Building a Collaborative Research Data Collection Employing Automated Policy Based Management

David Fellinger
iRODS Consortium, Chapel Hill, North Carolina, USA, dave.fellinger@gmail.com

The Requirement to Enable Collaboration

There is a growing requirement for collaboration within and across universities and research institutions. Collaborative efforts are being funded by organizations across the world with initiatives like the "Big Data Hub" project of the NSF in the USA and dictates for research data preservation and maintenance by the EPSRC in the UK just to mention two. Collaboration, however, poses a requirement for long term repositories that are well indexed and searchable. Traditional techniques for building an indexed archive rely on the human intervention of a librarian to "check-in" research works as publications that may be accessed by a campus wide or nationwide network. The problem is that the volume of the data and the diversity of maintenance and distribution mechanisms will eventually rule out the use of human intervention to accomplish these tasks.

As data sets increase in size, number, and complexity the need for automation in data management will become more apparent. The Integrated Rules Oriented Data System (iRODS) is a flexible, open source, product which can be utilized to enable automation providing compliance to retention and access policies, a searchable index, and a reporting infrastructure useful to institutions, funding organizations, and individual researchers.

The first challenge of ingesting data for publication is format compliance. An institution should publish guidelines and a template for metadata to allow automated ingestion so that the metadata can be extracted to a searchable data base. There must be a policy based process in the ingestion server that can check for compliance and extract the required metadata. This process should have the capability to accept some infractions to the required template but should also have the capability to reject the publication based on non-compliance notifying the researcher of the required changes. The publication could be placed in a temporary holding cache so that it can move into the repository once the small changes are made by the researcher. Notifications of acceptance and publication should be automatically sent to the researcher, advisors, and funding bodies based on the compliance requirements.

Another challenge is the establishment of access rights to both the publications and the index in whole or part. It is possible that some specific research work was funded by a commercial entity so the access privileges would be entirely different from a publically funded work. In fact, the index listing for some research works may have to be hidden from some users based upon the requirements set by the funding body. This is especially true for institutions that may have competitive commercial research contracts.

Data Retention as a Mandate

Data retention policies have become increasingly more important as funding bodies like the EPSRC have mandated that publically funded research work must be kept online for 10 years. At first glance, this policy seems easy to maintain with a simple incremental timer value in a database. The mandate is complicated by the fact that researchers cite the work of predecessors and a citation produces a link to the original work. The time must then be reset for the original work or the end result will be a broken link in 10+ years. Adherence to this mandate from the funding body may mean that a research document must be available for decades if it is a primary work in a specific area.

In some cases access tracking may also prove be a requirement and a rule could be written triggering an iRODS "microservice" which sends an email to a researcher when his work is cited or even examined. This service provides a report and essential data for a researcher applying for an additional grant based on the popularity and usefulness of his work.

A well designed index should also be updated with changes in a researcher's position and institution. This is important for maintaining both provenance and a chain of custody of the work and must follow the legal requirements of the institutions involved.

The entire concept of collaboration infers data distribution as well. If a joint study generates papers at several institutions, all of the papers will have to be distributed to all of the sites involved in the work to fulfill the policies of local content management maintained by most universities.

Long Term Hierarchal Storage Management

A key question will be, "is 10 years enough for critical research retention?" Institutions or funding bodies may establish longer term tape or optical archives so that the data can always be recovered. Moving data to this long term archive should also be an automated function.
Finally the entire process from ingestion through storage, distribution, and retention, must be verifiable through the use of audit tools. This audit will become a requirement of the funding bodies to verify that the guidelines are being maintained. Regular reports can be generated and automatically sent to various agencies without human intervention provided that the process is policy driven.

Enabling the Past to Communicate with the Future

Clearly, the establishment of a research repository goes far beyond the acceptance of papers and the incremental change to an index. Ingestion of a research work must launch an automated workflow that defines properties which, based on enforcement points, dictate policies and launch procedures which not only store the data but build a persistent data structure including an index with reporting criteria.

In a complex long term endeavor it is entirely possible that a researcher may be eventually collaborating with a colleague who has not yet been born. The entire collaborative process can only work if the younger scientist can easily access the work of his predecessors. As the volume of research increases, it is essential that the entire process of curation is automated and traceable.

ABOUT THE AUTHOR

Dave Fellinger is a Data Scientist with the iRODS Consortium. He has over three decades of engineering and research experience including film systems, video processing devices, ASIC design and development, GaAs semiconductor manufacture, RAID and storage systems, and file systems.

As Chief Scientist of DataDirect Networks, Inc. he focused on building an intellectual property portfolio and representing the company to conferences with a storage focus worldwide.

In his role with the iRODS Consortium Dave is working with users in research sites to assure that the functions and features of iRODS enable fully automated data management through data ingestion, security, maintenance, and distribution. He serves on the External Advisory Board of the DataNet Federation Consortium and was a member of the founding board of the iRODS Consortium.

He attended Carnegie-Mellon University and has been awarded patents in video processing, pattern recognition, motion control, data manipulation, and file systems.