Department of Energy Strategic Roadmap for Earth System Science Data Integration

Dean N. Williams
Lawrence Livermore National Laboratory
7000 East Avenue
Livermore, CA 94550 USA

Giri Palanisamy
Galen Shipman
Thomas A. Boden
Oak Ridge National Laboratory
1 Bethel Valley Road
Oak Ridge, TN 37831 USA

Jimmy W. Voyles
Pacific Northwest National Laboratory
902 Battelle Blvd.
Richland, WA 99354 USA

***Abstract*- The U.S. Department of Energy (DOE) Office of Biological and Environmental Research (BER) Climate and Environmental Sciences Division (CESD) produces a diversity of data, information, software, and model codes across its research and informatics programs and facilities. This information includes raw and reduced observational and instrumentation data, model codes, model-generated results, and integrated data products. Currently, most of these data and information are prepared and shared for program specific activities, corresponding to CESD organization research. A major challenge facing BER CESD is how best to inventory, integrate, and deliver these vast and diverse resources for the purpose of accelerating Earth system science research. This paper provides a concept for a CESD Integrated Data Ecosystem and an initial roadmap for its implementation to address this integration challenge in the "Big Data" domain.**

I. Introduction

A. *Objective*

Rapid advances in experimental, sensor, and computational technologies and techniques are driving exponential growth in the volume, acquisition rate, variety, and complexity of scientific data. This wealth of data offers tremendous potential for scientific discovery. However, to achieve scientific breakthroughs, these data must be made scientifically meaningful to a diverse community of researchers, who must be able to effectively and efficiently analyze the data, and share and communicate results.

These solutions must facilitate (and where feasible, automate and record) every stage in the data lifecycle (shown in Fig. 1), from collection to management, documentation, assessment, annotation, sharing, discovery, analysis, and visualization. The mission of CESD's Data and Informatics Program is to accelerate understanding of the Earth system by integrating all existing and future distributed CESD data holdings into an environment that is unified but flexible enough to accommodate its diversity. Toward that end, a new BER Virtual Laboratory Infrastructure (VLI) [1] (see Section III, Fig. 2) will be established, which will include services and software connecting the heterogeneous CESD data holdings, and constructed with open source software based on industry standards, protocols, and state-of-the-art technology.

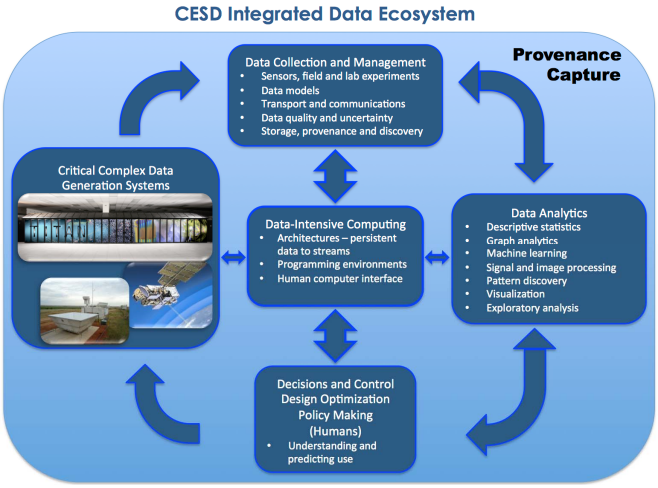

Fig. 1. The diagram depicts the proposed components of the CESD Integrated Data Ecosystem, where provenance capture is pervasive throughout.

B. *Primary goals*

The primary goal for CESD's data infrastructure and BER's VLI is to integrate CESD's diverse data holdings and to provide data and information technology resources that can optimally deliver scientific data and models.

C. *Background and motivation*

As stated in numerous articles and reports [3, 6, 9, 10], one of Earth science's most difficult challenges is managing and understanding massive amounts of global atmospheric, land, ocean, and sea-ice model data generated by complex computer simulations and driven by ever-larger qualitative and quantitative observations [2].

Many CESD-funded projects handle large and diverse data collections. The following are some of the key existing CESD data centers and portals.

C.1 Earth System Grid Federation (ESGF) [3]

This community-driven effort, heavily funded by DOE, was critical to the successful archiving, delivery, and analysis of the Coupled Model Intercomparison Project (CMIP), phase 3 (CMIP3) data for the International Panel on Climate Change (IPCC) Fourth Assessment Report (AR4). It was equally

important in meeting the data management needs of the subsequent CMIP, phase 5 (CMIP5), which produced petascale data used for the 2013 IPCC Fifth Assessment Report (AR5). Although the ESGF has been indisputably important to CMIP, its current and future impact on climate is not limited only to this high-profile project. ESGF has been used to host data for over 40 projects so far [1].

The ESGF enterprise system is a worldwide collaboration that develops, deploys, and maintains software infrastructure for the management, dissemination, and analysis of model output and related observational data. The core management capabilities of ESGF include a software stack to publish data, search services, federated security, and large-scale data transfer interconnected via international network organizations [5].

C.2 ARM Climate Research Facility [6]

Designated a national user facility in 2003, the ARM Facility provides the climate research community with strategically located in situ and remote sensing observatories designed to improve the understanding and representation in climate and Earth system models of clouds and aerosols as well as their interactions and coupling with the Earth's surface. The scale and quality of the ARM Facility's approach to climate research has resulted in ARM setting the standard for ground-based climate research observations. The ARM Data Center now provides over 4,000 data products along with data quality information. These include observational data, PI data products, and value-added products.

The ARM Data Center provides a wealth of data management tools and services, such as the Online Metadata Editor (OME), Data Discovery Tool, ARM data integration tool, data quality assessment and distribution, data monitoring tools, digital object identifiers, ARM radar data processing and visualization clusters, and interactive Web data visualization (NCVWeb) [7].

C.3 Carbon Dioxide Information Analysis Center (CDIAC) [4]

As the DOE's primary climate-change data and information analysis center, CDIAC provides scientific and data management support for projects sponsored by a number of agencies. These include: the AmeriFlux Network, providing continuous observations of CO_2, water, energy, and momentum at different time scales for sites in the Americas; the Ocean CO_2 Data Program of CO_2 measurements taken aboard ocean research vessels; DOE-supported Free-Air Carbon dioxide Enrichment (FACE) experiments, which evaluate plant and ecosystem response to elevated CO_2 concentrations; and the HIAPER Pole-to-Pole Observations (HIPPO) project.

In addition to the vast experience gained through these data center activities, over the past several years, workshops and reports have highlighted the increasing size and complexity of scientific data produced by modern science in general [9, 10], DOE facilities [11], and the Earth science community—in particular, BER . BER conducted a data workshop on June 26, 2012, with the purpose to develop specific use cases, as follows:

> *Use Case I:* A student is developing an Earth System Model of intermediate complexity for a particular application and needs to test the results of her model against available CMIP and observational data. What data and model results are available and where can she find them?
>
> *Use Case II:* A scientist is generating data sets and model outputs and needs to perform integration and analysis, potentially with access to other BER data resources in other programs. What platform and tools could be used to find these resources and perform the analytical functions?
>
> *Use Case III:* A computational scientists wants to develop a simplified workflow for candidate users. First, a user issues a scientific question on specific properties of data contained in any (or all) of the data centers. The system will then distribute partial queries to the participating individual data centers. The results are then augmented by specific domain rules input by domain experts and fused together over common attributes. The result is presented to the user as a reusable data product.

To address Use Case I, a multi-lab team developed the BER Data Gateway prototype (http://berdata.ornl.gov/cesd/index.jsp) where users can discover and access select BER-funded and related data sets. The initial prototype included data sets from ESGF, ARM and CDIAC. The gateway provides metadata publishing capabilities for BER data projects and various data search capabilities for end users. This tool also provides seamless access to visualization, sub-setting, and data-download tools presently served by the participating data centers and projects [1].

This white paper provides a high level road map for Use Case II and Use Case III.

II. Scientific Data Landscape

Earth science is an example of a discipline in which scientific progress is critically dependent on the availability of a reliable infrastructure for managing and accessing large and heterogeneous quantities of data on a global scale. Advancing Earth science is inherently a collaborative and multi-disciplinary effort that requires sophisticated modeling of the physical processes and exchange mechanisms among multiple Earth realms (atmosphere, land, ocean, and sea ice) and comparison and validation of these simulations with observational data from various sources, possibly collected over long periods of time.

For the past decades, the climate community has worked on concerted, worldwide modeling activities led by the Working Group on Coupled Modeling (WGCM) and sponsored by the World Climate Research Programme (WCRP), leading to

successive reports by the IPCC. Similarly, observational facilities such as ARM sites, AmeriFlux sites, and Earth observing satellites are continually collecting and disseminating very complex and diverse observational data to improve the scientific understanding of global climate change and promote the advancement of climate models.

III. Data Integration

The overall integrated architecture will evolve proven CESD technologies and domain knowledge that are already in use by the broader community, such as ESGF, ARM data services, and Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT) [8]. The work will progress along two main directions: (1) extend and integrate the system to support all of CESD's research activities and external collaborations, and (2) continue to support data centers and other data intensive facilities to improve their data collection, processing, archival and distribution capabilities and facilitate further data integration (see Fig. 2).

Figure 2 depicts a scalable and flexible approach for managing CESD's data archives and services across Advanced Scientific Computing Research (ASCR) and BER data centers and smaller facilities. Disseminating data, software, and computer services to the greater community, the integrated cyber-infrastructure represents an environment of interoperable services designed to integrate diverse data holdings and process extreme-scale data, heavily leveraging and advancing Office of Science resources.

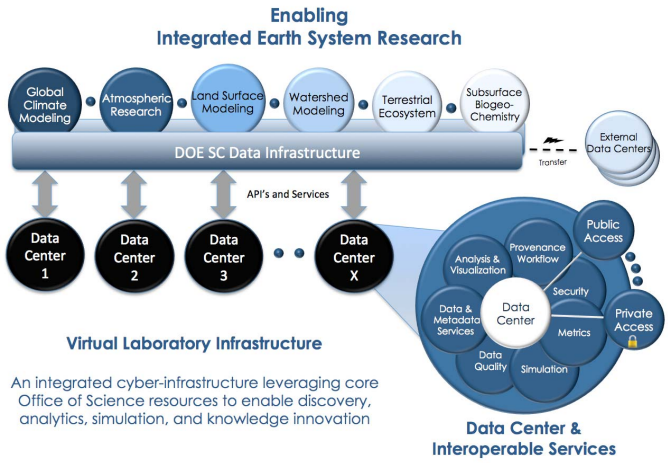

Fig. 2. CESD Data Integration Architecture

For each CESD scientific domain included in the BER VLI, both data and metadata will be archived and accessed from existing data centers. Each of these participating data centers will be part of one or more virtual scientific focus groups, thus allowing for sharing of data and metadata services with other data centers in the same scientific domain. A software stack will co-evolve to share data, metadata, data quality information, ontologies, visualization, and analysis services between the data centers. In order to support a powerful and flexible access model, each service hosted on a data center will be exposed though a simple and well-documented service Application Programming Interface (API) (layered with security when appropriate). Through this API, clients of different kinds can easily execute invocations and possibly chain requests in complex scientific workflows.

The architecture shown in Fig. 2 for the BER VLI promotes the convergence of high-level service APIs towards discipline-neutral standards. For example, the same client can be used to search and download data from "Global Climate Modeling" or "Subsurface Biogeo-Chemistry" data centers. Whenever possible, the system will use or extend existing standards developed by the community, such as the OpenSearch specification for metadata querying and the Web Processing Service API for remote job execution. The general goal is to promote reuse of modular software components on the server- and client-side across multiple fields, while retaining attention to each BER CESD science domain's specific needs and requirements.

A. Data and metadata collection capabilities

DOE data centers have unique expertise in managing their data, and these capabilities will be preserved in the new architecture and possibly used by other upcoming data intensive projects. The BER VLI proposed architecture would have an exemplary system for sharing these data and metadata records based on various community-developed standards such as International Organization for Standardization (ISO) 19115, Federal Geospatial Data Committee (FGDC), OAI-PMH, Thematic Real-time Environmental Distributed Data Services (THREDDS) and Open Geospatial Consortium (OGC). The architecture will also reuse some of the metadata creation tools such as OME, which is currently used by many DOE projects. Using OME will not only allow users to register their data sets, but also to use consistent keywords using standards such as the Climate and Forecast (CF)- and Global Change Master Directory-controlled vocabularies.

The BER VLI architecture will also have a common resource registry, which will allow projects to register their resources—such as tools, web services, and domain expertise—using common protocols and standards.

B. Data quality

Understanding data quality is a hierarchical process beginning with documenting systematic and random errors and the measurement environment. When working with scientific data across domains, this understanding is ever more important because users must interpret and synthesize information for their analysis.

Data quality services must be designed to generate a standard view of data product quality across BER data centers and projects. The data quality service must also incorporate a feedback loop from the user back to the CESD science domains to document and resolve data quality deficiencies. Similarly, data quality-related findings in synthesis and modeling activities need to channeled back to CESD science domains.

C. Uncertainty quantification

Researchers test hypotheses that are framed and bounded by uncertainty. To gain or improve knowledge in any area of research, the data product uncertainty must be documented and the associated measurement errors should be less than the anticipated uncertainties of the physical processes being analyzed. The propagation of uncertainties must be quantified and communicated to the scientific data user. Standards-based approaches currently in place or under development should be adopted to facilitate the communication of data product uncertainty of synthesis data products.

In addition, uncertainty quantification and data assimilation techniques are required to analytically model the subsystems of the end-to-end Integrated Data Ecosystem (Fig. 1) process and workflow and for predictive data infrastructure (Fig. 2) behavior and corrections.

D. Ancillary information

This data roadmap does not focus solely on model output, primary observational measurements, and derived products. Ancillary data and information are essential to understanding and characterizing the CESD data collection and permitting this collection to be used to the fullest. Ancillary data and information, which help bound data types, prevent misuse or misapplication of data; inform models; permit synthesis studies; advance science; and include site characterizations, measurement heights and depths, sampling times, model parameterizations, carbon stock estimates, calibration standards, station histories and land-use histories, statistical summaries, web camera imagery, and so on.

E. Data preparation for archival

As synthesis data products are developed through analysis and research activities, the associated data product groups will be organized, registered, and archived for sharing within the CESD virtual collaboration environment. Requirements for standard file naming conventions, versioning, and provenance will be defined and implemented [9], as will rules to verify data integrity. An extensible database structure will be implemented and governed by a complete data model for records and registration. This structure will be developed to facilitate exploration and relational associations across product holdings.

F. Data discovery and access

F.1 Collaboration and sharing policy

CESD data are freely available and are furnished by individual scientists who encourage their use or by projects committed to making data available to wider scientific audiences.

Users are expected to acknowledge the original data source as a citation in publications. Recommended data citations will be furnished to users as digital object identifiers assigned to data products and data streams. To foster collaboration, data contributors will be notified when fellow scientists retrieve their data.

F.2 Authentication and security

Data authentication and security architecture is essential to support Use Case II (mentioned in Section I), where the focus is on leveraging interoperable capabilities to examine constituent data in a regimented but free-flowing way. This will require modifications and extensions to existing Earth system software security architecture that fit into CESD's overarching strategic data roadmap. In addition, secure user identity formats and mapping for various resources, from services to high-performance computing (HPC) resources, will also need to be created.

F.3 Data analytical and visualization capabilities and services

Analysis and visualization tools should be sufficiently flexible and scalable to incorporate existing and future software components. In general, data analytical and visualization capabilities and services must incorporate the following minimal requirements: interactive and batch operations; workflow analysis and provenance management; parallel visualization and analysis tools (exploiting parallel I/O); local and remote visualization and data access; comparative visualization and statistical analyses; robust tools for re-gridding; and support for unstructured grids and non-gridded observational data.

F.4 Data downloading and sub-setting services and capabilities

CESD data centers currently use various data downloading and sub-setting services such as FTP download, Globus, Web and OGC services, and Open-source Project for Network Data Access Protocol (OPeNDAP). In addition, some have customized data sub-setting and extraction capabilities as part of the data delivery options. The new architecture will identify and support sets of data downloading protocols, which could be used for effectively sharing the data.

F.5 User interface, portal (gateway), and APIs

CESD data centers currently use data portals customized to effectively serve their respective user community. The proposed architecture recognizes the need to preserve such customized portals, but it will also build common Web services and a higher-level data discovery portal to discover and access data that are managed in distributed systems [10]. The search results will contain the matched metadata records and an option to do faceted search refinements based on the logical grouping of various themes and keywords.

The architecture will explore the possibilities of reusing the CESD Clearinghouse system as a data discovery tool. The CESD Clearinghouse currently allows users to search data from four different data centers (ESGF, ARM, CDIAC, and NGEE) and this could be expanded to support other DOE projects.

F.6 Ontology

Users attempting to answer broader scientific questions, such as factors inducing climate change, require

interdisciplinary data [10]. Proper ontology architecture is critical for discovering the data they need, confirming its usability, and integrating the data into their analysis. The proposed architecture will effectively use community-developed ontologies, such as Semantic Web for Earth and Environmental Terminology (SWEET), OBOE (extensible observation ontology), ARM, ESGF, and CF-controlled vocabularies, to annotate the keywords found in the metadata records.

G. Integrating the data services

G. 1 Data integration and advanced metadata capabilities

One important strategy to date, which has proven successful, has been tight specification of metadata standards. However, to accommodate a broader variety of environmental and scientific data, we will require a more extensible metadata infrastructure, more powerful processes for handling diverse data formats, and scalability up to billions of objects. We must develop mechanisms to describe and organize a wide variety of data not fully supported in any one community. In climate research, flexible data-format support will include mechanisms that allow users to work with complex data sets (e.g., Ameriflux, NGEE, ARM, CMIP), high-resolution global models, regional models, and observational data sets. The commonly used metadata standards include CF, ISO 19115, FGDC, and Ecological Metadata Language (EML).

G.2 Performance of model execution

A significant challenge facing the climate science community is the extremely large data sets that are produced today from coupled model simulations and the projected increase in data set sizes as higher resolution models (T341) are commonly used. Scientists will require scalable tools and technologies for analysis of multi-terabyte data sets. These tools must be easy to use while scaling to HPC environments in which parallel analysis will be required due to processing time requirements and per-node memory capacity.

Emerging services can couple the capabilities of HPC environments with web-based service delivery mechanisms. These multi-tiered "applications" provide users with access to high performance parallel analysis routines hosted on HPC platforms using common technologies such as MPI and OpenMP, coupled with a web-services framework using standard RESTful interfaces.

G.3 Analysis services when multiple data sets are not co-located

To enable truly integrative research, many analysis tasks will need to fuse data sets from multiple locations. In some cases, this analysis can be accomplished by running the constituent parts of the analysis on computational infrastructure co-resident with the data. An example of this type of analysis would be calculating an average temperature over a given set of months at a specific grid cell across multiple climate model outputs hosted at different data centers. This task can be distributed across the computational infrastructure at the data centers and then aggregated back at the original computational resource. This example requires very little data movement, and integration of the resultant analysis can easily be accomplished on a typical workstation.

G.4 Advanced product services (i.e., exploratory, specialized, etc.)

The product services provide users with custom visualization, sub-setting, and basic analysis and exploratory capabilities applied to the underlying data collection via a browser-based interface or via command-line. These services are essential to serving a diverse user community. It is essential to keep in mind scientists not accustomed to working with complex output domains and who need to quickly discern which data are suitable for their needs. In the BER VLI, non-HPC experts will have ease of access to pre-built analysis products based on both smaller-scale as well as extreme-scale data sets.

G.5 Advanced networks as easy-to-use community resources

A primary goal of the BER VLI will be to provide CESD scientists with tools to manage and analyze extreme-scale Earth system data using the first-ever Energy Sciences Network (ESnet) 100-Gbps backbone [11] and large- and mid-range LCF and NERSC computing resources. DOE, and indeed the world community, is making significant investments in hardware, network, and software services and resources; however, researchers and non-researchers do not yet know how best to work with them. A goal should be to ensure that all services and resources, including documentation, are easy to access and use by all.

H. Provenance and workflow

Throughout the BER VLI infrastructure containing model runs, diagnostics, production cycles, and rich research, provenance captures will enable unprecedented levels of reproducibility and collaboration. Remote services, including analysis and visualization, will be connected through a provenance API to automatically capture meaningful history and workflow.

I. Compute and data services

The integrated cyber-infrastructure (Fig. 2) will build upon core computing and data services supported by the DOE Office of Science (SC). These facilities include large-scale computing and data storage resources available at NERSC and the LCFs at Argonne National and Oak Ridge National Laboratories, advanced networking infrastructure from ESnet, CESD data lifecycle management capabilities at ARM, CDIAC, NGEE-Arctic and CMIP, and emerging data services in the DOE SC.

As DOE SC continues to expand the set of available computing and data services, the BER VLI will incorporate and build upon them to deliver end solutions to CESD. It is envisioned that as DOE SC increasingly moves towards providing robust Infrastructure-as-a-Service (IaaS), the proposed software development and integration will target these technologies, allowing broad adoption of the innovative

software. Similarly, should DOE SC begin to deploy Software-as-a-Service (SaaS) offerings such as Map-Reduce or NoSQL storage, it will rapidly expose these technologies to other science domains.

J. Dashboard and system monitoring

J.1 User and data product usage metrics and reporting

The SC national user facilities and associated research components have processes and rules for registering scientific users and recording their requests. These processes are designed to create a cooperative and consistent view of user identity, information, research purpose, data product requests, and demographics. On a quarterly basis, the SC national user facilities track and report "unique scientific users" by number and affiliation, and the data product usage patterns are available for analysis. The virtual collaboration environment for BER data centers will be designed to incorporate SC user and data product usage rules and processes, and to provide feedback about user and data product usage statistics to data centers served through the virtual environment.

J.3 Network monitoring (tracking network speeds and usability)

The emergence of globally distributed, large-scale science demands a network fast enough to meet the needs of the new, highly distributed data model. With this comes the need to monitor the network speed and usability. With the help of ESnet and other network resources and expertise, the movement of data will be tracked for performance and for greater acceleration of scientific knowledge. Achieving this capability on production systems will help to prepare the CESD infrastructure for the demands of future large-scale data activities such as CMIP6, ARM, etc. and set the stage for continued scientific productivity in other critically important areas of CESD Earth system science.

IV. System Evolution and Operational Requirements

The BER VLI integrated data infrastructure will never be a static system. As the platforms on which it operates—server hardware, networks, operating systems, and browsers—evolve, the data centers and interoperable services must be adapted. The infrastructure will also be a collaborative system with components from several quasi-independent projects. As one component advances or modernizes, adaptations in others are inevitable, despite best attempts to isolate functionality through interface definitions. Therefore, continued operational and maintenance support will be necessary for the DOE BER CESD data infrastructure.

V. Summary

To improve research ability and productivity, an integrated data infrastructure must be in place to help make vital and quick strategic decisions reflecting the future of Earth's climate and energy. To address these challenges, the authors envision the establishment of a BER VLI (Fig. 2) to help integrate disparate community software tools for the discovery, examination, and intercomparison of coupled multi-model and observational climate data sets. The BER VLI will bring together top climate institutions, computational organizations, and other science communities to provide proven data management, analysis, visualization, diagnostics, network, and hardware capabilities to CESD scientists.

REFERENCES

[1] Janet Braam, Judith A. Curry, et al., BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges. Office of Biological and Environmental Research, Office of Science, Department of Energy, http://genomicscience.energy.gov/program/beracvirtuallab.shtml.

[2] Overpeck, J.T., G. A. Meehl, S. Bony, and D. R. Easterling, 2011: Climate Data Challenges in the 21st Century. Science, vol. 331, no. 6018, pp. 700-702, dx.doi.org/10.1126/science.1197869.

[3] The Earth System Grid Federation home page. http://esgf.org/.

[4]Carbon Dioxide Information Analysis Center home page. http://cdiac.esd.ornl.gov/.

[5] Luca Cinquini, Daniel Crichton, Chris Mattmann, Gavin M. Bell, Bob Drach, Dean Williams, John Harney, Galen Shipman, Feiyi Wang, Philip Kershaw, Stephen Pascoe, Rachana Ananthakrishnan, Neill Miller, Estanislao Gonzalez, Sebastian Denvil, Mark Morgan, Sandro Fiore, Zed Pobre, Roland Schweitzer, "The Earth System Grid Federation: An Open Infrastructure for Access to Distributed Geospatial Data", IEEE Future Generation Computer Systems, http://dx.doi.org/10.1016/j.future.2013.07.002, 17 September. 2013.

[6] ARM Climate Research Facility home page. http://www.arm.gov/.

[7] Interactive Web-based tool for viewing Atmospheric Radiation Measurement (ARM) data website [Accessed April 2014]; Available from: https://ams.confex.com/ams/annual2003/techprogram/paper_55288.htm.

[8] Dean N. Williams, Timo Bremer, Charles Doutriaux, John Patchett, Sean Williams, Galen Shipman, Ross Miller, David R. Pugmire, Brian Smith, Chad Steed, E. Wes Bethel, Hank Childs, Harinarayan Krishnan, Prabhat, Michael Wehner, Claudio T. Silva, Emanuele Santos, David Koop, Tommy Ellqvist, Jorge Poco, Berk Geveci, Aashish Chaudhary, Andy Bauer, Alexander Pletzer, David Kindig, Gerald L. Potter, Thomas P. Maxwell, "Ultrascale Visualization of Climate Data", IEEE Computer Magazine, September 2013, vol. 46 no 9, pp. 68-76.

[9] The Federal Geographic Data Committee Geospatial Metadata Standards home page. http://www.fgdc.gov/metadata/geospatial-metadata-standards.

[10] Pouchard, Line C., Marcia L. Branstetter, Robert B. Cook, Ranjeet Devarakonda, Jim Green, Giri Palanisamy, Paul Alexander, and Natalya F. Noy. "A Linked Science Investigation: Enhancing Climate Change Data Discovery with Semantic Technologies." Earth Science Informatics 6, no. 3 (September 1, 2013): 175–85. http://dx.doi.org/10.1007/s12145-013-0118-2.

[11] Eli Dart, Brian Tierney, Editors, "Biological and Environmental Research Network Requirements Workshop, November 2012 - Final Report", November29,2012, LBNL LBNL-6395E http://www.es.net/assets/pubs_presos/BER-Net-Req-Review-2012-Final-Report.pdf.