Conceptual Study of Intelligent Data Archives of the Future

H. K. Ramapriyan, Steve Kempler, Chris Lynnes, Gail McConaughy, Ken McDonald, Richard Kiang
NASA Goddard Space Flight Center
Greenbelt MD 20771

Sherri Calvo, Robert Harberts, Larry Roelofs
Global Science and Technology, Inc.

Donglian Sun
George Mason University

Ramapriyan@gsfc.nasa.gov
Tel: +1-301-614-5356
Fax: +1-301-614-5267

Abstract

A conceptual architecture study is under way to address the problem of getting the most scientific value from the large volumes of Earth and space science data that NASA expects to accumulate in the future. This involves efficient storage and access, but goes beyond that to facilitate intelligent data understanding and utilization through modeling realistic virtual entities with predictive capabilities. The objective of the study is to formulate ideas and concepts and to provide recommendations that lead to prototyping and implementation in the period from 2010 to 2020. The approach consists of the definition of future scenarios and needs for data usage in applications (in consultation with scientific and applications users), projection of advances in technologies, and an abstraction of an intelligent archive architecture. Strategic evolution is considered in various areas such as storage, data, information and knowledge management, data ingest and mining, user interfaces, and advances in intelligent data understanding algorithms.

1. Introduction

NASA's collections of Earth science data have more than quadrupled in volume since the launch of the Terra satellite in December 1999. At the end of September 2001, NASA's Earth science archives contained over 1,000 terabytes of data and are currently growing at the rate of about 2.8 terabytes per day. Other agencies (e.g., NOAA and USGS) also have large and growing archives of Earth science data. The volumes of Earth science data held by NASA, NOAA and USGS are expected to exceed 18 Petabytes by 2010. Significant increases are expected in the data volumes in space science as well. For example, planned synoptic sky surveys in astronomy could produce 10 Petabytes data per year.

In addition to the large data volumes, there are multiple challenges in managing and utilizing them:

- Data acquisition and accumulation rates tend to outpace the ability to access and analyze them.
- The variety of data implies a heterogeneous and distributed set of data providers that serve a diverse, distributed community of users.
- Human-based manipulation of vast quantities of archived data for discovery purposes is intellectually overwhelming and certainly cost prohibitive.
- The types of data access and usage in future years are difficult to anticipate and will vary depending on the particular research or application environment, its supporting data sources, and its heritage system infrastructure.

Increased hardware capabilities partially mitigate the data access problem. However, adding "intelligence" to the data management and utilization process is essential to automating the end-to-end data lifecycle in order to reduce the burden on data producers and archivists and provide the greatest value to the nation for the data collected. Thus, Intelligent Data Archives here are viewed not just as a set of permanent repositories of data, but also as a suite of services that facilitate the use of data and deriving information and knowledge from them. Therefore "intelligence", in various embedded roles, means the computational transformation of bits into information and knowledge (processing sensory data into models), the ability to automatically act appropriately to complex dynamic conditions (operations automation), and ability to facilitate human interactions with digital resources (semantic management).

A conceptual architecture study is under way to address the problem of efficient access to and effective utilization of the large volumes of data that NASA expects to accumulate in the future. The study is sponsored by NASA's Intelligent Systems Program, and specifically the Intelligent Data Understanding technical area within the program. The intention of the study is to develop ideas and concepts and to provide recommendations that lead to prototyping and implementation in future years. As such, it is not constrained by the need for operational implementation in the near future (e.g., two to five years).

The approach to this study consists of the characterization of future scenarios and needs for data usage in applications (in consultation with scientific and applications users), projection of evolutionary/revolutionary advances in technologies, and an abstraction of an intelligent archive architecture. These steps will lead to a strategy toward the formulation and development of conceptual architectures for intelligent archives. The analysis is used to identify what kinds of intelligent processes are both desirable and feasible, and determine where their application might most effectively drive down costs and enable new applications and research, given anticipated advances in technology. Strategic evolution is considered in various areas such as storage, data, information and knowledge management, data ingest and mining, user interfaces, and advances in intelligent data understanding algorithms.

The following section provides a brief discussion of the preliminary abstracted architecture obtained using this approach. Section 3 presents a description of scenarios and user needs. Section 4 covers projections of evolutionary and revolutionary changes in technology. Section 5 provides a set of recommendations in the form of a road map leading towards intelligent archives supporting intelligent data understanding and utilization.

2. Abstracted Architecture

The abstracted architecture represented here is defined without regard to distributed or centralized nature of implementation and is considered purely from the point of view of the functions that need to exist to support the types of usage scenarios analyzed in section 3. It is possible that with a broader set of scenarios, we will need to identify additional functions in a later version of this abstraction. The functions of an intelligent archive are more stable than the architectures and technologies used to implement them. By abstracting elements and processes into functional elements, we can explore application strategies of technologies and system resources for future intelligent archives.
However, it is first useful to provide our definitions for **data**, **information** and **knowledge**, as these entities are key to the abstraction of the architecture. These are not general definitions, but rather somewhat specific to the domain of scientific research.

- **Data**: output from a sensor, with little or no interpretation applied.
- **Information**: a summarization, abstraction or transformation of data into a more readily interpretable form.
- **Knowledge**: a summarization, abstraction or transformation of information that increases our understanding of the physical world.

Future intelligent archive architectures (see Figure 1) manage these entities with such functional elements as:

- Models and Intelligent Algorithms
  - Consist of models of sensors, resources, data, information, knowledge, and application domain entities (e.g., farm)
  - Include models that exist at multiple levels, ranging from detailed sensor models to models of an entire application domain (e.g., global models in the case of Earth science)
  - Support human understanding of the objects and processes that make up a virtual digital entity and allow users to update the knowledge about the domain as new discoveries are made
- Flow and Feedback Loops
  - Control performance of all other functional elements
  - Include mechanisms that construct, organize, store, update, manage, and provide essential operational services
  - Support self-optimizing operations
- Virtual entity
  - Consists of a representation of the data, knowledge, and processes involved in an application domain

  - Provides a context for ingesting, organizing, and managing data and information for the real world entity it represents
  - Allows interrogation of past, present, and projected future events, as well as "what if" analyses
- Intelligent information and knowledge extraction
  - Facilitates the transformation of data into information and useful knowledge
  - Automates mechanisms that extract meaning from data and therefore leverage the value of all data in the process
  - Supported by models in the knowledge base, which provide a basis for understanding the data
- Intelligent data production, management and archiving
  - Consists of production, persistence, and active management of valued massive data assets
  - Automates efficient data management mechanisms supporting knowledge-building enterprises in the face of an overwhelming "tidal wave" of data
  - Must dynamically manage high volume inputs from a diversity of observational sensors, converting them into quality usable data products
  - Manages storage close to sensors such that data can be processed locally and passed on to the virtual entity as needed
- Intelligent sensors
  - Are responsible for observations and measurements taken from nature and are the raw ingredients for data
  - Operate from various platforms such as satellites, aircraft, balloons, and in situ constructs
  - Have capabilities for performing autonomous functions and also interact with other sensor systems and external functional elements
  - Include storage, management and processing resources that are part of the overall archive
  - Are modeled in the context of the knowledge base and can support collaborative operation by supplying processing and storage resources when they are not needed locally
  - Are expected to become an integral part of an archive as the architecture becomes more distributed. Here the archive would control sensor data collection based on data needs and would use sensor resources to perform its functions

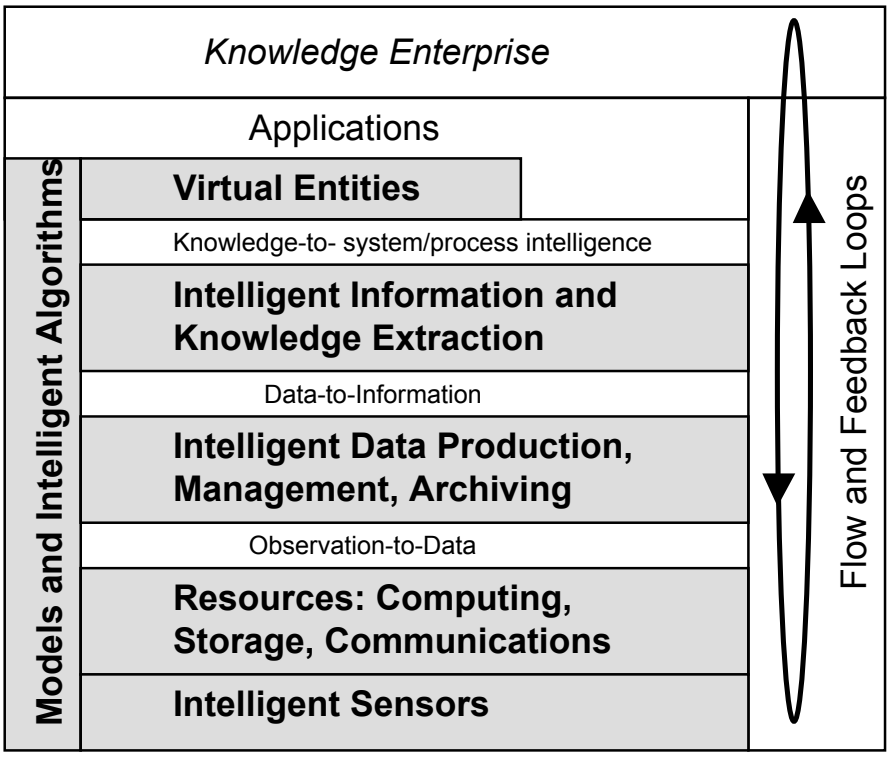

Figure 1: Abstracted Architecture for Intelligent Archives

Using this abstracted architecture to construct an intelligent data system would require a number of design decisions regarding how these elements and entities are represented, such as whether data are represented as bit streams, files, database records, or some other entity. Other decision points concern the relationships among entities, infrastructure to support and connect the various elements, and various optimization schemes. There is much ongoing development in the area of data system intelligence today, such as grid computing, distributed data mining, mobile agents etc. However, because one of the main goals of the abstracted architecture in this study is to aid future research programmatics, the key challenge is to devise an architecture that can be "mapped" into ongoing research and development without being limited to a single architecture evolutionary path.

3. Scenarios and User Needs

We are using a scenario-based approach to the development of futuristic conceptual architectures that enable intelligent data understanding of massive data volumes. Scenario-based approaches are used to drive clear and complete pictures of end-to-end interrelationships among data and information, consumers, data providers, value-added information services, data archives, and data acquisition missions [1,2]. Also, scenario development uncovers a range of requirements for services and capabilities that can be mapped to existing and future technology application. Consequently, forward looking,

tangible and imaginative Intelligent Data Archive (IDA) system application scenarios can be factored into an architectural framework with descriptions of supporting technology.

The scenarios are oriented to an end-user perspective. Scenario descriptions identify "actors" or involved stakeholders and illustrate dependencies among them within an enterprise context. By extension this helps to clarify requirements for corresponding system components and in identifying challenges to be addressed.

Applications scenarios lead to requirements, requirements have implications on technology, and advances in technology affect the evolution of applications. By observing this feedback process, we can characterize several futuristic scenarios. In addition, such a strategy allows the architectural process to adapt quickly to new and evolving scenarios and technologies.

A variety of contexts for possible scenarios have been identified with which to explore, understand, and refine requirements for the IDA architecture. Examples of candidate scenario contexts are:

- Ecological forecasting
- Precision agriculture
- Natural events and hazards (e.g., volcanoes, earthquakes, hurricanes, floods, fires)
- Skilled (10 – 14 day) weather forecasting
- Space weather
- National Virtual Observatory

Of these, in the initial phase of this study, we have used the precision agriculture and precision weather forecasting contexts and developed two scenarios.

3.1 Precision Agriculture

The precision agriculture scenario is concerned with the scope and parameters of a farm employing high-resolution Earth science data. The farm, which constitutes the virtual digital entity in this scenario, is characterized as a relatively small spatial area (considered in acres) for agricultural products suited to regional ecological, weather, and growing constraints.

The “digital farm” concept interrelates ideas about digital technology, digital information, GIS, and human-machine interfaces. We explored potential future requirements and uses of quality high-resolution geo-spatial data employed in precision agriculture. The information resources needed represent the consequence of interoperating services, value-chain processes, automation, and filtering of data of specific relevance to the farmer.

Information-intensive support services helpful for crop planning, cultivation and harvesting include current conditions monitoring, histories and time series studies, trends/risks analysis, prediction, and forecasts, “what-if” investigations, and outcome comparisons. Detailed information about land, weather, water, agriculture markets, prior

yields, agri-chemical options, seeds, etc. are useful for crop planning and planting. High-resolution information is helpful to monitor, assess risks, and make decisions about appropriate interventions to maintain crop health. Similarly, to maximize yields, decisions about harvest timing require information about current and future conditions (e.g., local weather, soil moisture, crop maturity). Remotely sensed information about farm assets, including information collected from the farm about outcomes of plans, cultivation techniques, and harvests, is integrated within a digital farm for long-term use. In all cases it is important that the information be provided to the end-user with confidence estimates.

To make sense of all this information, the digital farm concept includes a digital assistant that works on behalf of the grower and is very intuitive and simple to use. The digital assistant is available from any interface (workstations, mobile devices) from within the house, farm buildings, vehicles, or even the combine. Interaction with the digital assistant can be conducted by natural language either via voice or keyboard.

The digital assistant can interpret, broker, and fulfill requests for information and services from the virtual entity both dynamically and autonomously. In this scenario, the virtual entity is a digital wheat farm that contains encyclopedic farm-relevant information ontologically, spatially, and temporally organized. The digital farm keeps its information stores about soil, crop, weather, and moisture conditions constantly updated. It interfaces with external inputs of data and information sources as well as with farm-specific sensor inputs. These functional interfaces are crucial to pooling farm-relevant data from raw data sources such as primary archives and agricultural services.

The digital assistant can produce different views of this information by summoning an array of functional services. These services can be invoked and combined with an existing farm state model to produce a virtual 4-D representation of the entire farm that the grower can inspect from his or her office or combine cab. The virtual farm serves as an interactive reference of farm-specific assets integrated with historical, current, and modeling information. Views of the farm can be summoned to within a square meter with variable time series. Types of information range from historical to actual current conditions to what-if scenarios cast into the future. Because the grower's digital farm can "learn" from his or her queries and interests, the content and services it provides adapt with change and specificity over time.

Most of the machinery on the farm also interacts with the digital farm information services. Autonomous and semi-autonomous machines that plant, cultivate, and harvest crops are precisely controlled with a combination of GPS, distributed functions, and data from the digital farm. Optimal applications of seeds, fertilizers, and chemicals can be controlled and recorded via wireless digital farm services. Similarly, data taken from the field during cultivation and harvesting can be relayed to the digital farm as input for archiving and further use. Together, the estimated levels of data usage in this scenario approach 650 GB/year for a 1000 acre farm (275GB/year for subset data). Extrapolating the subset data volumes to 600,000 acres of Central Valley agriculture zone in California implies a potential distribution of 165TB/year.

3.2 Skilled (10-14 day) Weather Forecasting

Predicting future weather conditions over a particular region requires accurate data and knowledge about atmospheric forces, physical parameters, boundary conditions, and the interrelated nature of the atmosphere to the physical Earth system. While future knowledge will remain incomplete, scientific processes and visionary methods for improving that knowledge promise more accurate forecasts of atmospheric behaviors as technologies and sensing systems evolve. However, the accuracy of weather predictions tend to decay rapidly as a function of time due to the inability of prediction systems to compensate for noise generated by the chaotic nature of the science, a lack of precise initial conditions and the non-linear complexities of weather.

The weather prediction scenario we considered involves testing a 4-D model of the mid-Atlantic region of North America while studying a developing weather condition. The archival system includes the forecast model and the sensor systems used as input. The strategy used in the forecast scenario is to link the sensor systems with the model such that the archive drives the sensor data collection process. . These sensor systems act in concert, as a web of connected, inter-communicating sensors ("sensor web") [3].

As the system collects data, it creates an initial forecast state that it uses at a future time to compare against actual sensor data. The forecast from the model and the sensor data are compared, and model errors identified. The forecast model is then corrected and a new future state created. This cycle occurs periodically based on forecasting requirements. Employing this closely coupled sensor model process allows short term and long range forecasting with minimal error.

From the scientists' perspective, planning sessions are conducted with an interactive visualization interface equipped with collaborative and immersive human-machine technology. Team members have the option to meet virtually via their workstations or in one of the research center's hypermedia tele-immersive conference rooms. In the tele-immersive room, the scientists plan their research forecasts by summoning a vivid holographic 3-D projection of the Earth, pointing to the region of interest, zooming in, and accessing projections of scaled real-time weather conditions.

The scientists cycle through several current satellite views of the region selected from a list and scan each view. Next they request views of the latest graphics and values for temperature, pressure, humidity, and winds superimposed over the satellite image slightly above the defined region on the global reference projection. In order to assess the whole virtual picture of the weather condition, the team requests that the system detach the selected region from the reference globe and project it as a cube presenting a 3-D visualization of the weather conditions to an altitude of 25,000 meters. By rotating the cube the researchers inspect the sensor grid sensitivities over the region from every angle.

The team next decides to run one hour, one day, five day, and ten day forecasts of weather for this region using the current operational model, adding some custom-selected

inputs from a sensor array pick-list. After a minute, the results are ready to be displayed in the same virtual region cube space. The team studies each forecast display by a variety of interactive real-time commands (by voice, gesture, and keyboard). They explore the 4-D visualizations by varying the temporal resolution, zooming spatial areas/volumes to inspect details, requesting displays of simultaneous analysis result visualizations, and selecting predicted parameters for further comparative analysis. Some team members perform dynamic what-if prediction scenarios comparing what the system generates with their own hypotheses.

With this experience the team then formulates a test of their beta version model using insights gained from the immersive collaborative session. Several on the team notice that higher resolution remote sensing values are needed in certain areas of the region to accurately predict future changes of the pending weather condition. This might accord with the deviation of the standard model from theoretical expectations after one day. Furthermore, there is team consensus that coupling their beta model with selected components of the standard model would elucidate new dependencies and parameters crucial to accurate predictions. Scientist-provided specifications for this new research configuration are then interpreted, translated, brokered, and automatically tasked by the system.

In the final episode of this scenario the team studies the emerging weather phenomenon through virtual projections of real-time information and various combinations of modeled predictions. For the modeling portion of the research, the team observes how the standard model self-adjusts its forecasts as a function of near-real time automated comparison of actual versus predicted parameters. When the predicted varies too much from the actual, new initial conditions are set. This continually keeps the predictive accuracy on track for the near term, but progressive adjustments of the model are required. The standard model in this scenario has intelligence applied so it monitors its own performance. With access to a knowledge base, the model may also pinpoint components to be modified either automatically or by human intervention.

In parallel with this modeling activity, the team custom-configures its beta version model. The team includes a system request that re-tasks the sensor web to gather highly detailed inputs for a critical area of the study region, to generate new forecasts. The sensor web schedules and promptly complies with the request, providing critical detailed data for the beta model to process.

Ten days after the start of the research event, the team is able to conclude from their findings that new knowledge was gained about the rare weather condition. Furthermore, comparisons of performance and outcomes between the beta and standard models identify strong points in the beta model responsible for improving the accuracy of overall forecasts. Validation of these findings leads to the promotion of specific beta version components and two external model linkages to the standard model, adding a new phenomenon to the knowledge base with additional predictive power.

Making the above vision possible obviously involves developing new observation sensor systems as well as innovative techniques for data management and utilization. It is anticipated that improvements to existing capabilities combined with evolving infrastructures and innovative research technologies can enable skilled weather forecasts of ten to fourteen days by 2025 (current forecast predictive skill is five to seven days) [3]. Skilled forecasting goals such as this require quality, mixed-resolution observations and data acquisition systems; very rapid processing of observations; complex data assimilation strategies; predictive modeling strategies and algorithms; and powerful technology infrastructures for archiving, distribution, and interactive visualization. An initial assessment of expected optimized global data volumes covering required parameters, temporal/horizontal/vertical resolutions, and vertical measurement layers yields an estimate of about 20TB/day by 2025.

3.3 Empirical Observations

While futuristic scenarios project the needs for research and applications, empirical observations of data access and usage patterns provide a base state and historical trends. They also give hints on how these patterns may change in the future. The access patterns are a function of the requirements of various users and applications as well as the state of technology. The term technology here includes both hardware and software. For example, existence of faster hardware promotes the use of data mining software, which in turn allows different and more useful forms of access from the archives than is currently possible. As visualization tools, network bandwidths, and desktop computing capabilities increase, new requirements may emerge in accessing archived data.

In the initial phase of this study, we have studied patterns of users' access at the Goddard Distributed Active Archive Center (DAAC) since a record exists starting from the DAAC's inception in 1994. More observations at other DAACs and other types of data centers would be useful to provide a broader insight to access patterns. Some of the questions to be addressed by such empirical observations are:

- Should data products be processed routinely and stored for future distribution, or should they be produced only when a user or an algorithm requests them?
- For data-intensive algorithms, should the data be moved to the software, or the software to the data?
- Should architectures be developed primarily based on average data access requirements or peak requirements, and how can peak requirements be characterized?

A key capability implicit in the term Intelligent Data Archive is an awareness that extends beyond the data. While we commonly think of this awareness in its "operational intelligence" context (e.g., resource management, autonomous data gathering), an intelligent archive should also have "scientific intelligence," i.e., the higher-level knowledge that is derived from the data. Clearly, intelligent archives that include models

have some higher-level knowledge about the data. Beyond that, a wealth of knowledge is published in scientific journals. Studying the connection between data in archives of today and the scientific knowledge derived from them will provide valuable hints for the design of future intelligent archives that embed knowledge with data. This initial phase of study includes a "proof-of-concept" attempt at closing the data-knowledge loop using automated (and semi-automated) methods to link datasets from the Goddard DAAC with scientific knowledge resulting therefrom as expressed in publications (limited to those available electronically). Some of the difficulties encountered here provide valuable lessons in current shortcomings in the world of data archives and electronic publication, which offer opportunities for future work.

4. Technology Evolution/Revolution

In the development of data and information systems over the last ten years, significant progress has been made in several areas. These areas include: handling large volumes of data at high rates, distributed computing, archiving and distribution, data and metadata standards to facilitate system interoperability and provision of services such as subsetting, and user interfaces.

In the Earth science domain, this progress is exemplified by NASA's Earth Observing System Data and Information System (EOSDIS) [4] with its distributed set of DAACs and Science Investigator-led Processing Systems (SIPSs), the NASA-initiated federation of Earth Science Information Partners (ESIPs) [5], and the international Committee on Earth Observing Systems (CEOS).

On a more general level, the Global Grid Forum and NASA's Information Power Grid [6] represent efforts to develop persistent networked environments that integrate geographically distributed supercomputers, large databases, and high-end instruments. These resources are managed by diverse organizations in widespread locations, and shared by researchers from many different institutions. Within the Global Grid Forum, the Jini activity [7] is chartered to address the need for a grid framework to support both resource and service discovery, in an environment in which these resources and service providers may enter and leave the grid dynamically, and where diverse protocols are expected to exist.

It is expected that near-term archiving systems will arise from these efforts as well as several commercial developments in hardware and software technologies. We envision that over the longer term, such "grid" infrastructures will evolve into a finer-mesh, perhaps self-organizing "fabric" as computing and communications become increasingly ubiquitous.

The evolution of (and revolutions in) technology over the last twenty-five years demonstrates the difficulty in predicting the technologies that may be available ten to twenty-five years from today. However, a study of existing forecasts by well-known scholars in various areas relevant to data access and management is useful in conceptualizing new architectures for IDA.

Potential technology drivers include processors, microelectronics, nanotechnology, biotechnology, sensors, intelligent systems, communications, and user interfaces. In each of these areas, advances are being made that will have a dramatic impact on future archive architectures and functionality. In the hardware technology areas, the cost per unit capability has been decreasing rapidly and is expected to continue to do so. The implication of this on the end-to-end data management process and data utilization is that it enables implementation of a number of services that have heretofore been limited by hardware costs and encourages experimentation and advances in software techniques. Advances in techniques resulting from research in intelligent systems (including intelligent data understanding) sponsored by NASA and other organizations become suitable for incorporation into the overall data management and utilization process.

4.1 Advances in Storage Technologies

Today we are witnessing the rapid progress and convergence of the fundamental technologies that make up archiving: storage, computing, and communications. Traditionally, digital storage demands have grown at or beyond 60 percent annually. Over the past several years, growth has exceeded 100 percent per year for Internet and e-commerce applications. Data storage functions have undergone an evolutionary change over the past ten years, and are now commonly performed by smaller high-performance disk drives implementing high-availability RAID storage coupled with more capable archiving software. In addition, magnetic tape technology is continuing to increase in capacity and speed. On the other hand, optical storage now seems more oriented toward the entertainment market. Both storage area networks and network-attached storage (SAN and NAS), along with high-speed optical communication, have fundamentally reshaped the traditional storage model. In addition, SAN and NAS archiving strategies have separated storage from being dedicated to any one server and refocused architectural strategies to implement a union of storage devices interconnected by high-speed optical networks.

Even in the near future (i.e., the next five years), the costs per unit of computing, storage, and bandwidth are expected to continue their rapid decline. The historic trend has been that increases in requirements have kept pace with the reductions in per unit cost to maintain roughly the same annual expenditures for hardware. However, in general, the value of an archive system will move from the hardware to the management and utilization of the data. These are what an intelligent archive should aspire to do as performance and functionality increase, especially in a distributed architecture.

Currently, NASA uses both magnetic disk storage (for on-line access to relatively moderate data volumes) and tape storage (for long-term storage and access to large volumes.) The amount of data available on disk has been increasing as disk storage capacity has increased exponentially over the past ten years (over 60 percent annually since 1992). Indeed, some predict that magnetic disk storage will become more cost effective in coming years even for large volumes, as magnetic tape densities have not been increasing so fast as disk. However, while online storage capacity has increased, our ability to access data has not kept pace because input/output performance has only increased linearly [8]. Magnetic recording for both disk and tape will continue to grow at

about 60% annually until the physical barrier (known as the super-paramagnetic limit) is reached.

4.2 Paradigm Shifts

It is also expected that as a result of scientific advances or fundamental limits of nature, paradigm-shifting revolutionary events are likely over the next twenty-five years. For example, quantum mechanics will play an ever-increasing role because it involves the performance of all microelectronic devices and the creation of molecular and atomic size tools. Today's smallest transistor etchings span a mere 130 nanometers. The expected quantum dimension limit for microelectronics is approximately 25 nanometers, where the laws of quantum physics allow electrons to transition across semiconductor gates even when the gates are closed. In other words, the basis for all modern computing technologies will run into a "brick wall."

The effects of these paradigm shifts are illustrated in Figure 2. In the pre-paradigm shift era, we may have extensions to the architectures of today, with increases in the speed and ability to serve data and information to users. However, in the post-paradigm shift era, the nature of the entire end-to-end system could undergo revolutionary changes. This implies that in conceptualizing IDA architectures, it is useful to think in terms of functional capabilities and their necessary interactions and interfaces without being constrained by today's limitations on the locations of such capabilities.

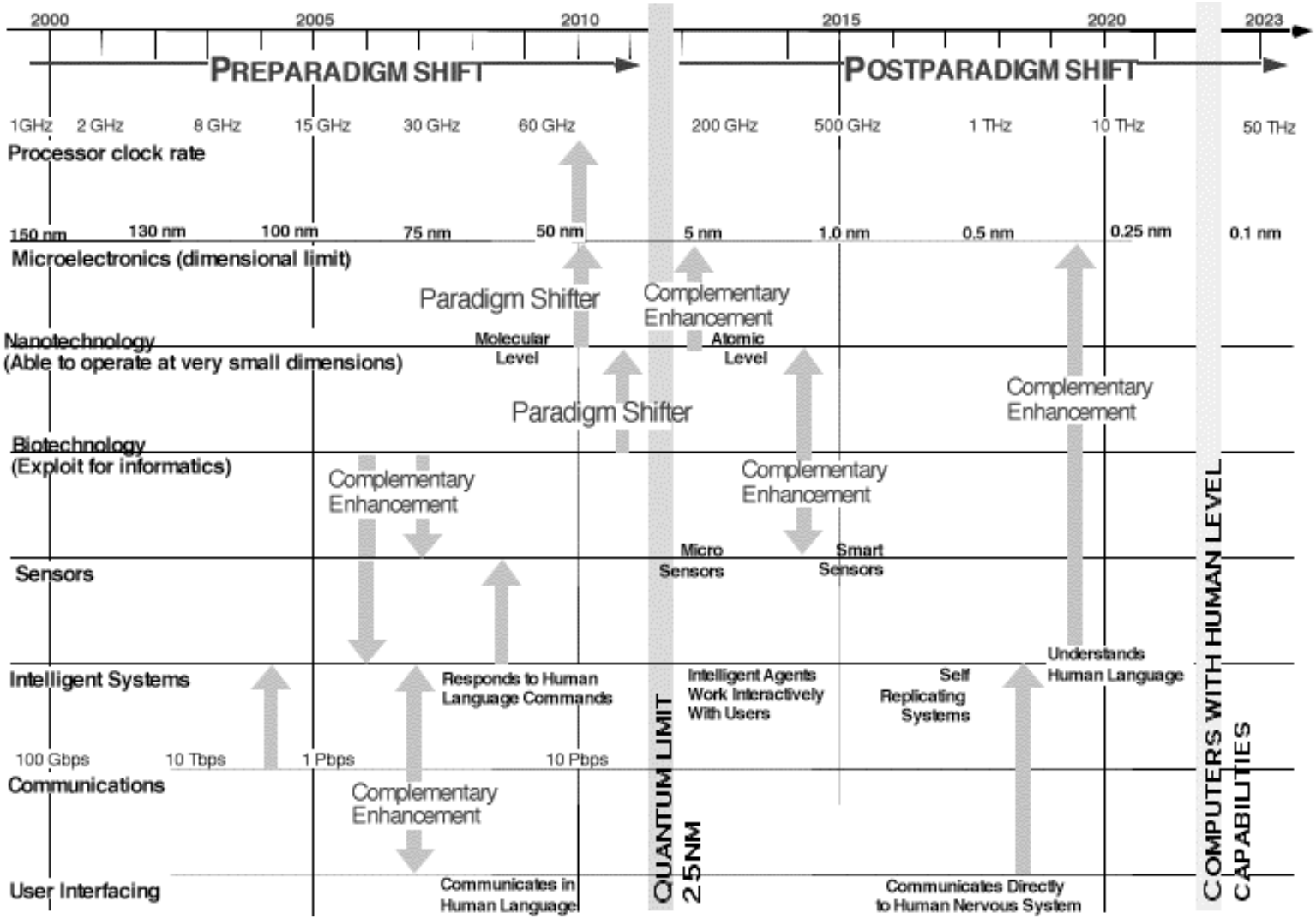

Figure 2: Technology Timeline

5. Recommendations and Future Work

At this stage in our study, we have a set of recommendations shown in the form of a preliminary roadmap to move from archive organizations for traditional data access to intelligent archives that facilitate and take advantage of intelligent data understanding. This roadmap is shown in Figure 3. As shown in this figure, the steps leading to an intelligent archive involve obtaining a better understanding of the following sequence of items:

- Current data access and archiving
- Future data access and archiving trends
- Future scientific applications
- Future enabling technologies
- Roadblocks involved in the formulation, development, and building of an intelligent archive
- Costs involved in formulating, developing, and building an intelligent archive.

There are several areas for further, more detailed, exploration as we continue this study:

- More Scenarios

- It is important to have sufficient scenario diversity to avoid biasing the architecture. Thus, we plan investigation of additional science and applications scenarios in the areas of space science, ecological forecasting, and natural hazards forecasting.

- Specialized Technology

- The initial investigation began with surveying general technologies, such as computing and networking, to determine how they might drive or enable intelligent data understanding. However, there are several areas of more specialized technology, particularly in the area of software, which may be equally important as drivers or enablers. These include areas such as IP-in-Space (enabling a seamless space-ground data system) as well as the various data mining, fusion and visualization technologies being developed as part of NASA's Intelligent Data Understanding program. Advances in science and modeling algorithms are another fertile area.

- Further Architectural Definition

- As the investigation of new scenarios and specialized technologies advances, these should allow further definition and clarification of the IDA architecture. This in turn should promote further definition of the key architectural issues, challenges and trades, which represent an important input into research directions.

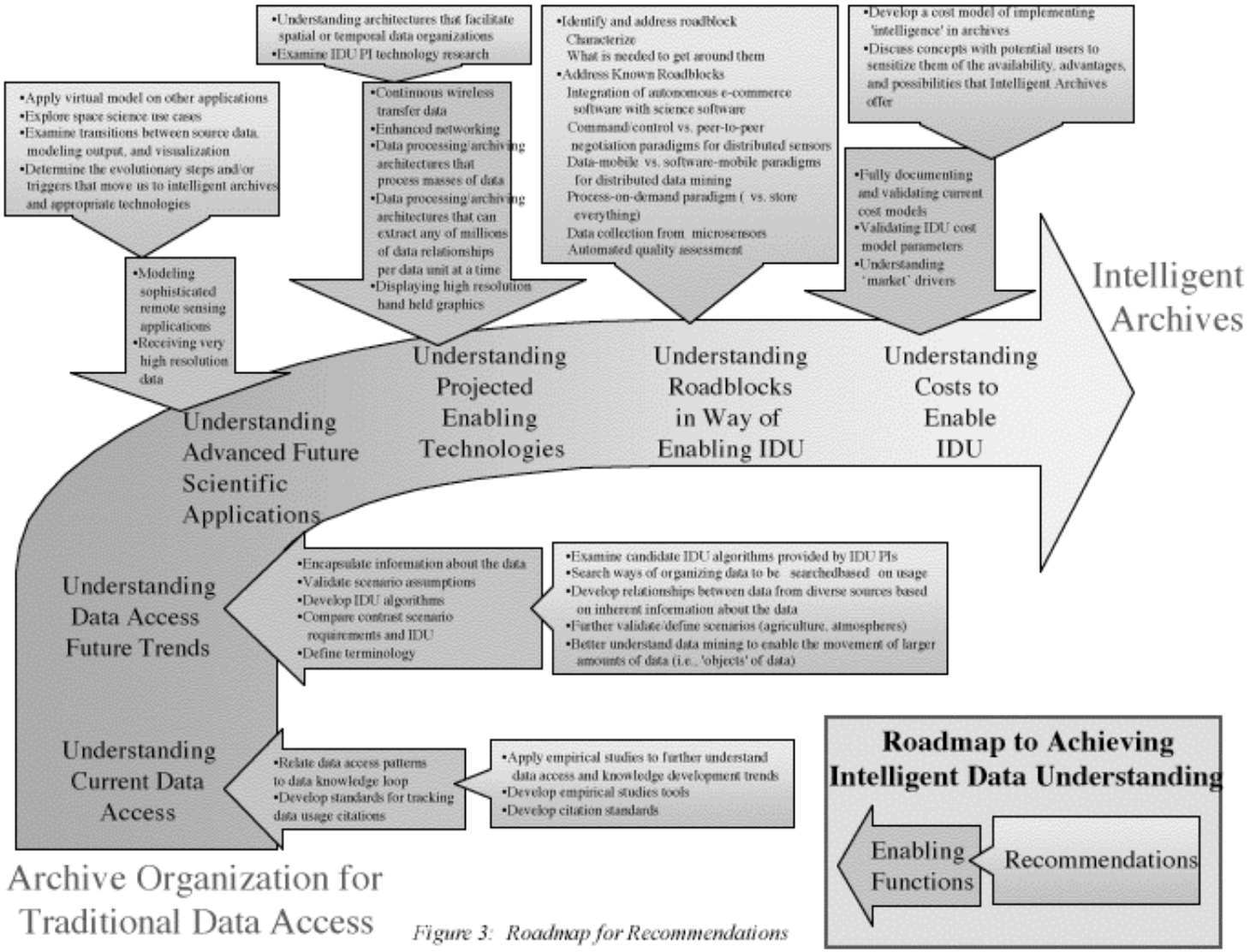

Figure 3: Roadmap for Recommendations

6. Acknowledgements

This study was funded by the Intelligent Data Understanding area of NASA's Intelligent Systems Program. Views and conclusions contained in this paper are the authors' and should not be interpreted as representing the official opinion or policies, either expressed or implied, of NASA or the U. S. Government. The authors would like to thank George Serafino of NASA Goddard Space Flight Center, Kwang-Su Yang of George Mason University, and Randy Barth and Jean Bedet of SSAI for their help with empirical studies, and Lara Clemence of GST for editorial assistance.

References

[1] T. Quatrani and G. Booch. *Visual Modeling with Rational Rose and UML*, (Boston MA: Addison-Wesley, 1998)
[2] Lockheed Martin Advanced Concepts Center and Rational Software Corporation. *Succeeding with the Booch and OMT Methods: A Practical Approach*, (Boston MA: Addison-Wesley, 1996)
[3] M. Steiner, R. Atlas, M. Clausen, M. Kalb, G. McConaughy, R. Muller, M. Seablom, "Earth Science Technology Office (ESTO) Weather Prediction Technology Investment Study," NASA Goddard Space Flight Center, October 5, 2001
[4] G. Asrar and H. Ramapriyan, "Data and Information System for Mission to Planet Earth," *Remote Sensing Reviews*, **13** (1995) 1-25.
[5] The Federation of Earth Science Information Partners, http://www.esipfed.org/
[6] W. E. Johnston, et al. "Information Power Grid," NASA Ames Research Center. Available at http://www.ipg.nasa.gov/aboutipg/presentations/PDF_presentations/IPG.AvSafety.VG.1.1up.pdf
[7] Global Grid Forum. "Charter of the Jini Activity Working Group." March 2001. Available at http://www-unix.mcs.anl.gov/gridforum/jini/charter.pdf
[8] Fred Moore. *Storage IN*fusion. Storage Technology Corporation, 2000.
[9] J. Gray and P. Shenoy. "Rules of Thumb in Data Engineering," Redmond, WA: Microsoft Research Advanced Technology Division, December 1999 (Revised March 2000).