The Data Center InterOperability initiative - DCIO

Christian Retscher[1,2], Yasjka Meijer[3,4], Martine De Mazière[5], Ian Boyd[6], Geoff Busswell[7], Rolf v. Kuhlmann[3], Sander Niemeijer[8], Aasmund F. Vik[9], Jeanette Wild[10] , and the DCIO partners

1 NASA Goddard Space Flight Center, Greenbelt, USA
2 University of Maryland Baltimore County, Baltimore, USA
3 ESA/ESRIN, Frascati, Italy
4 Rhea System S.A., Louvain-La-Neuve, Belgium
5 Belgian Institute for Space Aeronomy, Brussels, Belgium
6 NIWA - Environmental Research Institute, Amherst, USA
7 Logica UK Limited, Leatherhead, UK
8 S[&]T, Delft, The Netherlands
9 Norwegian Institute for Air Research, Kjeller, Norway
10 Wyle IS/NOAA, Camp Springs, USA

Introduction

Today many data providers, who operate instruments from ground, ships, aircraft or balloons, report the same data to multiple data centers. Often this has historic reasons: data centers actively pursue data submission from many Principle Investigators (PI), data are collected in the context of different projects which all have their own data centers, or data providers just want to increase the visibility of their data. This situation has many disadvantages. Often data are accepted with metadata and data formats specific to one data center or even one specific project. Moreover, there is a risk that different versions of the same data exist at different data centers. One also finds that each data center maintains its own data submission, quality control, data search, discovery and access logic as well as data exchange agreements. Accordingly, data providers need to cope with this variety and maintain data production routines capable of handling related formats. In the majority of cases this is an inconvenient and time consuming process. Data users on the other hand are confronted with similar issues of handling different data search mechanisms, and of designing and maintaining a variety of data access and read functionalities. For instance in a validation study it is often necessary to access data from all kinds of data centers. Of course, this may include a number of different datasets and formats, so the user needs to figure out a way to effectively interpret and merge the retrieved data. The data users are additionally confronted with the issue of retrieving the same measurement more than once and the need to de-duplicate data on their end in order to run a systematic and successful validation process. There is an explicit need for space agencies to have access to many ground-truth data sets for the validation of satellite instruments. Until internet search engines start indexing Earth Observation data systematically, the discovery and search of new data resources or the update of old data across the internet is a cumbersome undertaking.

In this article we present the Data Center InterOperability (DCIO) [1] initiative, which is dedicated to provide solutions to these issues and build a network of collaborating data centers for geophysical data validation.

Data Center InterOperability - DCIO

The Data Center InterOperability project is an initiative started by the European Space Agency (ESA) in December 2008. The DCIO was motivated by the Generic Environment for Calibration/Validation Analysis (GECA) project which supports existing and future ESA calibration and validation activities, and by the need to access correlative datasets from multiple Earth Observation (EO) domains and from multiple data centers.

The main objective of DCIO is to bring EO data centers closer together and foster future collaboration.

The specific objectives of DCIO are:

- Motivate collaboration of peer data centers
- Harmonize metadata standards for EO
- Provide data discovery and search capabilities for distributed resources
- Increase exposure of hosted data
- Develop joint data exchange agreements

- Enable uniform user authentication
- Allow systematic exchange of cal/val data
- Provide feedback on data use

DCIO has been developed by data centers in close contact with data providers. The group of data centers participating in DCIO is enabled to host data in a collaborative manner and to share metadata and data using agreed and harmonized interface standards. The DCIO will ensure the visibility of many data resources and enable access to databases across several EO domains. Principle investigators from networks like the NDACC can benefit from DCIO as their data will be available to a greater number of data users. This may enable a better collaboration between scientists, can result in an increased quality of data sources and will eventually lead to more publications.

As of October 2010 the DCIO consists of the following partners:

- AVDC - Aura Validation Data Center
- Ceilometer Network
- EARLINET - European Aerosol Research Lidar Network
- EVDC - Envisat Validation Data Center
- GECA - Generic Environment for Calibration and Validation Analysis
- GEOmon - Global Earth Observation and Monitoring of the Atmosphere
- GEOSS - Global Earth Observation System of Systems
- NDACC - Network for the Detection of Atmospheric Composition Change
- WEGC - Wegener Center for Climate and Global Change - Radio Occultation data base
- WIS - WMO Information System
- WOUDC - World Ozone and Ultraviolet Radiation Data Centre

Discussion has been initiated with new partners:

- AERONET - AErosol RObotic NETwork
- GlobWave
- MyOcean

The Generic Earth Observation Metadata Standard - GEOMS

Each data center maintains a database of any kind, where data can be discovered and found through information stored in a catalog. In order to enable a collaborative access to data center resources, DCIO agreed on catalog metadata and harmonized file metadata. DCIO partners developed catalog metadata based on Dublin Core® [2], an international and widely accepted metadata standard which has been broadened by GEOMS elements in order to fit all DCIO needs.

To enhance the usability of the diverse correlative datasets collected for satellite validation activities, metadata definitions, covering a broad range of instrument types and geophysical parameters have been established. The Generic Earth Observation Metadata Standard (GEOMS) guidelines and templates [3] describe the standard metadata definitions adopted for the correlative, experimental and model data archived for the Aura validation program, the Envisat calibration and validation campaign, and the GECA project, as well as for NDACC data. The definitions have been carefully chosen to allow applicability to other scientific endeavors. This development was initiated in 1998 through the European Commission (EC) project COSE, Compilation of atmospheric Observations in support of Satellite measurements over Europe [4], and extended in collaboration with ESA, NASA, PIs of the Envisat and Aura validation campaigns, and selected PIs from NDACC, for the implementation of a uniform data exchange standard. GEOMS compliant data can be implemented on any kind of hierarchical data format, though GEOMS compliant data centers currently only support the HDF4 and HDF5 data formats. There are discussions on a future support of netCDF.

Data centers participating in DCIO provide a variety of GEOMS compliant online and offline tools for data download, access, conversion, comparison, and collocation. Data conversion routines from various kinds of data formats and metadata (e.g., NASA-Ames, netCDF, WOUDC) into GEOMS are available. Some of these tools are published as open source code and/or compiled versions. Any data provider or data user can adapt these source codes and is encouraged to feed them back to the community.

Provide data discovery and search from distributed sources

DCIO chose the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [5] for interchanging catalog metadata. Some peer data centers are fully OAI-PMH compliant, and some have chosen to provide metadata through a custom-built mechanism. OAI-PMH provides structured access to the catalog metadata of peer data centers. OAI-PMH runs as an application on the data centers' web-servers. Metadata can be retrieved by any enabled data center or data user. The system also allows automated retrieval of metadata through regular harvesting. DCIO only regulates the metadata server side. The OAI-PMH service responds with an XML file, which needs to be further processed by the client data center. The client side needs to be designed by each peer data center itself. For example, if AVDC retrieves the complete EVDC catalog, then AVDC needs to process this information and feed its own database tables. AVDC data search and discovery functionality have to be enhanced by EVDC catalog information in order to allow data users to find EVDC data. Actual access to data is then regulated by data exchange protocols. Metadata on data variables is based on GEOMS, so, for example, if somebody looks for ozone column data, the search term is 'O3.COLUMN'. Nonetheless, not all datasets shared through DCIO carry file metadata that is compliant to GEOMS: the DCIO motivates all partners to allow for a mapping of their native metadata into GEOMS.

When DCIO centers support commonly shared metadata, users can find data resources across the centers. All DCIO centers can also participate in broader shared-catalog 'libraries' such as the WMO Information System (WIS), the Global Change Master Directory (GCMD), and the Global Earth Observation System of Systems (GEOSS), among others.

Develop joint data exchange agreements

Every DCIO peer data center has some sort of data exchange agreement regulating data submission, access and subsequent data use. A central point is that the ownership of data archived at data centers remains with the data's principal investigator or data originator.

The DCIO is investigating the harmonization of the various used data exchange agreements amongst the partner data centers. As of October 2010 DCIO still operates with data protocols controlled by the individual data center, though the group is working towards a joint data access protocol. This is a tedious and delicate process, but is crucial for the success and reliability of the DCIO service.

Allow systematic exchange of cal/val data and enable uniform user authentication

The DCIO enables contributing parties, data providers and data users to have access to a large number of datasets with a harmonized access. Datasets, which are found through DCIO, mostly reside on the originating data center with only limited tasks to mirror data onto peer data centers. **Data users can select their data center of choice and use it's functionality to find and access data from collaborating data centers.**

A significant problem in the collaboration of the current peer data centers is that many original measurements are hosted on more than one data center and sometimes have different file names or data formats. Data users have to deal with the issue of de-duplication of datasets. If datasets have the same file name, the identification of duplicates is trivial, but if the data file container, format or metadata is different the search for unique datasets may be an exhausting process. The DCIO wants to solve this and is currently working on the introduction of unique Digital Object Identifier (DOI) for hosted files. This will ensure that one geophysical measurement will have one and the same DOI independent of used data format or hosting location.

Once a data source has been discovered and a specific dataset has been found through a data center's web interface, the corresponding file can be accessed by clicking on it. Currently the data user still needs user credentials at the data center that hosts the file. DCIO plans to setup a single sign-on (SSO) portal based on OpenID, which will allow data centers to handle user authentication and data access in a straightforward manner. The user will have individual user credentials with every data center of interest, but will also have an OpenID, which allows systematic access to all enabled data centers, by only providing a general username and password combination.

How to join the DCIO group and become a peer data center

DCIO development and maintenance tasks are shared between ESA and NASA. DCIO meets in regular telephone conference calls (1 – 2 months) organized by ESA and infrequent meetings every 1 – 2 years. ESA and NASA with AVDC are currently responsible for the update of DCIO catalog metadata and further provide support for new data centers to become compliant. If you are a data center manager, who wants to pursue collaboration with DCIO please contact either the DCIO responsible at ESA, Rolf von Kuhlmann (rolf.von.kuhlmann@esa.int) and Yasjka Meijer (yasjka.meijer@esa.int) or at NASA, Christian Retscher (christian.retscher@nasa.gov).

References

[1] Data Center InterOperability (DCIO) at AVDC, http://avdc.gsfc.nasa.gov/DCIO/.

[2] Dublin Core® Metadata Initiative (DCMI), http://dublincore.org/.

[3] Generic Earth Observation Metadata Standard (GEOMS) at AVDC, http://avdc.gsfc.nasa.gov/GEOMS/.

[4] De Mazière, M., “Final Report of the EC-COSE Project (contract ENV4-CT98-0750)”, BIRA-IASB, Brussels, Belgium, (2001).

[5] Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/pmh/.