Romanus Grütz[a]
[a] Institut für Medizinische Informatik, Universitätsmedizin Göttingen

D3.2: Evaluation des OAIS[1]

	D3.2: Evaluation des OAIS
Autor(en)	Romanus Grütz, Thomas Franke, Frank Dickmann
Editor(en)	
Datum	30.09.2013

A: Status des Dokuments

Version 1.0.0

B: Bezug zum Projektplan

M3.2: Evaluation des Open Archival Information System (OAIS)
Dieses Dokument wurde als Konferenzbeitrag bei der MedInfo2013 eingereicht und ist daher in Englisch verfasst.

[1]Dieses Dokument wurde im Rahmen des Projekts LABIMI/F erstellt. Das Projekt LABIMI/F wird gefördert von der Deutschen Forschungsgemeinschaft (DFG) unter dem Förderkennzeichen RI1000/2-1.

C: Abstract

There are no sufficient solutions for the preservation and reuse of research data, especially for genome and biomedical imaging data. The LABIMI/F project of the German Research Foundation (DFG) addresses this gap by establishing an infrastructure for preservation, retrieval and reuse of biomedical research data based on grid / cloud computing technology. The paper describes the proceedings and current state of the project: The previous work determined the requirements via workshops with the relevant stakeholders and evaluated software products/solutions. The paper shows the level of fulfillment of the technical components of the infrastructure concept regarding the requirements and reveals some constraints regarding 'Data Sovereignty', 'Audit Trail' and 'Data migration'. Furthermore a mapping of the technical components to the functional entities of the Open Archival Information System (OAIS) could successfully be conducted, which proves the OAIS conformity and thus the completeness of the concept.

D: Changes

Version	Datum	Name	Kurzbeschreibung
1.0.0	28. March2013	Romanus Grütz	Firste Version

E: Content

F: Figures

G: Tables

1 Introduction

In biomedical research, especially in genome and brain imaging research, data is complex and of great size and quantity. Many institutes organize their research data locally by themselves. Applied solutions are often semi-professional, e.g. storing data on external hard drives.

The German Research Foundation (DFG) recommends to preserve research data that has been used for publications for at least 10 years [1]. During this period, the data should be stored in an accessible, readable and in an understandable manner for later reuse [2].

To improve the retrieval and reuse of primary research data (payload) within these two areas, it has to be enriched with additional describing information (metadata). Metadata represents the context of the payload (e.g. the gender of the subject) and is a key factor for understanding the payload. [3]

To comply with these recommendations, biomedical scientists need to be supported by information technology. However, currently there are no solutions for genome and brain imaging research data.

The DFG funded project ‘long-term preservation of biomedical research data’ (LABIMI/F) addresses this gap and wants to establish an infrastructure for preservation, retrieval and reuse of biomedical research data based on grid / cloud computing technology. To demonstrate the functionality, the developed infrastructure will exemplarily be used for two applications: the processing of a) genome and b) biomedical imaging data.

This paper describes the developed technical concept for a digital preservation infrastructure and discusses whether it is an adequate solution that meets all requirements.

2 Materials & Methods

At the beginning of the project several workshops with domain experts and relevant stakeholders [4] where conducted in order to determine domain specific requirements. Beside the domain specific requirements we collected further common archive and data management requirements in a workshop with other disciplines and from literature (see requirement analysis).
Based on these requirements we conducted product evaluations [5] to determine the most fitting components for both use cases.
The further proceeding describes the developed overall concept while explaining the technical components and their interactions. Furthermore the technical components are evaluated regarding the fulfillment of the requirements (2 = fulfilled, 1 = fulfilled with restrictions and 0 = not fulfilled).
Afterwards, we map our components onto the functional entities of the Open Archival Information System (OAIS) in order to show the completeness of the concept.

2.1 Requirements analysis

The LABIMI/F project adapted the use cases that were developed in the KoLaWiss project [4] according to the context of a) genome research and b) biomedical imaging research. Both use cases were analyzed and the stakeholders were consulted in order to determine functional requirements for the digital preservation infrastructure. The functional requirements were discussed with IT- and domain experts in three workshops: [6] with biomedical imaging experts, [7] with genome experts and [8] with other disciplines.
The investigated key functional requirements are

R1. Data sovereignty/privacy
Only the owner of the payload should be able to define user access control to his data.

R2. Persistent references
Scientists should be able to reference payload in publications via the established digital object identifier (DOI) infrastructure.

R3. Content sharing
Depending on the privacy level of the payload, the infrastructure should facilitate data sharing between scientists/institutes a) global b) in GER/EU c) in Univesities d) in the project e) in the working group.

R4. Reliability
A failure of one (key) component must not result in a downtime for the whole service.

R5. Metadata schemas
The infrastructure has to deal with multiple metadata schemas for different kinds of payload and context.

R6. Provenance
Any production and usage of research data has to be monitored / comprehensible documented to prove validity as well as to provide understanding.

R7. Data migration
The infrastructure should provide mechanisms / extension points to migrate a) the file format and b) the storage technology of the payload and metadata.

R8. Integration in the scientific workflow
To improve the acceptance of the infrastructure, the user interfaces for ingest and retrieval of payload and metadata should be well integrated into the scientist's (daily) workflow.

2.2 Evaluation

After determining the requirements, product evaluations have been conducted regarding the key components: data management, data transfer and ingest system. Each evaluation is based on a use-value analysis (UVA), in which the products/solutions were rated [5].

2.2.1 Data management

The evaluation of data management tools focused on metadata and payload management and included three products: a) Fedora Commons, b) DSpace, and c) ISA Tab Tools.
The evaluation revealed that the ISA Tab tools are not qualified for the usage in the LABIMI/F infrastructure; Fedora Commons implements the most fitting 'metadata and payload management' and DSpace got in total the highest application score [5].

2.2.2 Data transfer

The evaluation of data transfer tools focused on reliable, secure data transfer and access via wide area network (WAN). The investigated products are a) PowerFolder, b) iRODS, c) CryptShare, and d) Globus Online. [5]
The evaluation revealed that none of the investigated products meets all requirements of the LABIMI/F infrastructure sufficiently. Further research revealed XtreemFS. Thus XtreemFS was taken into account (and selected). XtreemFS is developed by the LABIMI/F project partner Zuse-Institute Berlin (ZIB).

2.2.3 Local ingest system

The key functionalities that the local ingest system should contain are a) GUI to collect metadata depending on custom metadata schemas, b) partly automated metadata documentation of genome respectively biomedical imaging data, c) plausibility checks depending on the metadata schema, d) interface for retrieving DOIs, e) customizable SIP packaging, and f) interface to deliver SIP to an custom archive.
An evaluation has revealed no adequate software solution to ingest genome data. To ingest biomedical imaging data, the Extensible Neuroimaging Archive Toolkit (XNAT) was chosen.

2.3 Open Archival Information System (OAIS)

The Consultative Committee for Space Data Systems (CCSDS) developed and recommends a generic reference model for an open archival information system (see Figure 1) [9].
The open archival information system (OAIS) became the standard (ISO 14721:2012) of the International Organization for Standardization (ISO) and describes the components and functionality of a digital archive.

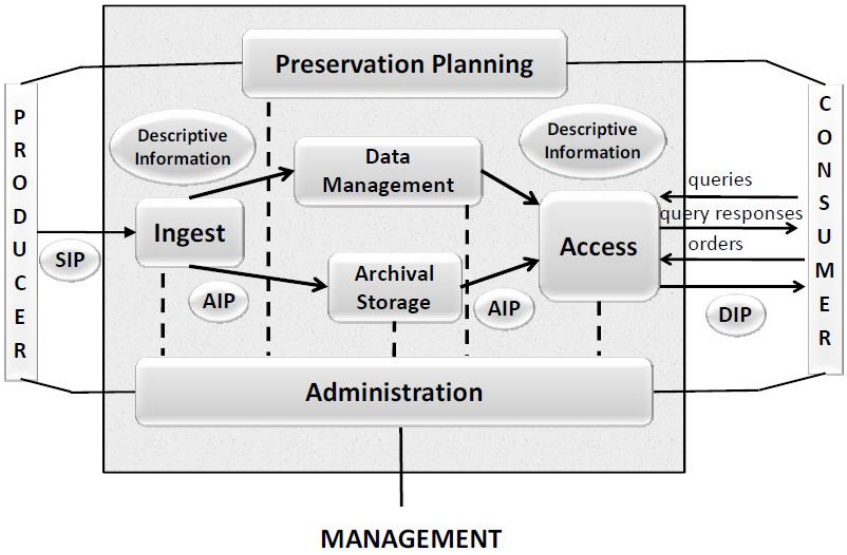

Figure 1: OAIS functional entities [9] (see OAIS for details)

The OAIS determines six functional entities and their related interfaces (see figure 1). The 'ingest' A1) accepts Submission Information Packages (SIP) complying with the archive format from the producer, A2) converts them to Archival Information Packages (AIP), A3) extracts descriptive information, and A4) conducts plausibility checks to include them into the archive database. [9]
The 'archival storage' B1) accepts AIPs from functional entity of the ingest, B2) stores them B3) manages storage resources incl. refreshing storage media, performing error checks, managing the storage hierarchy B4) provides AIPs to the access functional entity. [9]
The 'data management' provides functionality to maintain the data holding of the archive including C1) updating the archive database incl. schema, and C2) presentation of definitions and referential integrity. [9]
The 'administration' provides functionality regarding a) negotiation submission agreements, b) system configuration, and c) maintaining archive standards and policies. [9]
The 'preservation planning' should ensure the long-term accessibility of the data holding and includes a) technology watch to predict relevant format or technology changes, b) migration planning. [9]
The 'access' a) supports a consumer to determine information about existence and describing information of managed data holding as well as controls and coordinates access to the data holding, converts AIPs to Dissemination Information Packages (DIP), and delivers them to the customer. [9]

3 Results

Based on the previous work, we developed the following technical concept (see figure 2) and analyzed the concept regarding the fulfillment of the requirements (see table 1). Afterwards we mapped the components to the functional entities of the OAIS.

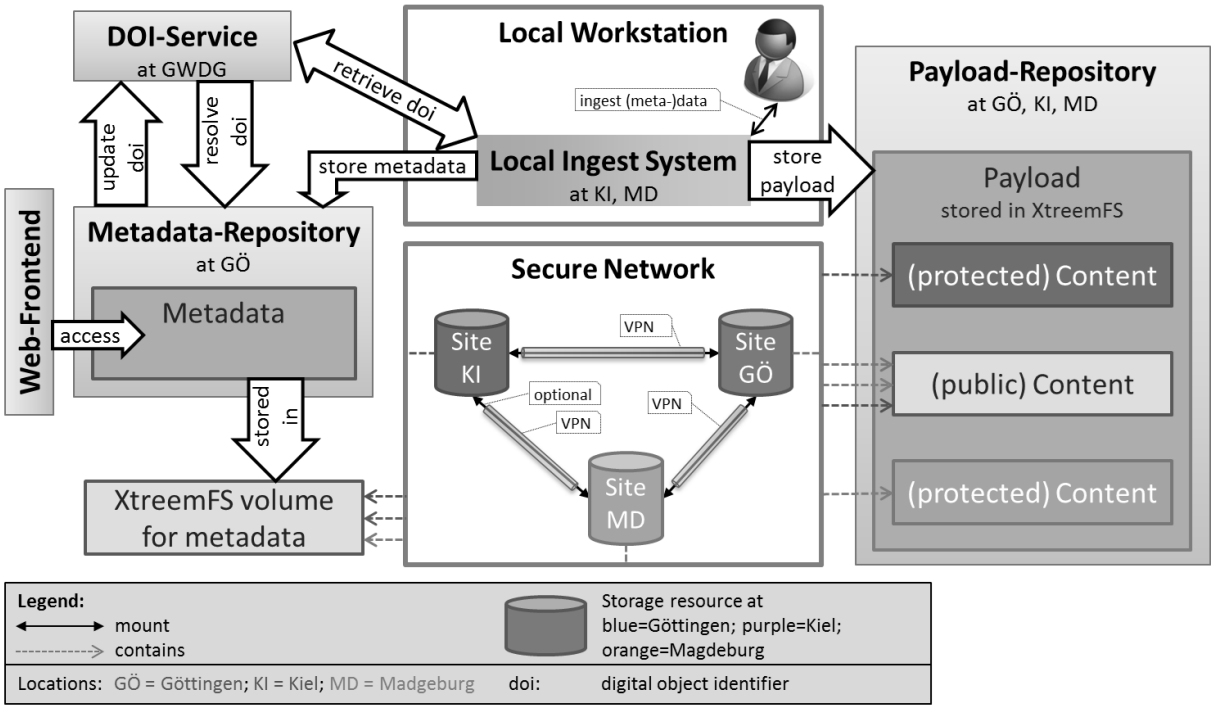

Figure 2: Concept of the LABIMI/F infrastructure (see results for details).

3.1 Technical components

The LABIMI/F infrastructure is distributed across several partner / sites. To ensure privacy the infrastructure has to use a secure WAN communication (see R1). To comply with this prerequisite, the whole communication is encrypted.

3.1.1 Ingest systems (GeMeCo)

The ingest system for use case a) genome data is the Genome Metadata Collector (GeMeCo). GeMoCo will be developed by the university medical center Göttingen (UMG) as part of the LABIMI/F project.

The ingest system for use case b) is XNAT. The XNAT installation is located in Magdeburg. XNAT is an open-source platform to ingest, manage, and access biomedical imaging and corresponding data developed by the Neuroinformatics Research Group (NRG Lab) at the Washington University.

To fit into the infrastructure concept and meet the requirements R2 and R8 the ingest systems GeMeCo and XNAT have to be capable of L1) processing multiple metadata schemas, L2) clear processing the metadata query from the scientist and L3) performing plausibility checks depending on the selected metadata schema, L4) partly automated metadata documentation e.g. via extraction tools, L5) retrieving a digital object identifier (DOI), L6) moving payload to XtreemFS, and L7) ingest metadata into DSpace.

XNAT already supports L1-L3, L5 and L6. L4 and L7 can be achieved by developing a XNAT plugin [10].

3.1.2 Distributed storage management (XtreemFS)

XtreemFS[2] is an open source distributed file system to establish a secure file system over WAN and inside of clouds developed by the Zuse-Institute Berlin (ZIB). It contains cross-site replication, no single-point-of-failure (SPOF) and auto-failover (see R4).
At least XtreemFS consists of a) one Directory Service (DIR), where every component has to register itself with its Universally Unique Identifier (UUID), b) one Metadata and Replica Catalog (MRC), which contains technical metadata about the location of data and how they are parted and replicated, and c) one Object Storage Device (OSD), where the data are stored physically. The XtreemFS file system can be accessed and used like a normal file system via a file system driver and uses POSIX[3] Access Control List (ACL) authorization. This driver is available for Windows, Linux and OS X and integrates XtreemFS as virtual volumes into the operating system. Storage media migration is also possible (i.e. replicating OSDs using old media to new OSD using new media). [11]
In the LABIMI/F infrastructure, only one DIR and MRC are used. In contrast OSDs are installed at each site (UMG in Göttingen, UKSH in Kiel[4], and UMM in Magdeburg[5]). Public payload and metadata is stored in volumes distributed (striped or replicated) across every site (see R3, R4), non-public payload just within the site of the owner (see R1).

3.1.3 Metadata repository

The metadata repository component is realized with DSpace. DSpace contains a role-based access control (RBAC) and thus meets requirement R1. It is capable to handle persistent identifiers (see R2), multiple metadata schemas (see R5), to divide payload and metadata storage (see R1), and to provide an item history (see R6) as well as a web front-end for a metadata based retrieval (see R8). [12]

3.1.4 DOI Service

To provide persistent identifiers (PID) in form of DOI it is necessary to integrate a Handle-/DOI-System. Thus the concept of the LABIMI/F infrastructure uses the handle-system of the GWDG (see R2).

[2] http://www.xtreemfs.org/
[3] Portable Operating System Interface
[4] http://www.uni-kiel.de/medinfo/institut/
[5] http://www.med.uni-magdeburg.de/ibmi.html

3.2 Mapping technical components onto OAIS

Table 1: Requirements of the concept and their fulfillment (2=fulfilled, 1=fulfilled with restrictions, 0=not fulfilled).

Requirement		Technical component	Score
R1	Data sovereignty	RBAC of DSpace and XNAT; XtreemFS uses POSIX ACLs and non-public payload volumes are stored only at the side of the owner and authorized partner Only one point because the infrastructure contains two different authorization mechanisms.	1
R2	External references	DOI-Service of the GWDG, DSpace is able to manage DOIs	2
R3	Content sharing	Cross-side striping and replication via XtreemFS volumes	2
R4	Reliability	Striping and replication via XtreemFS volumes and multiple DIRs and MRCs without SPOF	2
R5	Metadata schemas	GeMeCo, XNAT, DSpace support multiple metadata schemas	2
R6	Audit trail	Item history of DSpace and XNAT, but no seamless item history in one place / system.	1
R7	Data migration	Migration of storage media: XtreemFS, but no special mechanisms for data format migration.	1
R8	Workflow integration	GeMeCo, XNAT: L1 ... L7 DSpace: web front-end	2

In the concept of the LABIMI/F infrastructure the scientist who generates and archives new payload, represents the OAIS producer. The OAIS reference model shows that the producer has to submit the data as SIP to the archive (see A1). To enable the scientist to submit the payload and correlated metadata, the scientist has to use an additional ingest tool. In use case a) the Genome Metadata Collector (GeMeCo), which will be developed by the university medical center Göttingen and b) XNAT is used as ingest tool (see table 2).

The ‘ingest’ step from OAIS is represented by the ingest tool and the metadata repository / archive DSpace (see table 2). The ingest tool moves the payload into the storage management system XtreemFS and the metadata including information to the payload location to DSpace. DSpace accepts the metadata as SIP (see A1) and extracts the inherent information of the SIP into its own database structure (see A2, A3).

To improve the consistency and responsibility of the infrastructure, the plausibility is checked during the ingest step at the ingest system and in DSpace (A4).

The ‘Archival Storage’ OAIS functional entity is represented by XtreemFS. XtreemFS stores the payload received from the local ingest tool as well as the data holding of the metadata archive (see B1, B2). It is capable to change storage resources incl. media and underlying file system (see B3). Furthermore it grants DSpace and data owner access to its data holding (see B4).

The ‘Data management’, ‘Administration’ and ‘Access’ OAIS functional entity is combined in DSpace. DSpace is an open digital repository and influenced by the OAIS reference model [13] and is not further explained.

The access to research data is realized by searching metadata for relevant primary data. According to access rules the selected primary data can then directly accessed by the user via XtreemFS.
The OAIS ‘Preservation Planning’ entity is not considered here because it is an operational process, which has to be described in the operational model of the infrastructure.

Table 2: OAIS functional entities and corresponding technical components.

OAIS func. entity	Technical component
Producer	Payload and metadata generating scientist. To generate the SIP the scientists using GeMeCo in use case a, and XNAT in use case b.
Ingest	DSpace, GeMeCo, XNAT
Archival Storage	XtreemFS
Data Management	DSpace
Administration	DSpace
Access	DSpace, API
Consumer	Data seeking scientists

4 Discussion

Every functional entity of the OAIS could successfully be mapped onto technical components of the LABIMI/F infrastructure concept. Therefore, the results show that the infrastructure concept is in conformance with OAIS and thus a complete archival information system.

The ‘producer’ is mapped to two (GeMeCo, XNAT) and the ‘ingest’ to three technical components (DSpace, GeMeCo, XNAT). At first view, this seems to be redundant and the alternative solutions are a) using XNAT also for genome data, and b) using GeMeCo also for biomedical imaging data. However, using XNAT for genome data would require a great deal of customization regarding X1) file formats, X2) metadata schemas, X3) GUI including X4) the applied vocabulary, as well as X5) the data management including its APIs. In addition, the data management is highly specialized to biomedical imaging data (e.g. DICOM compatibility). This effort is estimated to be higher than a development from scratch with GeMeCo.

The second option, to use GeMeCo also for biomedical imaging would include the same customization steps. However, it would be easier to implement them because they could be considered during the development phase of GeMeCo. But it would be even more time consuming.

XNAT has already a user as well as a development community. Furthermore, it provides additional functions to archiving, view and accessing biomedical imaging data while user rights adhere to RBAC and organizational structures. All these features support the decision to use XNAT for the case b) biomedical imaging.

The distinction between public and non-public payload regarding its physical storage location is necessary due to ‘reliability’ and ‘data sovereignty/privacy’ reasons. The distribution of payload across several sites improves the reliability. In case of a failure of one site other sites still provide the payload and could be used to recover payload at the failed site if necessary.

Scientists are unwilling to use an archive if they lose control of their payload in general [14]. Therefore, the focus of non-public payload is to provide the data owner with a maximum of control and safety of his data. This has to include additional system administrators with potential full access. Consequently, non-public payload is stored at the site of the owner only.

The infrastructure concept gains only the score one regarding ‘data sovereignty/privacy’, which seems not to be sufficient if privacy is that important. One scoring point is missing due to two different authorization mechanisms (RBAC, POSIX ACL). Nonetheless, two different authorization mechanisms are not an obstacle because this enables distinguishable user rights for payload and metadata. The additional authorization step (via ACLs of XtreemFS) before a user is able to access payload could also be understood as additional security mechanism to improve ‘data sovereignty/privacy’.

Besides the importance of privacy, it is crucial for the success of an archive that some metadata about the payload gets published, e.g. the owner and his contact information. Without these information an interested scientist, using the infrastructure to search relevant payload would just determine the existence of the payload, but no responsible person to ask for access for.

The infrastructure achieves another one point score regarding ‘audit trail’. The audit trail is not seamless because the payload and metadata will not be imported directly from the technical machines where they are produced and GeMeCo will not implement unforgeable logging/reporting functions. An audit trail is important to prove

provenance and detect illegal/ unauthorized data access. Although the infrastructure should only contain research data, which are less critical than healthcare data, a seamless audit trail should be intended.
The last one point score goes to ‘data migration’. This is due to missing mechanisms for migrating data formats. If the data curator wants to migrate data formats, it is only possible to select the corresponding payload, download, convert and upload it e.g. with remarks to the original version / format. The item history would not contain an entry for the migration process. The relation between original and migration has to be described in the metadata of both (original and migration) data objects. Furthermore, a recursive concatenation of the item history of these data objects would be necessary to implement a complete audit trail. If this constraint could be solved appropriately is to be shown in the future operational model.

5 Conclusion & Outlook

The analysis of the concept regarding the requirements shows that it could be used for both use cases. The mapping of the OAIS functional entities onto the technical components of the LABIMI/F infrastructure concept reveals its OAIS conformity and proves that the concept includes all necessary functions.
The next steps in the LABIMI/F project will be to implement and configure all technical components, the development of GeMeCo and the conception of an operational model to work with and maintain the infrastructure.

6 References

[1] German Research Foundation. Recommendations of the Commission on Professional Self Regulation in Science 1998.

[2] Alexander von Humboldt Foundation, German Academy of Sciences Leopoldina, German Research Foundation, Fraunhofer-Gesellschaft, Helmholtz Association, German Rectors' Conference, Leibniz Association, Max Planck Society, German Council of Science and Humanities. Principles for the Handling of Research Data 2010.

[3] Dickmann F, Grütz R, Rienhoff O. A „meta"-perspective on „bit rot" of biomedical research data. Stud Health Technol Inform 2012;180:260–4.

[4] Dickmann F, Rey S. Stakeholder Analysis for Digital Preservation in Biomedical Research. Proceedings of the 13th World Congress on Medical Informatics, Berlin: IOS Press; 2010.

[5] Gruetz R, Brodhun M, Loehnhardt B, Dickmann F. Evaluation of data management and transfer tools for the biomedical community. 2012 6th IEEE International Conference on Digital Ecosystems Technologies (DEST), 2012, S. 1–6.

[6] Hertel F. D4.1: Ergebnisse aus dem Workshop zum Use Case Bilddaten 2012.

[7] Herrmann A, Hampe J. Workshop Genomdaten - Ergebnisportokoll 2012.

[8] TMF e.V. D038-01 Langzeitarchivierung. Langzeitarchivierung 2012.

[9] Consultative Committee for Space Data Systems. Reference model for an open archival information system (OAIS) 2012.

[10] Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The extensible neuroimaging archive toolkit - An informatics platform for managing, exploring, and sharing neuroimaging data. Neuroinformatics 2007;5:11–33.

[11] Hupfeld F, Cortes T, Kolbeck B, Stender J, Focht E, Hess M, Malo J, Marti J, Cesario E. The XtreemFS architecture—a case for object-based file systems in Grids. Concurrency and Computation: Practice and Experience 2008;20:2049–60.

[12] Smith M, Barton M, Branschofsky M, McClellan G, Walker JH, Bass M, Stuve D, Tansley R. DSpace. D-Lib Magazine 2003;9.

[13] Strodl S, Rauber A. Preservation Planning in the OAIS Model. New Technology of Library and Information Service 2008;1:61–8.

[14] Nelson B. Data sharing: Empty archives. Nature 2009;461:160–3.