ADVANCES IN SPATIAL DATA INFRASTRUCTURE, ACQUISITION, ANALYSIS, ARCHIVING & DISSEMINATION

Hampapuram K. Ramapriyan[1], *Gilbert L. Rochon*[2],
Ruth Duerr[3], *Robert Rank*[4], *Stefano Nativi*[5], *Erich Franz Stocker*[1]
[1]NASA Goddard Space Flight Center, USA; [2]Purdue University-Purdue Terrestrial Observatory, USA;
[3]University of Colorado-National Snow & Ice Data Center, USA; [4]NOAA NESDIS, USA; [5]National
Research Council & University of Florence, Italy

ABSTRACT

The authors review recent contributions to the state-of-the-science and benign proliferation of satellite remote sensing, spatial data infrastructure, near-real-time data acquisition, analysis on high performance computing platforms, sapient archiving, multi-modal dissemination and utilization for a wide array of scientific applications. The authors also address advances in Geoinformatics and its growing ubiquity, as evidenced by its inclusion as a focus area within the American Geophysical Union (AGU), European Geosciences Union (EGU), as well as by the evolution of the IEEE Geoscience and Remote Sensing Society's (GRSS) Data Archiving and Distribution Technical Committee (DAD TC).

Index Terms— Digital Data Archives, Geoinformatics, Data Distribution, Data Access

1. INTRODUCTION

Remotely sensed data streams and data sets derived from them often push available transmission, processing and storage technology to their limits, and thus require special techniques in their handling, distribution, application, rendering, fusing, mining, and compression. The DAD TC (originally called the Data Standardization and Distribution TC when it was established in 1994) considers all of these aspects of dealing with remotely sensed data. The DAD TC's charter, defined in 2001, is: "To provide recommendations and responses to issues related to the archival and distribution of remotely sensed geospatial and geotemporal data, and on how new media, transmission means, and networks will impact the archival, distribution, and format of remotely sensed data. Also, to study the impact of media, channel, and network scaling on the archival and distribution of data." A special responsibility of the DAD TC is to function as a liaison between the IEEE GRS-S and the International Standards Organization (ISO) on standards for geographic information (ISO TC211). The DAD TC serves as a clearinghouse for coordinating any GRS-S members' comments on ISO TC211 proposed standards. The DAD TC develops an agenda for research in data archiving and distribution via inputs from its members. Of course, the research agenda must evolve over time as technology advances and users' needs change. Through inputs from the DAD TC members, this paper provides a broadened perspective on the salient issues confronting scientists specializing in the analysis, manipulation, storage and applications of remote sensing data. In addition to data standardization, such issues cover the entire data life cycle: data architectures, data acquisition, validation, quality assessment, citation, tagging, heterogeneity, encoding, compression, security, archiving, search and access, distribution, evaluation, readability, integrity, availability, usability, identity, dynamics, visualization, analysis, algorithm development, provenance, modeling and end-user services. Clearly, this area of research is quite broad and growing, and it is beyond the scope of this paper to describe it fully. This paper focuses on some of the highest priority topics relevant to the DAD TC's research agenda. The topics are categorized under four major groups: 1. Archiving and Preservation, 2. System Interoperability, 3. Access Mechanisms, and 4. On-Line Services and Enabling Analysis. The next section briefly discusses the increasing number of publications in areas relevant to DAD TC in recent years. Section 3 considers research questions in each of the four major groups indicated above. This is followed by a concluding section.

2. RECENT DEVELOPMENTS

There is considerable evidence to show that, in support of Geoscience and Remote Sensing, the area of data and information systems in general and data archiving and distribution in particular has matured over the last decade. Indeed, several distributed systems for archiving and distributing data and derived products to a diverse community of users exist and continue to evolve to meet users' needs. In the U.S., NASA has developed and has been operating and evolving the Earth Observing System Data and Information System (EOSDIS) since 1994, with a growing multi-petabyte archive and distribution of about 7 terabytes (average of about 600,000 files) per day to users in several earth science disciplines [1]. Most of the data holdings in the 12 science data centers of EOSDIS are held on-line and available to users at no charge. Also, in the U.S., NOAA has been evolving its National Data Centers, and has developed the Comprehensive Large Array-data Stewardship System

(CLASS). CLASS is NOAA's enterprise-wide information system designed to support long-term, preservation and standards-based access to environmental data collections and information. The NOAA National Data Centers owned and operated system supports ingest, quality control, archival storage of and public access to data and science information. Together, the NOAA National Data Centers, utilizing the IT infrastructure of CLASS, will provide the necessary ingredients to fulfill NOAA's data stewardship mission. To ensure the preservation of these data, the distributed system replicates data and metadata holdings automatically to operational instances at the National Climatic Data Center in Asheville, NC, and the National Geophysical Data Center in Boulder, CO [2].

In Europe, there are several capabilities at the European Space Agency (ESA), the Centre National d'Etudes Spatiales (CNES), Deutsches Zentrum für Luft- und Raumfahrt (DLR), and other organizations performing data archiving and distribution and advancing the state of the art. The European initiative GMES (Global Monitoring for Environment and Security) aims to establish a European capacity for Earth Observation. This European Earth Observation program provides data useful in a range of issues including climate change and citizen's security. Each Earth component (i.e. land, sea and atmosphere) is observed through GMES, delivering information which corresponds to user needs. The processing and dissemination of this information is carried out within the "GMES service component" [3]. The thematic areas within the GMES service component comprise:

- land, marine and atmosphere information;
- climate change information;
- emergency and security information.

Also in Europe, a major recent development has been the entering in force of the INSPIRE Directive in May 2007 [4], establishing an infrastructure for spatial information in Europe to support Community environmental policies, and policies or activities which may have an impact on the environment. INSPIRE is based on the infrastructures for spatial information established and operated by the 27 Member States of the European Union. The Directive addresses 34 spatial data themes needed for environmental applications, with key components specified through technical implementing rules. This makes INSPIRE a unique example of a legislative "regional" approach. INSPIRE covers a wide range of themes representing the broad needs for fulfilling expected actions for sustainable development and the multipurpose needs for e-Government actions [5].

Recently there have been significant interagency and international efforts to increase interoperability among systems and promote broader exchange of data among organizations. The Global Earth Observation System of Systems (GEOSS) is the broadest example of this effort. GEOSS will be a "system of systems" consisting of existing and future Earth observation systems, supplementing but not supplanting their own mandates and governance arrangements [6]. GEOSS, collectively, has several functional components:

- To address identified common user requirements;
- To acquire observational data;
- To process data into useful products;
- To exchange, disseminate, and archive shared data, metadata, and products; and,
- To monitor performance against the defined requirements and intended benefits.

The implementation of GEOSS will facilitate the development and availability of shared data, metadata, and products commonly required across diverse domains (i.e. societal benefit areas). GEOSS will encourage the adoption of existing and new standards to support broader data and information usability. The success of GEOSS will depend on data and information providers accepting and implementing a set of interoperability arrangements, including technical specifications for collecting, processing, storing, and disseminating shared data, metadata, and products [6].

Each year since 2005, the American Geophysical Union's fall meetings have included several sessions in Earth and Space Science Informatics (ESSI). Many of these sessions address topics of interest to the data archiving and distribution community. In 2008, the European Geosciences Union (EGU) created a sister ESSI scientific division; it also organized several sessions on data management and distribution, at the EGU annual General Assembly meetings [7]. Sample session titles from these meetings are listed below as they apply to the major groups identified at the end of Section 1. Some of the sessions cover more than one group, but are included in the group where they best fit.

Archiving and Preservation: Data Stewardship in the 21st Century; Data Preservation and Long Term Access; Data Sources and Management for the 2007-2009 International Polar Year; Data, Metadata and Mark-up Languages; Emerging Issues in Science: Collaboration, Provenance, and the Ethics of Data; Making Earth Science data Records; EU - Africa Cyberinfrastructures.

System Interoperability: Standards-Based Interoperability among Tools and Data Services in Earth Science; Real Use of Standards and Technologies; Fostering Multidisciplinary Research via Interoperable Data Systems Based on Geospatial Standards for Earth and Space Sciences; Ontologies for Earth and Space Sciences; Community Approaches to Data Sharing; Service-Oriented Architecture solutions for Earth and Space Sciences; Earth and Space Science Cyberinfrastructure: Application and Theory of Knowledge Representation; Cyberinfrastructure to Support Large-Scale Long-Term Observation Experiments.

Access Mechanisms: Transforming Geoscience Access through Standards Implementation; Standardizing Fine-Grained Access to Geoscience Data; Challenges in Achieving Earth System Model Interoperability; Web-Based Service Oriented Earth Science; Grid Technologies and Associated Infrastructures; Unifying Discovery, Access and Knowledge Extraction from Space and Geoscience Virtual Data Repositories.

On-Line Services and Enabling Analysis: Geoscience Applications in Virtual Globes; Visualization of Large Datasets; Visualization and Portrayal Services and Tools; Intelligent and Adaptive Systems for Data Collection, Processing and Knowledge Discovery; Geosciences Applications on Grid and HPC; Collaboration Technologies, Social Networking and Web 2.0.

Similarly, each year since 2004, the IEEE International Geoscience and Remote Sensing Symposia (IGARSS), have had special sessions dealing with data system issues such as archiving, distribution, access, special services, web and grid services, and measurement-based systems as they specifically apply to geoscience and remote sensing. For the first time in history, in January 2009, the IEEE GRSS published a Special Issue on Data Archiving and Distribution of the Transactions on Geoscience and Remote Sensing [8].

In addition, NASA's Technology Infusion Working Group, one of four Earth Science Data System Working Groups, has articulated the following new information system capabilities as important for data archiving, distribution, access and increased utilization of remotely sensed data [9]: Evolvable Technical Infrastructure, Interactive Data Analysis, Seamless Data Access, Interoperable Information Services, Responsible Information Delivery, Verifiable Information Quality, Scalable Analysis Portals, Community Modeling Frameworks, Assisted Data and Service Discovery, Assisted Knowledge Building, and Collaboration Environments.

3. DAD TC RESEARCH AGENDA

Given the context of progress and interest in archiving, distribution, access, interoperability, visualization and other aspects of data and information systems, the DAD TC community has identified a few key questions to be addressed by research and development activities in the near future. These are identified below in four subsections corresponding to the major groups given at the end of section 1:

3.1. Archiving and Preservation

Data Readability and Integrity: How long can current data formats be expected to survive, and will they be readable after 2 or 3 updated versions of the format have been released or after other formats have become more popular? How can we insure ensure that critical data survive this process of technological evolution with integrity?

Data Availability: How can we insure ensure that data remain accessible for reasonable periods of time, irrespective of what happens to the site archiving them? How long a time period should be considered minimal for public access?

Data Identity: How do we know that two files contain the same data even if the formats are different? That is, how do we ensure that two data sets are "scientifically identical"? How do we find the data used in a particular publication? How can we uniquely and unambiguously identify a particular piece of data no matter which copy a user has? How can we provide online citation technology in a consistent and interoperable way?

Provenance: How do we define the appropriate levels of provenance information and ensure that they are included along with data during production and in the archive?

Data Encoding and Compression: How can we encode data in an interoperable, flexible, scalable, efficient way that preserves the likelihood that the data will be understandable decades into the future? This includes data compression issues for network (Web) exchange.

Validation of Data Properties: What are the appropriate methods and frequencies with which data object properties should be validated in an archiving system that is subject to hardware and software failures, operational errors, natural disasters, or malicious attacks?

Transparent Technology Refreshment: What techniques should be used to ensure "transparent technology refreshment", i.e., upgrading to new generations of hardware and software while maintaining high levels of operational availability, addressing the dynamic and evolving archive environment, and maximizing the application of limited resources?

3.2. System Interoperability

Data Discovery: How can we provide online discovery for disparate (i.e. heterogeneous and distributed) datasets?

Hardware Technology Trends: What are the continuing trends of technology evolution and cost (a la Moore's law) in processing, storage and network bandwidth? What are their implications on overall end-to-end systems' architecture?

Standardization: Are current standards adequate or are new standards needed to eliminate or reduce impacts of heterogeneity? Standardization is essential for interoperability and information heterogeneity management. It also facilitates evolvability and helps reduce costs. Standardization efforts apply to:

- people, primarily in the form of terminology standards
- information, primarily in the form of structural and semantic representation standards
- systems, primarily in the form of interface and communication standards.

Conceptual Composability (System of Systems): How do we introduce the necessary "interoperability arrangements" necessary to implement complex System of Systems collecting task-oriented, autonomous systems that pool their resources together to obtain more complex, 'meta-system' (e.g. GEOSS) ?

3.3. Access

Security: How do we strike a balance between open access to data and the need to protect data from malicious or inadvertent corruption? How do service providers protect their systems from "denial of service" attacks and other improper uses of the data and services?

Standards: What standards should be developed or adopted to facilitate access to data? Which basic processing functionalities should be included (e.g. domain/co-domain subsetting, transformations, etc.)?

3.4. On-Line Services and Enabling Analysis

Data Visualization and Analysis: How can we provide on-line visualization and analysis tools that can assist users in identifying meaningful data subsets within large sets?

Data, Algorithms, and Services: How do we associate data with the services that act on them and with the algorithms that create them? How do we make distributed data and associated services discoverable without requiring users to learn multiple search tools? In order to support the data and information needs of the application communities is it possible to determine what products have the most socio/economic value and what algorithms are needed in order to produce them? Can such lists be updated dynamically as sensors and applications continue to evolve? For example, how best can we make the user community aware that digital elevation maps (DEMs) can be produced accurately from Synthetic Aperture Radar (SAR) interferometry, or that sea ice surface temperatures are now available from infra-red (IR) channels, or that a new vegetation index has been produced from Moderate Resolution Imaging Spectroradiometer (MODIS) data?

Data Evaluation: How do we evaluate datasets, including their quality, content, and constraints? Data quality issues are especially important in the present Web era where global viewers help inexpert users fuse and visualize heterogeneous and distributed datasets, potentially in scientifically erroneous ways due to a substantial lack of information about data uncertainty and error propagation.

Standards: How will uncertainty and error propagation description and management affect existing standards?

4. CONCLUSION

Over the last decade, data and information systems in general and data archiving and distribution in particular have matured significantly. Evidence of this is the growth of conference and journal publications, formation of interagency and international organizations and efforts to promote increased exchange of data, information and services among them. In this paper, we have classified topics of high interest to the DAD TC into four major groups of: 1. Archiving and Preservation, 2. System Interoperability, 3. Access Mechanisms, and 4. On-Line Services and Enabling Analysis. We have identified several key research questions in each of these groups. It is hoped that these questions will catalyze further advancement of the state-of-the-art in data archiving and distribution in support of geoscience and remote sensing.

5. REFERENCES

[1] H. K. Ramapriyan, J. Behnke, E. Sofinowski, D. Lowe, M. A. Esfandiari, "Evolution of the Earth Observing System (EOS) Data and Information System (EOSDIS)", *Standard-Based Data and Information Systems for Earth Observation,* pp. 63-92, Springer-Veralg, Berlin, 2010.

[2] R. H. Rank, "Enterprise IT support for NOAA archives", *Geoscience and Remote Sensing Symposium, 2007*. IEEE International, pp. 4025 – 4028.

[3] Commission of the European Communities, "Global Monitoring for Environment and Security (GMES): Challenges and Next Steps for the Space Component", *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions,* Brussels, October 19, 2009, COM(2009) 589 final. http://ec.europa.eu/gmes/pdf/communication_589_en.pdf

[4] European Commission, *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*, http://inspire.jrc.ec.europa.eu/, March 2007.

[5] INSPIRE Drafting Team, *Data Specifications*, "Definition of Annex Themes and Scope", INSPIRE Deliverable D2.3, http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/DataSpecifications/D2.3_Definition_of_Annex_Themes_and_scope_v3.0.pdf, March 2008.

[6] GEO, *The Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan*, http://www.earthobservations.org/documents/10-Year%20Implementation%20Plan.pdf, 16 February 2005.

[7] S. Nativi, P. Fox, "Advocating for the use of informatics in the earth and Space Sciences", *EOS, Transactions, American Geophysical Union*, Vol. 91, No 8, pp. 75 – 76, 2010.

[8] L. Di, H. K. Ramapriyan and L. Bruzzone, "Guest Editorial of the Special Issue on Data Archiving and Distribution", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 47, No. 1, Part 1pp. 17-18. , 2009.

[9] Earth Science Data Systems Working Group on Technology Infusion, "NASA Earth Science Information Systems Capability Vision", 2009