



Crossing the digital divide: an interoperable solution for sharing time series and coverages in Earth sciences

F. R. Salas¹, E. Boldrini², D. R. Maidment¹, S. Nativi², and B. Domenico³

¹Center for Research in Water Resources, University of Texas at Austin, Texas, USA

²Earth and Space Science Informatics Laboratory of the Institute of Atmospheric Pollution Research, Italian National Research Council, Rome, Italy

Correspondence to: F. R. Salas (fernando.r.salas@utexas.edu)

Received: 1 April 2012 – Revised: 17 July 2012 – Accepted: 28 July 2012 – Published: 2 October 2012

Abstract. In a world driven by the Internet and the readily accessible information it provides, there exists a high demand to easily discover and collect vast amounts of data available over several scientific domains and numerous data types. To add to the complexity, data is not only available through a plethora of data sources within disparate systems but also represents differing scales of space and time. One clear divide that exists in the world of information science and technology is the disjoint relationship between hydrologic and atmospheric science information. These worlds have long been split between observed time series at discrete geographical features in hydrologic science and modeled or remotely sensed coverages or grids over continuous space and time domains in atmospheric science. As more information becomes widely available through the Web, data are being served and published as Web services using standardized implementations and encodings. This paper illustrates a framework that utilizes Sensor Observation Services, Web Feature Services, Web Coverage Services, Catalog Services for the Web and GI-cat Services to index and discover data offered through different classes of information. This services infrastructure supports multiple servers of time series and gridded information, which can be searched through multiple portals, using a common set of time, space and concept query filters.

etc.) that tries to explain and predict the complex processes that make up the Earth system as a whole. These studies require researchers to collect and synthesize large quantities of data stored in various formats by numerous organizations, agencies and research groups that represent a diverse array of scientific communities. For centuries, scientists in all areas of research have viewed data as a means for solving very specialized problems and have locked their results away. However, the emergence of the Internet supports transparency in science through the sharing of information of multiple types (e.g. sensors, observations, models, etc.). Users should have the ability to seamlessly search and discover data published by researchers across multiple disciplines and then readily access that information through a standard process.

Much of the disparity that exists between the sharing of hydrologic and atmospheric science data arises not only from the methods by which data are stored but also the methods by which data are discovered and utilized. The spatiotemporal domains of the two disciplines are very different from one another, thus making the perspective from which data are visualized also different. Hydrologic studies typically are based on data collected from spatially discrete observational gauges measuring a small set of variables in a limited area over an extended period of time. Conversely, atmospheric studies are typically based on spatially continuous data measuring or modeling large sets of variables over large spatial regions at discrete instances in time. Because of the inherent nature of each information domain, data in each field is stored differently, making it a challenge to provide a common data model for discovering, accessing and visualizing information (Nativi et al., 2004). As illustrated in Fig. 1, hydrologic data is stored in Geographic Information System(s) (GIS), tables

1 Introduction

Data interoperability in the study of Earth science is essential to performing interdisciplinary multi-scale multi-dimensional analysis (e.g. hydrologic impacts of global warming, regional urbanization, global population growth

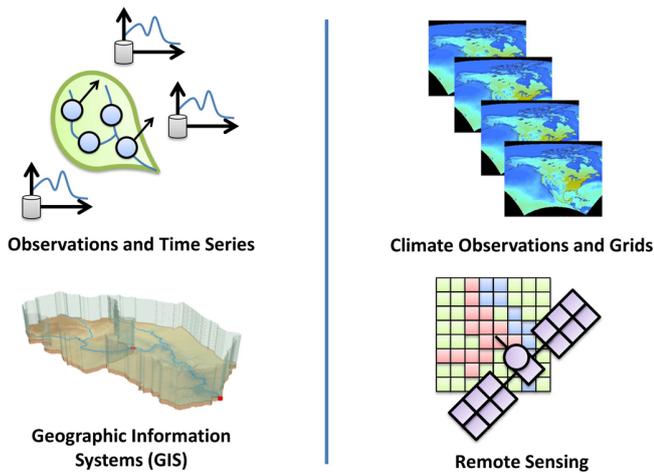


Fig. 1. The digital divide between hydrologic and atmospheric science data.

and relational databases, while atmospheric data is stored in large binary files in specialized formats as multi-dimensional arrays (e.g. netCDF, GRIB, etc.)

Three of the most prominent leaders in the sharing of hydrologic and atmospheric science data are the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI), the Unidata Program Center, and the Earth and Space Science Informatics Laboratory (ESSI-Lab) of the Italian National Research Council. Each group has been developing cyberinfrastructure to share data efficiently via the Web using standard Web services. CUAHSI has primarily focused on the distribution of time series acquisitions from in-situ gages, while Unidata and ESSI-Lab have focused on the distribution of atmospheric science grids and coverages. As of 2011, the CUAHSI Hydrologic Information System (CUAHSI-HIS) has compiled the largest catalog of hydrologic time series in the world (CUAHSI, 2012). These catalog records are stored on a collection of CUAHSI-HIS servers and databases and are indexed at the San Diego Super Computing Center (SDSC) in a catalog called HIS Central (Maidment, 2009). By indexing records available by the United States Geological Survey (USGS), National Oceanic and Atmospheric Administration (NOAA), Environmental Protection Agency (EPA) and other agencies and academic institutions, CUAHSI-HIS has cataloged approximately 23 million time series accounting for more than five billion data values (Tarboton et al., 2010). Similarly, Unidata has systematically indexed various models and remote sensing and satellite coverages from the National Centers for Environmental Prediction (NCEP), National Aeronautics and Space Administration (NASA), etc. on Thematic Realtime Environmental Distributed Data Services (THREDDS) servers (Nativi et al., 2006; Unidata, 2012). While ESSI-Lab does not provide tools to physically store gridded data, ESSI-Lab instead provides a broker

catalog service, GI-cat, which seamlessly integrates data stored across disparate catalog systems (ESSI-Lab, 2012). Although CUAHSI, Unidata and ESSI-Lab have developed successful information systems that connect data providers and users, each system was initially designed to respectively manage and share either time series or grids, not both together.

When sharing information across scientific communities, it becomes important to define a standard framework through which large quantities of multidisciplinary information can be shared, discovered and accessed. In 2010, CUAHSI-HIS demonstrated that hydrologic time series can systematically be shared and discovered across the Web using standardized Open Geospatial Consortium (OGC) Web services as opposed to its own customized WaterOneFlow Web services (Bermudez and Arctur, 2011; Seppi, 2010). Likewise, Unidata and ESSI-Lab have demonstrated that OGC Web services can be used to share atmospheric grids and coverages. Since 2000, the OGC has been fostering collaboration amongst researchers in Earth sciences by building a standard operational platform using Web services through which data users can readily access and ingest large quantities of geospatial metadata and data (OGC, 2012). This is just one of the many examples where the world is building a Web services framework for computers to communicate in an ad hoc manner (Vector, 2012).

Utilizing a collection of OGC Web services, CUAHSI designed a “services stack framework” that shares catalog data, metadata and data with the user (Seppi, 2010). The services stack framework identifies three types of services as essential to sharing water information across the Web: catalog services, metadata services and data services. These three services work together to completely index, describe and provide access to water information (e.g. time series). Catalog services provide users with an index of hydrologic metadata, metadata services identify collections of time series available over a domain of space and time, and data services provide the user with the raw data for a specified temporal period and spatial area. While this framework was originally designed to publish and distribute time series, it can be shown by implementing the OGC services infrastructure that this framework can be extended to include grids and coverages as well. While other interoperability studies have focused on implementing custom data streams (e.g. bridges, adaptors, etc.) between clients and server interfaces, this study will focus on a common data and metadata management model that leverages a suite of OGC standard Web services which can be applied to multiple scientific communities – in particular hydrologic and atmospheric sciences. Furthermore, it will be demonstrated that the aforementioned data management model can be integrated within existing data discovery frameworks (e.g. portals, gateways, etc.) by leveraging mediation and brokering services.

Although this framework is the basis of this study, the role of semantic mediation cannot be overlooked. In conjunction

with spatial and temporal filters, semantic filters aid the user in the data discovery process by systematically retrieving the data that matches or is related to a concept defined by the user (i.e. search term). Within the CUAHSI-HIS, concepts and relationships between concepts, are methodically organized using the CUAHSI ontology which have been developed and optimized for hydrologic time series data (Whitenack, 2010). In contrast, the atmospheric science community has most commonly relied on the Climate and Forecast (CF) Metadata Conventions to describe gridded data stored in netCDF files (Unidata, 2012). It is recognized that many semantic ontologies exist and are not limited to the two presented above (Bermudez and Piasecki, 2006); however, in this paper we focus on common search terms within the CUAHSI ontology and CF conventions.

This research demonstrates a collaborative effort between CUAHSI, Unidata and ESSI-Lab to provide a sound interoperable framework that systematically allows users to discover and access both hydrologic and atmospheric science data through a common interface that leverages standard Web services; clients that build upon this framework will be the focus of future work. In order to justify this claim, two interoperability experiments were conducted and implemented using a variety of tools, software and services. The first experiment utilizes the GeoPortal interface designed by the Environmental Systems Research Institute (ESRI), while the second utilizes a broker catalog service called GI-cat, designed by ESSI-Lab.

2 Background

2.1 Interoperability through network-based software architectures

The stateless Client–Server (C–S) is a commonly implemented architectural style in network-based systems and applications. Within C–S, two types of components are present: clients and servers. Clients request services from servers via their interface while servers listen for requests based on the services they offer (Fielding, 2000). A Service-Oriented Architecture (SOA) is a set of principles and methodologies enabling software interoperability through a C–S architecture style. SOAs typically include a third component that allows clients to search through available services and their providers; this third component is implemented as a service registry (Oasis, 2006).

When building connection streams between C–S components, systems must conform to the interfaces provided by each. The connection streams are usually described in terms of message payload models, encodings, protocols, bindings, etc. Together, they define the mechanisms through which messages via the Web are exchanged and data carried in them decoded. Within the context of interoperability, connection streams between interfaces are a focal point of many

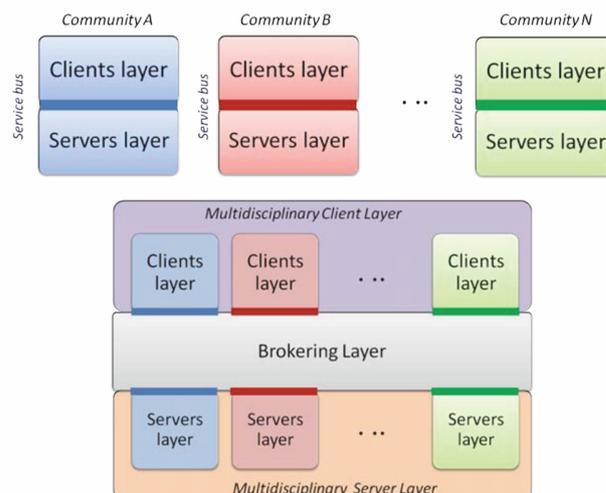


Fig. 2. Community Layered C–S architectures and the broker architecture for multi-disciplinary environments (Nativi et al., 2012).

standardization processes. The literature defines a collection of interface standards that characterize a network-based system as a “service bus”. A service bus can also be defined as the middleware glue between a client and service layer; a service bus enables communication within network-based systems (Ortiz, 2007; Schmidt et al., 2005).

In information science, communities, within their respective scientific disciplines, utilize single or several service buses to enable domain application and build disciplinary data exchange infrastructure. Both CUAHSI and Unidata have each established a service bus within their respective scientific communities, whereas ESSI-Lab has developed a mediation approach to interconnect system components across scientific disciplines, as can be seen abstractly in Fig. 2.

There are several important aspects that are considered when working towards an interoperable data solution. As communities define sets of standards for their respective service bus(es), the following ideas are typically considered in the drafting process (Nativi et al., 2012; Ramamurthy, 2006):

- data and metadata models
- encoding formats
- controlled vocabulary and ontologies
- service interfaces and binding protocols
- data policies

Because there are many unique disciplinary cyber-infrastructure in existence today, popular interoperability solutions have surfaced (e.g. ISO TC211, OGC standards for geospatial information, etc.). However, defining an individual service bus for a specific client and server within

or across scientific disciplines is often tedious and can result in a high entry cost; either clients must implement interface adaptors or bridges or servers must publish data through multiple service buses. To overcome some of the difficulties associated with imposing a single service bus for a particular data stream, mediation layers have been created to integrate across different data models (Nativi et al., 2009). Mediation was first used to map from existing and well-adopted specifications to the mandated federal specifications by implementing the mediators approach described by Wiederhold (Wiederhold, 1992). This strategy has been proven successful in federating existing and legacy capacities, while at the same time avoiding high level entry costs associated with implementing difficult and heterogeneous standards.

The introduction of mediation components establishes a Layered C–S (LCS) architecture in which each layer provides services to the layer above it and uses services from the layer below it (Garlan and Shaw, 1993). Some common LCS infrastructures include solutions that leverage proxy and gateway components. A gateway service publishes multiple interfaces, each taking requests from myriad clients and forwarding them (possibly after translation) to a single service component (realizing an M to one cardinality). A proxy service on the other hand appears as a single service to its clients, but is able to forward the incoming requests (with possible translation) to its “inner-layer” servers (realizing a one to N cardinality). Whereas proxies and gateways both limit their exposure to either a single client or server, a broker service reduces the interoperability burden on both the client and server. The middleware components within a brokering service mediate between multiple service providers and multiple service consumers (realizing an M to N cardinality). A broker can interconnect different service buses from different communities, mediating between their existing (and future) models and interface specifications. In addition, it works out all the necessary distribution and virtualization capabilities to lower the entry barriers for multidisciplinary applications, both for services and clients as seen in Fig. 2 (Nativi et al., 2011).

The following sections will expand upon both the CUAHSI and Unidata architectural frameworks so as to provide a basis for the interoperable solutions presented in this study. As leading data publishers within their respective scientific communities, these two systems provide a relevant use case where two disparate systems encounter interoperability obstacles. Several studies focus on adaptors, bridges and gateway technologies to overcome interoperability issues (Alameh et al., 2006; Giuliani et al., 2011; McDonald et al., 2006; Padmanabhan et al., 2011) but few explore common data management models across scientific disciplines (Rui et al., 2011). This paper will demonstrate (1) how a common data management framework can be built around existing infrastructures by leveraging standard web services and

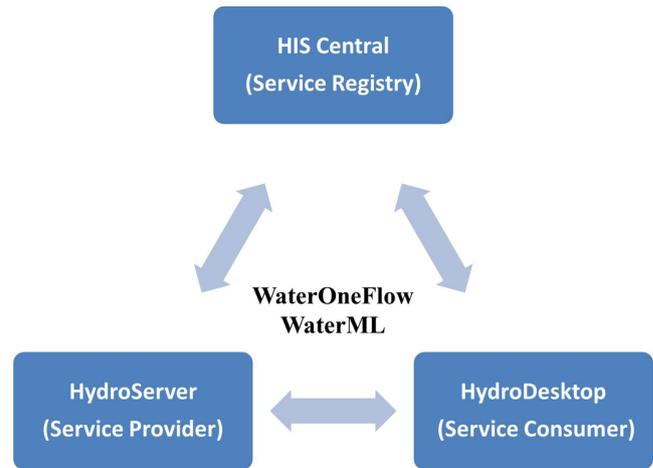


Fig. 3. Services-oriented architecture for CUAHSI-HIS.

(2) how this solution can be integrated with other scientific disciplines through a mediation approach.

2.2 The CUAHSI Hydrologic Information System

As part of the CUAHSI-HIS development, an SOA was identified as one of the key components to building a sustainable and reliable system that supports the sharing of hydrologic data (Tarboton et al., 2011). As with any other SOA, CUAHSI-HIS was built around two fundamental components: (1) service providers and (2) service consumers. Although service consumers directly connect to service providers to request and receive data, a third component, a service registry, is introduced to facilitate the discovery of different service providers (Tarboton et al., 2011); this can be done using various keywords, metadata and filters. As service providers introduce their services within CUAHSI-HIS, services are registered at the service registry. Service consumers can then search the registry to find available services of interest. Figure 3 outlines the SOA for CUAHSI-HIS.

CUAHSI-HIS can be defined as a collection of components which work together to store, index, access and distribute hydrologic information (Maidment, 2009). The system contains servers, catalogs and applications which communicate with one another through a set of WaterOneFlow web services. WaterOneFlow web services are the set of protocols and specified functions that exchange hydrologic metadata and data (i.e. time series) through the web using a common standardized language, Water Markup Language (WaterML) (Maidment, 2009; Zaslavsky et al., 2007). WaterOneFlow and WaterML were specifically designed for CUAHSI-HIS to provide the vehicle or service bus through which hydrologic data can be completely described and efficiently delivered via the Web. In conjunction, WaterOneFlow web services and WaterML support the infrastructure within CUAHSI-HIS to efficiently share hydrologic information.

The individual components of CUAHSI-HIS each serve an important role in the data discovery and fetching process. HydroServers function as the principal locations for storing large volumes of hydrologic data, specifically time series. Within the server itself, data and metadata are managed in a database and then exposed through a suite of Web services (e.g. a WaterOneFlow Web service) so that remote users can then access the data through the Web (Horsburgh et al., 2010).

Another component of CUAHSI-HIS is HIS Central or the hydrologic metadata catalog. HIS Central is the component of CUAHSI-HIS which facilitates the discovery of hydrologic data that has already been published on HydroServers. Within CUAHSI-HIS, HydroServers are the primary repositories for hydrologic data, while HIS Central is the primary repository for hydrologic data services (Maidment, 2009). HIS Central provides an interface where users can search registered HydroServers by specifying keywords and metadata which describe the hydrologic data of interest (Tarboton et al., 2010). HIS Central is like a Google for discovering hydrologic time series information. Data publishers can register their data on HIS Central and provide brief descriptions of the datasets they want to share. This is an important aspect of CUAHSI-HIS because it allows for data to be organized and discovered in an efficient, structured and methodical process.

The third and final component of CUAHSI-HIS is HydroDesktop. HydroDesktop is the component of CUAHSI-HIS that allows for the harvesting of hydrologic information at the locality of one's own computer or analytical system (Ames et al., 2010). HydroDesktop is a platform located on the user's machine and communicates with both HydroServers and HIS Central (Tarboton et al., 2010). Users can directly download hydrologic information from HydroServers if they already know of their existence or can search HIS Central for data that they might not know about (Ames et al., 2010). Once the data of interest has been discovered, users can download the information onto their local databases. With the information readily and locally available, users can take data they have harvested and combine it with other data already available on their machine and use it to perform insightful analysis and/or modeling. HydroDesktop is intended to synthesize hydrologic information in an environment that supports both time series and geographic visualization (Maidment, 2009). With this unique structure, HydroDesktop provides a method for users to efficiently manage and work with hydrologic information. Although HydroDesktop plays an important role in the CUAHSI-HIS SOA, this study will focus on the underlying services framework through which HydroDesktop can be modified to retrieve and synthesize both hydrologic and atmospheric science information.

As of 2011, CUAHSI-HIS contains the biggest water data catalog in the world. With 66 public services registered at HIS Central, 5.1 billion data values measuring 18 000 variables at 1.9 million sites are made accessible to the public

for quick and efficient use (Tarboton et al., 2010). Not only does CUAHSI-HIS allow data consumers to access small datasets used for research, it also allows data consumers to access large datasets published by federal agencies. The United States Geological Survey (USGS) and the Environmental Protection Agency (EPA) are both examples of federal agencies distributing time series data through CUAHSI-HIS. Although CUAHSI-HIS has demonstrated a successful web services approach to managing and sharing hydrologic information, it has yet to cross the digital divide and provide access to the plethora of gridded information collected by those in the field of atmospheric sciences.

2.3 Unidata and Thematic Realtime Environmental Distributed Data Services

The Unidata project, developed within the University Corporation for Atmospheric Research (UCAR), has been standardizing the manner in which atmospheric science information (e.g. satellite, radar, model outputs, lightning data etc.) is openly shared across the Web. Like CUAHSI-HIS within hydrologic sciences, Unidata has designed an SOA based service bus that enables users to efficiently publish, discover and access atmospheric science data through the Web, specifically grids and multi-dimensional arrays. Unidata has developed three main tools to help facilitate this process: the Network Common Data Form (netCDF), the Thematic Realtime Environmental Distributed Data Services (THREDDS) (Domenico, 2002), and the Integrated Data Viewer (IDV). These three components work together in a form similar to CUAHSI-HIS: THREDDS servers manage, store and publish gridded data in netCDF format via web services; Unidata builds a registry of THREDDS servers; and IDV discovers, synthesizes and accesses gridded data.

Within CUAHSI-HIS, WaterML was designed to facilitate the exchange of time series data across the Web. As an analogy, netCDF is the WaterML of Unidata. NetCDF is a data model that incorporates a set of interfaces, libraries and standardized formats that support the creation, access and sharing of gridded scientific data (OGC, 2010). As part of this effort, a netCDF binary encoding as well as an XML realization called NcML have been defined (Nativi et al., 2005). Studies have shown that the array-oriented structure of netCDF files provides the most efficient form of storing and retrieving gridded time series (Doraiswamy et al., 1999). Moreover, netCDF allows data to be visualized using GIS software, which has become a leading technological and analytical platform through which interoperability studies are performed. The efficient structure of netCDF allows for access to small subsets of large multidimensional arrays (OGC, 2011).

Like the HydroServer for time series, THREDDS servers were developed for storing and accessing multidimensional arrays and grids provided by multiple data sources. THREDDS servers are distributed inventory systems that

allow data providers to publish and completely describe gridded data through the utility of standard Web services (Domenico et al., 2006). THREDDS servers act as the intermediary between the data provider and data user by standardizing the format in which gridded data is made accessible regardless of the format the underlying data is stored in (Domenico et al., 2006). In this study, we focus on gridded data that is published as Web Coverage Services (WCS) because of its standardization within the OGC infrastructure. A WCS (comparable to a WaterOneFlow web service) is a standardized Web service that facilitates the exchange of coverage data (e.g. netCDF, GRIB, HDF datasets) across the Web (OGC, 2008). Because THREDDS servers distribute gridded data using standardized Web service interfaces, THREDDS has become a well-used and robust tool for managing and distributing large quantities of gridded information.

The final component within the Unidata SOA is IDV. IDV is the HydroDesktop for discovering and accessing gridded data and metadata. IDV enables data consumers to search and retrieve gridded information stored on remote THREDDS servers (Meertens et al., 2006). Data consumers can search for gridded information by filtering on keywords and metadata and then readily connect to the data provider to access the dataset of interest. IDV primarily functions to connect the data user to the data provider and facilitate the manner in which gridded data is discovered, transmitted and retrieved.

Although slightly different in paradigm, Unidata and CUAHSI have both developed SOAs and unique service buses that focus on efficiently delivering scientific information from data providers to data consumers. CUAHSI-HIS manages and distributes hydrologic data stored as time series, while Unidata manages and distributes atmospheric data stored as multidimensional arrays or grids. A framework which can tie together both the CUAHSI-HIS and Unidata systems will help promote interoperability among scientists.

2.4 Standardization process and initiatives in the geospatial Web realm

In recent years, a plethora of initiatives linked to GeoInformatics have rapidly emerged. More than forty organizations, agencies, initiatives and programs attended *The Earth & Space Science Informatics Summit*, held in Rome in 2008. This vast presence of multidisciplinary organizations enforced the need to promote interoperability, communication and cooperation across scientific disciplines within the Earth sciences (Fox, 2008). Amongst the initiatives and organizations that have committed to drafting guidelines and regulations supporting interoperability are the AGU ESSI group of interest, the EGU ESSI division, the OGC, the European directive INSPIRE, the Global Monitoring for Environment and Security (GMES) program, the ISO technical committee 211 and the Group on Earth Observations (GEO). There are several other groups working on interoperability issues as well; however, many of them are focused on the underlying

technological aspects of interoperability (e.g. OASIS, IEEE, IETF, W3C).

Within the context of this study, the OGC provides the services infrastructure that enables a common interoperable data model that conforms to standards. The OGC is a consortium of industry leaders from government, private and research sectors around the world that develops international open standards and interoperable solutions that “geo-enable” the web (OGC, 2012). As part of this effort, the OGC develops schemas and specifications for geospatial Web services. Some of these services include Sensor Observations Services (SOS), Web Feature Services (WFS), Web Coverage Services (WCS) and Catalog Services for the Web (CSW) (OGC, 2005, 2007a, b, 2008). Respectively, each of these services focus on transmitting different types of geospatial information across the Web: observations data; geographic features, multidimensional arrays and grids; and geospatial metadata. In recent years, many governments and international agencies (i.e. GEO, Federal Geographic Data Committee (FGDC), World Meteorological Organization (WMO), etc.) have endorsed several of the OGC’s Web service standards (OGC, 2012). It is becoming apparent that standardized Web services are a common practice amongst data providers and consumers worldwide by providing the building blocks for not only e-infrastructure but also spatial data infrastructure (Nebert, 2004).

3 Design concepts

Using the knowledge and experience gained by CUAHSI, Unidata and ESSI-Lab, we introduce the conceptual data model behind an interoperable solution that would allow users to readily access both hydrologic and atmospheric science data within a common interface. In the following sections we introduce the individual components of this solution and describe their role in fostering interoperability.

3.1 Conceptual data object model

One of the main differences that inhibit the sharing of hydrologic and atmospheric data is the conceptual model for which a data object in space and time is described. In hydrologic science, there is one common approach to describing a time series object: a time series object is a variable measured at particular point in space over a period of time. In atmospheric science, this is not the case. There are multiple approaches to describing a collection of grids as a single data object: a collection of grids can measure myriad variables over a period of time, myriad variables at a single instance in time, or one variable over a period of time. See Fig. 4.

In hydrologic sciences, scientists are interested in acquiring data over a period of time as time series. This conceptual framework is derived from the data cube model for describing a single data value within a space, time and

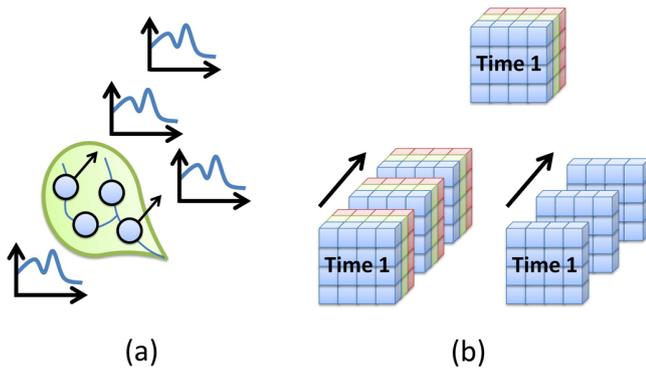


Fig. 4. Traditional data storage for (a) time series and (b) grids.

variable domain (Maidment, 2002). The data cube states that a particular data value measures a single variable at a location in space and position in time. If one were to extend this model to encompass many values over a particular domain, one can then describe a set of values instead of just a single value. This is what CUAHSI-HIS has done to describe time series within its SOA. CUAHSI-HIS conceptually describes a time series object as a set of values, sampled in time, describing a variable at a specific site within a given network provided by a data source (Maidment, 2009). Although this is the conceptual model, this does not limit a time series object from having additional metadata associated with it. In fact, there have been numerous studies that focus on metadata within the field of hydrology (Horsburgh et al., 2009; Piasecki and Beran, 2009; Whitenack et al., 2010).

Similarly, one can think of a data object as describing a collection of time varying grids. The OGC defines a coverage as a “space-time varying phenomenon” or more succinctly put, a collection of grids describing a variable(s) over a period of time within a dataset provided by a data source. With respect to time series, one might think of this as a spatially continuous 2-D or 3-D coverage as opposed to a discrete 1-D coverage (i.e. time series); 3-D refers to a coverage containing multiple variables. In this study, we extend the CUAHSI-HIS model and describe gridded data objects as 2-D coverages to facilitate the discovery of data; this will be expounded upon in the following sections. Although we choose to describe gridded data as a set of 2-D coverages, we recognize that gridded data can be stored on servers as either 2-D or 3-D data objects; it is possible to access subsets of 3-D coverages by leveraging OGC WCS. As with time series, additional metadata can be attached to a 2-D coverage to completely describe the object of interest. Figure 5 demonstrates the data object model for both a 1-D time series and 2-D coverage.

This conceptual model is the basis for the proceeding work and interoperability experiments. It is the conceptual framework for organizing data objects which allows data to be managed and published in a way that users can discover and

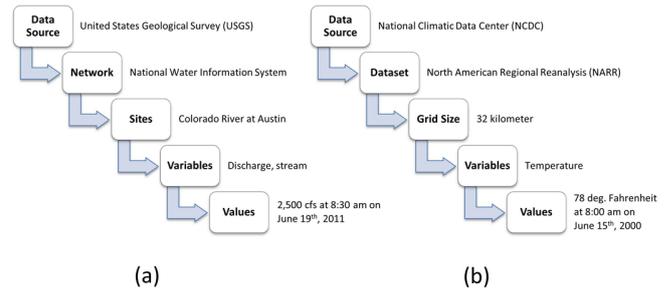


Fig. 5. Conceptual data object model for a (a) 1-D time series and (b) 2-D coverage.

access both hydrologic and atmospheric science information within a common interface. In many cases gridded data is not organized in the aforementioned manner, thus providing challenges in the data discovery process. Grids routinely published on THREDDS servers in real or near real-time often describe a collection of variables for a single (not multiple) time step. In these situations, large quantities of disjointed grid files are generated and hinder the ability for users to access temporal subsets of large datasets.

3.2 CUAHSI-HIS services stack

The services stack framework was initially developed by CUAHSI as a solution for sharing hydrologic time series metadata and data using OGC standard Web services (Seppi, 2010); however, it will be shown in this paper that this framework can be extended to share atmospheric coverage metadata and data as well. In this light, the services stack framework is an information model that facilitates interoperability among data providers and data users in hydrologic and atmospheric sciences. There are three components to the services stack framework which work together to provide a system in which data consumers can readily discover and access both time series and coverage data using spatial, temporal and semantic filters: Catalog Services, Metadata Services and Data Services (see Fig. 6). At the core of the services stack framework lie the metadata services which act as middleware between the catalog services and data services. Data services ultimately provide the user with the data they are searching for, whereas catalog services allow users to perform federated searches across multiple data providers. Metadata services link both these layers together by being registered at the catalog level and providing all the information needed to access information at the data level. Although there are various OGC specifications and schemas used within this architecture, this paper describes the general approach and concepts through which hydrologic and atmospheric science data providers can publish data and become a part of a common interoperable information system.

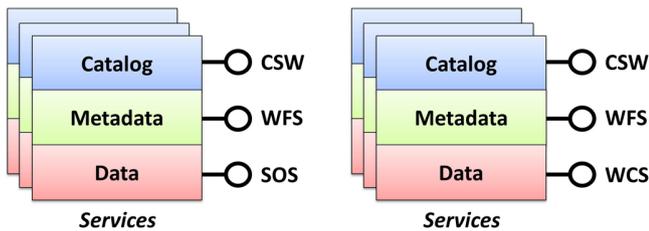


Fig. 6. CUAHSI-HIS services stack framework for time series and coverages.

3.2.1 Data services

Data services are at the bottom of the services stack framework and are responsible for providing users with the data sought after. These services follow the Sensor Observation Service (SOS) and Web Coverage Service (WCS) specifications outlined by the OGC. For hydrologic information, time series data is obtained through a SOS implementation encoded in WaterML 2.0 (Seppi, 2010; OGC, 2012, 2007b). Similarly, for atmospheric science information, coverages can be obtained through a WCS implementation as a netCDF file (Nativi and Domenico, 2009). Each service allows users to specify spatial, temporal and semantic filters to obtain subsets of the larger datasets supplied by the data service. Below is a sample SOS and WCS request:

<http://129.116.104.174:8080/KiWIS?>

```
service=SOS&
request=GetObservation&
featureOfInterest=30463&
ObservedProperty=Q&
eventtime=2012-01-00T00:00/2012-03-01T00:00&
responseFormat=text/xml;subtype=WML2
```

<http://nomads.ncdc.noaa.gov/thredd/wcs/aggs/narrmonA?>

```
service=WCS&
request=GetCoverage&
bbox=-145.54,0.7532,-2.3244,46.5161&
coverage=Total_precipitation&
time=1979-01-01T00:00/1986-01-01T00:00&
format=NetCDF3
```

3.2.2 Metadata services

Metadata is an essential component of the data sharing process. Not only does it facilitate the search and discovery of information within one's own research community but also helps foster interoperability between research communities. Metadata is used to describe a set of data that share a common ground to others who are not directly familiar with the information. With so many different research fields in existence, finding common approaches and structures for

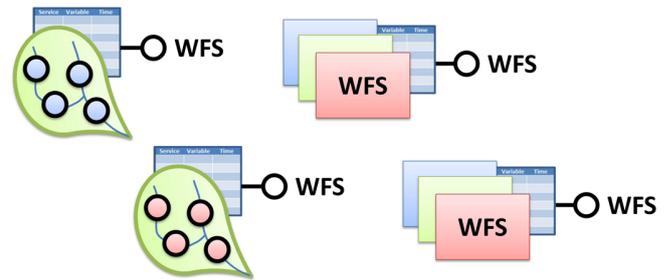


Fig. 7. CUAHSI-HIS metadata services for time series and coverages.

developing metadata is not yet clear. However, the metadata issue is very well documented and has led many scientific communities to adopt metadata standards created by the International Organization for Standardization (ISO). Specifically, the ISO-19115 standard provides a manner in which geographic metadata can be published across Web based information systems (Inspire, 2010).

Currently, the majority of the work that has been done with metadata standards focuses on the generic representation and description of geospatial data with specialties in features, coverages, etc. These standards provide a basis for building metadata but they lack some key functionality in application to hydrologic time series and atmospheric coverages. Metadata in the services stack framework is conveyed through a metadata service implemented as a Web Feature Service (WFS). These metadata services provide the user with a complete description of either the time series or coverage of interest. Geographically, time series are symbolized as point features representing the geographic location of a gauge, whereas coverages are symbolized as polygon features representing the geographic extent of a grid. See Fig. 7.

Within a given network or dataset, a data provider can describe a set of time series or coverages using a WFS. For example, one can imagine a network of four observation gauges, each measuring two variables. In this case, the data provider would describe eight time series objects within a single WFS. Similarly, a climate model could contain model outputs for six different variables. In this case, each variable would be described as a single data object; therefore, the data publisher would provide metadata for six coverages. By ingesting metadata within a WFS, time series and coverages can be described in a consistent format, thus allowing users to visualize the metadata information in a similar fashion. Hence, the fundamental data object in this architecture is a single variable described over a domain of space and time – in hydrology this means a time series measured at a gaging or sampling site, in atmospheric science it means a coverage observed or computed over a spatial domain for a period of time.

Table 1. Condensed CUAHSI-HIS metadata specifications.

Observation Metadata Catalog	Coverage Metadata Catalog
Source	Source
Site Name	Coverage Name
Location (Lat/Long)	Location (Lat/Long Box)
Variable Name	Variable Name
Variable Units	Variable Units
Elevation	Level
Concept	Concept
Ontology	Ontology
Start Date	Start Date
End Date	End Date
SOS URI	WCS URI

CUAHSI-HIS has developed similar metadata structures for describing time series and coverages. The metadata specifications contain fields that not only describe a data object in detail but also provide sufficient information for a client (user or computer) to directly access each dataset described in the catalog. All the information needed to make a complete HTTP-GET or POST request on the respective Web services (e.g. SOS and WCS) can be found in corresponding fields in the metadata specifications. These requests can be used to access full data objects and subsets of data objects filtered by space, time and variable. See Table 1.

One of the key components involved in providing metadata for hydrologic and atmospheric science is the semantic mediation that is resolved using the CUAHSI ontology (Whitenack, 2010). As part of the CUAHSI metadata specifications, semantic mediation or definition of search terms is addressed by providing a set of fields in the specification through which variables can be defined; these fields are defined as the *Concept* and *Ontology* fields. The *Concept* field represents the concept within the CUAHSI Ontology through which the data consumer can search. The *Ontology* field represents the version of the CUAHSI Ontology that is being utilized. If so desired, a data provider can also use a different semantic ontology to describe a series as long as the search client is locally aware of the ontological mapping. In general, the semantic ontology associated with each series allows custom clients to search for similar information over myriad data sources (i.e. precipitation, evaporation, etc.). Figure 8 shows metadata implemented as a WFS for a single data object hosted on a THREDDS server. The metadata describes monthly evaporation data coming from the North American Regional Reanalysis (NARR). A sample WFS request is shown below:

```
http://129.116.104.176/arcgis/services/NARRMonthly/
MapServer/WFSServer?
```

```
service=WFS&
request=GetFeature&
TypeName=Climate_NARRMonthly:NARRMonthly
```

3.2.3 Catalog services

Catalog services aid in the management, discovery and distribution of metadata describing geographic datasets and services (Nativi and Bigagli, 2009). Within the CUAHSI-HIS services stack framework, catalog services function as the interface through which data consumers discover indexed metadata services published as WFS. As part of the standard suite of OGC services, Catalog Services for the Web (CSW) are the Web services that focus on the management and indexing of geographic metadata. The OGC designed CSW to help data consumers search through a set of matching resources. As such, CSW allow data publishers to register and index a set of metadata services with a variety of different metadata profiles as defined by the ISO (e.g. ISO 19115/19119). Using a CSW interface, data publishers are able to share their catalog of metadata with search clients as well as other catalogs. Some CSW implementations even permit the federation of other remote catalogs; this functionality allows search clients to perform federated searches across multiple catalogs. Although this framework provides one conceivable solution, it is also possible to eliminate the metadata service layer within the services stack framework and directly register a set of data services within a CSW catalog. Below is a sample CSW request:

```
http://hydroportal.crwr.utexas.edu/geoportal/csw/discovery?
```

```
request=GetRecords&
service=CSW&
resultType=results&
elementSetName=full
```

By organizing hydrologic and atmospheric science information in this manner, data publishers have the ability to maintain and manage their own metadata and data while still conforming to standards and participating within a larger interoperable information system. In order to leverage this distributed approach, CUAHSI-HIS has created an experimental meta-catalog, called HydroPortal, which functions as a catalog of catalogs (or catalog gateway). See Fig. 9.

Within the CUAHSI-HIS infrastructure, HydroPortal is the top layer of the data discovery process. Data publishers can register their own service stacks by registering a CSW within the HydroPortal system. This approach is fundamentally different than that of the current HIS Central system. HIS Central serves as the centralized metadata hub for all data services which have been registered; it contains a harvested catalog of all the time series within the

```

</Climate_NARRMonthly:Shape>
<Climate_NARRMonthly:ServName>NARRMonthly</Climate_NARRMonthly:ServName>
<Climate_NARRMonthly:CovName>NARR-A Monthly Complete PoR Aggregation</Climate_NARRMonthly:CovName>
<Climate_NARRMonthly:VarCode>EVP</Climate_NARRMonthly:VarCode>
<Climate_NARRMonthly:VarName>Evaporation</Climate_NARRMonthly:VarName>
<Climate_NARRMonthly:VarUnits>kilogram per meter squared</Climate_NARRMonthly:VarUnits>
<Climate_NARRMonthly:Concept>Evaporation, actual</Climate_NARRMonthly:Concept>
<Climate_NARRMonthly:Ontology>CUAHSI Ontology - March 2010</Climate_NARRMonthly:Ontology>
<Climate_NARRMonthly:StartDate>1979-01-01T00:00:00Z</Climate_NARRMonthly:StartDate>
<Climate_NARRMonthly:EndDate>2010-08-01T00:00:00Z</Climate_NARRMonthly:EndDate>
<Climate_NARRMonthly:Level>Surface</Climate_NARRMonthly:Level>
<Climate_NARRMonthly:IsRegular>1</Climate_NARRMonthly:IsRegular>
<Climate_NARRMonthly:TimeUnits>Month</Climate_NARRMonthly:TimeUnits>
<Climate_NARRMonthly:TimeStep>1</Climate_NARRMonthly:TimeStep>
<Climate_NARRMonthly:DataType>Cumulative</Climate_NARRMonthly:DataType>
<Climate_NARRMonthly:GridSize>32</Climate_NARRMonthly:GridSize>
<Climate_NARRMonthly:GridUnit>kilometer</Climate_NARRMonthly:GridUnit>
<Climate_NARRMonthly:GridType>Lambert Conformal Conic</Climate_NARRMonthly:GridType>
<Climate_NARRMonthly:Method>Modeled data</Climate_NARRMonthly:Method>
<Climate_NARRMonthly:QCLevel>Interpreted products</Climate_NARRMonthly:QCLevel>
<Climate_NARRMonthly:XLL>-210.8600000000001</Climate_NARRMonthly:XLL>
<Climate_NARRMonthly:YLL>0.59882901899999996</Climate_NARRMonthly:YLL>
<Climate_NARRMonthly:XUR>-4.5</Climate_NARRMonthly:XUR>
<Climate_NARRMonthly:YUR>85.359088060000005</Climate_NARRMonthly:YUR>
<Climate_NARRMonthly:WFSURI>
  http://129.116.104.176/arcgis/services/Climate/NARRMonthly/MapServer/WFSServer
</Climate_NARRMonthly:WFSURI>
<Climate_NARRMonthly:WCSURI>
  http://nomads.ncdc.noaa.gov/thredds/wcs/aggs/narrmonA
</Climate_NARRMonthly:WCSURI>
<Climate_NARRMonthly:WMSURI>
  http://nomads.ncdc.noaa.gov/thredds/wms/aggs/narrmonA
</Climate_NARRMonthly:WMSURI>
<Climate_NARRMonthly:NCSSURI>
  http://nomads.ncdc.noaa.gov/thredds/ncss/grid/aggs/narrmonA
</Climate_NARRMonthly:NCSSURI>

```

Fig. 8. Example Web Feature Service (WFS) response containing coverage metadata.

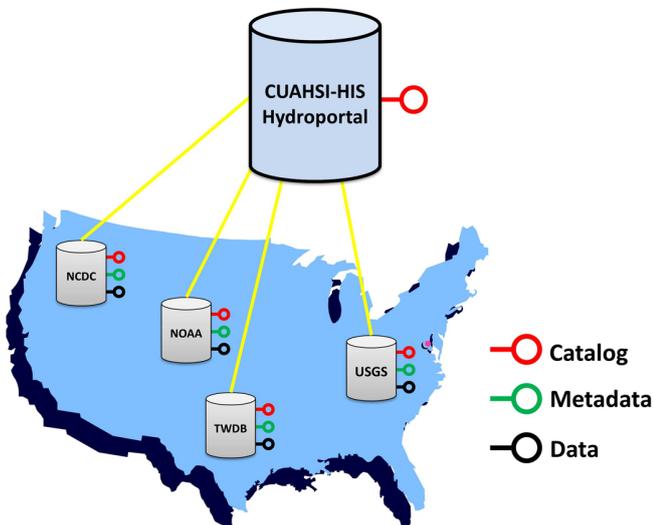


Fig. 9. Distributed CUAHSI-HIS HydroPortal.

system. In contrast, the services stack framework is a distributed approach where metadata, time series or coverage, can be harvested and indexed by multiple systems instead of one. Furthermore, because HydroPortal is consistent with the OGC framework, a user would have access to all the underlying metadata and data via the CSW interface. This could be demonstrated by registering the HydroPortal within the Global Earth Observation System of Systems (GEOSS); however, this has yet to be tested.

3.3 Comparison and relations between CUAHSI and Unidata frameworks

In the previous section, a framework for sharing both hydrologic and atmospheric data was presented and described. This framework conforms to OGC standard Web services and allows data providers to publish and manage their own data while giving data users the ability to readily discover and access that data. Within this system, hydrologic metadata and data, stored on CUAHSI-HIS HydroServers, are published using the suite of OGC standard Web services described in the services stack framework. Similarly, atmospheric science metadata and data stored on THREDDS servers are published in the same manner. As of 2011, CUAHSI has begun to migrate the existing CUAHSI-HIS to the OGC standard framework described above. Although this is a promising first step, there are currently hundreds of THREDDS servers worldwide that contain atmospheric science data and are not a part of the CUAHSI-HIS system.

In the previous sections, two similar service stacks have been designed and implemented by CUAHSI and Unidata for time series and coverages, respectively. Each stack is composed of different service types: at the top there are services working at a high abstraction level (i.e. catalog services, used for data discovery); at the bottom there are services operating at a low abstraction level (i.e. data services, which enable the actual downloading of data). Figure 10 depicts the mapping from the CUAHSI-HIS services stack to the corresponding

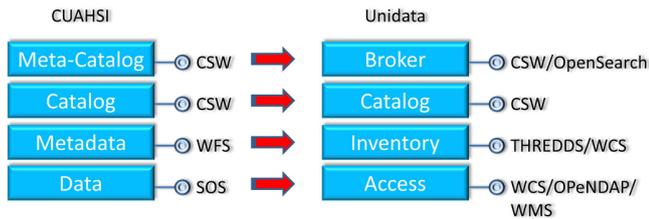


Fig. 10. Correspondences between existing services in the CUAHSI-HIS and Unidata frameworks.

Unidata services stack. The figure describes a terminology divide that exists between the two frameworks, and at the same time helps cross it.

At the topmost level of the CUAHSI-HIS framework, a meta-catalog service is found. This service, implemented as a CSW, is used to discover metadata services by distributing incoming queries to a set of federated catalogs (implemented as CSW as well), realizing a catalog gateway or clearinghouse system. Each of the federated catalogs can be queried to find one or more time series published by the metadata service, implemented as WFS. Each time series contains in its metadata a pointer to the data service, implemented as a SOS. This last service can be used to obtain the raw data, as acquired by the sensors.

Within the Unidata framework there lies a similar system. At the topmost level of the Unidata framework there is a broker service, implementing discovery interfaces such as CSW and OpenSearch. A broker distributes user queries to a set of heterogeneous services (i.e. catalogs services, but also inventory and access services), realizing also a distributed infrastructure functioning as a resources registry. Beneath, the catalog service type is shown, implemented as a CSW. A catalog service is able to harvest the available metadata offered by THREDDS services, and executes complex queries against the available metadata. Metadata can also be harvested directly from WCS services in a fashion similar to the CUAHSI-HIS framework. THREDDS services work as an inventory (or listing) service, being able to hierarchically organize and publish a local collection of multi-dimensional arrays (e.g. netCDF, GRIB files), as well as publish auxiliary standard services to realize the actual data access and visualization (e.g. WCS, OPeNDAP, WMS).

4 Interoperability experiments

Based on the design concepts presented in the previous sections, two interoperability experiments were performed using a variety of clients, tools and interfaces. These experiments were conducted to demonstrate data interoperability by enabling users to search, discover and access both hydrologic and atmospheric science data through an implementation of a standard set of OGC compliant web services.

The first interoperability experiment in this study was based on the CUAHSI-HIS services stack framework and was implemented using ESRI's GeoPortal interface. GeoPortal is a free open source product designed by ESRI that empowers users to discover a collection of registered services via their metadata (ESRI, 2012). Along with its Web based GUI interface, GeoPortal also allows users to search for records through its CSW interface. The second experiment in this study was based on the Unidata framework and the GI-cat mediation software. GI-cat is an implementation of a broker catalog service designed by ESSI-Lab.

These experiments were chosen to answer the following questions:

1. Can one publish hydrologic and atmospheric science data (i.e. time series and coverages) in a common manner that would allow a client to systematically discover and access data by applying spatial, temporal and semantic filters?
2. Is it possible to integrate CUAHSI time series services into existing portals and gateways within the Earth science community?

The answers to these questions will be provided through the following experiments and will demonstrate a proof of concept for the proposed interoperability solutions. Not only is it important to demonstrate a framework which empowers interoperability but also one that is in line with other existing interoperable systems.

4.1 Experiment #1: GeoPortal

The first interoperability experiment was performed in accordance to the CUAHSI-HIS services stack framework. A set of hydrologic data services published on CUAHSI-HIS HydroServers (e.g. time series data services) were thematically organized using the CUAHSI metadata specification. Once organized within a table, the metadata catalog was ingested into ArcGIS as a point feature class and then published as a WFS using ArcGIS Server; see Fig. 7 for reference. The service was then registered in a central catalog via CUAHSI's version of ESRI's Geoportal, HydroPortal.

Similarly, a set of WCS published on Unidata THREDDS servers (e.g. gridded data services) were thematically organized using the CUAHSI metadata specification. Once organized in a table, the metadata catalog was ingested into ArcGIS as a polygon feature class and then published as a WFS using ArcGIS Server; again see Fig. 7 for reference. Once published as a WFS, the service was registered in HydroPortal and made available through its GUI and CSW interface (note that in these two circumstances HydroPortal acts as a *catalog service* not a *meta-catalog* as presented in Sect. 3.2.3). Figure 11 shows a returned search request within HydroPortal using "evaporation" as the keyword. Both metadata describing time series and coverages

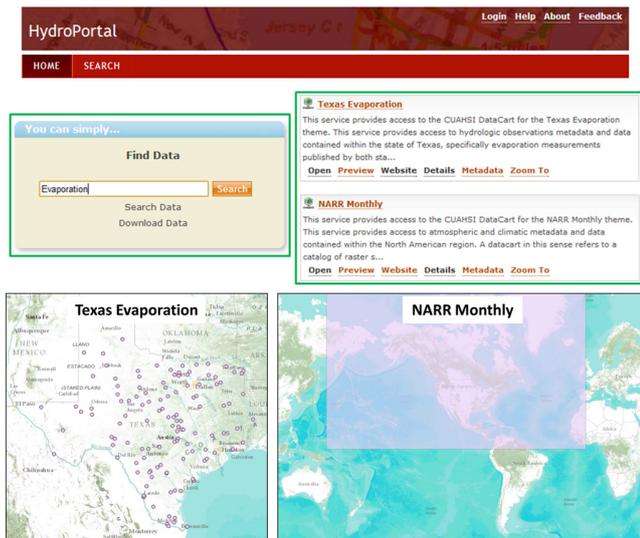


Fig. 11. List of metadata services, implemented as WFS, returned by experimental CUAHSI-HIS HydroPortal (keyword “Evaporation” was used as the semantic query).

are returned. Figure 12 shows a metadata record within HydroPortal implemented through its CSW interface.

As has been demonstrated by Seppi (Seppi, 2010), a client can be built to systematically discover and access information organized in this fashion. Clients can also apply filters (either at the CSW or WFS level) to spatially, temporally, and semantically sift through the returned metadata and then efficiently access the data of interest either through an SOS or WCS request. See Fig. 13 for an abstract layout of this process.

4.2 Experiment #2: GI-cat

The second interoperability experiment in this study was performed using ESSI-Lab’s GI-cat service (Bigagli et al., 2004). A GI-cat service implements a discovery broker within the publish, find, bind SOA.

GI-cat allows a client to query heterogeneous data sources and services through a common discovery interface by binding directly to the different service types and mediating between the multiple service providers and the client (Nativi et al., 2009). GI-cat supports several international and community standards and services: catalog services (such as CSW in its ISO and eBRIM profiles, OpenSearch engines, Degree and GeoNetwork); inventory services (such as THREDDS, OAI-PMH, Web Accessible Folders, FTP); access services (such as WCS, WMS, WFS); and local folders and databases (with support for different formats such as ISO19139, DIF, Dublin Core, netCDF).

GI-cat provides a flexible framework to interconnect heterogeneous resources (i.e. data repositories and services) by means of a mediation and adaptation approach (Nativi et al., 2007). For each resource type, the protocol and data model

mediation functionalities are implemented by a specific software component called an “Accessor”. New “Accessors” can be added to the system in order to support the discovery and access of a new resource type. This standard based approach allows to interconnect, in a loosely-coupled way, existing and even future resources.

At the same time GI-cat can be accessed by different discovery clients such as ArcGIS, the GEO/GEOSS Portal, GeoNetwork, GI-go Geobrowser and its own built-in Web portal. The software components which carry out the publication of specific catalog interfaces are called “Profilers”. They carry out mediation functionalities between the published interfaces and the GI-cat internal interface. Just like the Accessors, new Profilers can be created and plugged into the system in order to publish new discovery interfaces. Queries can be executed on the fly against the available sources or a local metadata collection, periodically updated by a specific component (called the Harvester); a mixed strategy can be easily configured in order to tailor the broker to the desired user scenario.

Within the context of this study, GI-cat was used to harvest approximately 400 000 metadata records describing gridded resources from the Motherlode THREDDS server hosted at Unidata (see Fig. 14). These resources include gridded data from the National Center for Environmental Prediction (NCEP), the Unidata Real-time Regional Model, Next Generation Radar (NEXRAD), satellites and others (nearly all resources are available via a WCS implementation). The metadata of those resources were then made available for discovery through a CSW ISO interface published by GI-cat. The optional link needed to inject metadata resources from HIS Central directly into GI-cat is also shown in Fig. 14 with a dotted line, however it was not used for the described tests; it may be subject of further tests.

Queries to the configured system can be issued by any CSW client by using a combination of the standard ISO queries, as defined in the CSW ISO AP specification (e.g. geographic extent, keywords, temporal extent etc.). In this experiment, different clients were used to retrieve metadata from the described system: GI-go GeoBrowser, GeoNetwork, ArcGIS Explorer and also a built-in Web portal within GI-cat. GI-cat can also publish other interfaces beside the CSW/ISO interface used for the tests (e.g. OpenSearch, OAI-PMH, etc.); these additional configurations may be subject to future tests.

In order to test the integration of the described system within the CUAHSI-HIS architecture, a custom CSW instance of HIS Central was created using ESRI’s Geoportal interface (approximately 61 services registered). Once this was achieved, a series of federated searches across multiple CSW instances were instituted through HydroPortal. As a test case, a search for “precipitation” keyword on the HydroPortal returned 43 648 hits through the GI-cat CSW interface, while 15 hits were returned through the custom HIS Central CSW interface, making available both the matching

```

- <SummaryRecord>
  <dc:identifier scheme="urn:x-esri:specification:ServiceType:ArcIMS:Metadata:DocID">
    {38D2D803-5A18-4370-875E-DDB6C20D85E7}</dc:identifier>
  <dc:title>Texas Evaporation</dc:title>
  <dc:type scheme="urn:x-esri:specification:ServiceType:ArcIMS:Metadata:ContentType">liveData</dc:type>
  <dc:subject>climatologyMeteorologyAtmosphere</dc:subject>
  <dc:modified>2010-11-13T16:12:53-06:00</dc:modified>
- <dc:abstract>
  This service provides access to the CUAHSI DataCart for the Texas Evaporation theme. This
  service provides access to hydrologic observations metadata and data contained within the
  state of Texas, specifically evaporation measurements published by both state and national
  agencies. A datacart in this sense refers to a catalog of hydrologic time series described by
  various metadata. The datacart in this theme is served as a Web Feature Service (WFS).
  </dc:abstract>
- <dc:references scheme="urn:x-esri:specification:ServiceType:ArcIMS:Metadata:Server">
  http://129.116.104.176/ArcGIS/services/Themes/TexasEvaporation/MapServer/WFSServer
  </dc:references>
- <dc:references scheme="urn:x-esri:specification:ServiceType:ArcIMS:Metadata:Onlink">
  http://129.116.104.176/ArcGIS/rest/services/Themes/TexasEvaporation/MapServer/
  </dc:references>
- <dc:references scheme="urn:x-esri:specification:ServiceType:ArcIMS:Metadata:Document">
  https://hydroportal.crwr.utexas.edu/geoportal
  /csw/discovery?getxml=%7B38D2D803-5A18-4370-875E-DDB6C20D85E7%7D
  </dc:references>
- <ows:WGS84BoundingBox>
  <ows:LowerCorner>-105.0833333 26.06666667</ows:LowerCorner>
  <ows:UpperCorner>93.56666667 34.85</ows:UpperCorner>
  </ows:WGS84BoundingBox>
- <ows:BoundingBox>
  <ows:LowerCorner>-105.0833333 26.06666667</ows:LowerCorner>
  <ows:UpperCorner>93.56666667 34.85</ows:UpperCorner>
  </ows:BoundingBox>
</SummaryRecord>
  
```

Fig. 12. Example record within HydroPortal via its CSW interface (Seppi, 2010).

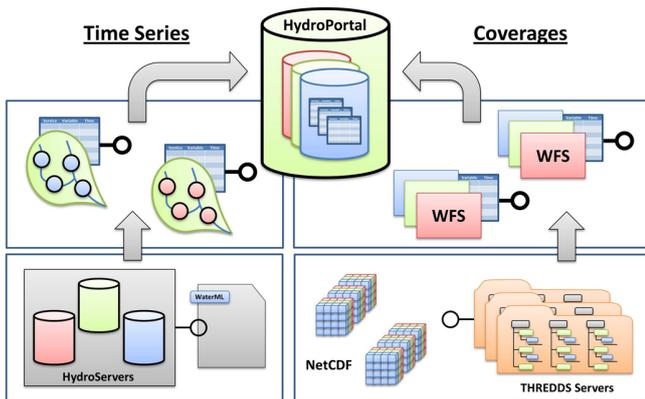


Fig. 13. Data and metadata services registered within CUAHSI-HIS HydroPortal.

metadata records through a common interface. Figure 15 shows a general representation of the implemented test case.

The sequence diagram for harvesting data services can be seen in Fig. 16. During the harvesting phase, the user first triggers the HIS servers, followed by the Motherlode servers (alternatively these can be triggered through an automatic

timer). The catalog services then store the data services incrementally through a loop (possibly after translation).

Figure 17 shows the sequence diagram and interaction between components during a typical query from HydroPortal. HydroPortal distributes the incoming query to HIS Central and GI-cat services at the same time, acting as a gateway. The results are then returned as shown in its GUI interface. The entire query process from user input to returned results takes a few seconds, thus making this test case a viable option for production scenarios.

5 Conclusions

This study has provided some insight and prompted some discussion for how to improve the sharing of scientific information and foster interoperability in research, private and government sectors around the world. It was the goal of this study to provide a standardized framework through which both time series and gridded data can be managed, discovered and accessed in a structured process that allows users to efficiently gather data of different types across scientific disciplines. It has become apparent in this study that currently no one system can adequately conquer the digital divide but instead a framework which leverages standards (i.e.

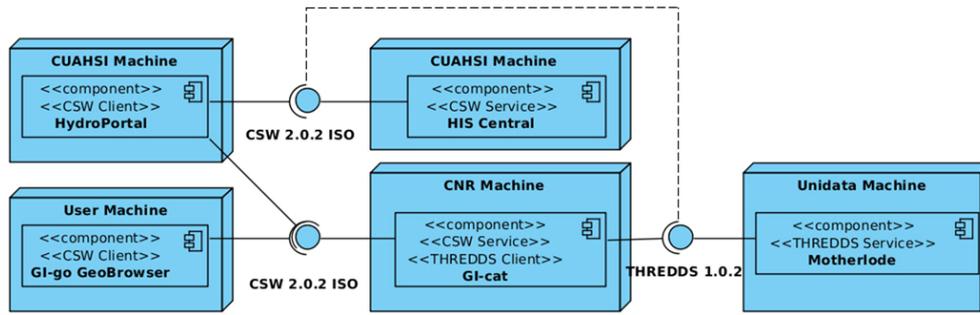


Fig. 14. GI-cat experimental deployment diagram.

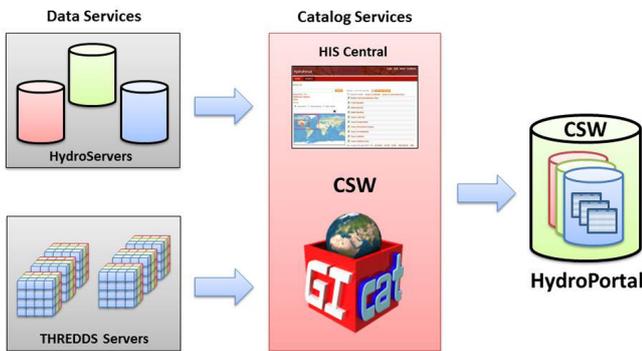


Fig. 15. Abstract diagram of interoperable data discovery system; hydrologic and atmospheric science data indexed within HydroPortal.

OGC compliant web services) can. It has been demonstrated that within hydrologic sciences, CUAHSI-HIS is a leading provider of time series data, while Unidata (within atmospheric sciences) is a leading provider of gridded data. This work has shown that through the use of standard Web services, federated catalogs can be built which can integrate data across multiple scientific domains. Within the context of this study, hydrologic data services published on HydroServers can be managed and indexed in a HydroPortal, while atmospheric data services published on THREDDS servers can be managed and indexed through a service mediator such as GI-cat. Because both of these interfaces allow data to be published and indexed through CSW interfaces, these individual catalogs can be aggregated within one meta-catalog to facilitate the discovery and access to interoperable data. However, it must be noted that no system can be successful without using a sound semantic ontological framework and complete metadata structure.

Although there have been significant advancements made in the sharing of scientific information, there still remain many unanswered questions. With respect to this research, it is still not clear how to deal with gridded data services in real-time. These data services provide access to copious amounts of gridded information that are continually updated. Both the high frequency of the updates and the high detailed

granularity of the datasets constitute present issues to the data consumer wanting to effectively search through the data using spatial, temporal and semantic filters. These issues may be overcome by providing services which aggregate and organize massive amounts of data streaming in real-time or by providing “granularity filters”. Furthermore, the question of how to deal with multiple ontologies has yet to be answered. Within hydrologic sciences, the CUAHSI Ontology has proven to be a successful approach to handling semantic filters; however within atmospheric sciences the CF Conventions are more widely used. Probably each discipline will prefer to use its own ontology and index the information provided by other disciplines in a consistent manner with that.

Future work will be dedicated to clients that seamlessly integrate the suite of interconnected services presented in this study. These clients will discover a set of metadata records (i.e. the existence of a dataset, what region it is in and what it contains) through a CSW interface. The expectation is that they will then utilize the WFS implementation to identify variables contained within a dataset (in addition to its spatial coverage, time extent, units etc.), and finally, the SOS and WCS implementation to directly access the dataset of interest. Moreover, with the integration of Web Processing Services (WPS), clients can have the ability to access value added information products such as indices. Some of this technology has already been demonstrated by Blodgett et al. (2011).

This work is one of the first steps towards building a sustainable cyberinfrastructure and distributed framework that meets the needs of data providers and data consumers across varying scientific disciplines. In today’s world, data is being captured at a rate far greater than ever imaginable. Data about global markets, national infrastructures, environmental systems, etc. are all being collected using sensors and computers on a network that stores, manages and distributes information. There is no doubt that the world currently is equipped with the technology to ingest oceans of data. However, there is still a struggle to use the resources we have to distribute information to individuals in a format that can be used to advance science and make prompt, smart, and progressive decisions.

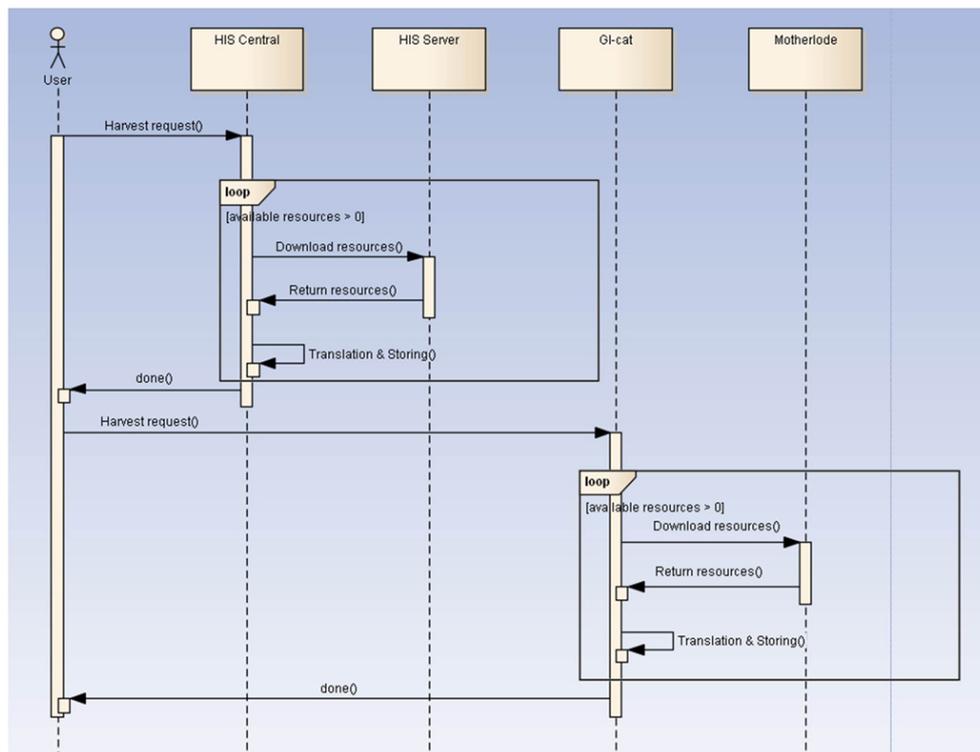


Fig. 16. Sequence diagram showing the harvesting of data services to the catalogs.

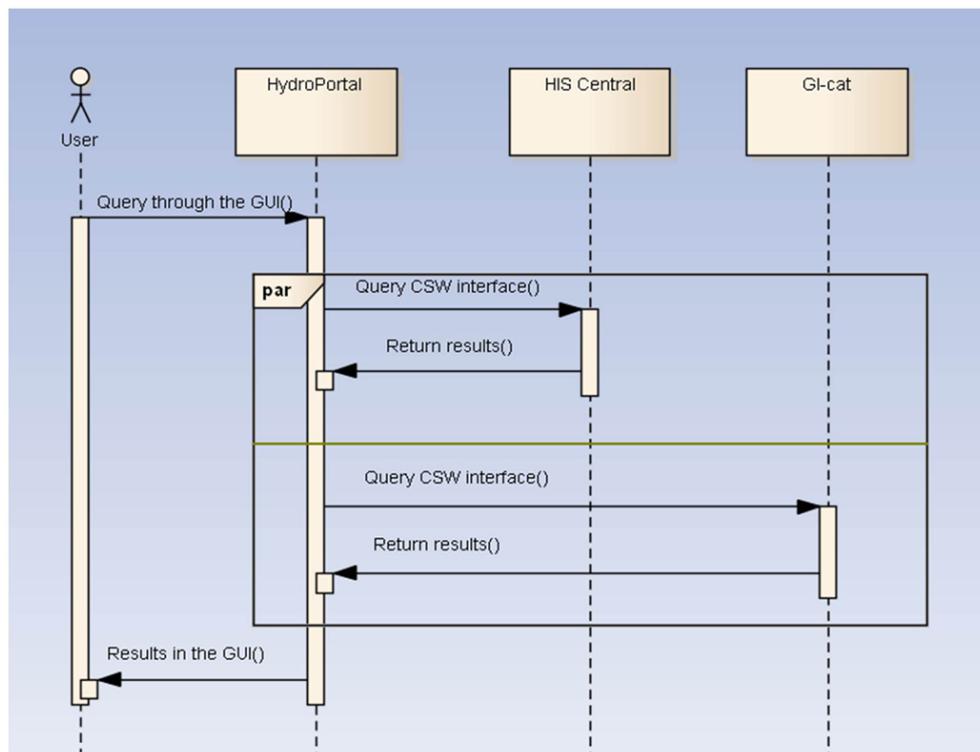


Fig. 17. Query time sequence diagram. HydroPortal, acting as a gateway, distributes the incoming query to the catalog services.

Acknowledgements. This work was supported by the National Science Foundation under grant #0622374. Any opinions, findings and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Edited by: A. Parodi

Reviewed by: two anonymous referees

References

- Alameh, N. S., Bambacus, M. J., Evans, J. D., and Marley, S. R.: NASA's Earth Science Gateway: A Platform for Interoperable Services in Support of the GEOSS Vision, Geoscience and Remote Sensing Symposium, IGARSS 2006, 2477–2480, 2006.
- Ames, D. P., Kadlec, J., and Horsburgh, J.: HydroDesktop: A Free and Open Source Platform for Hydrologic Data Discovery, Visualization, and Analysis, 2010 AWRA Spring Specialty Conference: Geographic Information Systems (GIS) and Water Resources VI, Orlando, Florida, March 2010, TPS-10-1, ISBN 1-882132-82-3, 2010.
- Bermudez, L. and Arctur, D.: OGC Engineering Report: Water Information Services Concept Development Study, OGC 11-013r6, available at: http://portal.opengeospatial.org/files/?artifact_id=44834, July 2011.
- Bermudez, L. and Piasecki, M.: Metadata Community Profiles for the Semantic Web, *Geoinformatica*, 10, 159–176, doi:10.1007/s/10707-006-7577-2, 2006.
- Bigagli, L., Nativi, S., Mazzetti, P., and Villoresi, G.: GI-cat: a Web service for dataset cataloguing based on ISO 19115, in: Proceedings of 15th International Workshop on Database and Expert Systems Applications, 846–850, August 2004.
- Blodgett, D. L., Booth, N. L., Kunicki, T. C., Walker, J. I., and Viger, R. J.: Description and Testing of the Geo Data Portal: A Data Integration Framework and Web Processing Services for Environmental Science Collaboration, US Geological Survey, Open-File Report 2011-1157, Reston, Virginia, 2011.
- CUAHSI HIS: Sharing hydrologic data, available at: <http://his.cuahsi.org>, last access: 31 March 2012.
- Domenico, B., Caron, J., Davis, E., Kambic, R., and Nativi, S.: Thematic Real-time Environmental Distributed Data Services (THREDDS): Incorporating Interactive Analysis Tools into NSDL, *Journal of Digital Information*, 2, 4, May 2002.
- Domenico, B., Nativi, S., Caron, J., Bigagli, L., and Davis, E.: GALEON: standards-based web services for interoperability among Earth sciences data systems, in: Proceedings of IGARSS'06, Denver, Colorado, July 2006.
- Doraiswamy, P., Pasteris, P., Jones, K., Motha, R., and Nejedlik, P.: Techniques for Methods of Collection, Database Management and Distribution of Agrometeorological Data, *Agr. Forest Meteorol.*, 103, 83–97, doi:10.1016/S0168-1923(00)00120-9, 1999.
- Earth and Space Science Informatics – Laboratory (ESSI-Lab): available at: <http://essi-lab.eu>, last access: 31 March 2012.
- Environmental Systems Research Institute: Geoportal Server, last access: 10 February 2012.
- Fielding, R.: Architectural styles and the design of network-based software architectures (Doctoral dissertation), University of California, Irvine, 2000.
- Fox, P.: Earth & Space Science Informatics Summit Report, Rome, March 2008.
- Garlan, D. and Shaw, M.: An introduction to software architecture, in: *Advances in Software Engineering & Knowledge Engineering*, edited by: Ambriola, V. and Tortora, G., World Scientific Pub Co., Singapore, Volume II, 1–39, 1993.
- Giuliani, G., Ray, N., Schwarzer, S., Bono, A. D., Peduzzi, P., Dao, H., Woerden, J. P., Witt, R., Beniston, M., and Lehmann, A.: Sharing Environmental Data through GEOSS, *Int. J. Appl. Geospatial Res.*, 2, 1–17, 2011.
- Horsburgh, J., Tarboton, D., Piasecki, M., Maidment, D., Zaslavsky, I., Valentine, D., and Whitenack, T.: Overcoming Semantic Heterogeneity in Publishing Environmental and Water Resources Data, *Environ. Model. Softw.*, 24, 879–888, 2009.
- Horsburgh, J., Tarboton, D., Schreuders, K., Maidment, D., Zaslavsky, I., and Valentine, D.: HydroServer: A platform for publishing space-time hydrologic datasets, in: Proceedings of 2010 AWRA Spring Specialty Conference Geographic Information Systems (GIS) and Water Resources VI, Orlando, Florida, 2010.
- INSPIRE, Infrastructure for Spatial Information in Europe: INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119, European Commission Joint Research Center, 16 June 2010.
- Maidment, D.: Arc Hydro: GIS for Water Resources, ESRI Press, Redlands, California, 2002.
- Maidment, D.: CUAHSI Hydrologic Information System: 2009 Status Report, Consortium of Universities for the Advancement of Hydrologic Science, Inc., October 2009.
- Maidment, D., Salas, F., Domenico, B., and Nativi, S.: Crossing the Digital Divide: Connecting GIS, time series and space-time arrays, AGU Fall Meeting, San Francisco, California, IN13A-1095, December 2010.
- McDonald, K., Enloe, Y., Di, L., and Holloway, D.: A Gateway to Support Interoperability of OPeNDAP and OGC Protocols, Geoscience and Remote Sensing Symposium, IGARSS 2006, 301–304, 2006.
- Meertens, C., Wier, S., Murray, D., McWhirter, J., and Memon, A.: The GEON IDV (Integrated Data Viewer) for data exploration and discovery in the geosciences. Eos Transactions AGU, Fall Meeting, IN31A-1306, 2006.
- Nativi, S. and Bigagli, L.: Discovery, Mediation, and Access Services for Earth Observation Data, *Appl. Earth Obs. Remote Sens.*, 2, 233–240, 2009.
- Nativi, S. and Domenico, B.: Enabling interoperability for Digital Earth: Coverage access services, *Int. J. Digital Earth*, 2, Supplement 1, 79–104, 2009.
- Nativi, S., Blumenthal, B., Habermann, T., Hertzmann, D., Raskin, R., Caron, J., Domenico, B., Ho, Y., and Weber, J.: Differences among the data models used by the Geographic Information Systems and Atmospheric Science communities, in: Proceedings of AMS – 20th IIPS Conference, Seattle, Washington, January 2004.
- Nativi, S., Caron, J., Davis, E., and Domenico, B.: Design and implementation of netCDF Markup Language (NcML) and its GML-based extension (NcML-GML), *Comput. Geosci. J.*, 31, 1104–1118, 2005.
- Nativi, S., Domenico, B., Caron, J., Davis, E., and Bigagli, L.: Extending THREDDS middleware to serve OGC community, *Adv. Geosci.*, 8, 57–62, doi:10.5194/adgeo-8-57-2006, 2006.

- Nativi, S., Bigagli, L., Mazzetti, P., Mattia, U., and Boldrini, E.: Discovery, query and access services for imagery, gridded and coverage data a clearinghouse solution, *Geoscience and Remote Sensing Symposium*, 2007. IGARSS 2007, IEEE Int., 4021–4024, 2007.
- Nativi, S., Bigagli, L., Mazzetti, P., Boldrini, E., and Papeschi, F.: GI-cat: A Mediation Solution for Building a Clearinghouse Catalog Service, *Advanced Geographic Information Systems & Web Services*, 2009, GEOWS '09, International Conference on, 1–7, 68–74, 2009.
- Nativi, S., Khalsa, S. J., Domenico, B., Craglia, M., Pearlman, J., Mazzetti, P., and Rew, R.: The Brokering Approach for Earth Science Cyberinfrastructure, *NSF EarthCube White Paper*, 2011.
- Nativi, S., Craglia, M., and Pearlman, J.: The Brokering Approach for Multidisciplinary Interoperability: A Position Paper, *Int. J. Spatial Data Inf. Res.*, 7, 1–15, 2012.
- Nebert, D.: *The SDI Cookbook*, GSDI, 2.0 Edn., January 2004.
- Oasis: Reference Model for Service Oriented Architecture 1.0, Oasis Standard, Oasis, 12 October 2006.
- Open Geospatial Consortium: Web Feature Service Implementation Specification, Open Geospatial Consortium, 3 May 2005.
- Open Geospatial Consortium: OpenGIS Catalogue Services Specification, Open Geospatial Consortium, 23 February 2007a.
- Open Geospatial Consortium: Sensor Observation Service, Open Geospatial Consortium, 26 October 2007b.
- Open Geospatial Consortium: Web Coverage Service (WCS) Implementation Standard, Open Geospatial Consortium, 19 March 2008.
- Open Geospatial Consortium: NetCDF Binary Encoding Extension Standard: NetCDF Classic and 64-bit Offset Format, Open Geospatial Consortium, 27 August 2010.
- Open Geospatial Consortium: OGC Network Common Data Form (NetCDF) Core Encoding Standard version 1.0, Open Geospatial Consortium, 5 April 2011.
- Open Geospatial Consortium: available at: <http://www.opengeospatial.org/ogc>, last access: 1 February 2012.
- Ortiz, S.: Getting on Board the Enterprise Service Bus, *IEEE Computer*, 40, 15–17, doi:10.1109/MC.2007.127, 2007.
- Padmanabhan, A., Wang, S., and Navarro, J.: A CyberGIS Gateway Approach to Interoperable Access to the National Science Foundation TeraGrid and the Open Science Grid, *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery*, Salt Lake City, Utah, doi:10.1145/2016741.2016786, 18 July 2011.
- Piasecki, M. and Beran, B.: A Semantic Annotation Tool for the Hydrologic Sciences, *Earth Sci. Inf.*, 2, 157–168, 2009.
- Ramamurthy, M. K.: A new generation of cyberinfrastructure and data services for earth system science education and research, *Adv. Geosci.*, 8, 69–78, doi:10.5194/adgeo-8-69-2006, 2006.
- Rui, H., Teng, W., Vollmer, B., Mocko, D. M., Beaudoin, H. K., Nigro, J., Gray, M., Maidment, D., and Hooper, R.: Bridging the gap between NASA hydrological data and the geospatial community, in: *Proceedings of ASPRS 2011 Annual Conference*, Milwaukee, Wisconsin, 2011.
- Schmidt, M. T., Hutchison, B., Lambros, P., and Phippen, R.: The Enterprise Service Bus: Making service-oriented architecture real, *IBM Systems J.*, 44, 781–797, doi:10.1147/sj.444.0781, 2005.
- Seppi, J. A.: A Services Stack Architectural Model for CUAHSI-HIS (Master's thesis), University of Texas at Austin, 2010.
- Tarboton, D., Maidment, D., Zaslavsky, I., Ames, D., Goodall, J., and Horsburgh, J.: CUAHSI Hydrologic Information System 2010 Status Report, Consortium of Universities for the Advancement of Hydrologic Science, Inc., November 2010.
- Tarboton, D., Maidment, D., Zaslavsky, I., Ames, D., Goodall, J., Hooper, R., Horsburgh, J., Valentine, D., Whiteaker, T., and Schreuders, K.: Data Interoperability in Hydrologic Sciences, The CUAHSI Hydrologic Information System, *Proceedings of the Environmental Information Management Conference 2011*, Santa Barbara, United States, 28–29 September 2011, 132–137, 2011.
- Vector 1 Media, Interoperability Science Requires a Cultural Dialog: available at: <http://www.vector1media.com/dialog/interviews/11878-interoperability-science-requires-a-cultural-dialog.html>, last access: 10 February 2012.
- Unidata: providing data services, tools and cyberinfrastructure leadership, available at: <http://www.unidata.ucar.edu/>, last access: 31 March 2012.
- Whitenack, T., Valentine, D., Zaslavsky, I., Piasecki, M., Tarboton, D., Horsburgh, J., Whiteaker, T., Ames, D., and Maidment, D.: Hydrologic metadata catalog and semantic search services in CUAHSI HIS, in: *2010 AWRA Spring Specialty Conference: Geographic Information Systems (GIS) and Water Resources VI*, American Water Resources Association, TPS-10-1, ISBN 1-882132-82-3, March 2010.
- Wiederhold, G.: Mediators in the architecture of future information systems, *IEEE Computer*, 25, 38–49, 1992.
- Zaslavsky, I., Valentine, D., and Whiteaker, T.: CUAHSI WaterML, Open Geospatial Consortium Discussion Paper OGC 07-041r1, 2007.