. Melbourne, Australia.

[5] Innocenti, P., McHugh, A., Ross, S.. 2009. Tackling the risk challenge: DRAMBORA (Digital Repository Audit Method Based on Risk Assessment), In *Collaboration and the Knowledge Economy: Issues, Applications, Case Studies*. Cunningham, P., Cunningham, M. (Editors). Stockholm, Sweden.

[8] EU legislature requires publically funded institutions to clearly separate economic and non-economic activities in financial reporting. Non-profit entities need to have a detailed auditing for all processes going towards services such as hosting.

[6] Vermaaten, S., Lavoie, B., Caplan, P. 2012. Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment. *D-Lib Magazine*. Volume 18, Number 9/10 (September/October 2012).

[7] Archives New Zealand. 2010. *Government Digital Archive – "Rosetta" Gap Analysis – Update*. Version 2.0.Technical Report. 18th August 2010.

[8] 4C Project. 2013. *D3.1 – Summary of Cost Models*. Technical Report. 4C Project.

[9] Middleton, S. 2015. *D2.8 – Curation Costs Exchange*. Technical Report. 4C Project.

[10] CRL/OCLC. 2007. *Trustworthy Repositories Audit & Certification: Criteria & Checklist*. Technical Report. CRL/OCLC. Chicago, IL and Dublin, OH.

[11] Data Seal of Approval Board. 2013. *Data Seal of Approval Guidelines version 2*. Technical Report.

[12] nestor Certification working group.2013. *Explanatory notes on the nestor Seal for Trustworthy Digital Archives*. Nestor-materials 17. nestor.

[13] Schaffer, H. Will You Ever Need an Exit Strategy? In *IT Pro*. 4-6. March/April 2014.

[14] Euler, E. 2011. *Digitale Bestandserhaltung und Distributed Storage in LukII.* Legal report. DFG

[15] Federal Republic of Germany. *Bundesdatenschutzgesetz (BDSG)*. In der Fassung der Bekanntmachung vom 14.01.2003 (BGBl. I S. 66). Zuletzt geändert durch Gesetz vom 25.02.2015 (BGBl. I S. 162) m.W.v. 01.01.2016