The ARCS Data Fabric

Shunde Zhang (shunde.zhang@arcs.org.au), Florian Goessmann, Pauline Mak

The Australian Research Collaboration Service (ARCS)

Overview

The Australian Research Collaboration Service (ARCS) Data Fabric was developed as a solution for the growing need by researchers to easily store their research data, and to share that 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, and any modern web browser. 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 to be use methods and technologies supported by the Australian Access Federation (AAF).

1. Architecture

The ARCS Data Fabric consists of two modules: the ARCS webDrive and the ARCS OPeNDAP Network and Digital Library. They are both based on iRODS, the Integrated Rule-Oriented Data System, where data is stored.

1.1. ARCS webDrive

The ARCS webDrive as two distinct layers, a back-end and a front-end. The back-end interfaces with the physical storage, whereas the front-end provides the different interfaces to the user.

The back-end of the ARCS webDrive is 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 system.

The front-end, Davis, of the ARCS webDrive is a development by ARCS Data Services. It provides two easy-to-use interfaces: a WebDAV server and web browser 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 uploading and downloading data, the web interface also offers access control mechanisms, metadata for files and collections, as well as the ‘trash can’.

1.2. 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 and 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, which makes all data stored in it immediately available to all group members.

1.3. ARCS OPeNDAP Network and Digital Library

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

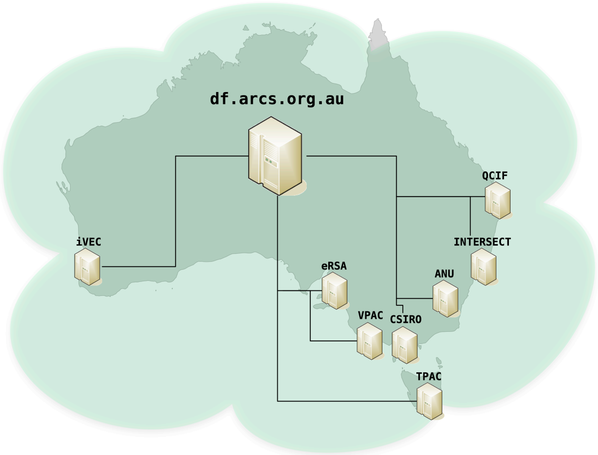
The data servers run THREDDS Data Server (TDS), an implementation of the DAP protocol. This protocol was designed for the delivery of scientific data over the web and is well established in the ocean, climate, and remote sensing sciences communities. At this stage, ARCS hosts five TDS servers, based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

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

The digital library component is provided by an instance of the Tasmanian Partnership for Advanced Computing (TPAC) Digital Library. The digital library provides a single front-end to all datasets available through any of the data servers, and hence enables researchers to discover datasets without prior knowledge of their physical location.

2. Current setup

Currently, seven iRODS nodes have been set up in each capital city of Australia. They are all in one iRODS zone, with one being the master zone and iRODS Metadata Catalog (iCAT), and others being slaves. Most sites have hierarchical storage, such as tape device, in the back-end. To date there are more than 18 TB of data stored and 280 users registered in the Data Fabric. Users and storage have nearly tripled in the last year.



3. Use Cases

While the ARCS Data Fabric is being used every day by individuals to store and share data, it is also integrated with eResearch service providers external to ARCS. For example, 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 that provides 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.

Another function has also been developed and will be deployed soon to enable Data Fabric users to create Persistent IDs for Data Fabric objects in the persistent Identifier Service (PIDS) of the Australian National Data Service (ANDS). This is achieved by invoking predefined rules from the Davis web interface.