Interoperability of Web Service-Based Data Access and Processing: Experience Using the DataFed System

Rudolf B. Husar, Stefan R. Falke and Kari Hoijarvi
Washington University, CAPITA
1 Brookings Drive, St. Louis, MO 63130

***Abstract*-This is a progress report on the federated data system, DataFed, partially supported by a NASA REASoN project. In DataFed, the distributed data sources are federated by applying a universal, multi-dimensional data model. The physical and semantic homogenization is accomplished by wrapper services. Data processing is performed through web services, which themselves can be distributed. Data processing applications are created by the composition of the distributed service components. International Standards and Protocols are used to establish interoperability of the service components. In this work we have adapted the OGC web services as a standard protocol for Air Quality data access. This report summarizes our experiences with the OGC W*S based service composition. It includes the results of our participation in the GALEON (Geo-interface to Atmosphere, Land, Earth, Ocean netCDF) Interoperability Experiment. We have created a WCS test server to deliver a wide variety of point, grid an image coverage data. Our goal was to evaluate the WCS protocol for accessing coverages of different types arising from a variety Earth observation and modeling systems. We are most encouraged by the possibilities but recognize that considerable work is to be done on the extending the describeCoverage schema to accommodate new coverage types and the returned data types. The GEOSS Services Network (GSN) is a persistent network of a publicly accessible OpenGIS-accessible services for demonstration and research regarding interoperability arrangements in GEOSS. The May 06 Beijing GSN demo included a collaboration chain between NOAA-NCDC, Unidata and DataFed using OGC standard interfaces.**

I. Introduction

Air quality data analysis requires considerable processing of raw observational/model data before these can be used for decision-making processes. The main processing operations are filtering, aggregation and fusion of multi-sensory data and/or models. The nature of the analysis, i.e. the choice of data as well as the sequence of these operations is highly dependent on the users needs. Hence, in an ideal data system, much of the analysis is to be performed by user-defined processing chains, using a flexible and agile information system.

The analysis of air quality data through agile information system requires (1) seamless data access (2) reusable data processing component to access, filter, aggregate, and fuse distributed data; (3) a service-oriented framework that facilitates the publish-find-bind (chain) web service model for application building. Such loose coupling can only be achieved if the service components adhere to strict standards-based protocols. Achieving such loosely coupled service oriented architectures has been the stated goal of numerous national and international Earth Science programs including GEO.

The most important step toward service oriented Earth Science information systems is the adaptation of strongly typed standards for *finding, describing, and accessing data*. Given such standards-based interface, providers of data and services can *publish* their data resources and users can *find* suitable data in formal catalogs. Most importantly formal protocols allow snap-like *binding* i.e. data access, between the server and the client operations. The publish-find-bind trilogy constitutes the basic operations that permit the building of agile, loosely coupled systems through web service chaining.

Current Earth Science data systems do not yet allow such flexible, user-defined data processing. However, considerable advances are being made in this direction. An attractive development in this regard is the emergence and the broad acceptance of geospatial standards coordinated internationally by Open Geospatial Consortium (OGC). In the Earth Sciences similar development led to a standard for binary-encoded Earth Science data, through the netCDF encoding, augmented by the CF Conventions. The combination of these standards to develop seamless Earth Science data flow is being pursued by several interoperability experiments (e.g. GALEON and GEOSS Services Network).

II. Abstract Spatio-Temporal Air Quality Data Model

The first requirement of interoperability is a common data model. The general data model for air quality data is that of a multi-dimensional data cube, with dimensions (X,Y,Z,T) in physical coordinates. Such a data model can be represented through **Views**, which are slices through the data cube organized, by latitude, longitude, elevation and time. The sub-cubes can be one dimensional, e.g. a time series at a specific location or 3-4 dimensional, depending on the view. These slices are shown below in Fig. 1.

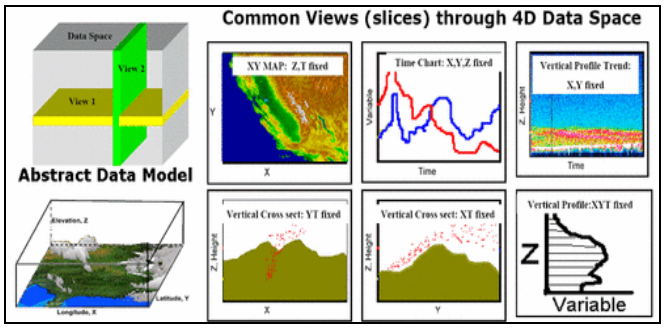

Fig.1. Multi-dimensional Data Model and Data Views

This is an abstract data model in a sense that the system response to queries that are addressed to this data model. Implicit in the use of abstract data models is that all the data are accessed through a well-defined interface rather than as physical files. In other words, the goal is to turn data into a service. The physical data storage and management is an implementation issue that is of no concern to the data user.

III. A Relational Data Model: Mapping to Standard Data Access Protocols

Most air quality monitoring data are stored and managed through Relational Data Management Systems (RDMS) using the SQL language for accessing and manipulation. Multi-dimensional data can be encoded in the relational model using a star schema in which the fact (data) table is surrounded by a set of dimensional tables as shown in Fig.2.

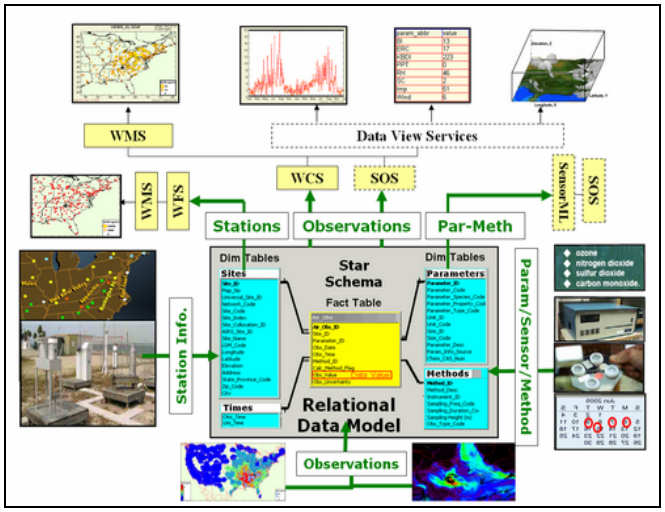

Fig.2. Typical Air Quality Relational Schema as a Data Source for OGC Web Services

The main air quality monitoring data are best accessed through the Web Coverage Service (WCS). The spatial pattern of sampling locations can be conveniently represented by the Web Feature Services (WFS). The description of sampling methods, sensors, and parameters can also be accessed through a variety of emerging OGC protocols.

IV. Data Access Protocols and Adapters

The rich structure and semantics of Earth Science data means that any given dataset can be accessed through multiple protocols. In general, each client and server is capable of communicating through a subset of protocols.
Thus, loose coupling between data access and processing services involves choices and negotiations. The main topics of client-server negotiation are the selection of a shared data access protocol and a choice of returned data format.

An example of a flexible data access interface is shown in Fig.3. It represents the data access module in the federated data system, DataFed.

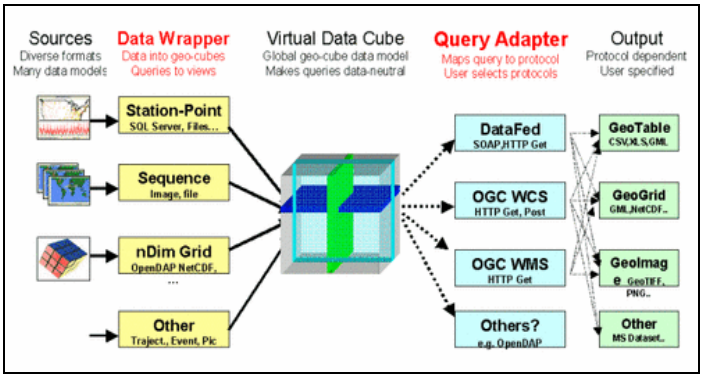

Fig.3. Data Access Protocols and Adapters. The electric adapter is a good analogue of the DataFed software adapters.

Individual, heterogeneous, distributed datasets are connected to the data system through *wrappers*, which homogenize the access. On the client side, the interface to the data system is through adapters that provide access through multiple protocols and data formats. The specific data access protocols offered through DataFed are shown in Table 1 for nine representative datasets.

TABLE I
EXAMPLE DATA TYPES AND ACCESS PROTOCOLS

Catalog		**Display/Discuss**		**Data Description**		**Spatial Data Access**					
Dataset	Registration	Viewer	Discuss	Sensor Type	Data Type	Data Access	WCS	WFS	WMS	Url	SOAP
AIRNOW	XML - Form	View	Wiki	In Situ	Point	Protocols	X	X	X	X	X
SURF_MET	XML - Form	View	Wiki	In Situ	Point	Protocols	X	X	X	X	X
VIEWS_OL	XML - Form	View	Wiki	In Situ	Point	Protocols	X	X	X	X	X
THREDDS_CDM	XML - Form	View	Wiki	Model	Grid	Protocols	X		X	X	X
THREDDS_GFS	XML - Form	View	Wiki	Model	Grid	Protocols	X		X	X	X
NCDC_AVG_WIND	XML - Form	View	Wiki	Model	Grid	Protocols	X		X	X	X
CIESIN	XML - Form	View	Wiki	Model	SeqImage	Protocols	X		X	X	X
OnEarth_JPL	XML - Form	View	Wiki	RemSens	SeqImage	Protocols	X		X	X	X
SEAWiFS_US	XML - Form	View	Wiki	RemSens	SeqImage	Protocols	X		X	X	X

Dataset. This column shows the names of the datasets selected for this demonstration. Each dataset has a unique name which can be used to access any dataset for browsing etc. The full list of the available data is in the DataFed catalog.

Registration. A dataset registration (rendered as XML and form) contains all the relevant information that is needed to find, and to access (bind) a dataset for purposes of web service chaining.

Viewer. Each dataset in DataFed can be accessed through a single generic viewer, which allows browsing through the multi-dimensional dataset and editing the service flow for each data layer. The viewer also provides access to the settings of each web service through the service flow diagram.

Discuss. Each dataset has a wiki page that can be modified by any user. Initially, the page only contains a brief description of the dataset along with a link to the viewer. The "talk" pages (accessible through discussion tab) are suitable for threaded discussion.

Sensor Type. This field identifies the nature of the "sensor" that is used to produce the dataset. In these examples in situ refers to point monitoring of air quality through surface sensors, model represents output from numerical simulation models and RemSens is typically from satellite sensors.

Data Types Used in DataFed. The output from sensors is structured into different data types. The three main data types in DataFed are point, grid, and seqimage, shown pictorially in Fig.4.

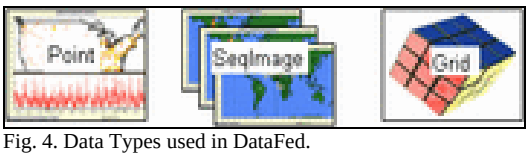

Fig. 4. Data Types used in DataFed.

Point data arise from monitoring sites at fixed geographic points. Typically these data have time series of multiple parameters at each station. *Grid* data arise typically from model simulations that have regular spacing and one of the standard coordinate systems (projections). Model grids are typically multi-dimensional, covering X,Y,Z, and T as well as parameter dimensions. *SeqImage* is a data type for time-sequenced georeferenced images such as satellite and radar images that are produced in fixed time intervals (hourly, daily). Sequential images are typically spatial, but they also vary in time. There are numerous other data types used in air quality that are not shown here including trajectory, multi-spectral satellite image etc.

Data Access. The data access protocols that are available for any given dataset are listed in a special form illustrated in Fig.5.

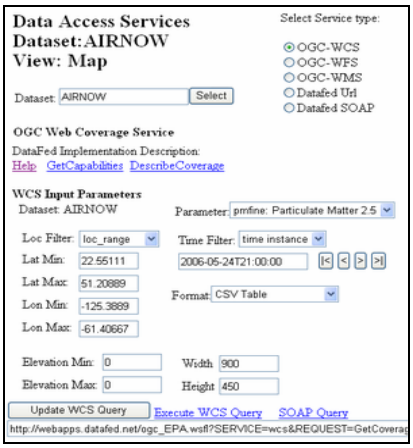

Fig.5. Typical Form Facilitating Access to Multiple Data Access Services

WCS, WFS and WMS are OGC protocols for Coverage, Features and Maps respectively. Each OGC service has an associated getCapabilities document, which lists the offerings. The data are also accessible through a DataFed-specific cgi interface using key-value-pairs, similar to the W*S OGC REST (**URL**) interfaces. Finally, the **SOAP** interface is offered to access data, through formal SOAP-based web services. The strong typing of this interface is assured by the WSDL for each service, which in turn is defined by a formal XML schema. The output formats for each data type and access protocol are listed in a separate table.

V. GALEON INTEROPERABILITY EXPERIMENT

GALEON (Geo-interface for Atmosphere, Land, Earth, and Ocean netCDF) is an OGC Interoperability Experiment supports open access to atmospheric and oceanographic modeling and simulation outputs. This is an active and productive group working on the nuts and bolts of ES data/model interoperability.

In phase I of GALEON, it was demonstrated that the DataFed WCS client can successfully access the WCS data sets from the U. Florence and from the UNIDATA THREDDS WCS test servers. The spatial queries to both servers yield the expected return as spatial grids. Within DataFed, the received WCS (NetCDF) data from both servers are easily transformed into data views and used in distributed web applications by chaining appropriate web services.

We have also created a WCS test server to deliver a wide variety of point, grid an image coverage data. Our goal was to evaluate the WCS protocol for accessing coverages of different types arising from a variety Earth observation and modeling systems.

We have demonstrated that for the air quality applications WCS is a well-suited protocol for point/station (Fig.6.), image and gridded data (Fig.7.). Fig.8. shows the WCS queries for Map, Time, and Elevation views for a 4-dimensional dataset. The strength of WCS is in the simplicity and universality of the BBOX, TIME data query. This vital query feature is common to WMS and WFS queries, which makes the OGC protocol compatible with the abstract multi-dimensional data model (and vice versa). It is clear, however, that there is considerable work to be done on extending the describeCoverage schema to accommodate these coverage types. Also, the data types returned need to be extended (see Fig.9.) to accommodate these additional coverage types. The discussion on the possible WCS extensions is given in a separate report to the GALEON IE group.

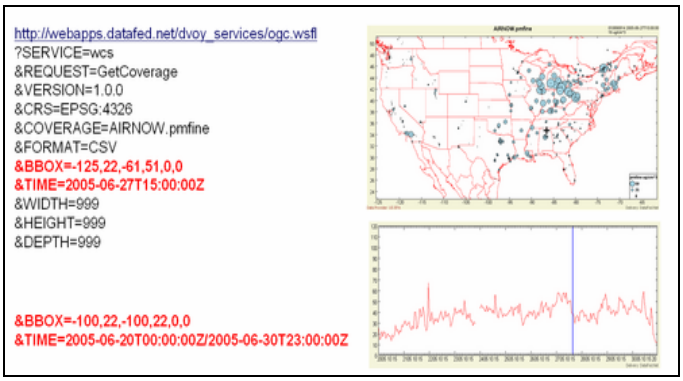

Fig.6. WCS Query for Point Data Type

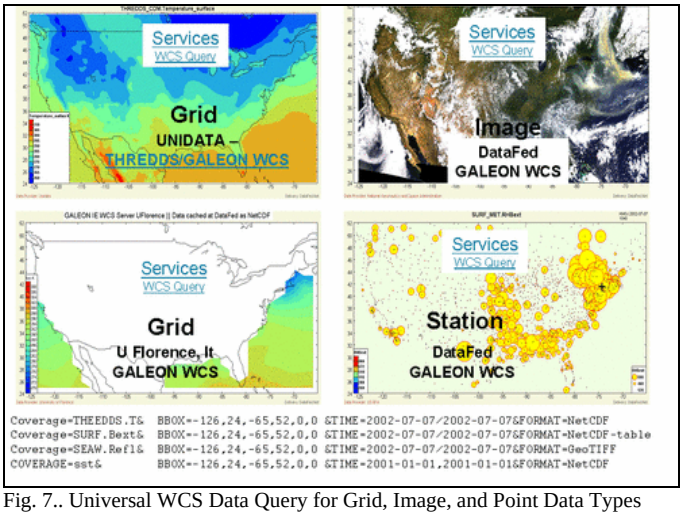

Fig. 7.. Universal WCS Data Query for Grid, Image, and Point Data Types

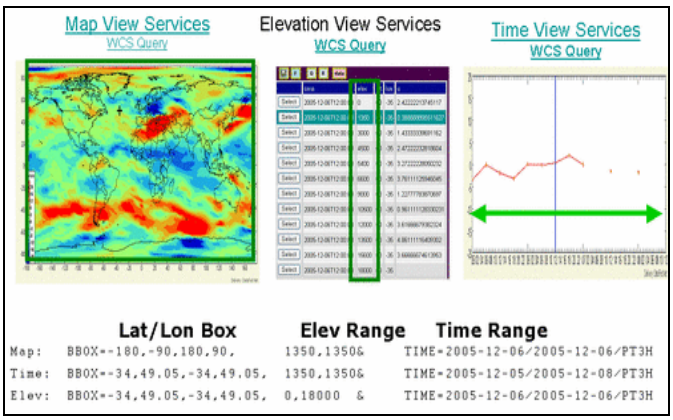

Fig.8. WCS Queries for Map, Time, and Elevation Views

Name	Definition	Data type	Multiplicity and use
BoundingBox	Unordered list of bounding boxes whose union covers spatial domain of this coverage offering [a]	ows:BoundingBox	One or more (mandatory)
Grid	Unordered (TBR) list of grids that describe internal grid structure of this coverage offering [b]	gml:Grid OR gml:RectifiedGrid	Zero or more (optional) Include one for each different grid used (TBR)
Polygon	Unordered list of polygons whose union covers spatial domain of this coverage offering [c]	gml:Polygon	Zero or more (optional) Include one for each polygon needed
Point	**Unordered list of points whose union covers the spatial domain of this coverage offering[d]**	**gml:Point OR gml:Position OR om: featureOfInterest**	**Zero or more (optional) Include one for each point**

a One bounding box could simply duplicate the information in the ows:WGS84BoundingBox, but the intent is to describe the spatial domain in more detail (e.g., in several different CRSs, or several rectangular areas instead of one overall bounding box).

b This element can help clients assess the fitness of the gridded data for their use (e.g. its native resolution, inferred from the offsetVector of a gml:RectifiedGrid), and to formulate grid coverage requests expressed in the internal grid coordinate reference system.

c Polygons are particularly useful for areas that are poorly approximated by a ows:BoundingBox (such as satellite image swaths, island groups, other non-convex areas).

d Points are useful for representing samples or observations from monitoring networks, such as soil samples or temperature measurements from weather stations, where the phenomena being measured is continuous over the spatial range but is measured only at a limited number of locations

Fig. 9. Suggested Changes to WCS 1.1. Table 17 (SpatialDomain)

Substantial hurdles are still ahead before we reach a 'snap'-like WCS client-server interoperability including: the temporal aspects of the WCS queries needs more work both on server and client side. Automatic (loosely coupled) registration of WCS services needs work. Geo-re-projection services (preferably web services) are needed to handle the homogenization of the mired projection used by the different data providers. We recognize that many of the current inadequacies are being addressed in other interoperability efforts and possibly by the continuation of this IE. We are looking forward enhancing our interaction with the group as well as individuals in this group while striving toward better interoperability of our data systems.

The second part of our participation involved designing and implementing WCS interfaces to our multitude of air quality data in form of point data (fixed monitoring networks), images and 4Dim grids. This effort was truly exploratory. Our goal was to evaluate WCS as 'universal' access protocol for Earth Science data. In other words, for accessing 'coverages' of different types arising from a variety Earth observation and modeling systems.

VI. GEOSS SERVICES NETWORK DEMONSTRATION – BEIJING

GSN is a persistent network of publicly accessible OpenGIS-accessible services for demonstration and research regarding interoperability arrangements in GEOSS. GSN is the basis for demonstrations in the GEOSS Workshop series, “The User and GEOSS Architecture”. The CAPITA group has participated in the workshop, Implementing the GEOSS Architecture using open standards, 22-23 May 2006, Beijing, China. The theme of the workshop was to show interoperability as applied to wind energy siting. The DataFed system was used to access aggregated, global, monthly wind vector fields provided by the NOAA National Climatic Data Center (NCDC). The NCDC server was upgraded by Unidata technologies to be a WCS server. The CAPITA group has provided the portrayal service that transformed the netCDF data files from WCS to wind vector maps and wind speed contour maps (Fig.10.). The resulting maps were then served as simple WMS maps. Illustrative examples of the vector and contour maps delivered through WMS are given in the (Fig.11.).

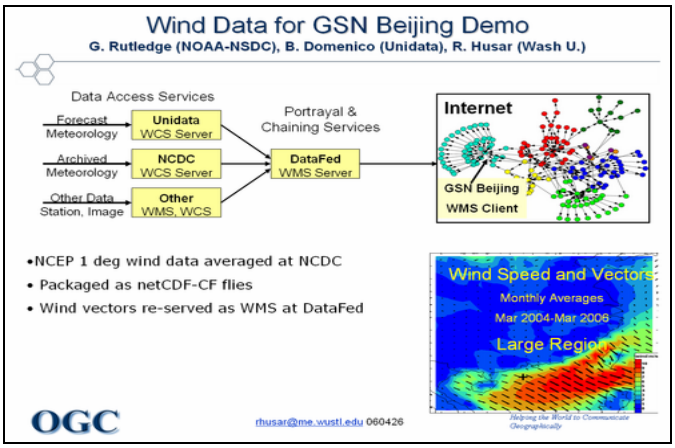

Fig. 10. Linking of NCDC, Unidata through DataFed Portrayal and Chaining Services

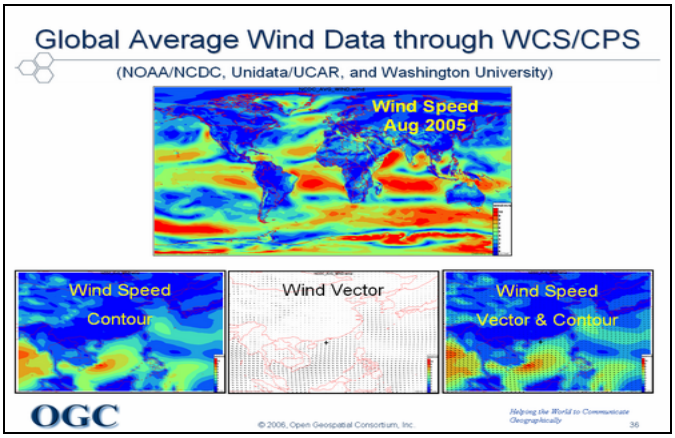

Fig. 11. Illustration of Web Service Chaining Product: Global Wind Vector Map

The Beijing interoperability experiment has demonstrated that data access, processing and delivery services can now be combined to create distributed web applications. However, GSN-Beijing has also shown that the execution of the linking is still tedious and requires considerable human interaction and iteration. It is hoped that the next GSN demo at Denver IGARRS 06 can benefit from the Beijing learning experience.

ACKNOWLEDGMENT

This research was supported by a NASA REASoN grant to Wasington University "Application of NASA ESE Data and Tools to Particulate Air Quality Management" and by grants from EPA and NESCAUM. The skillful support of Erin Robinson during the report preparation is gratefully acknowledged.