Defining and Driving Standards for Digital Repositories
OGF Digital Repositories Workshop
Wednesday 4 March 2009,
EGEE 4th User Forum/OGF25 & OGF-Europe's 2nd International Event, Catania, Italy

Digital Repositories (DRs) are essential assets in enterprise, educational and research institutions, assisting these organisations in capturing, identifying, storing and retrieving a wide range of objects, from audiovisual material to datasets, from presentations to learning materials and research works. EC-funded projects focusing on DRs and their development include D4Science; DRIVER-II; EuroVO-AIDA and Genesi-DR.

In enterprise, digital repositories are often referred to in different ways, such as a data warehouse, enterprise content management, or document management systems. Enterprise digital repositories are usually based on commercial databases with extensive customisation and highly centralised storage with distributed retrieval and processing.

Use cases in enterprise include CNN's Media Information Retrieval Application, or MIRA, is one example of DR use in enterprise. MIRA is built on top of IBM DB2 universal database and DB2 content manager VideoCharger, in order to manage over 200,000 hours of archived video. The system meets the Production Team's archival video needs by enabling the handling, ingesting, cataloguing, searching, selection and fulfilment of content.

OGF-Europe places strong emphasis on Digital Repositories through a series of focused events to define specific requirements and pinpoint best practices through use cases, with the aim of accelerating the adoption of standards for interoperability. With grids and cloud computing now driving highly distributed architectures, standards are needed to leverage existing work by catalysing the DR community and providing targeted support within OGF. To this end, issues around metadata are key.

Building on the outcomes of the Digital Repositories Seminar in June, which helped to connect user communities and developers from the field to explore key horizontal issues, last December, OGF-Europe supported the RECURSE (Repository Curation Service Environments) Workshop held during the Digital Curation Conference in Edinburgh together with DReSNET (EPSRC Digital Repository eScience Network) & OMII-UK.

David Martin, International Center for Advanced Internet Research, Northwestern University and Data Area Co-Director in OGF, provided insight into how OGF can work with the DR community by pinpointing specifications and recommendations relevant to digital repositories as a first-step approach.

Core OGF Work of Interest to the Digital Repository Community

- ❖ **SRM – Storage Resource Manager**-providing location-independent storage access.
- ❖ **OGSA Data Architecture** – providing grid data architecture & integrated into overall service-oriented OGSA Architecture
- ❖ **GIN – Grid Interoperability Now** – working to document interoperability. The Information Services & Data Sub-teams are of most interest to the DR Community.

OGF Proposed Standard from the Grid Storage Resource Manager GSM-WG (Working Group): GFD-129. The Storage Resource Manager is designed to manage very diverse storage systems and very large data volumes. SRM is used in gLite, the middleware developed by the EGEE project, and elsewhere. There are currently six different implementations, soon to become seven.

Connecting the DR Community

The OGF25 Workshop, which is co-sponsored by OGF-Europe, OGF, DReSNET and OMII-UK, is the third in a series of workshops that have successfully connected the DR community to explore how the adoption of distributed computing and OGF standards can respond to the specific needs of the community.

With the emergence of the knowledge-based economy, interoperability of digital repositories (DR) offers benefits not only for different types of users within the user community, but also eScience application developers. Through interoperability, e-Infrastructure providers can ensure the easy deployment and management of software distributions. eScience users have the freedom to choose services deployed in different infrastructures, based on functionality, with the potential to use a far greater amount of resources than is currently available to them. For eScience application developers on the other hand, interoperability ensures the portability of applications across multiple infrastructures to increased uptake.

The OGF25 Workshop explores these issues to highlight how the adoption of distributed computing and implementation of OGF standards can address the needs of the DR community and the specific challenges of using DRs in a grid or cloud environment. The Workshop aims to strengthen synergies between OGF and the DR community and define the scope and long-term goals of a dedicated OGF DR community/research group. Discussions will thus help identify the requirements for setting up such a group, as well as understand how to liaise with other standardisation development organisations (SDOs).

Further reading 1: OGF-Europe eAnnouncement, Issue 1, 2009, *Standards as key to Interoperability* on OGF-Europe's new Channel – eAnnouncements.

Workshop Agenda

The Workshop opens with a spotlight on **Policy & Standards**, chaired by Neil Chue Hong, OMII-UK and OGF with talks by Mark Hedges, King's College, London on a view on Fedorazon.

The Workshop then explores **Use Cases** and **Best Practices** from diverse perspectives. Matt Zumwalt from MediaShelf offers an enterprise perspective; Reagan Moore, University of North Carolina, brings insight into research perspectives through iRods use cases, while preservation of cultural heritage through a grid based digital repository will be examined by Antonio Calanducci, Italian Institute for Nuclear Physics (INFN Catania) through the gLibrary.

The session then continues with Mark Hedges chairing talks on standards perspectives by Jens Jensen, Science & Technologies Facilities Council (STFC); interoperability perspectives presented by Pasquale Pagano from the Institute of Information Science & Technology, Italian National Research Council (CNR-ISTI) and Roberto Cossu, European Space Agency (ESA), exploring environmental perspectives through the Earth Observation Group.

The Workshop concludes with a roundtable debate chaired by Neil Chue Hong with a deliberation of challenges in Data Repositories in a grid environment with particular reference to security; authentication/authorisation; data formats; and data management. The wrap-up will define next steps for DR community engagement with OGF.

Further reading 2: 'A workshop Series for Grid/Repository Integration', DLib Magazine, Jan-Feb '09 by Andreas Aschenbrenner, Tobias Blanke, Neil P. Chue Hong, Nicholas Ferguson and Mark Hedges:
http://www.dlib.org/dlib/january09/aschenbrenner/01aschenbrenner.html