ESCIENCE AND ARCHIVING FOR SPACE SCIENCE

Timothy E. Eastman[1], Kirk D. Borne[2], James L. Green[3], Edwin J. Grayzeck[4], Robert E. McGuire[5], and Donald M. Sawyer[4]

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

Abstract

A confluence of technologies is leading towards revolutionary new interactions between robust data sets, state-of-the-art models and simulations, high-data-rate sensors, and high-performance computing. Data and data systems are central to these new developments in various forms of eScience or grid systems. Space science missions are developing multi-spacecraft, distributed, communications- and computation-intensive, adaptive mission architectures that will further add to the data avalanche. Fortunately, Knowledge Discovery in Database (KDD) tools are rapidly expanding to meet the need for more efficient information extraction and knowledge generation in this data-intensive environment. Concurrently, scientific data management is being augmented by content-based metadata and semantic services. Archiving, eScience and KDD all require a solid foundation in interoperability and systems architecture. These concepts are illustrated through examples of space science data preservation, archiving, and access, including application of the ISO-standard Open Archive Information System (OAIS) architecture.

Keywords: Data archives, Distributed data systems, Archiving, active archives, Permanent archives, Data standards, Interoperability, Grid systems, Grid computing, Metadata, eScience, Cyberinfrastructure, Virtual observatories, Sensor web, Robotic missions, Adaptive design, Knowledge Discovery in Databases (KDD), Supervised and unsupervised learning methods, Data mining, Neural networks, Data registries, Ontologies, XML, OAIS, CCSDS, ISO, CODATA.

1 Data and Data Systems as Central to Science

A confluence of new technologies (internet, XML and Web Services, broadband networking, high-speed computation, distributed Grid computing, ontologies and semantic representation) is dramatically changing 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 (e.g., NVO, 2005), 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 1, 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 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 linkage of rich data sets, high-performance computing, models and sensors will lead to even greater scientific impact. Data-driven science is already advancing in numerous domains as a separate research discipline (e.g., Bioinformatics and Geographic Information Systems) in the same way that computational science has become an established research endeavor.

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

PROBLEMS (all associated with links in Figure 1)

- 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 fall within the "Data Grid" rubric.

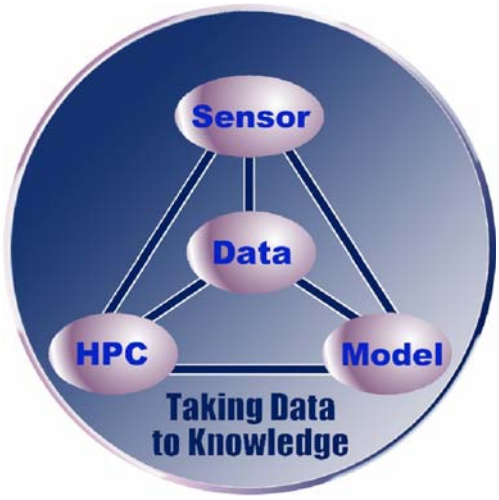

Figure 1. Taking data to knowledge – synergism of Data-Model-HPC-Sensor.

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 1. It is within this context that we can best ask about the appropriate relationship and tradeoffs between active and permanent archiving (see Appendix), between central and distributed data systems, and how best to 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.

2 Knowledge Discovery in Databases

An even greater challenge than managing this explosion of data, while the number of scientists remains roughly constant, is that of providing efficient harvesting of information and extraction of knowledge within this data avalanche. Knowledge Discovery in Databases (KDD) denotes "the nontrivial extraction of implicit, previously unknown, and potentially useful information" (Frawley, Piatetsky-Shapior & Matheus, 1991). Within the past decade there have been major advances in data mining, neural networks, pattern recognition, clustering, principal component analysis, Bayesian networks, Markov models and other tools, which are here referred to collectively as

KDD tools. KDD is particularly useful for the discovery of hidden relationships in large, complex databases where human means of pattern recognition or even model application may fail. In scientific databases, which may contain many hundreds of descriptive parameters, the possibility of discovering high-dimensional multi-factor dependencies is simply beyond the scope of human and brute-force computational analyses. In many problems, the combinatorial explosion (resulting from exponential growth in the number of possible parameter combinations) requires a non-traditional (i.e., KDD) approach, since computational horsepower alone cannot solve a problem that requires several hundred factorial model parameter tests.

Data selection, automating access through registries, translation and formatting, and data cleaning are just a few of the many data preparatory steps that are essential for successful KDD applications and which can consume up to 80% of a data-mining project (Pyle, 1999). With this investment and with sufficiently robust data sets, however, previously hidden facts can be discovered such as rare events, anomaly detection, patterns, correlations, linkages, complex multi-variable interdependencies and more (Borne, 2003; Bazell, Miller and Borne, 2002).

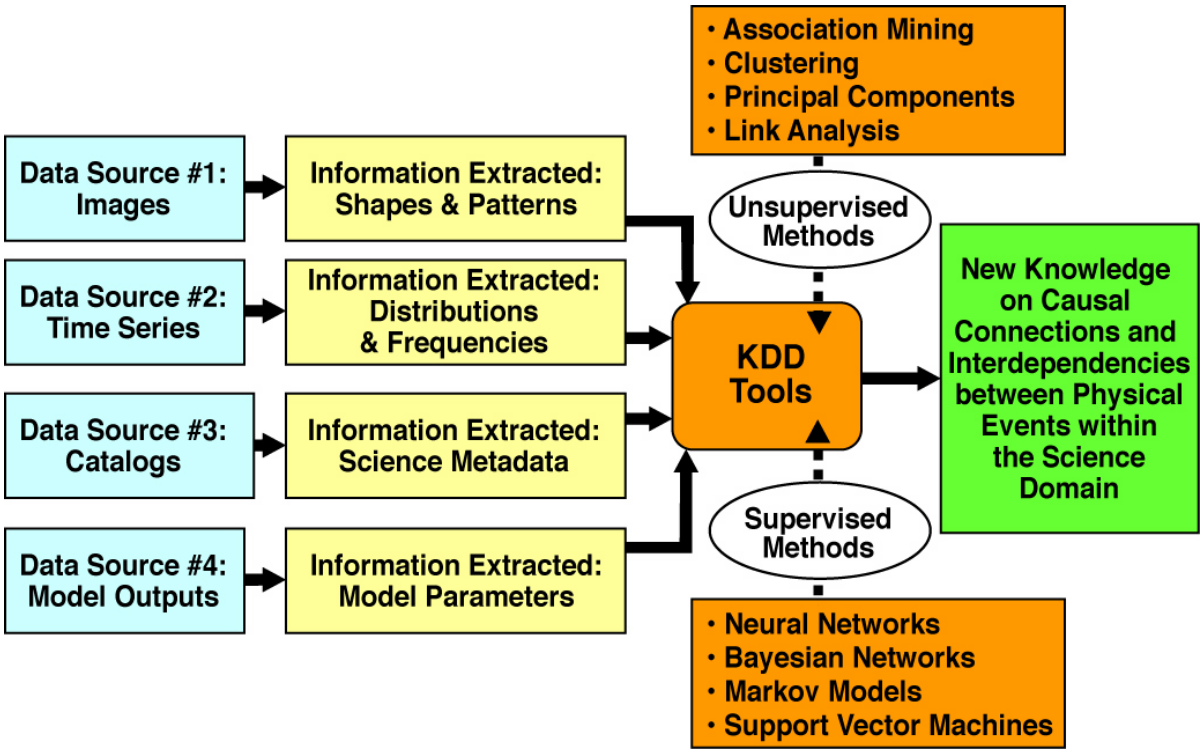

Figure 2. KDD tools can provide new knowledge of physical relationships.

For unprejudiced discovery of associations, connections, linkages, and relationships in data, it is best to use Unsupervised Learning methods (see Figure 2). These methods are applied to data parameters to discover new relationships and patterns. Such relationships may have complex multi-dimensional interdependencies that are beyond the scope of human analysts to discover. Unsupervised Learning methods include various Clustering techniques, Association Rule Mining, Link Analysis, PCA (Principal Components Analysis), and Independent Component Analysis. Unsupervised Learning is sometimes referred to as "Class Discovery" or "Building the Model" (Dunham, 2002). Supervised Learning techniques are applied to new data products to predict an outcome (or event) from among a set of possible outcomes (e.g., predicting different Solar energetic particle event classes from observed Solar Coronal Mass Ejection (CME) episodes; or predicting different CME event classes from Solar surface phenomena). Supervised methods include various Classification techniques – Hidden Markov Modeling, Bayesian Networks, Support Vector Machines, Nearest Neighbors, and Neural Networks (Hastie, Tibshirani & Friedman, 2001). Supervised Learning is sometimes referred to as "Applying the Model" (Bazell & Aha, 2001; Bazell et al., 2002).

The emergence of the International Virtual Observatory Alliance (IVOA) (www.ivoa.net) and many other venues for data access (see Appendix) provide important new opportunities for applying KDD tools, which will be powerful augmentations (not replacements) to traditional data analysis, modeling and simulation techniques. The application of KDD to modeling and simulation data output, in conjunction with sensor and archival data collections, will further magnify the significance of the synergies depicted in Figure 1.

3 Data-Intensive Missions and the New Data Environment

Adding to the Data Avalanche: 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" (NASA/JPL Sensor Webs Project, n.d.). 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" (Higgins, Kalb, Lutz, Mahoney, Mauk, Seablom & Talabac, 2003). This architecture can contribute to reductions in mission failure modes through optimal resource sharing among its nodes as depicted in Figure 3.

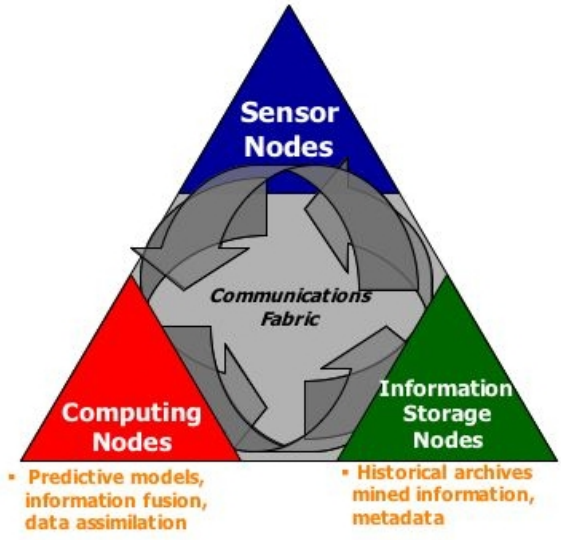

Figure 3. Sensor Web: Sensor, computing, and information storage nodes (provided by Stephen Talabac, NASA/GSFC).

The Sensor Web concept emerged recently in Earth science planning activities where mission success requires 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 (Solar Terrestrial Probes Program, n.d.). Combining sensor web concepts with new robotics and nanotechnology are leading to revolutionary new concepts for robotics missions (Autonomous NanoTechnology Swarm, n.d.). At the same time, multi-point sensor web systems and high-speed data sensors will be adding even further to the forthcoming data avalanche.

Towards a More Distributed Data Environment: Centralized data environments and data centers were dominant prior to the internet revolution. This provided more control over computer and support systems that were often not interoperable and had (without contemporary middleware) many hardware-level dependencies. Overcoming such limitations has greatly enhanced the power of data centers; however, such centers would be crippled without efficient access to off-site data resources. The other extreme in data architecture is that of radically distributed data sites with no centralized or primary nodes. This architecture is exemplified by wide-open aspects of the web with Google searches in place of any systematic index or catalogue. However, lack of any center would undercut efforts

towards interoperability, data preservation, or systematic data archiving. In addition, with emerging peta-scale datasets, "the datasets are so large, and the application programs are so complex, that it is much more economical to move the end-user's programs to the data and only communicate questions and answers rather than moving the source data and its applications to the user's local system (Gray, Liu, Nieto-Santisteban, Szalay, DeWitt & Heber, 2005).

Given the many drivers and constraints on data and data systems, the emerging data environment at NASA illustrates a hybrid model, or middle-ground balance between the above centralized and radically distributed scenarios for data handling. Similar to the evolving data environment for most scientific fields, science data preservation and access at NASA includes a complex mix of data sources at multiple levels with multiple, distributed active archives providing the highest level of user access support. And, for space science, a single permanent archive providing long-term preservation and back-up functions for the overall science data system (see Appendix). The rising ubiquity of computers and internet access has fundamentally changed data environments and the roles of data centers. For example, off-line requests to our data center have declined to less than one per day whereas on-line data served (now ~3 Tb per year for the Space Physics Data Facility) has steadily increased.

A panoply of such changes has led to a much more distributed data environment, but one that retains centralized features:

- 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
- Centralized data center for long-term data preservation; back-up to active archives
- Major active archives for user-oriented services and data access within major disciplines

  Examples: Planetary Data System http://pds.nasa.gov
  Global Change Master Directory http://gcmd.gsfc.nasa.gov/

Need for Reliable Foundations: All these wonderful new possibilities for future science may be like an elegant mansion at 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 to "my" data. The simple answer is to move towards open data environments – but how? What is needed is to ensure a solid foundation for future missions and science in open, distributed data environments. Key solutions are "interoperability" and "architecture."

Just as the current internet would lack a foundation without basic protocols and standards, the emerging distributed data systems require even greater attention than before to interoperability and architecture issues. Given 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.

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, comprises 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" refer to the combinations of hardware and software that embody 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,[†] VxO, grid computing and grid systems generally. Major monographs on these topics emphasize the foundational importance of interoperability and systems architecture (Foster & Kesselman, 2004; Berman, Fox & Hey, 2003; NSF, 2003).

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 reference model standard for Open Archival Information Systems (OAIS) concepts, 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/. The OAIS reference model has been widely adopted as a key framework in the discussion and implementation of digital preservation and management systems (Anderson, 2004).

The OAIS framework outlines three key roles (producer, consumer, management) and six functional entities: ingest, archival storage, data management, access, administration, and preservation planning. Relationships among the first four, which are the major operational functions, are depicted in Figure 4. Also shown are uses of the concept of "information packages" that come in 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, the OAIS conceptual framework can be taken as "best practice" guidelines that are essentially simple and logical. The information package concept is fully flexible and allows for a mix of born-digital, analog or physical data entities, including laboratory specimens of whatever form.

The National Space Science Data Center (NSSDC) has implemented several OAIS concepts in its latest version of software and system architecture, albeit with some stray legacy functions at the fringes. A concrete implementation of an Archival Information Package (AIP) was developed as a single file; however, it contains multiple metadata and data objects originally existing as individual files. This has provided the flexibility to capture attributes about each file and about the collection of files as a group in a form that is easily migrated across media and systems while retaining sufficient content to be a useful unit. It employs embedded standards-based pointers to refer to external metadata, such as format information, that is common across many files. In this way, updates to the common metadata can be made without having to retrieve and update very large numbers of AIPs, while attributes closely associated with individual files, like original file names, file sizes, checksums on each file, etc., are kept with file objects in the AIP. Experience has shown that it is also useful to separately capture metadata attributes within each AIP for storage in a database so that they can be readily searched if data quality or related issues arise. This follows in the NSSDC case as there are over a million AIPs now stored on robotic super-digital linear tapes (DLTs), and scanning these to extract attribute values would be very time consuming.

Another key feature of the OAIS that is being incorporated into NSSDC architecture is the clear separation of the Archival Storage function from other functions. AIP implementation has facilitated this, as now Archival Storage can concentrate on preserving AIPs and "pointed to" metadata, without being concerned about how AIPs are created

[†] We use eScience here to refer broadly to all grid system, virtual observatories (VxO), and related distributed scientific collaborations enabled by the Internet; one example is the UK e-Science Centre http://www.nesc.ac.uk/.

or how they are disseminated to end users. This provides a strong preservation focus and helps the staff to focus on what is actually being preserved, thus improving quality control. For example, new efforts are being made to capture provenance information associated with a migration that is moving older data into AIPs, with the result that when problems arise they can be easily documented and associated with the AIPs for greater data reliability and understanding.

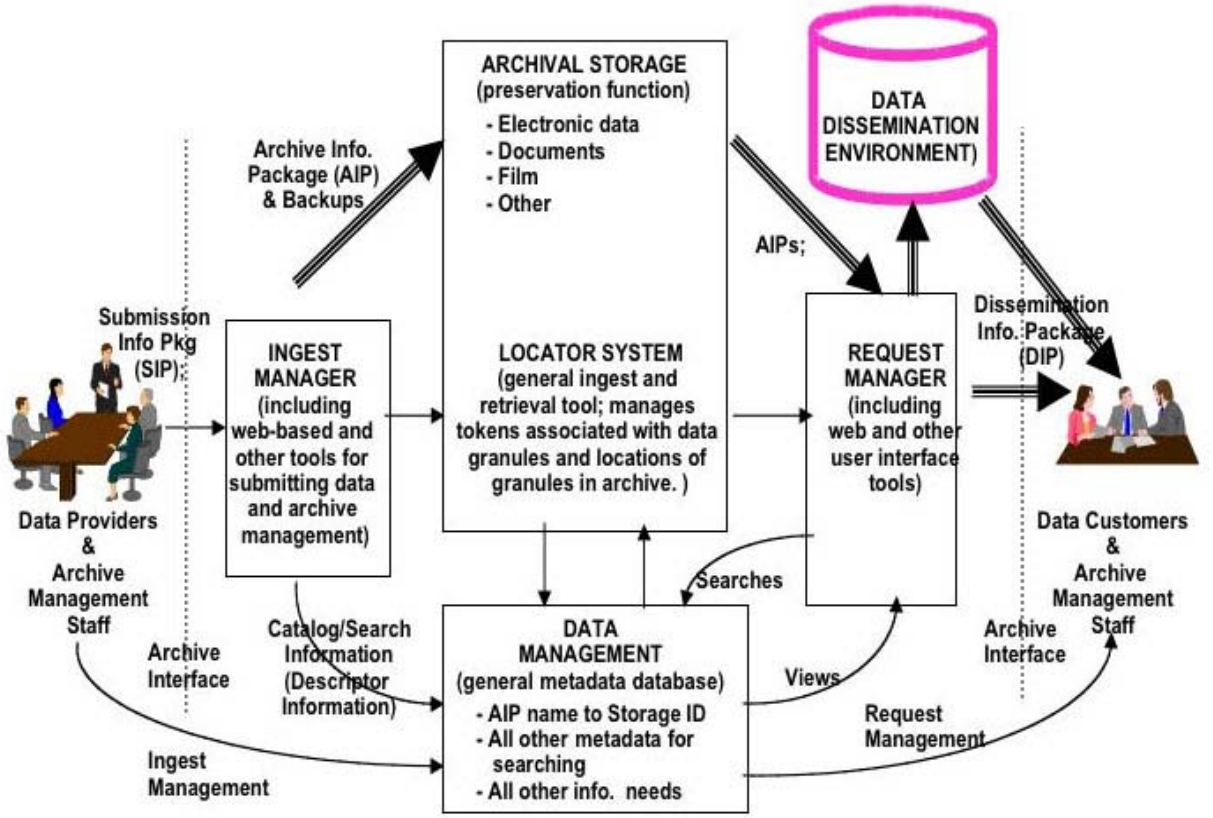

Figure 4. Open Archival Information Systems (OAIS) architecture as adopted by NSSDC.

The longevity of the NSSDC as a digital archive, being some 40 years in operation, has provided an opportunity to see first-hand a number of issues associated with long-term preservation and the role of migration in this preservation effort. For the current, or third major, migration that is moving 9-track and 3480 cartridge data to DLTs, a significant effort is being made to highly automate the process by creating a database at the core of the migration management. This effort is impeded by the lack of consistency and completeness in related operational databases, which is being overcome by manual checking with updates to the core database. Lessons learned include the need for thorough, documented, reviewed, and updated plans. It is easy for an organization to succumb to cost pressures and to cut corners. While taking shortcuts may look cost-effective, it turns out to be more costly in the long run. Another lesson learned is to use automation wherever and whenever possible. This reduces human mistakes and workloads, and provides a more consistent result.

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 (CCSDS, 2003). They are also working on a draft standard called XFDU that uses XML as the basis of a manifest document within a general data packaging scheme. This has wide applicability as a general container for exchanging data, metadata, and their relationships, and is an obvious candidate as the basis for future AIP implementations.

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).

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

6 Conclusion

The basic question that motivates new eScience initiatives is not about the preferred grid model or some tradeoff between distributed and centralized resources. Instead, the basic question is how best to support science endeavors in this new era of Knowledge Discover in Databases (KDD) and an enhanced synergism of Data-Model-HPC-Sensor within which we consider new eScience and grid systems, followed by analysis and modeling (boosted with KDD tools) to create the new knowledge that we seek. Critical to this support is the development of the core infrastructure (interoperability, architecture) that makes this new synergism possible, which includes stable and extensive archives covering all scientific fields with continued work in standards and interoperability issues, and cross-discipline tools that support distributed data systems.

Well-managed archives, eScience or VxO systems, and vigorous application of new KDD tools promise to be central to many, if not most, major science and technology advances in the coming century.

7 Acknowledgments

This work was partially supported under contract number NNG04EA43C to QSS Group, Inc. The authors wish to thank Robert Candey, Aaron Roberts, and Stephen Talabac of NASA/GSFC, Joseph King, Patrick McCaslin, Jane Russell and William Taylor of QSS Group, Inc., and John Garrett of Raytheon ITSS for comments and suggestions. Valuable suggestions from two reviewers, which resulted in substantial changes and improvements, are gratefully acknowledged.

8 References

Anderson, W. L. (2004) Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. *Data Science Journal, 2,* 191-202.

Autonomous NanoTechnology Swarm (n.d.) Homepage of Auonomous NanoTechnology Swarm. Available from: http://ants.gsfc.nasa.gov/

Bazell, D., & Aha, D. (2001) Ensembles of Classifiers for Morphological Galaxy Classification. *ApJ., 548*, 219.

Bazell, D., Miller, D., & Borne, K. (2002) Novel Approaches to Semi-supervised and Unsupervised Learning. *Conference ADASS XII* (pp. 427-430). Baltimore, MD, USA.

Berman, F., Fox, G., & Hey, T. (Eds.) (2003) *Grid Computing: Making the Global Infrastructure a Reality*. Chichester: John Wiley & Sons, Ltd.

Borne, K. D. (2003) Distributed Data Mining in the National Virtual Observatory. *SPIE Data Mining & Knowledge Discovery V, 5098,* 211-218.

Consultative Committee on Space Data Systems (CCSDS) (2003) Producer-Archive interface methodology documents. Retrieved July 15, 2005 from the NSSDC Web site: http://nssdc.gsfc.nasa.gov/nost/isoas/paim.html

Dunham, M. (2002) *Data Mining: Introductory and Advanced Topics*, Upper Saddle River, NJ: Prentice-Hall.
Foster, I. & Kesselman, C. (Eds.) (2004) *The Grid: Blueprint for a New Computing Infrastructure*, 2nd ed., Amsterdam: Elsevier.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. (1991) Knowledge discovery in databases: An overview. In G. Piatetsky-Shapiro, G. & Frawley, W. (Eds.), *Knowledge Discovery in Databases*, Menlo Park, CA: AAAI Press.

Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D., & Heber, G. (2005) Scientific data management in the coming decade. Technical Report MSR-TR-2005-10, *Microsoft Research*. Redmond, WA.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001) *The Elements of Statistical Learning*, New York: Springer.

Higgins, G., Kalb, M., Lutz, R., Mahoney, R., Mauk, R., Seablom, M., & Talabac, S. (2003) Advanced Weather Prediction Technologies: Two-way Interactive Sensor Web & Modeling System (Phase II Vision Architecture Study). Report of the Earth Science Technology Office, NASA Goddard Space Flight Center. Available from http://esto.gsfc.nasa.gov/files/2002/Weather-Forecasting/WeatherForecastingT_D19C3.pdf.

NASA/JPL Sensor Webs Project (n.d.) Homepage of the NASA/JPL Sensor Webs Project. Available from: http://sensorwebs.jpl.nasa.gov.

National Space Science Data Center (NSSDC) (2004) ISO Archiving Standards Overview, and the Open Archival Information Systems (OAIS). Retrieved July 15, 2005 from NSSDC Web site: http://nssdc.gsfc.nasa.gov/nost/isoas/

National Research Council (NRC) (1993) National Collaboratories: Applying Information Technology for Scientific Research. Report of the Committee on a National Collaboratory. Retrieved July 15, 2005 from the *National Academies Press* Web site: http://books.nap.edu/catalog/2109.html

National Virtual Observatory (NVO) (n.d.) Homepage of the US National Virtual Observatory. Available from: http://www.us-vo.org/about.cfm

National Science Foundation (NSF) (2003) Revolutionizing Science and Engineering Through Cyberinfrastructure. Report of the NSF Blue-Ribbon Advisory panel on Cyberinfrastructure. Retrieved July 15, 2005 from the *NSF* Web site: http://www.nsf.gov/cise/sci/reports/toc.jsp

Pyle, D. (1999) *Data Preparation for Data Mining*, San Francisco: Morgan Kaufmann Publishers.

Solar Terrestrial Probes Program (n.d.) Homepage of Solar Terrestrial Probes Program. Available from: http://stp.gsfc.nasa.gov/

9 Appendix. Science Data Preservation and Access at NASA

Space physics and Earth science data systems at NASA's Goddard Space Flight Center manage data from a wide variety of science missions. Long-term data archiving is provided by the 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), administration, data and metadata management, preservation planning, and data dissemination (principally as a supplement 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 (*NSSDC*, 2004). 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.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://nssdc.gsfc.nasa.gov/nssdc/obtaining_data.html. 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 from 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 (*NSF*, 2003; *NRC*, 1993). These virtual observatories are most often vertically integrated within a particular discipline (e.g., International Virtual Observatory Alliance http://www.ivoa.net/ - astrophysics focus). Numerous VxO systems are emerging in space science, Earth systems science and other fields (e.g., the Earth Observing System Clearinghouse (ECHO) system; and the NOAA Comprehensive Large Array-data Stewardship (CLASS) system).

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.