

Diversity and Interoperability of Repositories in a Grid Curation Environment

Andreas Aschenbrenner

Goettingen State and University Library,
Germany

Austrian Academy of Sciences, Austria
aschenbrenner@sub.uni-goettingen.de

Jens Ludwig,

Goettingen State and University Library,
Germany

ludwig@sub.uni-goettingen.de

Harry Enke,

Astrophysical Institute Potsdam, Germany

henke@aip.de

Thomas Fischer,

Goettingen State and University Library,
Germany

fischer@sub.uni-goettingen.de

Abstract

IT based research environments with an integrated repository component are increasingly important in research. While grid technologies and its relatives used to draw most attention, the e-Infrastructure community is now often looking to the repository and preservation communities to learn from their experiences. After all, trustworthy data-management and concepts to foster the agenda for data-intensive research [1] are among the key requirements of researchers from a great variety of disciplines.

The WissGrid project [2] aims to provide cross-disciplinary data curation tools for a grid environment by adapting repository concepts and technologies to the existing D-Grid e-Infrastructure. To achieve this, it combines existing systems including Fedora, iRODS, dCache, JHOVE, and others. WissGrid respects diversity of systems, and aims to improve interoperability of the interfaces between those systems.

1. Community Requirements

Adequate curation of digital data certainly improves - amongst other - the collaboration across fields (e.g. through interoperability), quality of research (e.g. through better validation of research results), and lowers overall costs (e.g. through re-usability). Initiatives like the Australian National Data Service (ANDS) [3], DataNet in the USA [4], and the nascent PARADE in Europe [5] aim to tap into these opportunities, and so does WissGrid inside the German digital infrastructure D-Grid [6].

One key objective for WissGrid is to foster sustainable organisational structure for academia within D-Grid and to support forming of new academic community grids. A complementary objective is to promote sustainability of scientific data management, its long-term curation and cross-disciplinary re-use. In this, WissGrid represents a growing number of disciplines, starting from astronomy, high energy physics, climate research, medicine, philology and includes now photon sciences, bio-statistics, and others. The requirements of its communities with regard to data management and curation vary considerably; to name but a few:

- some of the communities already have large-scale existing data management systems (e.g. the climate community), while others do not and can hardly muster the knowledge and resources to establish such systems alone (e.g. bio-statistics, social surveys);
- data is homogeneous in some communities (e.g. high energy physics, astronomy), while hugely heterogeneous in others (e.g. bio-statistics, medicine);
- in some contexts data must be immutable and its integrity must be ascertained (e.g. climate, astronomy), while others expect a data lifecycle where data can be changed in early phases (e.g. philology), or data must be erasable at any time for legal reasons (e.g. current German languages);
- digital rights management may need to accommodate an all-encompassing open access policy in some communities (e.g. climate), while others need to deal with licensing, to anonymise (personal) data, defining thresholds for private data, and similar issues.

Overall, the diversity between (and even within) the communities makes it impossible to aim for a single strategy or system of curation for technical, organisational and social (e.g. trust) reasons. Any approach that deals with the meaning and context of digital objects requires a more targeted approach adapted to the specific needs of the community. Therefore, WissGrid aims to support the communities in establishing their own curation strategies and systems, and supports convergence and exchange of experiences between them. The following sections (cf. sections "Common Curation Terminology" and "Common Curation Infrastructure") present the technology agenda for achieving this and the common terminology on which the different academic grid communities in the WissGrid project agreed. At the core of the technology agenda is the integration of repository systems into the existing research environments of the communities (cf. section "Grid-Repository Integration Patterns").

2. Common Curation Terminology

Even if no single strategy or system is possible for such diverse academic communities, it is necessary to settle on a common terminology and concepts. Especially concepts like archiving, preservation or curation are used very differently in different contexts. In WissGrid, a basic common terminology was derived from the three abstraction levels of digital objects suggested by Thibodeau: "Every digital object is a physical object, a logical object, and a conceptual object, and its properties at each of those levels can be significantly different. A physical object is simply an inscription of signs on some physical medium. A logical object is an object that is recognized and processed by software. The conceptual object is the object as it is recognized and understood by a person [...]" [7].

Since the properties of each of those levels differ significantly, very different measures have to be taken at each of those levels to ensure the reusability of research data and can therefore define different curation or preservation levels. The three corresponding curation levels are "bitstream preservation" for the physical object, "content preservation" for the logical object and "data curation" for the conceptual object (see figure 1). Although these different activities have only limited overlap, they usually depend on each other in practice. Without bitstream preservation it is impossible to curate the data on the long run. Each curation level is independent of the archiving duration; also a change of levels may occur in the lifetime of a digital object, e.g. if an object is no longer actively used.

While these curation levels can be considered complete regarding the challenges related to object properties, sustainable financing, organisational stability and legal certainty are examples of common task areas on any level.

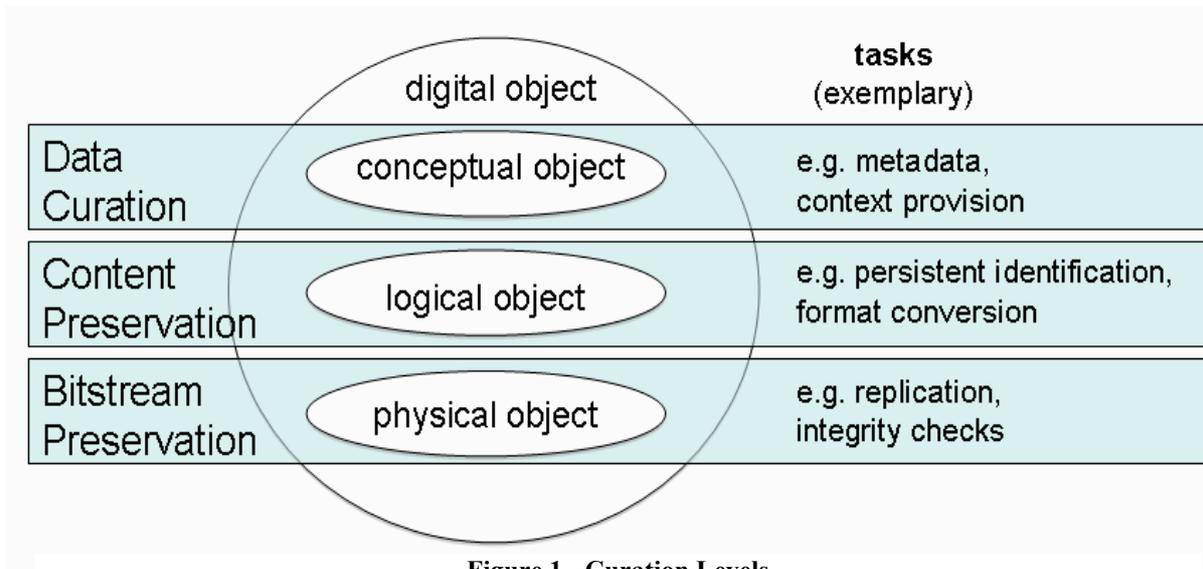


Figure 1 - Curation Levels

In detail the different curation levels are specified as follows:

Bitstream Preservation

Bitstream preservation ensures that every bit of a data object is retrievable without unintended modifications. It aims at basic technical stability and hardware evolution and addresses e.g. the decay of storage media or the obsolescence of storage technology. Primary factors of bitstream preservation are the number of copies, the distribution and independence of copies (geographic, but also organisational, financial, technological and political), the reliability and durability of the storage technology and regular integrity tests [8]. In Germany, the DFG requires since 1998 bitstream preservation for good scientific practice [9]. As primary service providers for bitstream preservation we envisage data centers, which can offer defined quality control levels, risk assessment or guarantees.

Content Preservation

For citations it is not sufficient that the bits of the object are still present. The used technology has to reproduce the content ensuring its authenticity even if the original technical environment might be no longer available. This is a shift of perspective from technical stability to technical reusability. Major factors for this curation level are continuous technology watch, technical quality assurance and a strategy for technical preservation measures. Typical examples for institutions, which aim to achieve this level of preservation, are cultural memory institutions like libraries, which are currently dealing mainly with published and static documents.

Data Curation

Content preservation or preserving the technical reusability may often be sufficient for finished and stable objects. But to ensure that research data is still of value for research does require also intellectual reusability. Without background information, e.g. of the configuration of an experiment, the data may be perfectly valid from the software perspective, but completely useless for researchers. This goes beyond the common approach of static conservation after production of an object, since the whole life cycle is considered. Also, later modification and enrichment is not precluded and may be necessary and allowed for objects, which are not considered historic. Data curation includes the use of data and appropriate

metadata, the integration of functionality in virtual research environments, versioning of objects, curation of access rights, appraisal, collection building, etc.

This is a more specific notion of data curation than the definition of the Digital Curation Centre (DCC) as "maintaining and adding value to a trusted body of digital information for current and future use" [10]. The WissGrid notion limits data curation on the conceptual object layer while the DCC also includes aspects of content preservation and bitstream preservation.

The data curation tasks are not accomplishable by community external service providers alone. The complexity and necessary background knowledge requires the participation of people with considerable acquaintance with the subject. Service providers on this level are therefore institutions working closely together with or part of the scientific community, like e.g. World Data Centers.

3. Common Curation Infrastructure

For a cross-disciplinary project like WissGrid neither bitstream preservation nor data curation are perfect matches for supporting a variety of communities. As mentioned the main service provider for bitstream preservation are data centers, which can exploit an economy of scale for offering storage. What a cross-disciplinary project can do and what WissGrid tries, is to articulate the need for storage with defined integrity requirements and foster their discussion with the scientific data centers in the D-Grid initiative.

On a data curation level the main infrastructure tasks are related to virtual research environments where the intellectual operations on research objects are performed. But these need to be dealt with on a user-specific level, and to be tailored to the individual requirements and context of the respective user community. As a cross-disciplinary project WissGrid sees the main value it can provide on a data curation level in best practice "blue prints" for data curation and data management checklists, which remain to be tailored to research communities in Germany.

The content preservation level seems to be the most appropriate curation level for which services spanning multiple disciplines can be developed. There are a couple of generic and discipline independent tools available from the preservation community. WissGrid integrates these tools in a grid environment, adds community specific modules and thereby not only fulfills the requirements of the target communities but hopefully also harnesses grid technologies and experience for scalability. The JHOVE2 tool [11] for file format validation and metadata extraction is one of these tools, which WissGrid regards as a strategically important tool for quality assurance. Conversion services to support interoperability are another.

Standing out from the WissGrid services for content preservation is its repository strategy. Other than e.g. format validation or migration tools, a repository encompasses all curation levels. They allow to manage digital objects instead of simple files and cover a range of functionalities from actual storage (and hence bit preservation) to metadata modeling and service provision (and hence data curation). The following section describes how WissGrid aims to achieve this while remaining generic and open to the diverse and changing requirements of the communities. (For an overview of the WissGrid architecture see figure 2)

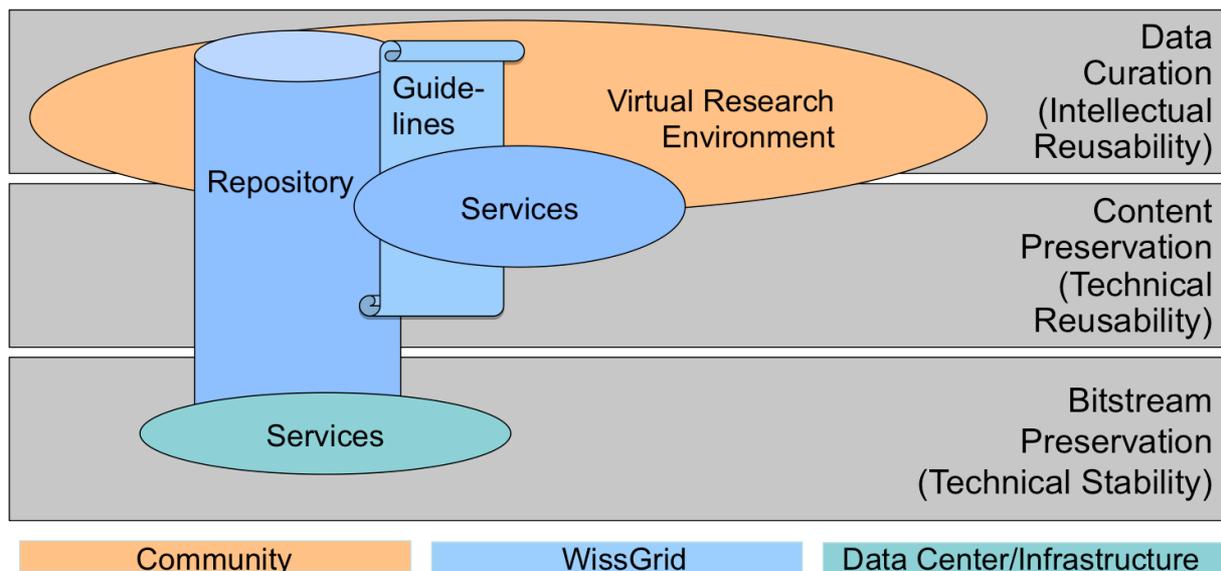


Figure 2: WissGrid Architecture

4. Grid-Repository Integration Patterns

Despite the heterogeneous requirements from the communities, there are three basic patterns with regard to the integration of repository systems into existing research environments. These three patterns may support and complement each other, each with a distinct set of standards, technologies, and research questions.

For each of these integration scenarios, WissGrid aims to provide a service package consisting of a technology stack and support. These packages allow communities with various requirements and different levels of expertise to establish their own curation systems that are interoperable with the grid environment or making their grid systems "curation ready".

Workflows in Data-intensive Science

This pattern involves the re-use and processing of digital objects, which are managed and preserved in digital repositories, in grid-based applications. This is of particularly importance in scenarios of data-intensive science, in which large amounts of data are managed, shared, and processed. David De Roure and Carole Goble [12] describe the properties of research objects for data-intensive science to be replayable, repeatable, reproducible, reusable, repurposeable and reliable. These properties require both reliable storage as well as powerful object modelling capabilities. Offering mechanisms for object modelling, various repository systems are capable of managing metadata with digital objects, rights management, versioning and weaving a network of relations between associated objects.

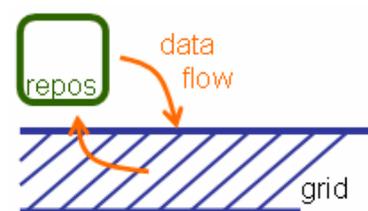


Figure 3: Grid Repository Interaction Pattern

There are two basic variations of the interaction between repository-based systems for data management and computational grid environments, which generate and process data.¹

Data to service - The e-Science community has been working on enabling scientific workflows, notably in the social workflow management platform myExperiment [13]. There is a multitude of workflow management systems and they often support interaction with

¹ In this context, we do not include ingest and access workflows. While they may involve or benefit from grid-repository-interaction, they are covered by the technologies mentioned in the interaction patterns "data to service" and "service to data".

repository-based systems, where SOAP- or REST-based interfaces are available for data access. However, while data files or data streams can be handled, scientific workflow engines are often unaware of metadata, object relations, or other aspects offered by repository systems.

Service to data - Due to bandwidth, large amounts of data are often better processed where they are stored and managed, rather than transferring the data before processing. However, while service workflows are well supported, data-flows often remain unaddressed. The prime reason for this is security issues when executing user-generated software within a trusted archive, and the potential performance impact this could have on the archive servers. OSGi [14] and other software container standards are key enablers for the execution of services close to the data; and e.g. Map/Reduce algorithms [15] demonstrate a different approach to this. Enabling the execution of services within or close to data repositories will become increasingly important in scientific environments, e.g. for text mining or other data analysis applications, but also for enabling administrative tasks such as format conversions for preservation.

Both patterns, "data to service" and "service to data", require vertical functionalities that bridge repositories and services. This includes e.g. authentication and rights management, which may need to be aligned across distinct technical environments. For example, repository systems often apply Shibboleth, OpenID or similar web-based authentication mechanisms, whereas grid environments employ Public Key Infrastructure (PKI). Gateways between those technical paradigms have been developed, e.g. Short-Lived Credentials as applied in D-Grid [16].

Another vertical functionality includes tracking provenance when data are being manipulated through human or machine agents. Provenance is essential for ensuring the authenticity of data and consequently the trust of researchers in data objects. The repository and preservation communities have developed provenance concepts in e.g. the PREMIS metadata framework [17], which is capable of describing events on objects. The e-Science community is e.g. developing the Open Provenance Model [18], which is a service framework for provenance information in service environments. While those concepts may be complimentary and may need to be combined in a trusted environment that enables grid-repository interaction, it remains an issue for research. WissGrid is collaborating with the D-Grid Integration Project (DGI) to implement provenance across D-Grid. [19]

Grid-based Repository Storage

This pattern introduces horizontal layers separating data storage (physical level) and object modelling (logical level). Repository storage is handled by a storage provider (e.g. institutional clouds, the national grid infrastructure), which transparently caters for all the functionalities needed for reliable storage. Storage providers may vary on the functionalities they offer, including the following:

- bitstream preservation - e.g. data replication (to distinct geographic locations), recurrent integrity checks, migration to fresh media
- data consistency - locking of files on access, ensuring reliable transactions (ACID - atomicity, consistency, isolation, durability)
- various technical interfaces - HTTP-based access for embedding into web environment; OAI-PMH [20] and other federation standards; other interfaces for direct access in grid or other technical platforms (e.g. through GSI-FTP)
- advanced functionalities



Figure 4: Grid Repository Storage Pattern

- provenance records and versioning
- rights management and licensing (e.g. a moving wall for publication)
- lifecycle management that supports phases of active changes to digital objects, freezing them on publication, as well as retention periods (e.g. resource can be disposed of after 10 years)

Various efforts for employing external storage paradigms have been undertaken in repository systems, for example for the DSpace repository system: the SRB storage handler [21], the Amazon S3 interface [22], as well as an SRM storage layer [23]. However, none of them became widespread and was adopted across repository systems, since they merely offered file storage and failed to offer some of the functionalities mentioned above. Ongoing discussions about a generic high-level storage for repositories [24] in the context of the Duraspace initiative (Fedora and DSpace) may change that.

WissGrid is contributing to these discussions and has recently implemented an Akubra storage module [25] for the iRODS data grid [26]. Using iRODS for storage of files managed by Fedora turned out to be fairly straightforward, however we intend to improve the interaction between iRODS and Fedora along several lines which are partly inspired by the PODRI project [27] which works on similar tasks:

1. direct ingest - WissGrid communities may be faced with situations where huge amounts of data need to be ingested into the repository directly from scientific instruments. In order to ensure performance in batch ingest, WissGrid aims to facilitate direct ingest into iRODS (e.g. using gsi-ftp) with a lazy update of Fedora. This requires a callback mechanism that informs Fedora about new files in the storage layer - a requirement that was previously raised in discussions about a high-level storage layer.
2. file structure defined by user - When directly ingesting data into iRODS, user communities may be unwilling to employ the specific storage structure of the repository (e.g. file naming conventions, folder structure). Rather, they may prefer to just deposit the data using the existing structure, and retrospectively annotate files with the required metadata.
3. rights management - Rights management schemes in Fedora and iRODS differ, with the former employing an XACML/SAML-based framework and the latter a proprietary mechanism. However, for enabling multiple, independent entry-points via iRODS and Fedora, rights management between the systems needs to be synchronised.
4. metadata in iRODS - In the current set-up, iRODS is mainly used as a blob-store without benefiting from its advanced functionalities in rule-based data management. The activation of the rule system requires the duplication of a subset of metadata into the iRODS metadata database iCAT. This will enable functions like policy-based data replication and rule-based preservation support [28].

Including those four additional properties in the iRODS-Fedora interaction go a long way towards enabling the functionalities listed above, and hence realises a repository storage handler that goes beyond a mere blob-store.

Repository Federation

This pattern aims to federate distinct data sources that exist within a single community or multiple communities. In doing this, it aims to capture all object properties including data, metadata and relations to other objects.

The repository community has been working on federation protocols for many years, including the widespread Z39.50 [29] and OAI-PMH [20]. Technically, these protocols are not related to grid technologies in any way, however the concept of virtualising repositories is related. Originally stemming from the library and repository community, protocols like OAI-PMH are increasingly being picked up by scientific communities as well. [30]

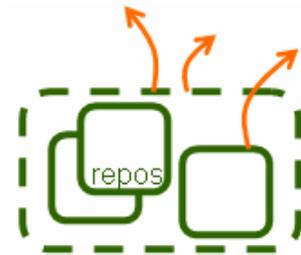


Figure 5: Repository Federation

In addition to basic metadata federation, other mechanisms may be needed in order to allow for processing actual content data, deal with heterogeneous research data, ensure consistency across repositories in the face of changing objects, enable other applications than mere "search" and others. A combination of mechanisms including CQL/OpenSearch, OAI-ORE, Sitemaps, and others may help achieve these requirements. We developed a respective pattern language for repository federation in [31].

5. Conclusion

While the key use case for repositories used to be that of a publication archive, there are now much more varied scenarios, incorporating them into a data management infrastructure component for research environments.

This paper presented the terminology and concepts developed in the cross-disciplinary WissGrid project and the grid-repository integration patterns to be implemented. For bitstream preservation, WissGrid supports the provision of a generic service to be established by D-Grid. However, for data curation, there is no single solution for the heterogeneous requirements of the diverse research disciplines. Rather than creating a single preservation system, WissGrid therefore aims to adapt existing preservation tools into the D-Grid infrastructure. Eventually this will lead to a pool of reference software that can be selected and customized for a specific preservation strategy and can be integrated into existing systems.

We are convinced that sustainable curation of cross-disciplinary research data can only be achieved by collaboration of the repository, the preservation and the e-Infrastructure communities. Where interoperability and re-usability can be achieved, diversity benefits the research community. However, it is at the same time a great challenge for the infrastructure.

6. Acknowledgments

This paper builds extensively on the work previously done in the WissGrid project and in the digital curation, digital preservation, digital repositories, grid, and adjacent fields. This work is funded by the German Federal Ministry of Education and Research (BMBF).

7. References

- [1] National e-Science Centre, Data-Intensive Research: how should we improve our ability to use data. e-Science Theme, March 2010. <http://www.nesc.ac.uk/esi/events/1047/>
- [2] WissGrid - Grid for the Sciences, a D-Grid project. Funded by the German Federal Ministry of Education and Research (BMBF). <http://www.wissgrid.de>
- [3] Australian National Data Service, ANDS. <http://ands.org.au/>
- [4] DataNet - Sustainable Digital Data Preservation and Access Network Partners. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141

- [5] Partnership for Accessing Data in Europe, PARADE.
<http://www.csc.fi/english/pages/parade>
- [6] Heike Neuroth, Martina Kerzel, Wolfgang Gentzsch (eds.): German Grid Initiative. Universitätsverlag Göttingen: 2007. <http://tinyurl.com/36cmm2o> or <http://www.univerlag.uni-goettingen.de/content/list.php?details=isbn-978-3-940344-01-4¬back=1>
- [7] Kenneth Thibodeau.: Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. CLIR Report, 2002.
<http://www.clir.org/pubs/reports/pub107/thibodeau.html>
- [8] Baker et al., A Fresh Look at the Reliability of Long-term Digital Storage, EuroSys2006, <http://www.cs.kuleuven.be/conference/EuroSys2006/papers/p221-baker.pdf>
- [9] Deutsche Forschungsgemeinschaft, Proposals for safeguarding good scientific practice, Bonn 1998, <http://tinyurl.com/2w8juez> or http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf
- [10] Kevin Ashley, Curated databases and data curation, DCC 2009,
<http://www.dcc.ac.uk/news/curated-databases-and-data-curation>
- [11] JHOVE2, The Next-Generation Architecture for Format-Aware Characterization,
<http://bitbucket.org/jhove2/main/wiki/Home>
- [12] David De Roure, Carole Goble: Research Objects for Data Intensive Research. Submitted to eScience 2009.
http://wiki.myexperiment.org/index.php/Research_Objects_for_Data_Intensive_Research
- [13] David De Roure, et. al: Designing the myExperiment Virtual Research Environment for the Social Sharing of Workflows. In: Proceedings of IEEE e-Science and Grid Computing, 2007.
- [14] OSGi - Open Services Gateway initiative. <http://www.osgi.org/Main/HomePage>
- [15] Jeffrey Dean, Sanjay Ghemawat: MapReduce: simplified data processing on large clusters. In: Communications of the ACM volume 51, issue 1. January 2008.
- [16] DFN-PKI: SLCS - Short Lived Credential Service.
<http://www.dfn.de/index.php?id=76085>
- [17] PREMIS preservation metadata standard. Maintained by the Library of Congress.
<http://www.loc.gov/standards/premis/>
- [18] The OPM Provenance Model (OPM). <http://openprovenance.org/>
- [19] Stefanie Rühle, Sven Vlaeminck: Template für Metadaten-Policies in den Grid-Communities. Die Dokumentation von Provenienz im Grid-Umfeld (in German).
<http://tinyurl.com/ybr3rxb> or http://dgi.d-grid.de/fileadmin/user_upload/documents/DGI2-FG4/FG4-6-metadaten/Template_Metadaten_Policies.pdf
- [20] Carl Lagoze, Herbert Van de Sompel (eds.), The Open Archives Initiative Protocol for Metadata Harvesting, 2002, <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- [21] Chaitanya Baru, Reagan Moore, Arcot Rajasekar, and Michael Wan: The SDSC storage resource broker. In Proceedings of the 1998 conference of the Centre for Advanced Studies on Collaborative research (CASCON '98), Stephen A. MacKay and J. Howard Johnson (Eds.). IBM Press 5-.
- [22] David Flanders: Fedorazon - Final Report. Project Report. 2009.
<http://ie-repository.jisc.ac.uk/426/>

- [23] Andreas Aschenbrenner, Flavia Donno, Senka Drobac: Infrastructure for Interactivity -- Decoupled Systems on the Loose. In: Proceedings of the IEEE Digital Ecosystems and Technologies Conference (DEST) 2009, Istanbul, Turkey. 1-3 June 2009.
- [24] Fedora Repository Development Wiki: High Level Storage. (Viewed December 2010) <https://wiki.duraspace.org/display/FCREPO/High+Level+Storage>
- [25] Fedora Commons Technology Roadmap V0.9, February 2008. <http://fedora-commons.org/pdfs/FedoraCommonsRoadmapDraft.pdf>
- [26] Arcot Rajasekar, et. al: iRODS Primer: Integrated Rule-Oriented Data System. In: Synthesis Lectures on Information Concepts, Retrieval, and Services. 2010. (doi:10.2200/S00233ED1V01Y200912ICR012)
- [27] David Pcolar, Daniel Davis, Bing Zhu, Alexandra Chassanoff, Chien-Yi Hou, Richard Marciano: Policy-Driven Repository Interoperability: Enabling Integration Patterns for iRODS and Fedora. iPRES 2010, <http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/pcolar-41.pdf>
- [28] Perla Innocenti, Brian Aitken, Adil Hasan, Jens Ludwig, Elena Maciuvite, José Barateiro, Gonçalo Antunes, Martin Mois, Gerald Jäschke, Wolfgang Pempe, Tom Wilson, Andreas Hundsdoerfer, Alfred Krandstedt, Seamus Ross, SHAMAN Requirements Analysis Report (public version) and Specification of the SHAMAN Assessment Framework and Protocol, SHAMAN Project 2009, <http://tinyurl.com/2wofa35> or https://shaman-ip.eu/shaman/sites/default/files/SHAMAN_D1_2Requirements Analysis ReportSHAMAN Assessment Framework.pdf
- [29] Clifford A. Lynch: The Z39.50 Information Retrieval Standard - Part I: A Strategic View of Its Past, Present and Future. In: D-Lib Magazine, April 1997. <http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/april97/04lynch.html>
- [30] Wolfgang Hiller, Kerstin Fieg: Daten-Lebenszyklus-Management - Anwendungsbeispiel C3Grid. D-Grid All-Hands-Meeting, September 2007. <http://tinyurl.com/38jenz6> or http://www.d-grid.de/fileadmin/user_upload/images/D_Grid_AHM_0907/presentations/D-Grid-AHM_workshop-knowledge_3-anwendung-daten-lebenszyklus-management-fieg.pdf
- [31] Andreas Aschenbrenner, Tobias Blanke, Marc W. Küster, Wolfgang Pempe: Towards an Open Repository Environment. In: Journal of Digital Information (JoDI), Vol 11, No 1 (2010).