CESNET Technical Report 2/2012

Long Term Preservation of Digital Data—Background Research

Jiří Kremser, Božena Kováčová, Marta Pavlovská, Lukáš Hejtmánek, and David Antoš

Received 28. 3. 2012

Abstract

With increasing number of digitized documents as well as the content born digital, the need for preserving data for next generations is becoming more and more important. For this purposes, LTP (Long Term Preservation) systems are being developed. The main part of this report describes the outcome of our questionnaire study on LTP systems that was performed during the second half of 2011. The study discusses types of systems deployed in memory institutions and their main features. The report also briefly discusses properties of selected LTP systems and gives a literature search related to the systems as a reference, as well as a general literature search in the LTP field.
Keywords: long term preservation, data preservation

1 Introduction

Long Term Preservation (LTP) of digital data is usually described as a set of processes and tools to ensure availability and readability of the data over time. LTP is not limited to preserving the data bit-stream free of errors, it also includes ability to interpret the data. By long term, we mean time span long enough to observe significant changes of technology and data formats. Long term can be understood as "indefinitely."

As time goes by, many risks are threatening data integrity. Some of the main threats are: deterioration and/or obsolescence of storage media, obsolescence of the data format or in general the software needed for running the application for accessing the data, e.g., an old operating systems. The same risk is related to hardware architectures.

Ensuring long term data preservation is a difficult problem which is handled by non-profit memory institutions, such as libraries, museums, archives and other cultural heritage institutions.

The purpose of this survey is to give an overview of systems deployed in memory institutions to handle the LTP problem. The study is divided into three main parts. The first one was intended to obtain relevant information from the memory institutions themselves by means of a questionnaire, that was

completed and processed during the second half of 2011. The questionnaire studies deployed systems, expected features of such systems, and the overall architecture of LTP processing used by the institutions.

The second part of the study briefly describes the most often used systems in order to give and idea about their structure and functionality. Finally, the third part is a literature search collecting general sources that are related to the LTP area, as well as literature on the systems of interest discussed in previous parts of this work.

2 The Survey

Our survey shall provide answers to following questions:

1. *What LTP (backend) systems are used by librarian institutions?*
2. *What are the expected features of such systems?*
3. *Which of the systems obtained from the question 1) is the most suitable and satisfies all the demands from the question 2)?*

This pieces of information were gathered by means of asking experts of librarian institutions around the world.

2.1 Questionnaire Processing

We have prepared a questionnaire with 19 questions discussed in Section 3. This questionnaire was sent to the libraries around the world: to the North America, South America, Australia, Africa, Asia and Europe. The distribution list for the questionnaire was prepared mainly through searching the web—using official contacts of libraries, either email addresses or online contact forms. The list also contained several direct personal contacts recommended by colleagues from the National Library of the Czech Republic, taking into account that contacting institution representatives based on personal recommendation increases the chance of retrieving relevant information.

2.2 Institutions

We have contacted 109 libraries of several types in total. We have received 20 relevant responses, the responding institutions are summarised in Table 1. 13 responses were from National Libraries, 5 from university libraries and 2 from other institutions (see Fig. 1). Most of the participants are from Europe, then America and Australia. We should note that the MetaArchive is a special case, not being a classical library. It is a group of 48 members forming the distributed network of archives.

We have also received several short replies stating that a particular library does not operate an LTP system, being obviously irrelevant for our survey. Those answers are not included in the results.

We are very thankful to all respondents.

Table 1: Institutions acronyms and types

Full name of the institution	**Acronym**	**Type**
German National Library	GERNL	national
National and University Library, Slovenia	NUK	university
National Library of New Zealand	NLNZ	national
National Diet Library, Japan	NDLJ	national
The Church of Jesus Christ of Latter-day Saints	CJCLS	other
University Library, Ghent, Belgium	UGENT	university
The Royal Library in Copenhagen, Denmark	RLCD	national
The Library of Congress	LOC	national
National Archives of Australia	NAA	national
Stanford University Library	STANFORD	university
State Library of Victoria	SLV	national
Educopia Institute, MetaArchive Cooperative	METAARCH	other
Harvard University Library	HARVARD	university
Columbia University	COLUMBIA	university
Bibliothèque nationale de France	BNF	national
German National Library of Science and Technology	TIB	national
State Library of South Australia	SLSA	national
State and University Library, Aarhus, Denmark	SULAD	national
Centre Informatique National de l'Enseignement Supérieur	CINES	national
Biblioteca Nacional de España	BNE	national

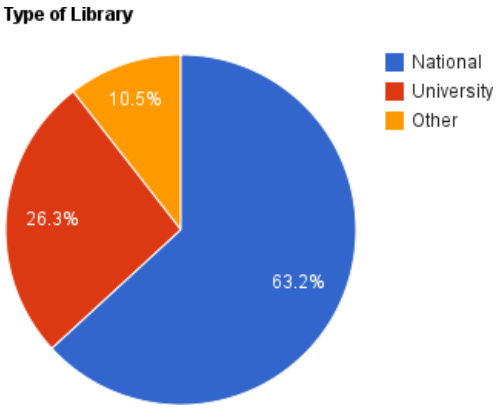

Figure 1: Types of Libraries

3 The Questionnaire and Its Outcome

This section discusses the questions and the responses in detail. The questionnaire consists of 19 questions, the full text of the initial query mail is given in Appendix A.

Many questions are related so the answers are meaningful to be discussed together. Results of Q1, Q2 and Q3 (asking for the basic information of the

system itself) will be presented together, as well as Q9, Q10 and partly Q5 (OAIS and other standards).

As we are aware of the situation that many libraries do not have LTP experts, the questionnaire is constructed to guide the respondents, even slightly compromising accuracy. Particularly, Q4 asks whether the institutions uses the LTP system also for the exposing the stored data to the public. Strictly speaking, such a system is not directly related to the LTP. Q4 also discusses performance in order to get the idea of required quality of service of such a system.

3.1 Q1–Q3

The first three questions are strongly related, and some institutions responded all the questions in an identical manner.

1. *What kind of systems do you use to ensure long-term data preservation?*
2. *Does your institution use any LTP systems (Rosetta, Tessella, Archivematica)? If not, are you planning to deploy some form of an LTP system in the future?*
3. *If so, what technical solutions stays behind it (home developed, iRODS, etc.)?*

See Figure 2 for the overview of results.

Many institutions have their own in-house developed solution (7 out of 20). Of them, the BNF utilizes the iRODS for the data storage. SULAD, COLUMBIA and UGENT use system built on top of Fedora Commons repository. NLNZ, CJCLS and TIP use pure Rosetta or systems built on top of it. STANFORD and METAARCH use LOCKSS, and the others have not deployed the system for long term preservation yet or are in the phase of planning or developing their in-house systems.

3.2 Q4

Do you use your LTP system directly for serving the user copies to the public or is there any system for accessing the user copies in the middle and your LTP stores only master copies? If the later option corresponds to your situation, how often do you synchronize the content of your LTP with the system for exposing the digital objects to the public? Is the performance (throughput/access time) of the LTP system a key quality in your infrastructure?

Libraries using a repository for user access are GERNL, NUK, HARVARD and SULAD. Libraries using LTP system directly are NLNZ, NDLJ, CJCLS and CINES. NAA, METAARCH, STANFORD, TIB, and SLSA have no public access component related to their LTP systems.

UGENT responded: "LTP system is currently a backend system containing master copies. From this system public datasets can be exported to our image databases and public full-text repositories. The metadata is synchronized with the Aleph 500 catalogue on a weekly basis."

RLCD responded: "The current DOMS system is separated in a LTP Master archive and an On line Master presentation archive. The two archives [haven't] been synchronized yet. Mainly because no one of the presentation formats have

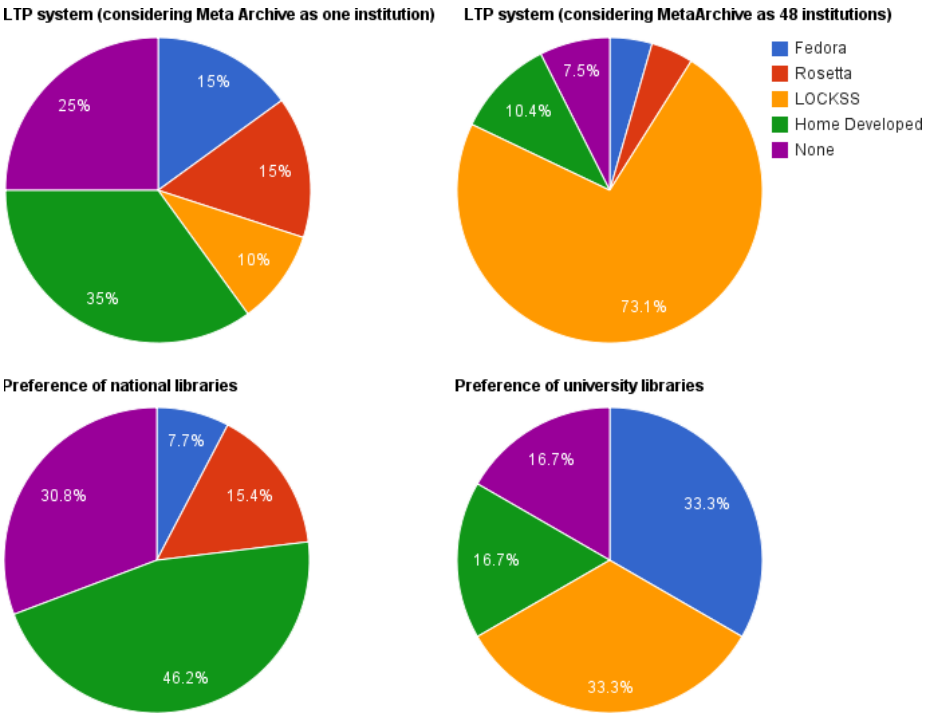

Figure 2: LTP systems (considering Meta Archive as one and/or 48 institution(s))

been changed and because no one of the objects on the On line Master archive has been lost."

LOC responded: "Our goals are to provide both content delivery and content management services that meet the needs of the users and provide secure data management without requiring the users to have to be technically knowledgeable of the specific technologies or components underlying the services. Business and content owners determine the workflows between content delivery and content management, depending on their specific functional requirements."

COLUMBIA responded: "The LTP only stores master copies. If a new version is produced, then a new master copy is produced for the new version. Therefore, synchronization is not necessary and the performance of the LTP is not an issue."

BNF responded: "User copies are stored separately. The master copies of our digital documents usually go through two different processes to preserve them on the one hand, and to make them available to end users in a more appropriate format on the other hand."

BNE responded: "Our LTP system will be strictly related with preservation (for master files and a copy of dissemination copies which will not be accessed by end-users)."

3.3 Q5

Do you have your LTP system certified as a trusted repository (TRAC, NESTOR) or do you plan a certification?

15 of 19 respondents (79%) do not have a certification, 3 (16%) do. One institution has no LTP system (5%). Also compare from Fig. 3. Among the 15 institutions with no certification, there were 2 which wrote that had done an internal self-audit with DRAMBORA [17]. 7 institutions are making plans to get any kind of certification: 2 of them are planning NESTOR [16]; 3 will have TRAC [69], and other 2 did not clarify. There was only one institution (CINES) having both certifications. Other 2 institutions (MetaArchive, Columbia University) have TRAC certification.

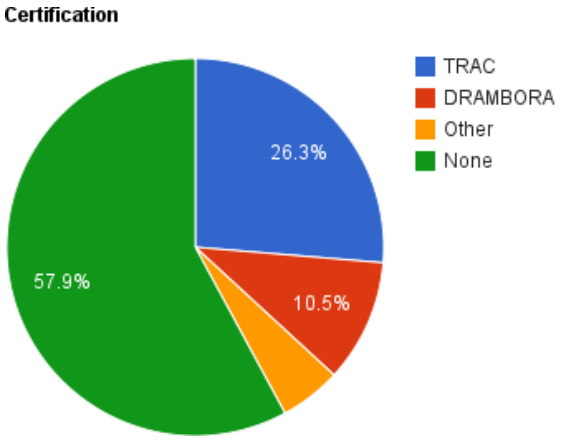

Figure 3: Certification

3.4 Q6

What technical solutions stays behind the system for preservation (home developed, SAN, etc.)? What kinds of HW technologies do you use for storing the master copies (disks, tapes, hybrid solutions, etc.)?

Eight respondents use both disks and tapes (43%; namely NLNZ, RLCD, BNF, TIB, SULAD, CINES, BNE, SLSA). Disk-only solutions are used by 5 institutions (26%; namely NUK, STANFORD, SLV, METAARCH, HARVARD), only 3 institutions use just tapes (16%; namely GERNL, CJCLS, LAA), one institution (5%; namely COLUMBIA) uses only optical discs (DVD/Blu-ray). One institution combines disks and network storage (5%, namely UGENT) and one institution combines optical disks and tapes (5%, namely NDLJ). For more details, see Figure 4.

3.5 Q7

Would you prefer one geographic location where the actual data is stored or some kind of more geographically distributed approach keeping in mind risks of physically destroying the site, e.g., by a natural disaster?

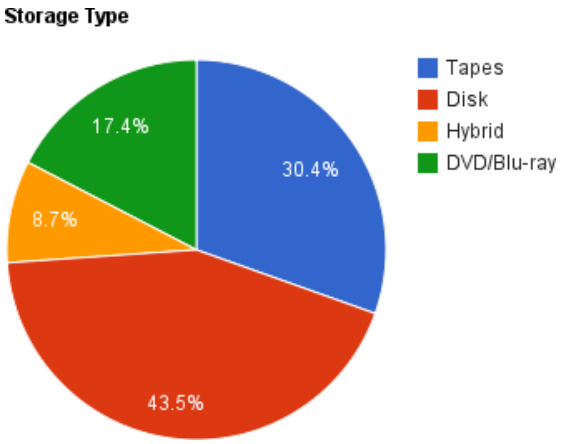

Figure 4: Storage type by institutions

The majority of respondents (18 institutions; namely GENRL, NDLJ, CJCLS, UGENT, RLCD, LOC, STANFORD, METAARCH, HARVARD, COLUMBIA, BNF, TIB, SLSA, SULAD, CINES, BNE, NLNZ and NUK) prefer geographically separated locations to store their data.

One of these 18 institutions (UGHEND) stated they prefer geographically separated locations, but not because of fear for natural disaster but more because of technology/operator failures.

NAA states that their systems are designed so that the two sets of storage can be geographically separated, but currently both are in the same room. SLV replied that "[the] [i]ssue [is] not discussed at this stage".

3.6 Q8

What are the main pros and cons of your LTP infrastructure (rather HW infrastructure questions than functional requirements of the LTP system)?

Answers to this question are very varying. Most institutions, however, agree as follows. As pros, the most repeated reasons are security. Flexibility and customization are also important. In addition, there are answers: automated ingest; tamper-resistant fixity checking; decentralized without single points of failure; use of low-cost simple servers; versioning capability; scalable and flexible storage; redundancy; scalability; geographically distributed infrastructure; making changes on the infrastructure without impacting the software or even stopping the system.

The main cons are mainly cost and knowledge requirements of work (preservation knowledge), difficult integration of the whole LTP process; problems with adapting to sudden changes or very high demand on the whole storage infrastructure.

Three (namely TIB, SLSA, BNE) institutions did not provide a detailed answer, since it is still too much early for them to be able to provide an evaluation.

3.7 Q9, Q10

Is your LTP system OAIS (ISO 14721:2003) compliant? How much is this important for your institution? How would you categorize this feature ("nice to have", "should have", "must have")? Do you have a document that maps your system to OAIS? Do you have any services/processes beyond OAIS? Do you miss some important functions/processes of OAIS and why?

The two questions are summarized in Fig. 5. It is worth noting that none of the in-house developed systems declares any kind of OAIS compliance.

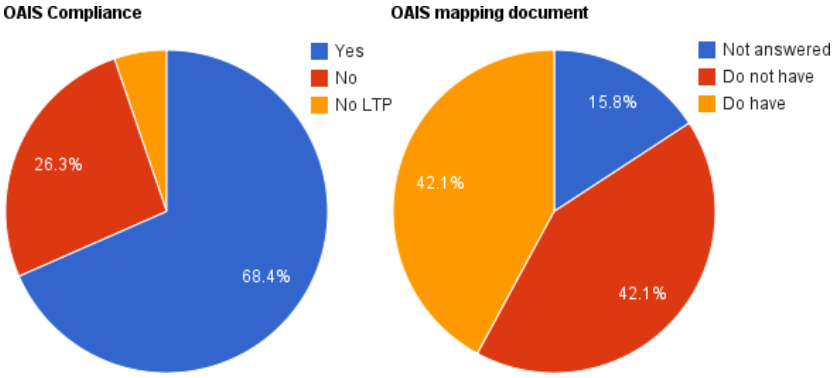

Figure 5: OAIS Compliance

3.8 Q11

Have you done a similar study before you decided to use an LTP system? If so, would it be possible to get its results?

Twelve institutions have done similar study before, but only five of them can provide the study to general public. However, rather than rigorous comparison of many LTP systems one with another, the documents are more or less advocating a solution chosen a priori or describing the design of the system deployed in particular institutions.

It is obvious that our question did not specify what should be considered a similar study, so the responses significantly vary.

For further information, see Table 2.

3.9 Q12

Do you have any documents describing your solution at the technical and/or architectural level?

Fourteen institutions responded positively, namely GENRL, NLNZ, CJCLS, UGENT, RLCD, NAA, STANFORD, METAARCH, HARVARD, COLUMBIA, BNF, TIB, SULAD and CINES.

Four institutions responded negatively, namely NUK, NDLJ, SLSA and BNE.

The SLV does not have a LTP system.

Table 2: Similar study before deploying LTP system

Library	**Sim. study**	**Avail. of results**
GERNL	yes	no
NUK	no	no
NLNZ	yes	yes
NDLJ	yes	no
CJCLS	yes	no
UGENT	yes	yes
RLCD	yes	yes
LOC	?	?
NAA	yes	yes
STANFORD	yes	yes
SLV	no	no
METAARCH	no	no
HARVARD	yes	no
COLUMBIA	yes	no
BNF	yes	no
TIB	no	no
SLSA	yes	no
SULAD	no	no
CINES	no	no
BNE	no	no

3.10 Q13

What is the approximate number of objects already stored? And what are the expected final (maximum) numbers?

The outcome is summarised in Table 3.

3.11 Q14

And how are the objects (data and metadata) structured? In other words, how is a periodical/monograph/map represented in the LTP? How does your API for various data types look like? What type of identifiers is used?

This question has very heterogeneous answers. Some patterns can be found in the answers, though. It is necessary to keep in mind that the identifier standards overlap. Most institutions use a combination of more identifiers. Following identifiers appeared in the answers:

- URN (Uniform Resource Name)—appeared four times among the answers
- UUID (Universally Unique Identifier)—3 times
- Handle—twice
- PURL (Persistent URL)—twice
- ARK (Archival Resource Key)—twice
- DOI (Digital Object Identifier)—once

Table 3: Number of stored objects and storage size

Institution	Approximate number	Size
GERNL	n/a	n/a
NUK	3 million	50 TB
NLNZ	n/a	(48 TB)
NDLJ	6,200	n/a
CJCLS	4,091 AIPs	other
UGENT	211,577	n/a
RLCD	less than 500,000	(6 TB)
LOC	n/a	2 PB (disks), 5 PB (tapes)
NAA	(millions)	n/a
STANFORD	n/a	(20TB)
SLV	n/a	n/a
METAARCH	n/a	n/a
HARVARD	28.5 million	125 TB
COLUMBIA	400	n/a
BNF	219,000	90 TB
TIB	n/a	n/a
SLSA	n/a	(20 TB)
SULAD	6 billion	1.2 PB
CINES	200,000	n/a
BNE	22.5 million	n/a

- barcode—once

Another interesting topic emerging from the answers is packaging of the data. As we mentioned, the systems are heterogeneous. Some of them are OAIS compliant and work with Information Packages and some do not. UGENT uses BagIt standard [6] for packaging their data and file names conventions utilizing NARA [47]. The ARC format is used by BNF and RLCD.

Many institutions use contemporary formats such as PDF/A or JPEG 2000 and METS/ALTO for OCR.

Finally, the first part of the question "And how are the objects (data and metadata) structured?" can be classified into three types of answers. The first type (Rosetta users and few others) was: "We are OAIS compliant and have rigid structure of SIPs and AIPs incorporating METS as a container metadata format including DC or other types, and PREMISE for administrative metadata". The second type of answers was: "We store our metadata in the catalogue separated from the actual data" where the "catalogue" means an application for searching the assets by the public users. LOCKSS users typically answered "it depends" because LOCKSS is format-agnostic, preserving all formats and genres of web-published content.

3.12 Q15

What extent of in-house customization was needed? Was the system delivered as an "out of box" vendor solution, did a contract include local customization or were you the major architects and developers?

The question was answered by 17 institutions. SLV does not have an LTP system and BNE did not respond to that question.

6 (31%; namely GERNL, NLNZ, METAARCH, UGENT, RLCD, TIB) institutions have a local customization of an off-the-shelf system or the institution created only part of the system and the other part has been completely delivered.

4 (21%; namely CJCLS, COLUMBIA, SLSA, CINES) institutions have a "out of the box" solution.

Finally, 7 (39%; namely NUK, NDLJ, NAA, STANFORD, HARVARD, BNF, SULAD) institutions use an in-house developed solution.

3.13 Q16

Have you tried any form of distributed data storage? Provided you have tried some distributed data storage, how consistently is the application layer of the LTP system separated from the lower distributed data layer?

Three institutions (NDLJ, NAA, TIB) answered negatively. GERNL uses the HP XP and EVA arrays, NZNL uses Rosetta together with IBM XIV storage.

LOCKSS adopters have no reason using another level of abstraction to achieve storage distribution as the system is distributed by its nature. Fedora Commons users can also abstract from the data storage level as long as the storage is a plain file system and/or the data is stored as "externally managed", accessible via URL links.

HRVARD answer is worth citing: "We do have distributed data storage using discs and tapes—we have copies stored in 4 different geographic locations. The lower storage layer is abstracted from the application layer. The application sees the content as if it were in a local file system."

BNF uses data storage based on iRODS having 3 levels of abstraction. And finally, CINES, SULAD, COLUMBIA use also the separated layer for storing the data, but they did not specify the details.

3.14 Q17

Is it legally permitted to keep your data saved outside the country/institution? If not, would you like to use some form of on-premise or spread-over-a-few-institution distributed repository for a long-term storage instead of cloud storage services like Amazon S3 or Google Storage?

Obviously, the answer to this question may depend on the character of the data stored. For instance, RLCD and SULAD note that it is not permitted in Denmark when handling personal data. Similar law restrictions, forbidding storage of personal data outside the country, can be expected in most countries around the world.

Data stored in LTP systems are not of personal nature, most of respondents answered positively, with the following exceptions. NAA responded it was not legal in Australia to outsource the storage outside the country. Also two French institutions (BNF and CINES) admitted it might be an issue necessary to be cleared.

3.15 Q18

Is there any centralized instance (registry of digitization) for monitoring a digitization and subsequent or preventive deduplication of the digitized data in your country?

4 institutions answered positively, 8 institutions negatively, two institutions stated it was planned and the rest (6) did not answered or were not sure.

There is certain effort to manage the process of digitization by creating centralized instances responsible for deduplication. These registries could significantly lower the amount of data to be stored later on in LTP systems.

3.16 Q19

Does your institution participate on exchanging the metadata with other institutions through OAI-PMH or other protocols?

The answers show that 15 institutions (79%) participate on exchanging the metadata with other institutions through OAI-PMH or other protocols, 4 institutions (21%) do not and the only one of them (Educopia Institute) is making plans to join in the future.

Among 15 participating institutions, 4 are also participating in Europeana and one (Columbia University) exchanges the metadata with other institutions using the FGDC CSDGM standard (and have not adopted OAI-PMH).

4 Results and Analysis

Our survey collected interesting data in the field of LTP across many memory institutions. Despite the very heterogeneous and in some sense autonomous nature of such institutions and despite all the cultural differences including varying law and government systems, some patterns can be found. In this section, we analyze the answers and find some interesting correlations among them.

Our goal was to find out what backend systems (low level of LTP) are typically used in the LTP area. In general, there are two kinds of approaches. The first one is represented by systems like LOCKSS and iRODS where the system is naturally distributed (data is shared across many nodes all over the world). The second approach is represented by more "monolithic" (i.e., not distributed) systems running inside the institution delegating the lowest level of storage to some proven solutions like NAS, hierarchical storage, tapes, hybrid solutions, etc.

The first three questions give us the idea about preferences of university and national libraries. While universities naturally tend to use open source and free solutions (Fedora, LOCKSS), national libraries tend to operate bigger budgets and either use out-of-the-box proven solutions (e.g., Rosetta) or develop a system to satisfy their custom requirements in-house.

Q4 deals with data flows inside the system, the role of a preservation module in the system and the interaction between the module for accessing the assets and the module for preserving the assets. According the OAIS ISO standard [42], the dissemination and preservation modules should be strictly isolated, and, except a few institutions, all the participants respect this fact and

built their archives as a "dark archives," i.e., separated from the data presented to the public.

Surprisingly, according to Q6, the optical discs are still being used by some institutions, but the trend is clearly descending. Disk and tape storage is the mainstream. Many institutions are interested in storing their data in the cloud. This idea is nevertheless somewhat limited by law restrictions, by project budgets, and also by the fact that it could be considered risky to outsource this service to external vendors.

Despite the trend of outsourcing IT services, there are still many unanswered questions in the field of LTP regarding this idea. Prices of such solutions, as well as legal questions are discussed. We intended to obtain the information whether the institutions invested some effort in a survey on legal issues with outsourcing data. The answers are, however, mostly negative.

Let us now summarize the answers to questions stated in Section 2.

1. Generally speaking, the type of an LTP system depends on the type of the librarian institution. National libraries tend to use expensive all-in-one (Rosetta-like) solutions or have their in-house developed systems. University libraries and small to medium sized libraries tend to use open-source solutions (Archivematica, Fedora) and are more open to experiments and storage grids (LOCKSS, iRODS).

2. Those qualities can be obtained partly from the Q8 presented in Section 3.6. Reliability is definitely one of them, ability of integrity checks, versioning, geographical distribution; these features were mentioned as pros. As cons, mostly the cost and complexity were mentioned. Standardization is also very important for the institutions as described in Q9 and Q10 in Section 3.7.

3. It is not possible to pinpoint the one and only proper system. The demands are conflicting, therefore some trade-off has to be found. The possibilities are obviously limited by available budgets. Classifying libraries into the three basic types, i.e., national, university, and some small-to-medium sized ("regional") ones, the most often deployed systems for each of the types are:

   - National – Rosetta
   - University – Fedora
   - Regional – LOCKSS

   It should be noted that this study does not cover trends in system deployment, therefore those observations are of time-limited value.

5 Overview of LTP Systems

In this section, we discuss main features of most commonly used LTP systems.

As we have learnt from our survey, most commonly used systems for data preservation are LOCKSS, Fedora Commons and Rosetta. We also briefly mention other systems. Many institutions deploy in-house developed systems. As there is significant lack of information about such systems, we discuss them

mainly in the literature review in Section 6—the systems are often described in research and/or technical papers. Proper technical documentation is not often available.

5.1 The OAIS Model

Before we introduce the selected systems, let us discuss the OAIS (Open Archival Information System) model [43]. This standards is intended to be a referential design of digital archives with long term preservation capabilities. Three basic roles are defined in the model: producer, consumer and management, cf. Fig. 6. The model defines functional entities Ingest, Data Management, Archival Storage, Access, Preservation Planning, and Administration. For detailed description, we refer to [43].

More than a strict specification of referential model, OAIS is a set of vague advices how to build a digital archive. It identifies some basic actors/roles, describes the data and their flows in the systems. The central notion of OAIS is an information package. There are three types of packages SIP (Submission Information Package), AIP (Archival IP), and DIP (Dissemination IP) depending on the life cycle of the data. The content of those packages slightly differs, but in general there are some descriptive and structural metadata, the manifest describing the content and the data itself.

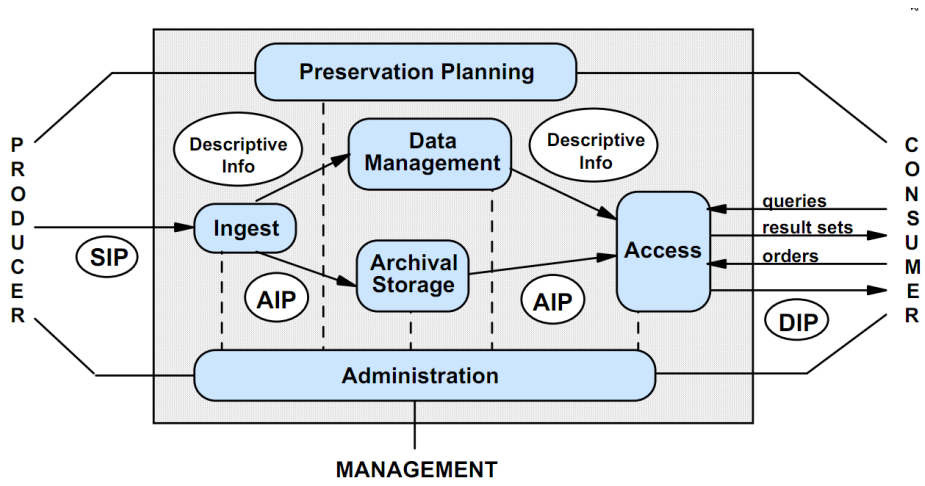

Figure 6: Functional model of OAIS, source: [43]

5.2 Commonly Used LTP Systems

In this section, we briefly describe major LTP systems as they appeared in our survey.

5.2.1 LOCKSS

LOCKSS (Lots of Copies Keep Stuff Safe), maintained by Stanford University Libraries, is an international community initiative that provides libraries with

digital preservation tools and support so that they can easily and inexpensively collect and preserve their own copies of authorized e-content [34]. LOCKSS is a prototype of a system to preserve access to scientific journals published on the Web. It is a majority-voting fault-tolerant system that, unlike other systems, has far more replicas than would be required just to survive the anticipated failures [53].

The LOCKSS technology has been undergoing increasingly stringent testing since 1999. The alpha test ran through 2000, and an early beta version was successfully deployed to 50 libraries worldwide from 2000 to 2002 [33].

LOCKSS Alliance library participants receive technical and collections support from the Stanford University LOCKSS team. Participants collect and preserve content from the over 500 participating publishers. They are building digital local and ensuring permanent access [32].

LOCKSS allows libraries to run web caches for specific journals. These caches collect content as it is published. They cooperate in a peer-to-peer network to detect and repair damaged or missing documents. This kind of continuous repairs is based on majority voting (opinion polls) among system nodes [37]. The caches run on generic PC hardware using open-source software and require minimum skill in administration, making the cost of preserving a journal manageable [48].

5.3 Fedora Commons

"Fedora (Flexible Extensible Digital Object Repository Architecture) was originally developed by researchers at Cornell University as an architecture for storing, managing, and accessing digital content in the form of digital objects inspired by the Kahn and Wilensky Framework." [19]

"The Fedora architecture is an extensible framework for the storage, management, and dissemination of complex objects and the relationships among them. Fedora accommodates the aggregation of local and distributed content into digital objects and the association of services with objects. The architecture is implemented as a set of web services, with all aspects of the complex management functions exposed through REST and SOAP API." [31]

Rather than a solution for long term data preservation, Fedora is a customizable repository with well defined interface. Fedora itself is definitely not ready for deployment "as is". There is a need for a quite complex settings and for making data accessible, it is necessary to develop a special application. Fedora is just a platform on which it is possible to build other standalone systems.

The main concept in Fedora architecture is the notion of a digital object represented in the format called FOXML (Fedora Object XML) containing a set of datastreams. There are two compulsory datastreams: DublinCore and RELS-EXT. The RELS-EXT datastream defines the relations of the object with the surroundings. The RDF is used for the description, therefore Fedora can be considered as prepared for Linked Open Data.

Fedora itself is not considered an LTP system, it can however be extended for LTP. Usually, such extensions are developed in-house by the institutions, general-purpose LTP extensions are also available, eg., the RODA system [18]. Current versions of Fedora also contain so called "high-lvl storage" subsystem serving as an abstraction of how the data is actually stored. Using such features, Fedora can cooperate with NAS and/or hierarchical storages.

Fedora can be integrated with systems like DuraCloud [7] which is used to replicate Fedora's data into the cloud-based content repository hosted on either Amazon S3 or Rackspace. Fedora does not provide any sort of functions ensuring data integrity and is not designed as a peer-to-peer system.

Fedora is used by two respondents of our survey, as the core repository in the Egyptian National Library and in the project of the Czech Digital National Library.

5.3.1 Rosetta

Rosetta [60] is a digital preservation system developed by Israeli company Ex Libris. The same company stays behind the famous widespread librarian catalogue Aleph. Rosetta is a commercial solution with closed source code. On the other hand, Ex Libris claims the Rosetta is an open platform in a way that [8] it provides a set of APIs for accessing the core functionality as well as the APIs for accessing the protocols like Z39.50, SRU/SRW, OAI-PMH [30], etc. Although the source code of the system is not publicly available, the source code escrow guarantees availability of the source code in case of Ex Libris bankruptcy.

The overall architecture consist of the following modules: deposit module, working area, delivery module, permanent repository, and database. Each of these modules can be scaled with the provision of additional servers, in order to achieve high availability or redundancy; or "all-in-one" strategy can be applied, in which all the modules are deployed on the same physical or virtual server.

Unfortunately, not much information can be found on the web pages of the vendor about the system itself, but from the information available and from the white papers it is obvious that Rosetta is a complex out-of-the-box LTP system and only a little customization is needed. Furthermore, it can cooperate with the other existing products from Ex Libris like Aleph or Primo. Prices of Rosetta are not low which makes it unreachable for many small institutions (especially taking into account that Oracle database license must be obtained in order to run Rosetta).

In our survey, Ex Libris Rosetta was used by two respondents (who are also mentioned in white papers [71], [41] available through Rosetta web page): NLNZ and CJCLS. To our knowledge, the Portuguese National Archives operate the Rosetta system and finally the German National Library of Science and Technology has recently adopted the system as well [61].

5.3.2 Other Systems

The systems mentioned only marginally in the answers are Kopal, Arcsys (Infotelsun) a DPSP (Digital Preservation Software Platform).

Other systems/distributed architectures that were not mentioned in the answers are, but we consider them interesting to mention, are Tesella[68], Archivematica [2], Protage[46], IBM Dias and iRODS[28].

iRODS [28] (Integrated Rule-Oriented Data System) is a inherently distributed system for storing arbitrary data. It is mostly used in the scientific area, however, it can be used by memory institutions as well. There is a basic support for metadata standards such as DublinCore, and this fact can be used for discovery services (iRODS Metadata catalogue).

5.4 Comparison of LTP Systems

We will now briefly compare the systems described above in properties important for data preservations.

It is obvious that there is a trade-off between accessibility and durability of the data.

With respect to the long term preservation, it is wrong to expect that data never change. Migration is one of the strategies for data preservation and all the systems provide a way to migrate (convert) the data from an old environment to a new one in a batch manner, LTP systems should provide some support for such processes.

Another aspect of the systems is the OAIS compliance. From this perspective, the Fedora is not OAIS compliant (there is however a way to satisfy the OAIS compliance described in [5]), LOCKSS and Rosetta follow the OAIS standard.

Distributed nature of data storage is another quality worth consideration. LOCKSS is similar to the iRODS in the sense that it is a peer-to-peer network of cooperating nodes. Replication is used as the main strategy for preserving and protecting data. Problems such as bit corruption are obviously less risky in the systems where more than one copy exist. Rosetta's specification claims to be scalable and there is a support for redundancy for keeping the collections safe. Fedora is the only one candidate not providing any out-of-the-box scalability. However, it is possible to plug more than one Fedora repository into cluster or to use Fedora together with DuraCloud.

5.5 Table comparison

We compare qualities of the solutions of interest in Table 4. In the following, we discuss the detailed meaning of the properties under consideration (unless they are obvious).

Scalability is the ability to handle increasing amount of records stored inside. Since the LTP system should not serve for exposing the data to the user, scalability does not mean the ability to handle many user requests at the same time.

Distributed storage is the ability to distribute data across many separated instances.

Preservation strategy describes the main strategies to keep the data unharmed.

Versioning is the ability to keep previous version of the data. It is strongly recommended to preserve also the original of a modified content.

Metadata standards describes metadata standards implemented in the systems.

OAI-PMH provider describes whether the system has the capability of exposing the metadata by OAI-PMH protocol, serving as an OAI-PMH provider. In that case any other OAI-PMH harvester can harvest all the metadata in order to allow federated search.

Table 4: System comparison

System	LOCKSS	Fedora	Rosetta
Cost	free	free	depends
Open source	yes	yes	no
Scalability	yes	no	yes
Distributed storage	yes	no	yes
OAIS compliance	yes	no	yes
Preserv. strategy	replication, migration	trustworthy storage	replication, migration
Versioning	yes	yes	yes
Metadata standards	DC, MODS, PREMISE	any	DC, MODS, PREMISE
OAI-PHM provider	yes	yes	yes

6 Related Journals and Magazines

The purpose of this section is to collect references to materials available in the area of LTP. Most of the literature available is concerned with digital libraries. LTP systems and topics are also mentioned, usually only marginally, making relevant material quite difficult to find. LTP features of systems are often hidden under general descriptions of digital libraries functionality. Many references can be also found in direct relationship to the systems, the papers are often written by LTP system authors.

The list of magazines and journals related to digital libraries is based on the list from the website of the DCC (Digital Curation Centre) [14].

6.1 Ariadne

"Ariadne is a Web magazine for information professionals in archives, libraries and museums in all sectors." [3] The magazine focuses on articles about projects and development, reports from events, and book reviews.

6.2 D-Lib Magazine

D-Lib Magazine [15] is focused on digital library research and development and also on technologies, applications, and contextual social and economic issues and conference reports. A lot of articles deal with trends and innovations in the field of digital libraries.

6.3 First Monday

First Monday [40] is an openly accessible, peer–reviewed Internet journal. The magazine is targeted very broadly, ranging from social media, e-commerce, online gaming, to libraries and data preservation.

6.4 International Journal of Digital Curation

The IJDC [25] is devoted to articles and news on digital curation and related issues. The journal is published twice a year. Main topics include education,

software preservation, data recovery, emulation, information requirements of various user groups and others.

6.5 International Journal on Digital Libraries

The International Journal on Digital Libraries [26] is aimed at the theory and practice of acquisition, definition, organization, management, and dissemination of digital information via global networking. Themes of each issues are very different, including education, semantic web, different types of evaluation, digitizing or various types of digital libraries (including audio, video).

6.6 Journal of Digital Information

"JoDI is a peer-reviewed electronic journal about the management, presentation and uses of information in digital environments. It is also covering research, technical, design and practical issues." [38] There are mainly the following topics: digital libraries, hypermedia systems, hypertext criticism, information discovery, information management, social issues of digital information and usability of digital information.

6.7 ECRIM News

ERCIM News [10] is the magazine of ERCIM (The European Research Consortium for Informatics and Mathematics). The focus of this magazine is quite wide, the issue of January 2010 [64] was specialised into digital preservation.

6.8 iRODS

- Article [13] introduces iRODS.
- In article [22], an approach to automation is described in which digital curation policies and strategies are represented as rules, which are implemented in data grids based on the iRODS middleware.
- In book chapter [23], an implementation approach that combines the Fedora digital repository software with a storage layer implemented as a data grid using the iRODS middleware are described.
- In poster [1], a set of services to address a problem in the metadata catalogueue of the iRODS data grid, strengthening that platform for digital preservation purposes are presented.
- In paper [70], the provenance needs of iRODS and an architecture that can be used to manage provenance in iRODS (and other systems) in a fault-tolerant way are described.
- In paper [62], a first study exemplary for the SHAMAN Integrated Project is presented. This project will be based upon the iRODS grid infrastructure using so called rules and micro-services to establish an archive environment.

- In paper [24], scenarios implemented in a benchmark tool to measure the performance of an iRODS environment are discussed as well as results of measurements with large datasets. The scenarios concentrate on data transfers, metadata transfers and stress tests.
- In paper [11], a system for data management based on iRODS is described both from system and user viewpoints.
- In paper[27], assessment of the SHAMAN demonstrators for the memory institution is discussed from the point of view of user's needs and fitness for purpose. An innovative use of TRAC criteria, DRAMBORA risk registry and mitigation strategies, iRODS rules and information system models requirements has been designed, with the underlying goal to define associated policies, rules and state information, and make them wherever possible machine-encodable.
- In paper [4], main threats to digital preservation are discussed, which are used to identify a central point of failure in the metadata catalogue of the iRODS data grid solution.

6.9 LOCKSS

- In [49], [50], [48], [57], [51], the LOCKSS project is presented.
- Article [56] was presented to the NIST Digital Preservation Interoperability Framework Workshop. In this article LOCKSS project is presented.
- In [20], several defenses for the LOCKSS peer-to-peer digital preservation system are described.
- In article [21], integration of LOCKSS, link resolvers and OPAC is presented.
- In article [63], a social model along the lines of the open source software community is presented. LOCKSS is building a community base and has over seven years of experience with archiving electronic journals.
- In article [52], LOCKSS technology that enables non-technical administrators to create and manage their own preservation network via a "Private LOCKSS Network" (PLN) is presented. A PLN is a scaled down version of the public LOCKSS network, which comprises two hundred library members and preserves over a thousand titles from more than three hundred publishers.
- In [59], designing, implementing a proof-of-concept and demonstrating transparent format migration on access for the LOCKSS digital preservation system are presented.
- In article [9], two basic approaches to address the problem of digital preservation using peer-to-peer systems are identified: conservation and consensus.

- In article [35], a set of defenses that systems can deploy against data loss and potential synergies among them are presented. There is also illustrated the application of these defenses in the context of the LOCKSS digital preservation system.

- In [36], a design and simulation of a new peer-to-peer opinion poll protocol that addresses scaling and attack resistance issues is presented. The new protocol is based on experience with the deployed LOCKSS system and the special characteristics of such a long-term large-scale application.

- In [37], a design and simulations of a novel protocol for voting in systems of this kind is presented.

- In article [58], a peer-to-peer digital preservation system for e-journals, a set of techniques that enable a large population of autonomous peers to resist attack by a substantial minority of malign peers endowed with unlimited computational resources are described in the context of LOCKSS.

- In [55], design, implementation and deployment of a network appliance based on an open source operating system is described. There is also provided an overview of the LOCKSS application and described the experience of deploying and supporting its first version.

- In [39], a set of protocols that achieve data resilience for the long term using a peer-to-peer network, where mutually untrusted peers are loosely organized are presented. There is also a brief overview of how LOCKSS performs Opinions, improve the current algorithms and evaluate our protocols in terms of their performance and security against adversary attacks.

- In [54], some techniques that have been developed in the context of building a peer-to-peer system whose design requirements make keeping long-term secrets impractical are described.

6.10 Fedora Commons

- In articles [67] and [31], the Fedora project is presented.

- In [65], a project of The University of Virginia to build a digital object repository system based on the Flexible Extensible Digital Object and Repository Architecture is presented.

- In [44], a digital object and repository architecture for storing and disseminating digital library content are described.

- In [45], a form of digital object known as the Value-Added Surrogate that can enhance the functionality of digital content that is not in one's direct control is presented. The Fedora Digital Object model is the basis for the V-A Surrogate design.

- In [66], The University of Virginia Library and their implementation of the Fedora system is discussed.

6.11 Rosetta

- In presentation [12], several slides mention why the library decided for Rosetta.
- In paper [29], digital preservation at the National Library of New Zealand is presented.

7 Acknowledgements

We are very thankful to Jan Hutař, Marek Melichar, and Leoš Junek of the National Library of the Czech Republic, and Petr Žabička of the Moravian Library for valuable comments on the survey and providing personalised contacts to cooperating institutions worldwide.

This work is partially supported by the OP VaVpI project eIGeR (registration number CZ.1.05/3.2.00/08.0142).

References

[1] ANTUNES, G. – BARATEIRO, J. Securing the iRODS Metadata Catalog for Digital Preservation. In *Research and Advanced Technolgy for Digital Libraries*, pages 412–415, Berlin, Germany, 2009. Springer-Verlag. ISBN 78-3-642-04345-1, `http://www.springerlink.com/content/97587724187x2n55/fulltext.pdf`.

[2] Archivematica – open archival information system, November 2011. `http://archivematica.org/`.

[3] Web Magazine for Information Professionals in Archives, 2011. `http://www.ariadne.ac.uk/information/#about`.

[4] BARATEIRO, J. et al. Using a Grid for Digital Preservation. In *Digital Libraries: Universal and Ubiquitous Access to Information 11th International Conference on Asian Digital Libraries*, Bali, Indonesia, 2008. Indonesia : ICADL. ISBN 978-3-540-89532-9, `http://www.springerlink.com/content/k71v8x6081738x18/fulltext.pdf`.

[5] BEKAERT, J. – SOMPEL, H. V. Access Interfaces for Open Archival Information Systems based on the OAI-PMH and the OpenURL Framework for Context-Sensitive Services. *CoRR*. 2005, abs/cs/0509090. `http://arxiv.org/abs/cs/0509090`.

[6] BOYKO, A. – KUNZE, J. The BagIt File Packaging Format, April 2011. `http://tools.ietf.org/html/draft-kunze-bagit-06`.

[7] BRANAN, B. – SMITH, C. Welcome to the DuraCloud, December 2010. `https://wiki.duraspace.org/display/duracloud/DuraCloud`.

[8] BREEDING, M. Ex Libris Sets Strategic Course on Open Systems. *Smart Libraries Newsletter*. 2008, 28, 8, p. 1–3. `http://www.librarytechnology.org/ltg-displaytext.pl?RC=13434`.

[9] BUNGALE, P. P. – GOODELL, G. – ROUSSOPOULOS, M. Conservation vs. Consensus in Peer-to-Peer Preservation Systems. In *4th International Workshop on Peer-To-Peer Systems (IPTPS 2005)*, pages 240–251, Ithaca, NY, USA, 2005. Springer-Verlag. ISBN 978-3-540-29068-1, http://www.springerlink.com/content/l51u627735834231/.

[10] CHAILLOUX, J. – KUNZ, P. ERCIM News, 2011. http://ercim-news.ercim.eu/.

[11] CHIANG, G.-T. et al. Implementing a Genomic Data Management System Using iRODS in the Wellcome Trust Sanger Institute. *BMC Bioinformatics*. 2011, 12. http://www.biomedcentral.com/1471-2105/12/361.

[12] CORRADO, E. M. – CARD, S. Implementing Rosetta at Binghamton University. In *ELUNA 2011*, Milwaukee, Wisconsin, USA, 2011. http://codabox.org/80/2/corrado_card_rosetta_eluna2011.pdf.

[13] DAVIES, K. iRODS: The Metadata Solution to Data Management. *Bio-IT World*. 2011, 10, 1, p. 54–56. http://search.proquest.com/docview/884228793/fulltextPDF?accountid=16531.

[14] Curation and Preservation Related Journals, 2010. http://www.dcc.ac.uk/resources/curation-journals.

[15] D-Lib Magazine, 2011. http://www.dlib.org/dlib.html.

[16] DOBRATZ, S. – NEUROTH, H. Network of Expertise in Long-term STOrage of Digital Resources – A Digital Preservation Initiative for Germany. *D-Lib Magazine*. April 2004, 10, 4. http://www.dlib.org/dlib/april04/dobratz/04dobratz.html.

[17] DRAMBORA: About – Digital Repository Audit Method Based on Risk Assessment, 2008. http://www.repositoryaudit.eu/about/.

[18] FARIA, L. – FERREIRA, M. – CASTRO, R. RODA, 2006. http://redmine.keep.pt/projects/roda-public?locale=en#home.

[19] Fedora Commons, Inc. Fedora Commons – About, 2009. http://www.fedora-commons.org/about.

[20] GIULI, J. et al. Attrition Defenses for a Peer-to-Peer Digital Preservation System. In *2005 USENIX Annual Technical Conference*, pages 163–178, 2005. http://www.usenix.org/publications/library/proceedings/usenix05/tech/general/full_papers/giuli/giuli.pdf.

[21] GUST, P. Accessing LOCKSS Content Through OPACS and Link Resolvers, 2010. http://www.lockss.net/locksswiki/files/Link_Resolver_Integration_White_Paper.pdf.

[22] HEDGES, M. Rule-based Curation and Preservation of Data: A Data Grid Approach Using iRODS. *Future Generation Computer Systems*. 2009, 25, 4, p. 446–452. http://www.sciencedirect.com/science/article/pii/S0167739X08001660.

[23] HEDGES, M. – BLAKE, T. – HASAN, A. Digital Library Storage Using iRODS Data Grids. In *Production Grids in Asia*, pages 129–139, New York, USA, 2010. Springer US. ISBN 978-1-4419-0045-6, http://www.springer.com/computer/communication+networks/book/978-1-4419-0045-6.

[24] HÜNICH, D. – MÜLLER-PFEFFERKORN, R. Managing Large Datasets with iRODS — a Performance Analysis. In *International Multiconference on Computer Science and Information Technology*, pages 647–654, Poland, 2010. ISBN 978-83-60810-27-9, http://proceedings2010.imcsit.org/pliks/80.pdf.

[25] International Journal of Digital Curation, 2011. http://www.ijdc.net/index.php/ijdc.

[26] International Journal on Digital Libraries, 2011. http://www.dljournal.org/.

[27] INNOCENTI, P. et al. Assessing Digital Preservation Frameworks: the Approach of the SHAMAN Project. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, New York, USA, 2009. ACM, New York. ISBN 978-1-60558-829-2, http://dl.acm.org/citation.cfm?id=1643899&bnc=1.

[28] Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems, December 2011. https://www.irods.org/index.php.

[29] KNIGHT, S. Early Learnings from the National Library of New Zealand's National Digital Heritage Archive project. In *World Library and Information Congress*, 2009. http://www.ifla.org/files/hq/papers/ifla75/146-knight-en.pdf.

[30] LAGOZE, C. – SOMPEL, H. V. The Open Archives Initiative Protocol for Metadata Harvesting, December 2008. http://www.openarchives.org/OAI/openarchivesprotocol.html.

[31] LAGOZE, C. et al. Fedora: An Architecture for Complex Objects and their Relationships. *CoRR*. 2005, abs/cs/0501012.

[32] LOCKSS – LOCKSS Alliance, 2008. http://www.lockss.org/lockss/LOCKSS_Alliance.

[33] Lots of Copies Keep Stuff Safe – About Us, 2008. http://www.lockss.org/lockss/About_Us.

[34] LOCKSS – What is the LOCKSS Program?, 2008. http://www.lockss.org/lockss/Home.

[35] MANIATIS, P. et al. Impeding Attrition Attacks on P2P Systems. In *11th ACM SIGOPS European Workshop*, Leuven, Belgium, 2004. http://www.eecs.harvard.edu/~mema/publications/shortDoS.pdf.

[36] MANIATIS, P. et al. LOCKSS: A Peer-to-Peer Digital Preservation System. *ACM Transactions on Computer Systems*. 2005, 23, 1, p. 2–50. http://dl.acm.org/citation.cfm?doid=1047915.1047917.

[37] MANIATIS, P. et al. Preserving Peer Replicas By Rate-Limited Sampled Voting. In *19th ACM Symposium on Operating Systems Principles*, Bolton Landing, New York, USA, 2010. ISBN 1-58113-757-5, http://dl.acm.org/citation.cfm?doid=945445.945451.

[38] MCFARLAND, M. – PHILLIPS, S. Journal of Digital Information, 2011. http://journals.tdl.org/jodi.

[39] MICHALAKIS, N. – CHIU, D.-M. – ROSENTHAL, D. S. Long Term Data Resilience Using Ppinion Polls. In *IPCCC: The 22nd International Performance, Computing, and Communications Conference*, Phoenix, Arizona, USA, 2003. http://labs.oracle.com/techrep/2002/smli_tr-2002-121.pdf.

[40] First Monday – Peer Reviewed Jurnal on the Internet, 2011. http://firstmonday.org/.

[41] The Ability to Preserve a Large Volume of Digital Assets – A Scaling Proof of Concept, 2010. http://www.exlibrisgroup.com/files/Products/Preservation/RosettaScalingProofofConcept.pdf.

[42] ISO 14721:2003 – Open Archival Information System – Reference Model, 2002. http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683.

[43] Reference Model for an Open Archival Information System, January 2002. http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683.

[44] PAYETTE, S. – LAGOZE, C. Flexible and Extensible Digital Object and Repository Architecture. In *Second European Conference on Research and Advanced Technology for Digital Libraries*, pages 41–59, Crete, Greece, 1998. ISBN 15508382, http://www.mendeley.com/research/flexible-and-extensible-digital-object-repository-architecture-fedora/.

[45] PAYETTE, S. – LAGOZE, C. Value Added Surrogates for Distributed Content: Establishing a Virtual Control Zone. *D-Lib Magazine*. 2000, 6, 6. http://www.dlib.org/dlib/june00/payette/06payette.html.

[46] Tessella – The Project, 2010. http://www.protage.eu/project.html.

[47] PUGLIA, S. – REED, J. – RHODES, E. Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images, June 2004. http://www.archives.gov/preservation/technical/guidelines.pdf.

[48] REIC, V. – ROSENTHAL, D. S. LOCKSS: A Permanent Web Publishing and Access System. *D-Lib Magazine*. 2001, 7, 6. http://www.dlib.org/dlib/june01/reich/06reich.html.

[49] REICH, V. Lots of Copies Keep Stuff Safe As A Cooperative Archiving Solution for E-Journals, 2002. http://www.istl.org/02-fall/article1.html.

[50] REICH, V. – ROSENTHAL, D. S. LOCKSS, A Permanent Web Publishing and Access System: Brief Introduction and Status Report. *Serials: The Journal for the Serials Community*. 2001, 14, 3, p. 239–244. http://select.ingentaconnect.com/uksg/09530460/v14n3/contp1-1.htm.

[51] REICH, V. Distributed Digital Preservation. In *Indo-US Workshop on International Trends in Digital Preservation*, Pune, India, 2009. http://www.lockss.org/locksswiki/files/ReichIndiaFinal.pdf.

[52] REICH, V. – ROSENTHAL, D. S. Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks. *Library Trends*. 2009, 57, 3, p. 461–475. http://muse.jhu.edu/login?uri=/journals/library_trends/v057/57.3.reich.pdf.

[53] ROSENTHAL, D. S. H. – REICH, V. Permanent Web Publishing. In *2000 USENIX Annual Technical Conference*, 2000a.

[54] ROSENTHAL, D. S. H. – SEIDEN, M. I. Is There an Alternative to Long-Term Secrets? In *Resilient & Active Defense Of Networks*, 2002. http://www.lockss.org/locksswiki/files/Santafe2002.pdf.

[55] ROSENTHAL, D. S. A Digital Preservation Network Appliance Based on OpenBSD. In *Proceedings of BSDcon 2003*, San Mateo, CA, USA, 2003. http://www.usenix.org/publications/library/proceedings/bsdcon03/tech/full_papers/rosenthal/rosenthal.pdf.

[56] ROSENTHAL, D. S. LOCKSS: Lots of Copies Keep Stuff Safe, 2010. http://ddp.nist.gov/workshop/papers/03_06_Dave_Rosenthal_NIST2010.pdf.

[57] ROSENTHAL, D. S. – REICH, V. Permanent Web Publishing. In *Proceedings of the USENIX Annual Technical Conference*, San Diego, CA, USA, 2000b. http://www.usenix.org/events/usenix2000/freenix/full_papers/rosenthal/rosenthal.pdf.

[58] ROSENTHAL, D. S. et al. Economic Measures to Tesist Attacks on a Peer-to-Peer Network. In *Workshop on Economics of Peer-to-Peer Systems*, Berkeley, 2003. http://www.eecs.harvard.edu/~mema/publications/P2P-Econ.pdf.

[59] ROSENTHAL, D. S. et al. Transparent Format Migration of Preserved Web Content. *D-Lib Magazine*. 2011, 11, 1. http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html.

[60] A New Way of Preserving Cultural Heritage and Cumulative Knowledge, 2011. http://www.exlibrisgroup.com/category/RosettaOverview.

[61] German National Library of Science and Technology selects the Rosetta digital preservation system, December 2011.

[62] SCHOTT, M. et al. Integrity and Authenticity for Digital Long-Term Preservation in iRods Grid Infrastructure. In *The 6th International Workshop for Technical, Economic and Legal Aspects of Business Models for Virtual Goods incorporating The 4th International ODRL Workshop*, pages

90–104, Poznań, Poland, 2008. Poznań University of Economics Publishing House. ISBN 978-83-7417-361-2, http://www.virtualgoods.org/2008/90_VirtualGoods2008Book.pdf.

[63] SEADLE, M. A Social Model for Archiving Digital Serials: LOCKSS. *Serials Review*. 2006, 32, 2, p. 73–77. http://www.sciencedirect.com/science/article/pii/S0098791306000414.

[64] SOLVBERG, I. – RAUBER, A. Digital Preservation: Introduction to the Special Theme. *ERCIM News*. January 2010, , 80. http://ercim-news.ercim.eu/en80/special.

[65] STAPLES, T. – PAYETTE, S. The Mellon Fedora Project: Digital Library Architecture Meets XML and Web Services. In *Sixth European Conference on Research and Advanced Technology for Digital Libraries*, pages 406–421. Springer-Verlag, 2002. https://esdora.ornl.gov/content/mellon-fedora-project-digital-library-architecture-meets-xml-and-web-services.

[66] STAPLES, T. – WAYLAND, R. Virginia Dons Fedora: A Prototype for a Digital Object Repository. *D-Lib Magazine*. 2000, 6, 7/8. http://www.dlib.org/dlib/july00/staples/07staples.html.

[67] STAPLES, T. – WAYLAND, R. – PAYETTE, S. The Fedora Project: An Open-source Digital Object Repository System. *D-Lib Magazine*. 2003, 9, 4. http://www.dlib.org/dlib/april03/staples/04staples.html.

[68] Tessella – Technology & Consulting, 2011. http://www.digital-preservation.com/.

[69] Trustworthy Repositories Audit & Certification: Criteria and Checklist, February 2007. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.

[70] WEISE, A. Managing Provenance in iRODS. In *Computational Science - ICCS 2009*, pages 667–676, Berlin, Germany, 2009. Springer-Verlag. ISBN 978-3-642-01972-2, http://www.springerlink.com/content/d125n3871h357301/fulltext.pdf.

[71] Preserving Digital Heritage at the National Library of New Zealand, 2010. http://www.exlibrisgroup.com/files/CaseStudy/NLNZ_RosettaCaseStudy.pdf.

A Questionnaire

Dear *name*,

As a part of our work on the development of a national data storage infrastructure in the Czech Republic, we are working on an extensive worldwide comparison of commonly used architectures of data storage and corresponding technical background of the Long Term Preservation (LTP) systems. This study is held under the auspices of CESNET, Center Cerit-SC, Institute of Computer Science at Masaryk University, and the Moravian Library in Brno.

We would appreciate if you will kindly forward this e-mail to the staff in your institution responsible for the technical foundations/basis of your long term data preservation systems.

Please, could you be so kind as to answer as many as possible of the following questions? The results of this "dialog survey" will be a part of a report that will be publicly available around end of this year. We will be pleased to send you an electronic or printed copy when the report is finished.

The information we would like to gather includes but is not limited to:

1. What kind of systems do you use to ensure long-term data preservation?

2. Does your institution use any LTP system (Rosetta, Tessella, Archivematica)? If not, are you planning to deploy some form of LTP system in the future?

3. If yes, what technical solutions stays behind it (home developed, iRODS, etc.)?

4. Do you use your LTP system directly for serving the user copies to the public or is there any system for accessing the user copies in the middle and your LTP stores only master copies? If the later option corresponds to your situation, how often do you synchronize the content of your LTP with the system for exposing the digital objects to the public? Is the performance (throughput/access time) of the LTP system a key quality in your infrastructure?

5. Do you have your LTP system certified as a trusted repository (TRAC, NESTOR) or do you plan a certification?

6. What technical solutions stays behind the system for preservation (home developed, SAN, etc.)? What kinds of HW technologies do you use for storing the master copies (disks, tapes, hybrid solutions, etc.)?

7. Would you prefer one geographic location where the actual data is stored or some kind of more geographically distributed approach keeping in mind risks of physically destroying the site, e.g., by a natural disaster?

8. What are the main pros and cons of your LTP infrastructure (rather HW infrastructure questions than functional requirements of the LTP system)?

9. Is your LTP system OAIS (ISO 14721:2003) compliant? How much is this important for your institution? How would you categorize this feature ("nice to have", "should have", "must have")?

10. Do you have a document that maps your system to OAIS? Do you have any services/processes beyond OAIS? Do you miss some important functions/processes of OAIS and why?

11. Had you done a similar study before you decided to use the LTP system? If so, would it be possible to get its results?

12. Do you have any documents describing your solution at the technical and/or architectural level?

13. What is the approximate number of objects already stored? And what are the expected final (maximum) numbers?

14. And how are the objects (data and metadata) structured? In other words, how is a periodical/monograph/map represented in the LTP? How does your API for various data types look like? What type of identifiers is used?

15. What extent of in-house customization was needed? Was the system delivered as an out-of-the-box vendor solution, did a contract include local customization or were you the major architects and developers?

16. Have you tried any form of distributed data storage? Provided you have tried some distributed data storage, how consistently is the application layer of the LTP system separated from the lower distributed data layer?

17. Is it legally permitted to keep your data saved outside the country/institution? If not, would you like to use some form of on-premise or spread-over-a-few-institution distributed repository for a long-term storage instead of cloud storage services like Amazon S3 or Google Storage?

18. Is there any centralized instance (registry of digitization) for monitoring a digitization and subsequent or preventive deduplication of the digitized data in your country?

19. Does your institution participate on exchanging the metadata with other institutions through OAI-PMH or other protocols?