

A Federated Repository Solution for Large Scientific Datasets

Steve Androulakis¹, Mark A. Bate², Anthony Beitz³, Wojtek Goscinski⁴ and Ashley M. Buckle⁴

¹Monash University, Melbourne, Australia, steve.androulakis@med.monash.edu.au

²Monash University, Melbourne, Australia, mark.bate@med.monash.edu.au

³Monash University, Melbourne, Australia, anthony.beitz@its.monash.edu.au

⁴Monash University, Melbourne, Australia, wojtek.goscinski@adm.monash.edu.au

⁵Monash University, Melbourne, Australia, ashley.buckle@med.monash.edu.au

THE INCREASING NEED FOR PUBLIC ACCESS TO SCIENTIFIC DATA

There is an increasing need for researchers to describe and share data sets associated with published scientific results with the wider scientific community. This need comes from scientific journals requesting data sets to be made available with results for verification purposes, but also from communities themselves. Having open access to data sets of publications allows researchers to learn more about others' findings, and process them to further their own research goals.

CHALLENGES OF SHARING DATA

Several challenges have arisen when attempting to share often-large data sets over the internet. There is usually no central location that has the capacity and funds to store the world's raw data for any one scientific discipline. Data sets, often several gigabytes, if not terabytes in size, are often too large to fit into the software constraints of traditional digital repositories which were created with the storage of papers and media clips in mind. The HTTP protocol itself isn't suited for the downloading of hundreds of files within a data set at once and needs to be countered.

Technological challenges aren't the only ones encountered when attempting to fill the data commons. Ease of data set deposition, description and citation are all considerations when trying to create a system that is widely adopted. Restricting the amount of mandatory metadata required to describe data sets is an important step towards reducing the need for time-consuming manual entry by the user, as is the automatic extraction of metadata from the data sets themselves, wherever possible. Many labs don't have direct access to computing expertise, thus data set deposition and repository setup must be as simple as possible.

A SOLUTION FOR THE STORAGE OF SCIENTIFIC DATASETS

We have created a solution that takes into account these challenges, using common web technologies, data annotation and deposition tools that can be deployed at any site cheaply and easily. TARDIS (<http://www.tardis.edu.au>) provides the protein crystallography community the means to easily describe their data sets, deposit them into a simple repository and have them displayed on the web. Data is shown alongside rich, searchable metadata and is freely downloadable. The TARDIS framework is designed to be easily adaptable to a wide range of scientific disciplines.

A FEDERATED APPROACH

TARDIS aims to eliminate the cost issues associated with a central repository for data by using a federated storage model. Through allowing labs and institutions to store data on their own servers and using TARDIS as a central index, the problem of storing large amounts of data and who will pay for such storage is eased.

REPOSITORY DESIGN BASED ON TECHNOLOGICAL STANDARDS

TARDIS' data storage is based on simple and free HTTP and FTP services. This eliminates the need for complex and difficult to configure digital repository software, and ensures storage of data for any lab is an easy task to achieve. All a lab or institution needs is an HTTP server, such as Apache and optionally an FTP server for easy mass downloading of files.

The storage of metadata follows the Metadata Encoding and Transmission Standard (METS) profile and the Science and Technology Facilities Council (STFC) data model, for easy communication and translation between different data management systems. As a result, TARDIS is well placed for integration with future systems.

Because the system is federated, and the metadata is purely XML-based, data is easily portable and consistently traceable by TARDIS.

PERSISTENT HANDLES FOR CITATION

Persistent identifiers provide permanent links to online data, ideal for citation in journal publications. The Australian National Data Service (ANDS) has provided TARDIS with a persistent identifier service, automatically giving users a handle in which to reference their data on experiment registration.

CUSTOMISED VIEWS FOR SCIENTIFIC DISCIPLINES

A modular back-end and template powered front-end ensures the TARDIS design can be extensively modified to suit the specific needs of individual scientific disciplines. We're currently working with different areas of research, such as Climate Mathematics to create rich views of the data they'd like published.