Feb 5 2009 – iRODS User Group Meeting – Lyon, France

Preservation Session:

Multiple topics were explored to understand whether the iRODS data grid provided the required capabilities. The topics included:

- Temporal scale for retention. Examples ranged from the Diamond Light Source which was required to hold observational data for 200 days, to the European Union laws that require all scientific data to be kept for ten years, to the TextGrid project which kept data for ten years. A required capability is a rule that automates disposition at a specified time interval. Typical disposition policies range from migration into an archive, to deletion of outdated data. A second required capability was the migration of records across different storage systems as more cost effective systems become available.
- Data caching. In many cases, active data sets were held on disk, and then migrated to tape archives after the data became less useful. A required capability is a rule that automates data movement between disk cache and tape archive. This is similar to Hierarchical Management System policies. A related question is whether a rule set can be designed for iRODS that implements the policies of a HSM.
- Data manipulation. For observational data, calibration data may need to be applied to the raw data for correct interpretation. A required capability is a rule that can be applied on each iget command. The rule might use the calibration data to return a calibrated version of the data.
- Persistent IDs. This is essential for long-term preservation. Four different name spaces were discussed: physical file name which is the actual storage location of the file; logical file name which can be organized into a collection hierarchy to support browsing; GUID globally unique ID; descriptive metadata attributes that can be queried. Each type of naming convention serves a different purpose.
- Multi-component preservation environments. The need to integrate multiple systems to build the preservation environment was discussed. Examples included the use of a separate metadata catalog to support browsing and discovery; the use of digital library services from Fedora or DSpace to support arrangement and retrieval of records; and the use of external identity and authorization management systems. A goal of the iRODS data grid is to support interoperability across the multiple components that are being integrated into the preservation environment.
- Preservation processing. A need was expressed for the automated processing of files on ingestion. Examples included the creation of derived data products, extraction of metadata, or conversion to an archival form on ingestion.
- Access. A stated requirement based on EU law was the need to publish scientific data on which publications were based. This implies the need to maintain links between publications in journals and data deposited into the data grid
- Federation. Many groups rely upon federation of independent data grids to enable sharing between project members. This was a primary requirement for the ARCS data grid.

- Medical patient records. Mechanisms are needed to encrypt data used in longitudinal studies. Two different requests were heard: one for encryption while the data was transmitted over the network and storage in an unencrypted form; a second for encryption at the client and the storing of encrypted data. The latter capability was provided with the SRB data grid. A similar capability could be created for iRODS, but would require submission to the US Export Control for review. It should be possible for open source software to support encryption.
- Medical patient records. An alternate approach was to store anonymized data, provided a mechanism could be implemented to go back to the original data and determine identity. This would require interaction with an external database that was under the control of the medical institution.
- Metadata replication. The ability to minimize risk of metadata loss was discussed. The two approaches of using vendor database replication mechanisms, and use iRODS-based federation were discussed. Between the two mechanisms, metadata could be backed up. A request was made for metadata replication between two different vendor database systems.
- Assessment criteria. Rules to validate properties of the preservation environment were discussed. Assessment criteria are being developed by multiple communities; RLG/NARA, TRAC, ISO MOIMS-rac, Drambora, Nestor/Kopal. The first three are closely related. A comparison of rules from the ISO MOIMS-rac standards effort with the TRAC criteria shows that very similar rule sets can be used to validate the preservation trustworthiness.
- Sustainability. Although this was discussed, no conclusions were reached. Economic arguments are needed that justify the expense of maintaining the archive. The examples that were discussed all relied upon demonstrating use of the material in the archives that influenced business policies or that served as reference collections for comparing with future research results.
- Access controls. The need for fine-grained access controls was reiterated for the DARIAH arts and humanities data grid.