"Advances in Spatial Data Infrastructure, Acquisition, Analysis, Archiving & Dissemination"

Gilbert L. Rochon[1], Hampapuram Ramapriyan[2], Erich Franz Stocker[2], Ruth Duerr[3], Robert Rank[4], Stefano Nativi[5]

[1]Purdue University-Purdue Terrestrial Observatory, USA; [2]NASA Goddard Space Flight Center, USA; [3]University of Colorado-National Snow & Ice Data Center, USA; [4]NOAA NESDIS, USA; [5]University of Florence, Italy

CONTACTS: rochon@purdue.edu; hampapuram.k.ramapriyan@nasa.gov

Overview Presentation:
IEEE IGARSS- July 25-30, 2010
Data Archiving and Distribution Technical Committee (DAD TC) Invited Session:
Data System Technologies for Improving Data Access and Usability - Challenges and Solutions

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), as well as by the evolution of the IEEE GRSS Data Archiving and Distribution Technical Committee (DAD TC).

Remotely sensed data streams and data sets derived from them often push available transmission, processing and storage technology to extreme 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 spatial 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 (short term as well as long-term preservation), search and access, distribution, evaluation, readability, integrity, availability, usability, identity, dynamics, visualization, analysis, algorithm development, provenance, modeling and end-user services.

The present topics of interest are summarized below with a few of the many questions that need to be addressed:

- *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 that critical data survive this process of technological evolution with integrity?
- *Data Availability*: How can we insure 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?
- *Data Discovery*: How can we provide online discovery functionalities 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?
- *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? 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 DEM can be produced accurately from SAR interferometry, or that sea ice surface temperatures are now available from IR channels, or that a new vegetation index has been produced from MODIS data?.
- *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.
- *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?
- *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.
- *Provenance*: How do we define the appropriate levels of provenance information and ensure that they are included along with data while they are being generated and archived?
- *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?
- 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

The nuances associated with each of the issues confronting the spatial data scientific community are formidable, especially given the increased demand for near-real-time access to high spatial resolution, high spectral resolution and high temporal resolution data from earth observing satellites, launched under the auspices of African, Asian, European, and Latin American countries, Australia, Canada, Russia and the United States. Moreover, the complexity of interdisciplinary ecological applications frequently require that satellite data from multiple sensors be parallelized, trends extrapolated, mosaiced, merged, fused, pan sharpened and/or integrated with *in situ* sensor data (e.g. NEXRAD Doppler Radar, hydrometers, seismic sensors), as well as with socio-economic or epidemiological data.

The probable future of informed decision-making within data-intensive and high performance computing environments is putatively depicted, wherein hindcasting, forecasting and visualization of alternative future scenarios emerges as the norm, rather than as the exception. The ultimate challenge for the spatial data scientific community will be the extent to which such formidable technologies and abundance of data can measurably contribute to effectively addressing the major problems facing the terrestrial biosphere: i.e. alleviating poverty, disease, hunger, biodiversity depletion, deforestation, ecosystem destruction and mitigating the impact of biogenic and anthropogenic disasters, of armed conflict, of climate change, and of industrial, urban and agricultural pollution.