

# Atmospheric Sciences and Informatics EarthCube Driver Whitepaper: Technical Infrastructure

**Beth Plale<sup>1</sup>, Rich Clark<sup>2</sup>, Craig Mattocks<sup>3</sup>, Keith Brewster<sup>4</sup>, Rebecca Barthelmie<sup>5</sup>,  
Kelvin Droegemeier<sup>6</sup>, Dennis Gannon<sup>7</sup>, Sara Graves<sup>8</sup>, Scott Jensen<sup>1</sup>, William  
Mahoney<sup>9</sup>, Sara Pryor<sup>5</sup>, Rahul Ramachandran<sup>8</sup>, Robert McDonald<sup>9</sup>, Mohan  
Ramamurthy<sup>10</sup>, Bob Wilhelmson<sup>11</sup>, Ming Xue<sup>4</sup>, Sepi Yalda<sup>2</sup>**

<sup>1</sup> School of Informatics and Computing and Data To Insight Center, Indiana University

<sup>2</sup> Dept. of Earth Science, Millersville University

<sup>3</sup> Rosenstiel School of Marine & Atmospheric Science, University of Miami

<sup>4</sup> Center for Analysis and Prediction of Storms (CAPS), University of Oklahoma

<sup>5</sup> Dept. of Geography, Indiana University

<sup>6</sup> School of Meteorology, University of Oklahoma

<sup>7</sup> Microsoft Research

<sup>8</sup> Information Technology and Systems Center (ITSC), University of Alabama

<sup>9</sup> National Center for Atmospheric Research (NCAR)

<sup>9</sup> Indiana University Libraries and Data To Insight Center, Indiana University

<sup>10</sup> Unidata

<sup>11</sup> Dept of Atmospheric Science, University of Illinois and NCSA

10 October 2011

## 1. Introduction

In response to the EarthCube call for transformative concepts and approaches to create the next generation of integrated data management infrastructures across the Geosciences, the above team of atmospheric, climate, emergency management, transportation, and informatics researchers has defined principles and technologies for technical infrastructure that are described in this document. We have defined a set of science scenarios as well, and these appear in an accompanying whitepaper by the same set of authors. The authors draw in part from the NSF large ITR Linked Environment for Atmospheric Discovery (LEAD) (2003-2009)(Droegemeier et al. 2005). LEAD was one of the pioneers of the Science Gateway, a portal serving a community of researchers and educators that was one of the first to bring high performance computing resources into the hands of users in an on-demand way. The science gateway concept became so successful that in 2010, it supported 30% of all TeraGrid users<sup>1</sup>. While the LEAD cyberinfrastructure was very successful in demonstrating key advances in technology, it is not being proposed as a solution here. This whitepaper does not advocate specific technologies per se. It represents new thinking, new people, new research questions drawn from our vast experience in data driven science, largely in the atmospheric sciences, but relevant to other geosciences as well.

## 2. Technical Infrastructure and Policy

We envision EarthCube as a federation of relatively independent repositories, governed as a virtual organization with participation by all relevant agencies at the state and national level as well as universities and others. An EarthCube repository should support human and programmatic access. To keep EarthCube vibrant, its governance must impose minimal standards, encourage new repositories to grow, help these new repositories conform to the minimal standards, and constantly incorporate feedback from the community. The governance

---

<sup>1</sup> Personal correspondence, 2010.

organization should be chosen to prevent a conflict of interest position, wherein the governance entity has a vested interest in their own tools and middleware being used.

## 2.1 General Features

The National Science Board developed a three-layer typology of data collections: research data collections, community data collections, and reference data collections. Data collections are broadly defined as the infrastructure, organizations and individuals needed to provide persistent access to stored data (National Science Board 2005). Research data collections refer to the output of a single researcher or lab during the course of a specific research project. This collection may or may not use the data standards of its community or have use beyond its own original purpose. Community data collections generally serve a well-defined arena of research. Often, standards are developed by the community to support the collection. At the highest level, reference data collections are broadly scoped, widely disseminated, well-funded collections that support the research needs of many communities (NSB, 2005). Specific to the geosciences, the OOI and NEON observatories, and Unidata's IDD network represent data collections that sit somewhere between the reference data collection level and community collection, while the archives of the National Climatic Data Center, NARCCAP <http://www.narccap.ucar.edu/> and CMIP3 and 5 <http://cmip-pcmdi.llnl.gov/cmip5/> represent reference data collections. EarthCube strives to be a community level data collection that is made up of all three kinds of collections. This is the vision and the challenge.

The EarthCube will have the following key features:

- 1) We advocate organizing EarthCube as a *federation of relatively independent repositories*. Through a model of independent organizations, the cost of maintenance of EarthCube can be amortized over multiple entities including the university who already has efforts underway to provide long term access to scientific research data. This model reduces the overall ongoing maintenance cost to the National Science Foundation.
- 2) Integrating data – from multiple sources for research still remains a significant challenge and should be adaptable to different user group needs. The NSF GEO directorate is diverse, with numerous communities and subcommunities. Communities have their own vocabularies, forms and formats of data. Some data products are fully described (“curated”), however many are not. A solution must work with this diversity while still enabling various use modes on the data.
- 3) Data citation – a uniform unique identifier scheme across EarthCube is necessary for healthy participation of GEO researchers within EarthCube. Data sets are assigned a unique, immutable, non-reusable ID that can be used to cite the data object or collection in a publication. While the Digital Object Identified (DOI) has limitations, it is emerging as the dominant ID scheme.
- 4) Monitoring – EarthCube will need to have its health monitored and managed. This is done in other networks of repositories, such as the Unidata IDD network (Unidata IDD 2011), by imposing a requirement of homogeneity. As a more flexible alternative, we advocate for a single notification system such as the Advanced Message Queueing Protocol (AMQP) for monitoring and require providers of EarthCube to periodically report their health to a monitoring and reporting service. AMQP is an industry standard, and is used in the Ocean Observatories Initiative (OOI).
- 5) Security topology – we advocate a single common authorization and authentication framework, such as inCommon, and an access control framework that supports different levels of proprietary be adopted by all repository members of EarthCube. As one of us commented, “I don’t know of a single US utility that has allowed real time access to

performance data” because of lack of demonstrated security. Commercial providers are willing to contribute data for research, yet want to restrict access to a subset of the data, often the real-time data sets, whereas they are willing to allow wider access for older data. A trust framework should be in place that provides a common mechanism by which the attribution and quality of a data set can be determined. Finally, anonymization or partial views on data may be needed for survey or human subjects kinds of data. Additionally, EarthCube should support restrictive access.

- 6) Licensing – EarthCube should recommend that creators of data license their data under something like a Creative Commons license for data, Open Data Commons Attribution License. It should additionally use its strength in numbers to advocate for different licensing agreements as needed, such as might arise when a scientist uploads data to a cloud service that has licensing terms that could change ownership of the results away from the scientist.
- 7) Industry involvement – emerging tools from the technology industry can enhance capability at client and server side. Industry cloud solutions, such as Azure, offer a platform-as-a-service functionality upon which repositories, caching, and analysis services can be easily deployed (Azure 2011). Companies may be contributors of data to the EarthCube as well.
- 8) Lifetime – EarthCube requires that policies be instituted for determining which data to retain over the long term, which to discard, and which will have multiple copies for fast access. A community advisory board could be established to periodically review community assets and make determinations as to level of continued availability of certain assets based on criteria such as use over time relative to lifecycle cost.

Technical components of our suggested solution are divided into three major groups: common data services on the server side, community support services on the server side, and community support tools on the client side. These layers, populated with demonstrative services and tools, is shown in Figure 1 and described briefly here and in more detail in subsections 2.2-2.4:

- 1) Common Data Services – federation of independent repositories that includes established community data collections (e.g., OOI, NEON, NARCCAP, CMIP3 and 5) and smaller research data collections of established value. The university or facility institutional repository is emerging as a strong player in hosting scientific data, and can be an ideal solution for the smaller- to mid-sized collection.
- 2) Community support services, server side – data discovery has two partners: the repository that works to make its contents easily discoverable, and client side tools that are smart enough to know how to look for content. Content dissemination network – distributes data in real time to serve communities carrying out time-sensitive research (e.g., time-/event-triggered weather forecasting). Unidata IDD is an example of a highly successful content dissemination network.
- 3) Community support tools – community support tools on the client side are rich. In addition to schemas, vocabularies, and conversion/subsetting workflows, client side tools support visualization, collaboration, analysis, calibration and verification.

## **2.2 Common Data Services: Institutional Repository**

We discuss here a solution that could target smaller, less well-organized community holdings or could be part of a virtual repository that relies on the university or facility institutional repository. Universities have a long history of existence. Harvard, for example, was founded in 1636 and the university library has an equally long history of keeping the scholarly record open and available. Today, scientific data is emerging as an artifact in need of preservation, and the institutional

repositories at university libraries will play a key role in long-term preservation of scientific data. There are two approaches emerging in long-term data preservation. One consists of universities capturing, cataloguing and serving the scientific data assets of their own researchers. In the other, universities will emerge as strongholds for domains. That is, the task of long-term preservation of scientific data will be taken on by a small set of research universities and these universities will each carve off a piece of the scientific domain that needs preservation. Getting universities to take responsibility of the long-term preservation of scientific data decentralizes the problem of long-term sustainability and distributes long-term costs.

A solution for community holdings is needed because the NSF AGS division, for instance, does not identify a single community repository as a destination for data products emerging from their NSF funded grants, so generators of these valuable resources will turn locally for the long term preservation of their scientific data if their data are of the kind that needs preserving. A virtual organization of institutional repositories is a cornerstone of the recently funded NSF DataNet project, Sustainable Environments – Actionable Discovery (SEAD) (2011) which targets the institutional repository as a long term solution, social networking techniques to reduce data curation costs, and new tools for discovery and use of data sets. The initial scientific domain for this infrastructure is environmental sustainability.

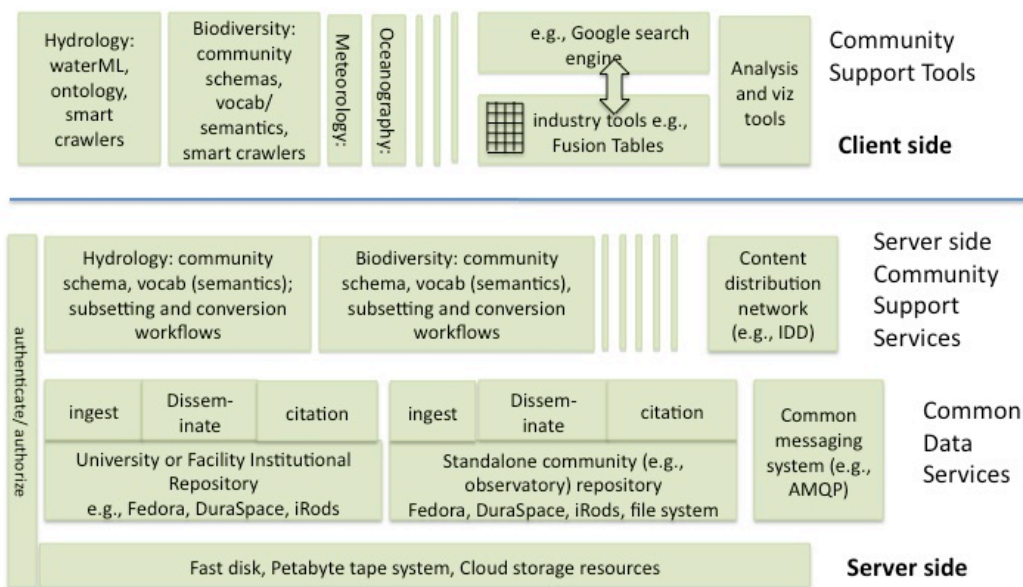


Fig. 1. This figure highlights key aspects of a technical architecture. It is not intended to be complete.

### 2.3 Implementation for Interworkability: Community Support Services on Client and Server Sides

Full interworkability is achieved when: 1) repositories are active participants in user discovery of content, 2) vocabularies and data needs of the various communities that make up the earth and atmospheric sciences are accommodated, 3) data is well curated at ingest to EarthCube, and 4) tools are smarter than they are today. Active participation can take the form of data dissemination, either through dissemination of data itself, such as IDD does, or through dissemination of metadata, which is the approach taken by Datacasting.

IDD is a well known data dissemination network from Unidata. A user sets up a LDM client at their institution and configures a filter that filters on data types. One could receive Level 3 radar

data from Doppler radars, but not Level 2 for instance. New content continually arrives at the LDM client. Datacasting (Wilson et al. 2010), on the other hand, has more flexible content selection criteria but disseminates only metadata. Datacasting uses RSS-based technology to distribute earth science content. In an example from [http://datacasting.jpl.nasa.gov/docs/read\\_more.php](http://datacasting.jpl.nasa.gov/docs/read_more.php), a user might subscribe to a Datacasting feed that contains information about files produced by an orbiting imaging satellite, but they may only be interested in data that contain wildfires in California. The user would therefore construct a filter that lists only the files that have been tagged in the Datacasting feed to contain data related to a wildfire and imaged within a bounding box. The user could further refine the filter to show only the files where the wildfire exceeds a specific size or within a certain distance of an urban area.

A repository could deliver data in the form a user wants, that is, the specific subset of data in the format and terminology the community understands (i.e., precipitation at surface for a certain catchment basin for January 2011). A repository should offer a dissemination support suite of tools for each community that has a need for its data. This suite is the set  $DIP_i = \{V/O, S, W\}$  where V/O is vocabulary/ontology of the community  $i$ , S the community XML schemas, and W is a set of workflows embodying knowledge of features and formats of a community needed to transform the data to a form familiar to the community requesting the data. Adding support for a new community to a repository then becomes the simple act of including a new instance  $DIP_j$  to the set of dissemination processes supported.<sup>2</sup> Relevant technologies include XMC Cat metadata catalog, RAMADDA content repository, OPeNDAP, CF vocabulary, and Trident scientific workflow workbench. See Tools and Services at the end of this document for URLs.

A repository will adhere to globally agreed upon EarthCube communication and discovery protocols while internally maintaining local governance and freedom over internal representation needed for scalable, sustainable and long-term growth of the EarthCube solution. Generally people put data on a webserver that are free for people to use, but in research one wants to establish a collaborative relationship with the users, and there may be restrictions on certain datasets for a time. In this case, an access protocol is needed, e.g., “These data are available to collaborators or registered users – register here or enter password or access using security certificate if registered”.

## 2.4 Community Support Tools

Community support tools on the client side are rich. In addition to schemas, vocabularies, and conversion/subsetting workflows discussed in 2.3 above, client side tools support visualization, collaboration and analysis. Existing community tools such as Unidata IDV, NCAR Graphics, NCL, etc, continue to play a valuable role. The new generation of numerical models will be based on unstructured (icosahedral, hexagonal, triangular mesh) grids; scientific visualization tools need to be upgraded to keep pace with the modeling technology. Google Fusion Tables (Gonzalez et al. 2010) allow a researcher to upload a csv file or spreadsheet into the Fusion Table service; the service will examine the contents looking for spatial and temporal fields. It will invoke a temporal time line tool for temporal data and spatial tools such as Google Earth for visual

---

<sup>2</sup> The OAIS model has a notion of the Dissemination Information Package (DIP) as the entity that is returned from the repository. Two ideas inherent in OAIS are relevant here. First, the DIP form is distinct from the Archival Information Package (AIP), or internal form in which the data object is stored. The second point about the OAIS model is that the DIP will take many forms. For example, an image might be made available as a thumbnail, as a jpg, or as a png file.

inspection spatially. There is even an extension that creates an html file from the table so the content is discoverable from a Google Search Engine.

**Smart Spider.** Data discovery is a formidable challenge in EarthCube and will remain so without focused attention and funding. Discovery must be able to peer inside the data set, whether the dataset is text or binary. We envision on the client side, for instance, a smart spider that executes on behalf of a user and crawls the Internet to discover content. A spider should be capable of discovering content that is not visible to a search engine such as Google or Bing. Issues about a spider's capabilities immediately arise: 1) What should be the architecture of a community data spider capable of crawling through, analyzing, and collecting relevant metadata in EarthCube datasets? 2) How should a scientific domain's "intelligence" (search rules/criteria, vocabulary, coherent feature extraction, pattern recognition algorithms) be represented in a customized spider to find the features/phenomena you are looking for? 3) Can these rules/algorithms be recorded in a domain-specific dictionary that other scientists can use/expand for their own spiders? 4) Will these spiders have the capability to "learn" and become more intelligent/efficient from previous searches? 5) If a researcher's spider becomes smarter than a numerical model at detecting gradients, spiral nebulae, simulated cloud/radar reflectivity patterns, 3D protein-folding topologies, etc., can the spider run while the model is running to either preemptively stop a simulation once a desired result/pattern is discovered (to avoid wasting computational resources) or activate a model's "zooming" capability (adaptive mesh refinement, dynamic grid adaption, employ a Google quad tree zooming algorithm) to better resolve/analyze the feature? 6) What language will the spider use to interact with/control the numerical model? 7) Once a spider finds the features/phenomena of interest, can it extract the relevant subset of data, generate visualizations/animations, and package them into iPhone/iPad-like apps that a researcher can scroll through and examine quickly via a touchscreen interface?

**Cloud Aggregators.** We envision data aggregators springing up in the Cloud that build off of repository push technologies such as data and service casting. These aggregators could become the stop of choice for popularly used content or themed content.

### **Sampling of Tools that Implement Needed Functionality**

Advanced Message Queueing Protocol (AMQP), <http://www.amqp.org/>

Internet Data Dissemination (IDD) <http://www.unidata.ucar.edu/software/idd/>

NetCDF – network Common Data Format, [www.unidata.ucar.edu/software/netcdf](http://www.unidata.ucar.edu/software/netcdf)

OPeNDAP, <http://opendap.org/>

RAMADDA content repository, <http://www.unidata.ucar.edu/software/ramadda/>

Trident Scientific Workflow Workbench, <http://tridentworkflow.codeplex.com/>

XMC Cat Metadata Catalog, <http://pti.iu.edu/d2i/xmccat>

Windows Azure <http://www.microsoft.com/windowsazure/>

### **References**

- K. Droegeleier, K. Brewster, M. Xue, D. Weber, D. Gannon, B. Plale, D. Reed, L. Ramakrishnan, J. Alameda, R. Wilhelmson, T. Baltzer, B. Domenico, D. Murray, A. Wilson, R. Clark, S. Yalda, S. Graves, R. Ramachandran, J. Rushing, E. Joseph 2005. Service-oriented environments for dynamically

interacting with mesoscale weather, *Computing in Science and Engineering*, IEEE Computer Society Press and American Institute of Physics, Vol. 7, No. 6, pp. 12-29.

Hector Gonzalez, Alon Y. Halevy, Christian S. Jensen Anno Langen, Jayant Madhavan, Rebecca Shapley, Warren Shen, Jonathan Goldberg-Kidony, Google Fusion Tables: Web-Centered Data Management and Collaboration, *ACM SIGMOD'10*, June 2010, Indianapolis, IN, USA.

Hedstrom, M., B. Plale, P. Kumar, G. Alter, J. Myers, NSF DataNet: Sustainable Repositories – Actionable Data, 2011

[http://nsf.gov/awardsearch/showAward.do?AwardNumber=0940824&WT.z\\_pims\\_id=503141](http://nsf.gov/awardsearch/showAward.do?AwardNumber=0940824&WT.z_pims_id=503141)

Jensen, Scott and Beth Plale, [Schema-Independent and Schema-Friendly Scientific Metadata Management](#), *4th International IEEE Conference on e-Science*, Indianapolis, IN Dec 2008.

Ocean Observatories Initiative (OOI) 2011. <http://www.oceanobservatories.org/>

Open Data Commons, Attribution License 2011. <http://opendatacommons.org/licenses/by/>

Wilson, Brian, Gerald Manipon, and Rahul Ramachandran. 2010. Lightweight Advertising and Scalable Discovery of Services, Datasets, and Events Using Feedcasts and Social Tagging. Paper read at IEEE Int'l Geoscience & Remote Sensing Symp, July 2010 at Honolulu, HI.