Data Science Journal (submitted 1-5-05)

eScience and Archiving for Space Science

Timothy E. Eastman[1], James L. Green[2], Edwin J. Grayzeck[3], Robert McGuire[4], and Donald M. Sawyer[3]

1 QSS Group, Inc., Space Physics Data Facility, NASA Goddard Space Flight Center (GSFC), Greenbelt, MD 20771 USA
2 Science Proposal Support Office, NASA/GSFC
3 National Space Science Data Center, NASA/GSFC
4 Space Physics Data Facility, NASA/GSFC
email: eastman@mail630.gsfc.nasa.gov
{James.Green,Edwin.J.Grayzeck,Robert.E.McGuire,Donald.M.Sawyer}@nasa.gov

Abstract

Our scientific meetings, IT publications, and even the media are awash with new terminology and possibilities about how "a new age has dawned in scientific and engineering research" made possible through distributed science collaborations enabled by the internet, viz., eScience, the grid, cyberinfrastructure, and virtual observatories. All of these new structures and tools depend critically on a solid and standards-based data archiving foundation. NASA's permanent archive for space science, NSSDC, and NASA's many active archives constitute an archive-grounded cyberinfrastructure. The ISO Open Archival Information Systems (OAIS) standard now guides the evolution of NSSDC. Protocols for the relationship of our permanent archive, active archives, PI sites, and scientific users are being enhanced. Working with the Consultative Committee for Space Data Systems (CCSDS), CODATA, and other organizations, new international standards are evolving for the producer-archive relationship and other key archive system needs. Initial steps towards supporting the semantic web of the future are being taken through new metadata systems, XML technologies, and interoperability tools. A new synergism of rich data sets, state-of-the-art modeling, high-performance computing, and high-speed sensor systems demonstrates how data and data systems are central to science. This Data-Model-HPC-Sensor synergism and linkages among programs supporting enhanced interoperability and systems architecture illustrate the importance of proactive data archiving efforts in the overall data environment of this new age.

Keywords: data archives, data warehousing, active archives, permanent archives, archiving, data standards, interoperability, grid systems, grid computing, metadata, metadata systems, eScience, cyberinfrastructure, virtual observatories, OAIS, NSSDC, CCSDS, ISO, CODATA.

1 NASA Science Data Archives

Space physics and Earth science data systems at NASA's Goddard Space Flight Center are designed to handle data from a wide variety of science missions as shown in Figure 1.

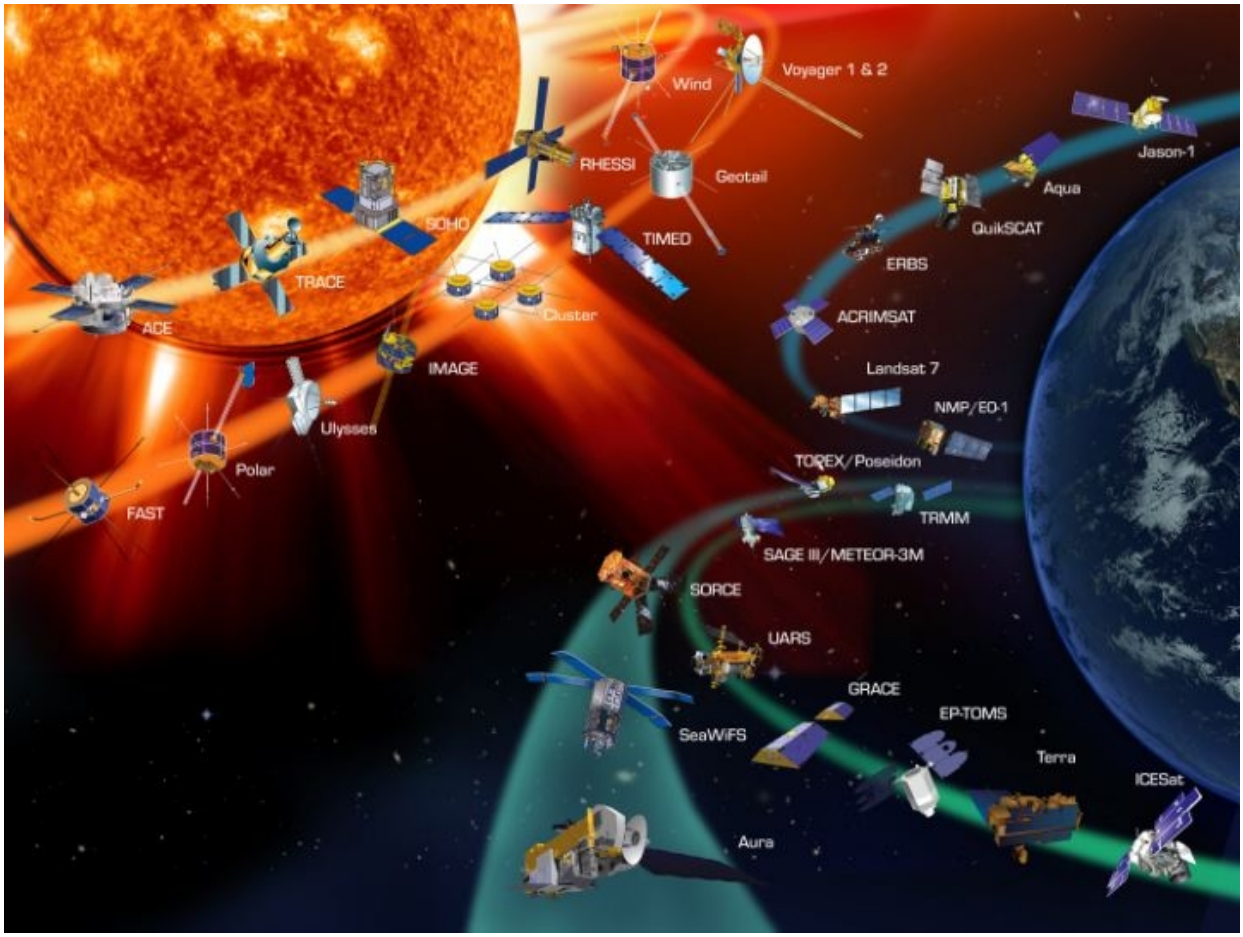

Figure 1. Space physics and Earth science missions.

Long-term data archiving is provided by the National Space Science Data Center (NSSDC), which is NASA's permanent archive for space science (http://nssdc.gsfc.nasa.gov). Permanent archiving of land remote sensing data is managed by the U.S. Geological Survey (USGS) (http://edc.usgs.gov/index.html) and, recently on an interagency basis for Earth systems data more generally, by NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) (http://www.nesdis.noaa.gov/datainfo.html). The archival storage focus of permanent archives includes data ingest (primarily from active archives), archival storage, data and metadata management, preservation planning, and data dissemination (principally as a backup to active archives – otherwise they function as an "active archive"). Permanent archives are concerned with the long-term independent meaningfulness and usability of data, which requires special attention to data migration, metadata, and standards issues.

Active archives provide more immediate, short-term or real-time data access. Data ingest, management, storage, and access functions are common to both active and permanent archives.[1] Data ingest is primarily from original data providers (missions and principal investigators (PIs)) but may include data from other active archives. The data access focus of active archives stresses interoperability, value-added services, and data dissemination, but they may also need to perform migrations.

Ideally, permanent archives communicate with active archives, and active archives communicate both with the permanent archive, original data providers, and the scientists, educators and others who are end users of the data. Examples of active archives in space science are the Planetary Data System (PDS) (http://pds.jpl.nasa.gov) and the Space Physics Data Facility (SPDF) (http://spdf.gsfc.nasa.gov). One overview of space science data systems is available at http://data.gsfc.nasa.gov/. For Earth science missions, a more integrated view is provided by the Global Change Master Directory http://gcmd.gsfc.nasa.gov/.

Numerous Project, Mission and PI web sites provide access to current data, some of which are not yet available through a centralized active archive; an example is the Polar/TIDE experiment site (http://satyr.msfc.nasa.gov/TIDE/). The simple "permanent archive – active archive – user" framework described above is augmented by a rapidly growing set of distributed systems functioning as virtual active archives or collaboratories.[2] These virtual observatories are most often vertically integrated within a particular discipline (e.g., National Virtual Observatory Alliance http://www.ivoa.net/ - astrophysics focus). Numerous VxO systems are emerging in space science, Earth systems science and other fields.

2 Rapid Change in the Data Environment

The rising ubiquity of computers and internet access has led to a rapid change in the data environment. For a six-year period from 1998 to 2003, Figure 2 shows the dramatic decrease in the number of off-line requests to the space science data center in contrast to the rapid increase in on-line data served. The same period experienced a similarly rapid growth in on-line, value-added services; e.g., CDAWeb, ModelWeb, and OMNIWeb provided by SPDF for the space physics field. All of these have emerged since 1995.

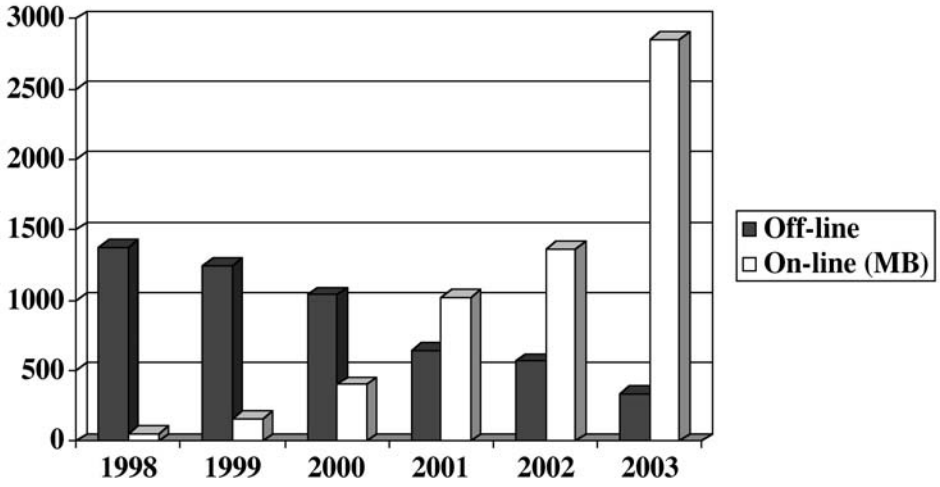

Figure 2. Rapid change in space physics data environment (number of off-line requests, and data transfer totals (MB)).

A panoply of such changes have led to a much more distributed data environment, which has the following characteristics, among others:

- Distributed, on-line, multi-source/media/format
- Web-based, machine/application-accessible data archives

- On-line registries of products and services
- Front-end applications and brokers to connect archives to front ends
- Diverse metadata, emerging standards and ontologies
- High-order search capabilities
- Data mining and other knowledge discovery tools
- Grid computing and broad-band networking

[Examples: CDAWeb (http://cdaweb.gsfc.nasa.gov/) NASA IPG (http://www.ipg.nasa.gov/) Space Physics Virtual Observatory http://vspo.gsfc.nasa.gov/websearch/html/VSPO.html) Virtual Solar Observatory (http://umbra.nascom.nasa.gov/vso/)]

3 Emerging New Science Missions

For certain robotics science missions, the traditional single-spacecraft is being replaced by multi-spacecraft, distributed, communications- and computation-intensive, adaptive mission architectures termed "Sensor Webs" (e.g., http://sensorwebs.jpl.nasa.gov). The traditional "stove-piped" approach tends to be a mere platform for an aggregate of independent instruments, and such missions are vulnerable to single-point failure modes. A Sensor Web architecture, by contrast, is an intrinsically adaptive design: "its constituent sensor, computing, and storage nodes coordinate, dynamically modify, and adapt their measurement modes, observing strategies, and processing states, to intelligently collect, exchange, and synthesize sensor data and other information in ways that tend to maximize useful science return."[3] The architecture can contribute to reductions in mission failure modes through optimal resource sharing among its nodes as depicted in Figure 3.

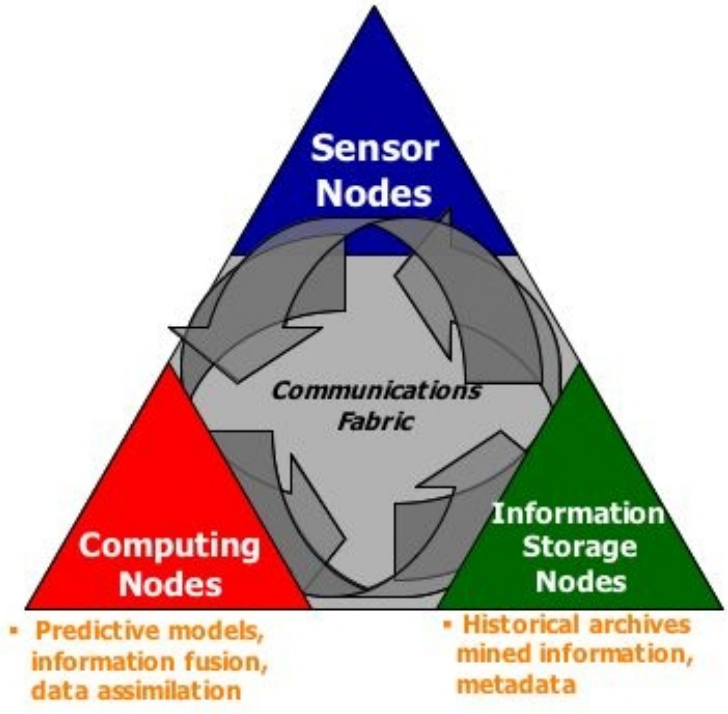

Figure 3. Sensor Web: Sensor, computing, and information storage nodes (provided by Stephen Talabac, NASA/GSFC).

The Sensor Web concept recently emerged in Earth science planning activities where mission success required multi-point remote-sensing space observation combined with coordinated, distributed ground-based in situ sensor networks. Concurrently, plans for spacecraft constellations in space physics have led to similar adaptive mission architectures (see Solar-Terrestrial Probes http://stp.gsfc.nasa.gov/). Combining sensor web concepts with new robotics and nanotechnology are leading to some revolutionary new concepts for robotics missions (e.g., http://ants.gsfc.nasa.gov/).

All these wonderful new possibilities for future science missions remind one of an elegant mansion pictured along a cliff-side location. While at first admiring its magnificent construction, one is suddenly struck by a gash of erosion cutting into the cliff, which reaches the building's foundations and will soon plunge the mansion into the sea.

The beautiful mansion of our dreams - virtual observatories, distributed, yet fully accessible data sets, sensor webs, etc. – is similar threatened by the erosion of business as usual and everyone holding on desperately to "my" data. The simple answer is to move towards open data environments – but how? What is needed to ensure a solid foundation for future missions and science in open, distributed data environments. Key solutions are "interoperability" and "architecture."

4 Foundations – interoperability and systems architecture

Data systems are comprised of four basic elements: data, metadata, software, and systems. Not including the data itself, "metadata" refers to all data about data, including location, ownership, and attributes, including bit representation, data format, and cataloging information. "Software" here refers to middleware, data analysis and modeling software. Middleware denotes software that mediates between application programs and the network; a principal example being "Web Services." And "systems" refers to the combination of hardware and software that embodies the architecture, data, metadata, and software that constitutes the overall data system.

Enhanced interoperability and systems architecture, based on best practices and standards, are key goals for the continuing improvement of data and data systems in our present transitional period from legacy analog data systems to hybrid or born-digital systems. Substantial investments are now going into new eScience, [4] VxO, grid computing and grid systems generally. Major monographs on these topics[5,6,7] emphasize the foundational importance of interoperability and systems architecture.

Working with scientists and data systems professionals internationally, NSSDC has been a leader in standards work. For space science data and communications, the principal standards body is the Consultative Committee for Space Data Systems (CCSDS) http://www.ccsds.org/. For example, the key architectural document for data systems, a guideline for Open Archival Information Systems (OAIS), was developed through CCSDS and is now an adopted standard with the International Organization for Standardization (ISO) http://www.iso.org. It should not be forgotten that if standards had not developed for file transfer protocol (FTP) and other foundational elements of the internet, we would still be communicating by "snail" mail. A guide to work done in digital archive preservation through CCSDS/ISO is provided at http://nssdc.gsfc.nasa.gov/nost/isoas/.

There are three key roles (producer, consumer, management) and six functional entities outlined in the OAIS guidelines: ingest, archival storage, data management, access, administration, and preservation planning. Their relationships are depicted in Figure 4, which further uses the concept of "information packages" of three types (submission, archive, dissemination). An information package is a conceptual container that includes content information and preservation description information. Although complex in full description or implementation, OAIS provides "best practices"guidelines that are quite simple and commonsensical at the core. Further, the information package concept is fully flexible and allows for a mix of born-digital, analog or physical data entities, including lab specimens of whatever form. NSSDC has implemented OAIS guidelines in its latest version of software and system architecture, albeit with some stray legacy functions at the fringes. We have found that architecture and interoperability work hand-in-hand; shortcuts in one of these will undercut efforts in the other. As with so much in life – (poor) good planning has its (penalties) rewards. To augment the OAIS effort, the standards team and CCSDS have fashioned and submitted to ISO a new proposed standard for the producer-archive interface, which helps to define data provider-to-archive relationships, such as agreements, standards, and quality assurance. [8]

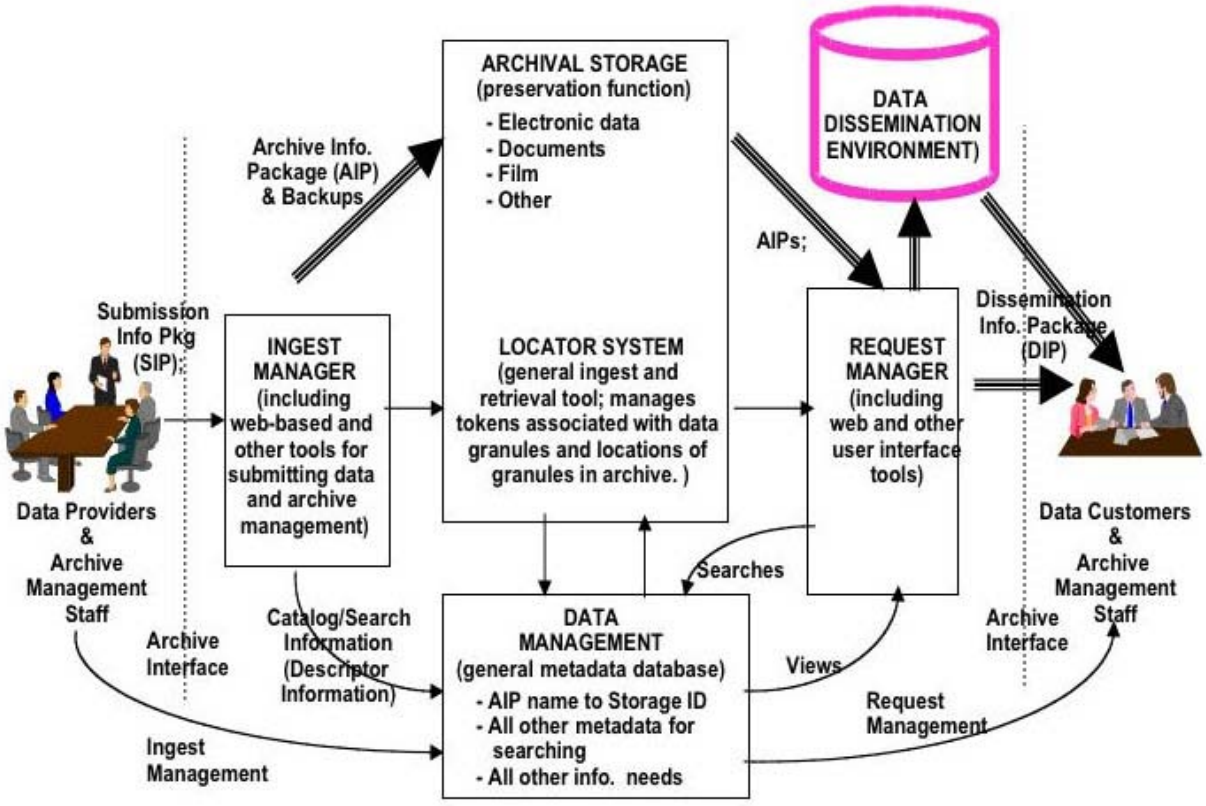

Figure 4. Open Archival Information Systems (OAIS) architecture guideline.

Figure 5. Data and data systems as central to science.

The relationship of archives to research and missions is illustrated in Figure 5, which places data standards technology and related interoperability needs at the intersection of all three circles (permanent archive, active archives, missions). Paired intersections of these three point to principal functions of data and data systems in support of science (planning, preservation, and data analysis or research).

5 Data and Data Systems as Central to Science

The recent confluence of new technologies (internet, XML and Web Services, broadband networking, high-speed computation) dramatically changes the data landscape. Distributed data and computing resources are more and more being linked together in virtual observatories and grid systems. Focusing only on possibilities emerging from virtual observatories, however, may distract us from the prime objective - support of science research.

This confluence of new technologies provides a greatly enhanced synergism, illustrated in Figure 6, between robust data sets (Data), state-of-the-art models and simulations (Model), high-data-rate sensors (Sensor), and high-performance computing (HPC). In the late 20th century, a major science revolution in chaotic systems and nonlinear dynamics arose because of a new coupling of models and high-performance computing. Similarly, we expect that the emerging coupling of rich data sets, and high-performance computing, models and sensors will lead to even greater scientific impact.

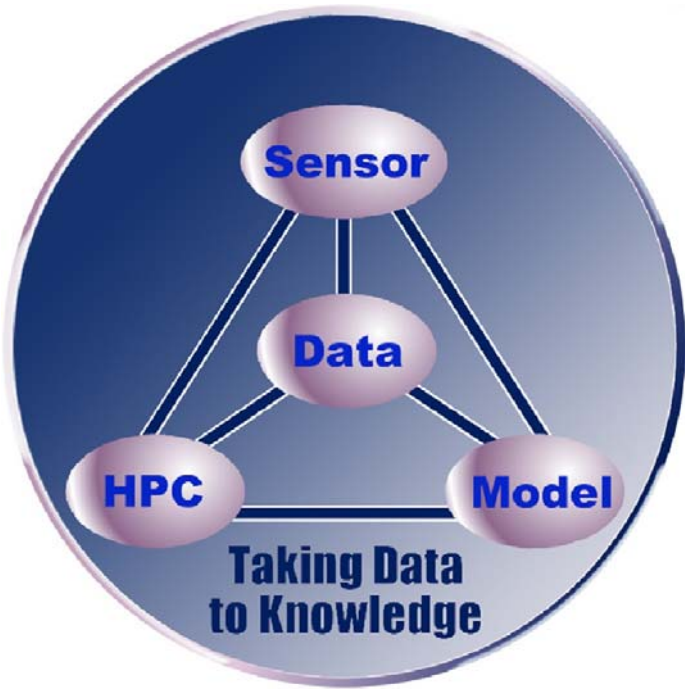

Figure 6. Taking data to knowledge – synergism of Data-Model-HPC-Sensor.

The need for this Data-Model-HPC-Sensor synergism derives from the following set of drivers (problems).

PROBLEMS (all associated with links in Figure 6)

- Information Explosion (Data-HPC)
- Understanding Multiscale Physical Systems (Data-Model)
- Solving Complex, Nonlinear Systems (HPC-Model)
- New High Data Rate Sensors (Sensor-HPC)
- Distributed, Intelligent Sensor Networks (Sensor-Model)

There is no single solution for these complex problems but probable contributors to a solution (see below) fall within the "Data Grid" rubric.

SOLUTIONS

- Distributed Data Environments
- Grid Services (interoperability; semantic web)
- eScience, Virtual Observatories, Data Grids
- Knowledge Discovery, Data Mining
- Data Archive Standards
- Sensor Web
- Sensor Development
- Scientific Modeling
- Advanced Visualization

Data and data systems are central to this new paradigm as indicated in Figure 6. It is within this context that we can best ask about the appropriate relationship and tradeoffs between active and permanent archiving, between central and distributed data systems, and how to best coordinate access to rapidly growing science data sets and support for eScience and grid systems. The data-Model-HPC-Sensor tetrahedron can have any vertex placed in the center, which symbolizes multiple important perspectives on this synergism; e.g., grid computing emphasizes the HPC vertex and Sensor Webs emphasize the Sensor vertex.

Just as the current internet would be without a foundation without its basic protocols and standards, the emerging distributed data systems require even greater attention than before to interoperability and architecture issues (see Section 4). With its international framework within the International Council of Scientific Unions (ICSU) combined with many institutions having strong data, data systems, and data archiving infrastructure, CODATA and its partners are uniquely positioned to provide leadership in these efforts for all science needs.

While distributed active archives can serve as the operational front line for scientific data access for most, if not all, scientific disciplines, this access is typically managed in discipline-specific ways as illustrated by most active archives and virtual observatories. In some cases, a major data center has been delegated a broader purview. For example, NSSDC is the designated permanent archives for all NASA space science disciplines. It carries out this responsibility with close attention to international data archiving standards and methodology to insure indefinite access and independent, well-documented usage of these data. In addition to providing leadership in data systems standards and interoperability, the NSSDC and its partners within NASA and in the science community have provided a clearinghouse role across all space science disciplines for research tools, models, and grid computing.

The basic question is not about some tradeoff between distributed and centralized resources. Instead, the basic question is how best to support science endeavors in this new era of an enhanced synergism of Data-Model-HPC-Sensor within which we consider new eScience and grid systems. What is most critical is to develop the core infrastructure (interoperability, architecture) that makes this new synergism possible, which includes stable and extensive permanent archives covering all scientific fields, just as NSSDC does for NASA space science with continued work in standards and interoperability issues, and cross-discipline tools that support the new distributed systems.

In conclusion, both well-managed archives and eScience or VxOs are essential to enable the best possibilities for 21st century science and technology.

6 Acknowledgments

This work was partially supported under contract number NNG04EA43C to QSS Group, Inc. The authors wish to thank Kirk Borne of George Mason University, Robert Candey and Steve Talabac of NASA/GSFC, Joseph King, Jane Russell and William Taylor of QSS Group, Inc., and John Garrett of Raytheon ITSS for comments and suggestions.

7 References

1 *ISO Archiving Standards Overview, and the Open Archival Information Systems (OAIS)* (2004) Webpage from the National Space Science Data Center, NASA Goddard Space Flight Center. Available from: http://nssdc.gsfc.nasa.gov/nost/isoas/

2 *National Collaboratories: Applying Information Technology for Scientific Research* (1993) NAS/NRC Computer Science and Telecommunications Board, National Academy of Science Press, Washington, D.C. Available from: http://www.nas.edu/

3 Stephen Talabac (2004) *Sensor Webs: Maximizing useful science return using dynamic measurement techniques and adaptive observing strategies.* Available from BU College of Engineering webpage: http://www.bu.edu/mfg/programs/outreach/etseminars/2004may/

4 We use eScience here to refer broadly to all grid system, virtual observatory, and related distributed scientific collaborations enabled by the Internet; one example is the UK e-Science Programme. Available from: http://www.rcuk.ac.uk/escience/

5 Ian Foster and Carl Kesselman, eds. *The Grid: Blueprint for a New Computing Infrastructure*, 2nd ed. Amsterdam: Elsevier, 2004.

6 Fran Berman, Geoffrey Fox, and Tony Hey, eds. *Grid Computing: Making the Global Infrastructure a Reality*, Chichester: John Wiley & Sons, Ltd., 2003.

7 *Revolutionizing Science and Engineering Through Cyberinfrastructure* (2003) Report of the NSF Blue-Ribbon Advisory panel on Cyberinfrastructure, National Science Foundation. Available from: http://www.cise.nsf.gov/sci/reports/toc.cfm

8 *Producer-Archive interface methodology documents* (2003) Consultative Committee on Space Data Systems (CCSDS). Available from: http://nssdc.gsfc.nasa.gov/nost/isoas/paim.html