

The ARCS Data Services

Florian Goessmann, Pauline Mak and the ARCS Data Services Team

¹Australian Research Collaboration Service (ARCS), Florian.Goessmann@arcs.org.au , Pauline.Mak@arcs.org.au

INTRODUCTION

The Data Services Team of the Australian Research Collaboration Service was established with the mandate to provide researchers with robust tools and services that help them face the ever-growing challenge to deal with the increasing data volumes researchers are confronted with.

To achieve this, ARCS Data Services has developed three core products:

1. *the ARCS Data Fabric*

1.1. *ARCS webDrive*

1.2. *the ARCS OPeNDAP Network and Digital Library*

2. *the ARCS Database Service*

4. *the ARCS Data Transfer Service*

THE ARCS DATA FABRIC

OVERVIEW

The ARCS Data Fabric was developed as a solution for the growing need by researchers to easily share data across institutional boundaries. As such, it is a generic service that is not tied to any specific kinds of data or research disciplines and is available to every Australian researcher and their international collaborators.

Access to the ARCS webDrive is possible through either a WebDAV client such as Windows Explorer and Mac OS Finder, any modern web browser or command line tools. Dedicated areas of the ARCS webDrive can be accessed using the OPeNDAP protocol through the ARCS OPeNDAP Network and Digital Library. The authentication mechanism of the ARCS webDrive has been designed around methods and technologies supported by the Australian Access Federation (AAF).

The ARCS OPeNDAP Network and Digital Library was developed in particular for the research communities connected to the Integrated Marine Observing System (IMOS) and the Terrestrial Ecosystems Research Network (TERN) in order to provide these communities with a national framework that provides unified access to their datasets.

THE ARCS WEBDRIVE - ARCHITECTURE

The ARCS webDrive has two distinct layers, a backend and a frontend. The backend interfaces with the physical storage whereas the frontend provides the different interfaces to the user.

The backend of the ARCS webDrive is the Integrated Rule Oriented Data System (iRODS) which sits on top of physical, large-scale storage infrastructure hosted by and provided through the Members of ARCS (MARCS). This setup allows the ARCS webDrive to be expandable and fault tolerant, as it does not have to rely solely on one physical storage facility.

The frontend, Davis, of the ARCS webDrive is a development by ARCS Data Services. It provides two easy to use interfaces: a WebDAV server and web access. The WebDAV server allows researchers to access and store data in the ARCS webDrive with any WebDAV client including those built into operating systems such as Windows XP and Mac OS X. The web access is available through most modern web browsers. In addition to upload and downloads of data, the web interface also offers access control mechanism, metadata for files and collections as well as the 'trash can'.

THE ARCS WEBDRIVE - DATA SHARING AND ACCESS CONTROL

Giving researchers the ability to share data was the main drive for the development of the ARCS webDrive. As a result, the ARCS webDrive puts sophisticated access control mechanisms at the disposal of the researcher. It is possible to assign access of different levels (read, write, own) to single files or whole collections to individuals or groups.

If a group of researchers frequently shares files, they can request for a group to be created for them. This further simplifies sharing of data as each group owns a group collection to which makes all data stored in it immediately available to all group members.

THE ARCS OPeNDAP NETWORK AND DIGITAL LIBRARY - ARCHITECTURE

The system consists of two distinct parts: a network of data servers and a portal that harvests and catalogues information on all datasets handled by all data servers in the network.

The data servers run the THREDDS Data Server (TDS), an implementation of the DAP protocol. This protocol was designed for the delivery of scientific data over the web and is well established in the ocean, climate and remote sensing sciences communities. At this stage, ARCS hosts five TDS servers, based at MARCS closest to IMOS facilities.

The TDS servers are co-located with the servers for the ARCS webDrive and access a shared storage system. This setup allows data stored in the ARCS webDrive available through the ARCS OPeNDAP Network.

The digital library component is provided by an instance of the TPAC Digital Library. The digital library provides a single frontend to all datasets available through any of the data servers and hence enables researchers to discover datasets without prior knowledge of their physical location.

USE CASES

While the ARCS Data Fabric is being used everyday by individuals to store and share data, it is also integrated with eResearch service providers external to ARCS. The Australian Synchrotron's Virtual Beam Line data portal was developed to give users of the synchrotron an easy way to transport the results of experiments off the facilities to storage which provide access to data from their home institution. ARCS Data Services and the developers at the synchrotron have successfully worked together to integrate the ARCS Data Fabric as a storage selectable target for the data transport mechanism.

NCRIS IMOS/eMII is using both, the ARCS webDrive and the ARCS OPeNDAP Network and Digital Library to deliver datasets collected and hosted across Australia.

THE ARCS DATABASE SERVICE

After consulting with its user base, ARCS Data Services identified the need for a database hosting service. This has led to the development of the ARCS Database Service. This service enables ARCS to offer researchers hosting capabilities for MySQL and PostgreSQL databases in a fully redundant setup with multiple master and slave servers located in datacenters of several MARCS across Australia.

USE CASE

The Australian pilot of Human Variome Project, an international effort to collect information on all variations of the human genome that can cause disease, has been hosting its central database on the ARCS Database Service since its inception.

THE ARCS DATA TRANSFER SERVICE

ARCS Data Services does, on behalf of research groups, facilitate large-scale data transfers, both nationally and internationally on a manual, case-by-case basis. However, with the growing demand for such transfers, ARCS Data Services recognizes the need for an easy to use, automated approach. The ARCS Data Transfer Service is currently in the early stages of development.

CLOSING REMARKS

In order to meet its mandate of providing Australian researchers with first-class tools and services around data storage, transport, access and sharing, ARCS Data Services has been consulting closely with research communities to define core products aimed at covering a wide range of real-world requirements. However, where a particular need can't be addressed by one of these central products, ARCS Data Services will work with research groups to develop custom or customized solutions wherever possible.