Diversity and Interoperability of Repositories in a Grid Curation Environment

Repository-based environments 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.

Community Requirements

There is increasing recognition that adequate curation of digital data potentially 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.

WissGrid is part of the German digital infrastructure D-Grid [6]. One key objective is to further organisational sustainable structure for the academic world within D-Grid and to support forming of new academic community grids. A complementary objective is to foster sustainability of the scientific data, 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 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-statistic, the social survey);
- data is homogeneous in some communities, while hugely heterogeneous in others (e.g. bio-statistics);
- 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, 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-embracing open access policy in some communities (e.g. climate), while others need to deal with licensing, data de-personalization, moving walls for private data, and similar issues.

Overall, the diversity between (and even within) the communities makes it impossible to aim for a single strategy and system of curation for technical, organisational and social (e.g. trust) reasons. Any approach that goes beyond mere bit preservation and deals with the meaning and context of digital objects requires a more targeted approach that is 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 section (cf. section "Curation Infrastructure") presents 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").

Curation Infrastructure

The Digital Curation Centre in the UK defines digital curation to involve "maintaining and adding value to a trusted body of digital information for current and future use" [7]. When looking closer at digital curation, we distinguish three curation levels (see figure 1). These follow the abstraction levels of digital objects suggested by Thibodeau [8] and they are recognised in the preservation community. Each level may change depending on the context and evolve over time. The levels build upon each other and may influence and support each other.

- Bitstream Preservation - ensuring the integrity of each bit, by monitoring the physical stability of data, and moving data to fresh and up-to-date carriers.
- Content Preservation - maintaining the citability and accessibility of data through e.g. format conversion
- Data Curation - capturing the meaning and intellectual context of an object over time, fostering comprehension and reusability when the original context has disappeared

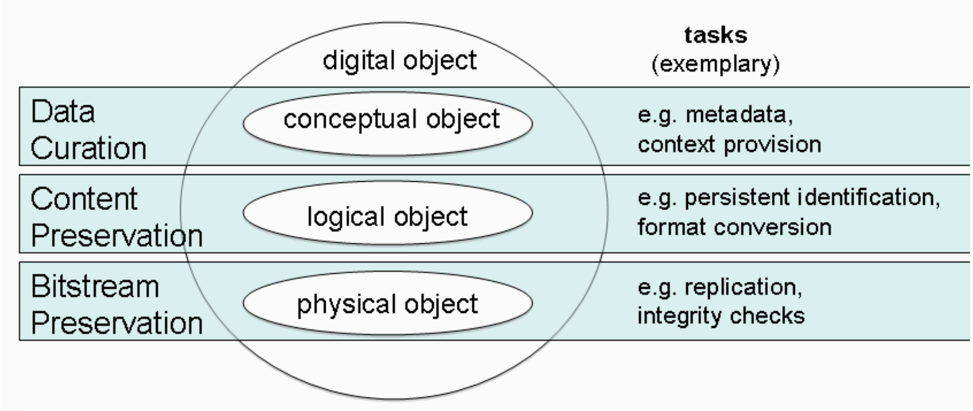


Figure 1 - Curation Levels

WissGrid advocates the creation of storage trust zones offering Bit Preservation in digital infrastructures such as D-Grid. Its focus area for Content Preservation is developing reference software and support. Data Curation needs to be dealt with on a user-specific level, and tailored to the individual requirements and context of the respective user community.

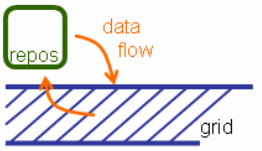
For Content Preservation , WissGrid is currently working with a range of preservation services including JHove and CRiB. WissGrid's key contribution is integrating these services into grid infrastructure and repository systems, while the actual services stem from the preservation community.

Standing out from the WissGrid services for Content Preservation is its repository strategy. Other than e.g. format validation or migration tools, a repository covers 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.

Grid-Repository Integration Patterns

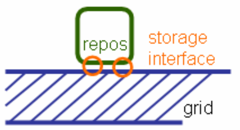
Despite the heterogeneous requirements from the communities, there are three basic patterns with regard to the integration of repository systems into existing research environments.

1.



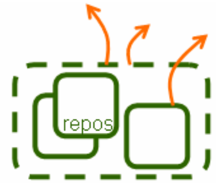
Digital objects, which are managed and preserved in digital repositories, are re-used and processed in scientific applications in the grid environment. In this scenario the repository (resp. multiple repositories) need (standard) interfaces for the data to be searched/filtered, extracted and written back in scientific workflows. Rights issues may need to be addressed (e.g. Shibboleth vs. PKI certificates through short-lived credentials) particularly in the case where over a long period of time numerous objects are automatically processed and written back to the repository recording their provenance and linking them to the initial objects.

2.



Repository storage is handled by a storage provider (e.g. institutional clouds, the national grid infrastructure), which transparently caters for all the functionalities needed to ascertain bit preservation (e.g. data replication, recurrent integrity checks). We are currently testing storage interfaces between Fedora and iRODS or dCache [9]. However, more work needs to be done with regard to security, so that digital objects can be accessed directly through grid mechanisms for processing (e.g. GSI-FTP).

3.



Federation of distinct data sources that exist within a single community or multiple communities. The federation protocol OAI-PMH is a starting point for achieving this. However, other mechanisms may be needed in order to allow for processing the content in addition to the metadata, deal with heterogeneous research data, ensure consistency across repositories in the face of changing objects, and other scenarios. A combination of mechanisms including CQL/OpenSearch, OAI-ORE, Sitemaps, and others may help achieve these requirements.

For each of these integration scenarios, WissGrid aims to provide a service package consiting of 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".

Conclusions

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

For bit 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. Convergence with regard to interfaces and formats (e.g. OAI-ORE, OpenSearch/CQL, GSI security, provenance) is often perceived as insufficient in the community. OpenRepositories may be the right forum to foster convergence in the field.

[1] 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). 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://www.univerlag.uni-goettingen.de/content/list.php?details=isbn-978-3-940344-01-4&notback=1

[7] Digital Curation Centre, http://www.dcc.ac.uk/

[8] 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

[9] 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.