Abstract

Scientific databases have recently become a challenging research area for a number of reasons: 1) the amount of data stored in scientific databases is rapidly increasing, with orders of magnitude increases on the horizon, 2) the data are becoming increasing complex, as more complicated data structures and data relationships must be captured, 3) there is a need to integrate incompatible data formats, commercial databases, and analysis tools into a seamless environment, and 4) scientific databases are becoming distributed, i.e., no single site can archive all the data potentially required to conduct some experiments. Unless these challenges can be met, the scientific researcher will spend an inordinate amount of time manipulating bits and bytes, instead of focusing on the scientific problems of most interest.

In this paper we discuss these database issues in more depth and then describe the Gaea system, a spatio-temporal database management system under development at Worcester Polytechnic Institute. Gaea's main objectives include: 1) providing DBMS support to all phases of scientific investigations by management of scientific data and meta-data, 2) extending database technology with an intrinsic class of operators which is extensible and responds to the growing needs of scientific research, 3) integrating spatial and temporal domains involving very large amounts of data, 4) allowing a clean extension to a distributed computing environment containing heterogeneous, specialized database and computing resources.

KEYWORDS: Distributed databases, Meta-data management, Scientific databases

Distributed Database Management for Scientific Data Analysis*

Nabil I. Hachem, Michael A. Gennert, Matthew O. Ward
Department of Computer Science
Worcester Polytechnic Institute
Worcester, MA 01609, USA
e-mail: {hachem,michaelg,matt}@cs.wpi.edu

January 8, 1999

1 Introduction

As man's influence on the global environment becomes increasingly substantial, there is a growing concern that the ramifications of this influence must be carefully studied [5, 6]. Global warming, the greenhouse effect, the ozone layer, acid rain, and tropical deforestation are just some of the issues requiring detailed analysis. In order to better understand how all of the factors determining the environment interact, many diverse types of information must be gathered and processed. By studying the history of change in particular areas, researchers are better able to form a model of causes and effects in similar regions. Sometimes change is determined primarily by one or two factors; other times the relationships are not so readily apparent, and may include the correlation of climate, hydrology, ecology, biogeochemical systems, solid earth processes, and human interactions [38]. Studies which attempt to correlate a subset of these factors include the analysis of vegetation change after a forest fire [27], changes in land use in Nepal based on deforestation and population increase [28], and estimating the effects of global warming on natural vegetation patterns [30].

Data from a wide range of sources can be brought to bear in the study of global change [38]. Spatial information can be in the form of point locations, lines, regions, grids, and rasters. Potential data sources

*This work is supported by the National Science Foundation under Contract IRI-9116988.

include Advanced Very High Resolution Radiometer (AVHRR), Landsat Thematic Mapper (TM), Coastal Zone Color Scanner (CZCS), Digital Line Graphs (DLG), Digital Elevation Maps (DEM), Land Use and Land Cover (LULC), vegetation indexes, meteorological information, surface weather and hydrologic stations, soil measures, solar radiation, and output of numeric models of ground-water flow or weather forecasting.

As the data can appear in diverse formats, conversion between formats is a necessary component of analysis. In addition, data are available at a variety of spatial scales, thus requiring registration to a common reference point and scale. Much of the data are necessarily temporal in nature as well, since change analysis requires at least two instances of a data set separated in time. Temporal registration (adjustments for seasonal changes) may also be required. Finally, non-spatial information, such as economic data, agricultural statistics, and farm management practices, may be useful to the analysis.

At present, no single database management system can provide the support necessary to perform the sophisticated data analysis experiments described above. The problem is not a lack of data; it is the inability to represent the diverse data types and conduct analyses integrating them. What is needed is an integrated data management and analysis system that can free the scientific researcher from concerns about data types and file manipulation, focusing instead on scientific inquiry.

The Gaea Project is developing the technology to meet this need. We begin by presenting an operational scenario for a hypothetical global change analysis system. We identify research issues that arise, and note where current database technology fails to meet user requirements. We then describe the Gaea system and show how we plan to extend it to meet these requirements, with an emphasis on managing and analyzing distributed scientific data.

2 Scenario of Operation

To clarify the operational characteristics of a scientific data analysis system, we present a hypothetical example based on [39]: A scientist wishes to investigate the foreseeable effects of deforesting Amazonia. Those effects include losses of nutrients, decreased water percolation, increased surface erosion, lower river levels, reduction of evapotranspiration and annual rainfall, and increase of the CO_2-content of the atmosphere (Observation 1 below).

The scientist uses a windowing environment to invoke the visual environment with which she is most comfortable. She then calls a browser

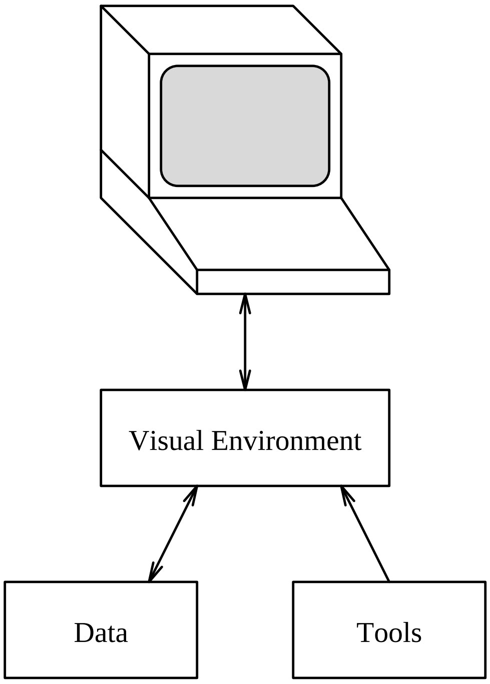


Figure 1: User's View of a Scientific Data Analysis System.

function to find a region of interest, selecting a spatio-temporal "region" or extent, namely the Amazon basin for the last two decades (Observations 2 and 3). In response, the screen displays a set of object classes which represent the data available for the study: such a set will include AVHRR, Landsat TM, Spot and EOS MODIS-N remote sensed data sets in the form of raw images and derived data products. Additional in-situ data sets stored as tables about rainfall measurements, carbon dioxide levels, and rain acidity data are shown to be available (Observations 4 and 5).

The scientist selects Landsat TM data, rainfall data and carbon dioxide data. The display shows, for each set, various available data products and their associated derivation networks (the set of procedures used to derive them). A set of analysis operators, classified by function, is selectable whenever a data type is manipulated. The scientist then applies specific analysis operators to process the selected images, generating for example land-use land classes (LULC) using unsupervised classification techniques. Then, she pulls in rainfall data and computes, using statistical analysis operators, the average rainfall and its standard deviation; for the region over the period of study

(Observation 6).

From the resulting data sets, trend analysis functions are applied to both LULC and rainfall averages to compute vegetation change levels and trends of change in rainfall data. At each step of the process, results are visualized using the Visual Environment (Observation 7). Finally, results are correlated visually and the scientist enters her annotations and conclusions (Observation 8). The whole process is committed and stored as an experiment which another scientist may later recall, replicate, or reuse with a different set of data.

3 Observations

1. In this application domain, the scientist could define the focus problem, which acts as the header of the investigation, as "Is there any correlation between deforestation and CO_2, rainfall levels and acidity of closed water bodies within the Amazon region over the last two decades?" Note that the problem, as defined, is independent of any particular dataset.
2. This study has a natural spatio-temporal extent which is assumed by the scientist.
3. Browsing, the rapid display of low-resolution or aggregate data, may be performed along any of the three data dimensions: space, time, object class (data type).
4. This type of study involves the use of different data types and formats. For example, rainfall levels can be acquired using in-situ measurements at sensing stations within the spatial extent of the study. While carbon dioxide, rainfall data, and acidity can be reported in tabular formats and stored in the data base, deforestation data are not directly available. Deforestation is derived based on satellite remote sensing data. Landsat TM images, with a resolution of 30 meters, are suitable for small-scale analysis. Alternatively, AVHRR data, with a resolution of 1 Km, are suitable for large-scale analysis. The replacement of one data set by another is made difficult, however, by the variety of data formats encountered.
5. Data may be stored at distributed locations, however, the exact location of the data is of no interest to the scientist; her view of the system is shown in Figure 1. A more accurate view of the data is shown in Figure 2. Having access to the data within this framework is important. For example, in the upcoming Earth Observation System [15, 16, 17], satellite imagery will be acquired

and archived at NASA sites within the EOS Distributed Active Archive Centers (DAAC).

6. The derivation of new data is based on other acquired or derived data. In this study, the level of deforestation can be evaluated using AVHRR or Landsat TM imagery for the last two decades. This image data is processed to perform land use land cover classification using, for example, unsupervised classification [13]. The portions of the images that relate to tropical rain forest classes are identified. Finally, changes in coverage of these classes are computed. Alternatively, one can compute the change component within the time series using principal component analysis [11] to qualitatively measure land cover changes. The "forest" component of these images, centered around the Amazon region, is then studied to determine the amount of deforestation.
7. A variety of visual environments may be provided, including versions of commercial or publicly available VEs.
8. Interaction between the scientist and the system will be captured so that necessary parameters, which may have been set interactively, will be available for later use, including the duplication and verification of the experiment elsewhere.

4 Issues

Several research issues arise from the above observations.

1. The amounts of data required are potentially enormous. With EOS expected to generate up to a terabyte of data daily [8, 9], the data requirements of time series analysis can be staggering. No single site can store more than a fraction of the data, necessitating remote access. Much of the data will be archived, leading to long delays; rapid browsing will be critical.
2. The scientific researcher does not care about the details of network access (or even if there is a network!); the network should be transparent. The only noticeable effect of the network should be the delay incurred in running an experiment. Good human factors engineering dictates that the user be made aware of the expected time to complete a given experiment. This is not as difficult as it may at first seem. Every operation should be able, as part of its meta-data, to estimate its execution time. The main reason for doing so is to optimize queries based on expected time to completion, but providing users an estimate of experiment completion time is a useful by-product.

3. Significant optimizations can be made in using the network to reduce the data traffic. For many queries, the required output may only be a single image whose size is relatively small (1 megabyte), whereas the source dataset may be much larger. Rather than extracting a large dataset over the network and massively subsampling it for display, network capacity would be more efficiently utilized if the subsampling could occur at the remote data site.
4. The database and computing environments will both be distributed and heterogeneous. Rather than viewing this as a difficulty, it present an opportunity to improve performance by scheduling computing resources on the most appropriate machines. In particular, specialized hardware, such as supercomputers or array processors can be utilized by an optimizing query processor.
5. A variety of analysis and visualization tools may need to be integrated. Rather than re-write existing packages, it would be better to encapsulate them so that they may accessed from the system.

5 The Gaea Architecture

The Gaea system architecture is designed to meet the needs of scientific research. Our view of a scientific data management and analysis environment can be layered along three levels (Figure 3): 1) The visual frontend, which allows the user to pose visual queries, apply analysis operators to data, and visualize data, including analysis results; 2) the Gaea Kernel, which provides support for meta-data, that is data about the data, and converts simple queries from the visual frontend into a complex series of database accesses and operations; and 3) The Database Backend, which actually stores the data, providing network and archiving functions. We describe each subsystem in turn.

The **Visual Frontend** mediates all interaction with the user. Our objective is to provide sufficient flexibility so that a variety of popular visual environments can be interfaced to the Gaea Kernel. There exists many such packages, either commercial (e.g., AVS [3]) or publicly available (e.g., Khoros [29]). These visual environments come with complete analysis subsystems; we would like to make use of the frontends and analysis operators separately, as shown in Figure 3. In addition, we have written our own visual frontend tailored to the Gaea Kernel [43]. One challenge on which we are currently working is the definition of a query and analysis language in which any visual query can be expressed. When that language is defined and implemented, any visual environment may be incorporated into Gaea by converting commands into the common query and analysis language.

The most important function of the **Gaea Kernel** is the management of meta-data and the semantics of derived data. This semantics is elaborated upon in Section 6. Users can query meta-data to obtain the meaning of derived data. Furthermore, capturing a data object's derivation process information enables the user to repeat that process and derive new data, given different input data. The kernel will include a schema manager which manages the meta-data and the associated derivation semantics and analysis operators (Figure 3). The Query/Analysis Processor (QAP) is responsible for processing queries, deriving new data whenever necessary, and using meta-data. The kernel includes a semantic and meta-data browser to allow a user to find relevant data without knowing specific file and path names. There is also a Data Abstraction Generation and Recall module which allows previously generated data to serve as a template for additional queries, i.e., queries can be abstracted. Finally, generic interfaces to the frontend visual environments and backend distributed computing and distributed databases and archives are provided.

The **Backend System** consists of distributed and archive databases such as Postgres, Object Store and Gemstone ([1], pp. 34–93). The distributed computing environment consists of scientific analysis operators which are available within commercial or public domain software systems. Examples are the analysis tools available within AVS, Khoros, and GRASS. These tools may be imported into Gaea because the meta-data manager will have registered information about analysis operators, their domains of application, data types and formats they apply to, among other meta-data. The Gaea kernel will be able to chose from these available tools and use them to provide a seamless integration between analysis and data management for scientific environments.

6 Meta-Data Management in Gaea

We view scientists as manipulating objects in three orthogonal extents: space, time, and data derivation (object classes). For example, in global change studies, objects have spatial as well as temporal extents. Although these two extents may be correlated, scientists retrieve and manipulate "scientific objects" by viewing those extents as orthogonal. The semantics of the spatial [14, 19, 20, 32] and temporal [2, 10, 31, 34, 36, 40] dimensions have been the subject of much research over the last decade. The third dimension, which has not received much attention so far, is the data derivation dimension, dealing with the derivation procedure followed in the generation of new or existing complex objects.

Meta-data management in Gaea extends semantic modeling technology [25, 33] with the following constructs [21]: concepts, processes, and tasks. Concepts are used to capture entity sets with imprecise definitions. A process captures the derivation procedure of a specific object class, while a task is the instance representing the derivation of a specific scientific data object. We believe that this framework, useful for GIS and global change studies, generalizes well to other scientific fields. The actual meta-data are viewed by the system at three semantic levels (Figures 4 and 5):

6.1 High Level Semantics and the Experiment Level

This level records the information that is necessary for the understanding of a specific experiment. In global change research, it is difficult to agree on carefully designed experiments. The Gaea kernel supports experiments through the experiment manager module of the metadata manager. The experiment manager is capable of manipulating conventional semantic modeling constructs [25]. In addition, we introduce the notion of *concepts*, which may either be base data or data derived from other data according to any of several well-defined algorithms.

A general definition of a **concept** is a representation of a spatio-temporal entity set, extended with an imprecise definition. Concepts are very common in scientific databases. In the context of geographical information systems and global change research, one can effectively cite many examples of concepts.

PERSON is an entity set as defined by the ER model [7], and may be considered a concept with a well defined and agreed upon meaning. But can we define what a DESERT or DESERTIC REGION is? According to [4], an acceptable definition of a desert must include consideration of the following factors: the amount of precipitation received, the distribution of this precipitation over a calendar year, the amount of evaporation, the mean temperature during the designated period, and the amount and utilization of the radiation received. Furthermore, every one of those factors may have different metrics: for example dryness, related to precipitation, can be measured by the Aridity Index, a Quotient of Dryness or the Radiational Index of Dryness [4]. So a DESERTIC REGION is an entity set whose definition may differ from one user to another.

The full semantic model for experiment management is based on *concepts* and the specialized relationship *derived-by*. Derived-by relationships connect concepts to those that are potentially used to derive them. Concepts are arranged in hierarchies, which map into sets of classes at the derivation semantics layer (Figure 5). We are in the process of extending this view to include specialized semantic entities such as scientific investigations, experiments, observations, hypotheses, conclusions, and others (Figure 4). A *scientific investigation* describes the case study being performed. In the described scenario, it captures the essence of the header and is a meta-level aggregation of entities, concepts and relationships such as: *hypothesis* confirmed by *observations* based on *experiments*. *Conclusions* are drawn from *observations*. An *experiment* is the set of tasks that are performed on data representing concepts to derive new data of other concepts. The specialized relationship *derived-by* maps into a set of processes. As there is one process that maps to a specific derived class in the derivation semantics layer, a *derived-by* relation is associated with one concept at the high level semantics layer (Figure 5). The model will include other conventional semantic constructs [25, 33], such as conventional entities, ISA hierarchies, associations and aggregation. This high-level model will provide a desktop manager for scientific investigations, based on a schema-centric view similar to [26]. Our long term objective is to provide an interface so that other high level managers could be integrated.

6.2 Derivation Semantics Layer

This layer provides for the management of (scientific) derivations of data. Concepts map to a set of object classes in the derivation semantics layer. Each class represents a different definition of a concept, based on a specific derivation procedure.

The derivation semantics layer records the derivation relationships among classes of data. Such relationships can also be used for the generation of new data objects for a class. Typically, when data are not stored in the database, we generate the needed data with the help of such derivation relationships. The basic constructs used are: 1) a **Process**, which captures the description of a scientific procedure used for the generation of new concepts from other concepts and 2) a **Task**, which is the instantiation of a process with input data objects. Every task will generate a set of objects (most commonly just one) for the output class.

An example of a process for the derivation of LULC is illustrated in Figure 5. Process P4 derives class C4 which has four attributes: the spatial extent `C4.spatialextent`, the temporal extent `C4.timestamp`, the number of land cover classes `C4.numclass`, and raster image data `C4.data`. The extents are invariantly transferred from the input classes, while the image data is derived using the functional application of the image operators: `unsuperclassify()` and `composite()`[12]. The assertions using the rule `common()` make sure that the spatio-temporal extents of the input classes are the same or overlap. The process definition in Gaea is illustrated in Figure 6.

6.3 System-Level Semantics or the ADT Level

This layer is used to manage the abstract data type (ADT) view of the system. The mapping between the derivation semantics layer and the system layer consists of the mapping of a process as a transformation of a set of input classes to an output class using operators that are applied to primitive classes. This is captured using a data flow network of functional operators that are applied on primitive classes, such as spatial coordinates, temporal attributes, and raster images. The mapping of Process P4 is illustrated in Figure 5.

The mapping we just described captures the structural aspect of the scientific procedure used to derived a concept (LULC in the example). Gaea provides for the actual application of such derivation procedure by dynamically interpreting each operation at run time. Whenever a user requests an LULC output, the Gaea kernel parses the query using the mapping from the top layer to the system level layer. Within that layer, operators are classified according to multiple taxonomies as illustrated in Figure 7. A user can specifically select an operator based on the function it performs; for example a classifier which performs clustering, or a multispectral image function similar to principle component analysis [35]. The specific operator can be implemented in different GIS systems such as Grass and IDRISI, or analysis software such as AVS and Khoros. Furthermore, such operators may have binaries for different architectures. Once a process is edited and specific operators selected, the user does not have to worry about the low level details of which version is used and on which platforms it is applied. The Gaea kernel through its meta-data manager and QAP takes care of this task.

We are currently designing and implementing the low level operator taxonomies in Gaea; GRASS operators are now accessible from within Gaea. To be able to use the taxonomies for distributed data management and analysis, the following types of meta-data must be included:

- Input and output data object types.
- Programming language used, necessary to determine the appropriate run-time environment.
- Experimental/computational. Is the input to the operator intended to be observed data (experimental operator) or a process model (computational operator).
- Hardware platform.

- Software environment needed. Library or other executable packages needed for proper execution.
- Invocation mode. Flags, switches, etc. must be supplied.
- Type of parallelism. SIMD, SISD, MIMD.
- Timing information. Allow query optimization based on expected time to complete an operation.

7 Seamless Integration of Data and Analysis in Gaea

In this Section we discuss how the Gaea architecture is designed to address issues of distributed data analysis and management.

1. By registering meta-data about the very large amount of distributed and archived data, Gaea will be able to present an integrated view of that data.

   Through the schema centric view of the high level semantics and experiments layer, scientist will be able to seamlessly browse data in various formats, select concepts and determine the spatio-temporal extents for which they are available. This is achieved without the knowledge of the location of the data by the scientist.

2. Within the derivation layers, scientists can select an existing process used for the derivation of data and incorporate it within their experimental framework, create a new proces or edit existing scientific procedures which they will use to derive new data; without regard for the data location, actual scientific procedures' code and implementation, thus concentrating on the concepts being used and the functionality of the analysis operators for the experiment on hand.

3. Within the system level semantics layer, users have the option of adding new analysis operators, or selecting existing ones from a multitude of operators available within well known GIS packages such as GRASS. For example, for unsupervised classification a scientist may use the `i.cluster` operator followed by the `i.maxlink` operator in GRASS, or alternatively select commands that produce unsupervised classification implemented within IDRISI or analysis tools such as AVS and Khoros. In this case, users will not have to worry about the hardware platform for which these operators have been compiled and made available. The QAP module, within the Gaea kernel, will use the operator taxonomies (Figure 7) to select and schedule suitable hardware platforms to run the requested analysis in a way which is

transparent to the user. Registered meta-data about operators is further used by the QAP module to optimize the completion of query and analysis in Gaea.

Based on the architecture of Figure 3, the overall effect is to provide a seamless view of the various "legacy" software [13, 42] or more recently developed analysis tools [3, 29]; as illustrated in Figure 1.

8 Conclusions

In the first phase of this project we identified the needs of the global change research community in the areas of data management, operators, and experiments [44]. This led to an extensible design for the Gaea system architecture composed of the visual interface frontend, Gaea kernel, and database backend [23, 24]. We also developed a model of temporal data [34], which is presently being incorporated into Gaea.

A prototype system is nearing completion [22]. The prototype includes spatio-temporal data and operator models, data language, and the visual environment [43]. Preliminary versions of the frontend and kernel have been written and serve as bases for future work. A major accomplishment was the formalization of the semantics of meta-data in a non-ad hoc manner [21].

Ongoing work includes the development of the visual browser and simplified object editors. One challenge on which we are currently working is the definition of a query and analysis language in which to express visual queries. When that language is defined and implemented, any visual environment may be incorporated into Gaea by converting commands into the common query and analysis language. At present, the query and analysis language is an extension of Postquel [41], but functional languages are under consideration.

The third phase will primarily address usability issues, including a history mechanism to cache the current spatio-temporal objects of interest and an entry processor to facilitate the importation of new data types into Gaea. We will also expand the operator set by including operators from other GIS, image processing, and scientific visualization packages.

Future phases will focus on query optimization, network access, and the application of Gaea technology to real-world scientific database problems such as NASA's Earth Observation System Data Information System and NSF's High Performance Computing and Communications Program. Based on our view of the scientific database landscape, the integration of very large databases with powerful analysis systems and intuitive interfaces will prove to be a development of immense importance, ultimately enabling us to answer questions such as, "Is there any correlation between deforestation and CO_2, rainfall levels and acidity of closed water bodies within the Amazon region over the last two decades?" and others.

References

1. "Next-Generation Database Systems," Comm. of the ACM, Special Issue, Vol. 34, No. 10, Oct. 1991.
2. J.F. Allen, "Maintaining Knowledge about Temporal Intervals," CACM, Vol. 26, No. 11, pp. 832-843, Nov. 1983.
3. Advanced Visual System Inc., 300 5th Avenue, Waltham, MA 02154. *AVS Technical Overview*, Oct. 1992.
4. G. Bender, "Reference Handbook on the Deserts of North America," Greenwood Press, p. 594, 1982.
5. Committee on Earth and Environmental Sciences, "Our Changing Planet: The FY 1991 Research Plan of the U.S. Global Change Research Program," 1990.
6. Committee on Global Change of the Commission on Geosciences, Environment, and Resources, National Research Council, "Research Strategies for the U.S. Global Change Research Program," National Academy Press, 1990.
7. P.P. Chen, "The Entity-Relationship Model: Towards a Unified View of Data," ACM Trans. on Database Systems, Vol. 1, No. 1, pp. 9–36, 1976.
8. J. Dozier, "Access to NASA's Earth Observing System," Keynote Address at the 1992 SIGMOD Int. Conf. on Management of Data, San Diego, CA, June 1992.
9. J. Dozier, "Looking Ahead to EOS: The Earth Observing System," Computers in Physics, pp. 248–259, May/June 1990.
10. R. Dugan Jr., "Temporal Abstraction in Global Change Research." M.S. Thesis, Worcester Polytechnic Institute, Aug. 1993.
11. J.R. Eastman, "Time Series Map Analysis Using Standard Principle Components," Proc. of the ASPRS/ACSM/RT 92, pp. 195–205, 1992.
12. J.R. Eastman and J.McKendry, "Change and Time Series Analysis in GIS," UNITAR, 1991.
13. J.R. Eastman, "IDRISI—A Grid-Based Geographic Analysis System (User's Manual)," Clark University, Worcester, MA, Nov. 1990.

14. M. Egenhofer and J. Herring, “A Mathematical Framework for the Definition of Topological Relationships,” Proc. 4th Intl. Symp. on Spatial Data Handling, pp. 803-813, Zurich, Switzerland, 1990.

15. EOS Ground Systems and Operations Office, “Earth Observing System Output Data Products and Input Requirements,” Volumes 1 and 2, NASA Goddard Space Flight Center, 1991.

16. EOS Ground Systems and Operations Office, “Draft EOSDIS Version 0: Architecture and Operations Concept,” NASA Goddard Space Flight Center, 1992.

17. EOS Ground Systems and Operations Office, “Draft EOSDIS Version 0: Functional Requirements,” NASA Goddard Space Flight Center, 1992.

18. C. Fisher and P. Sanderson, “Exploratory Sequential Data Analysis: Traditions, Techniques, and Tools,” SIGCHI Bulletin, Vol. 25, No. 1, pp. 34–40, January, 1993.

19. A.U. Frank, “Properties of Geometric Data: Requirements for Spatial Methods,” In Advances in Spatial Databases, 2nd Symp., SSD’91, Spring-Verlag, Zurich, Switzerland, Aug. 1991.

20. O. Guenther and A. Buchmann, “Research Issues in Spatial Databases,” SIGMOD Rec. **19**-4, pp. 61–68, 1990.

21. N.I. Hachem, K. Qiu, M.A. Gennert, and M.O. Ward, “Managing Derived Data in the Gaea Scientific DBMS,” Proc. Int. Conf. Very Large Databases, to appear Aug. 1993 (also WPI-CS-TR 92-08, Dec. 1992).

22. N.I. Hachem, M.A. Gennert, and M.O. Ward, “An Overview of the Gaea Project,” IEEE Data Engineering Bulletin, Vol. 16, No. 1, pp. 29–32, Mar. 1993.

23. N.I. Hachem, M.A. Gennert, and M.O. Ward, “The Gaea System: A Spatio-Temporal Database System for Global Change Studies,” Proc. NSF Track on Advancement in Data Management for the Scientist and Engineer, Am. Assoc. Advancement of Science Annual Meeting (AAAS ’93), Boston, MA, Feb. 1993.

24. N.I. Hachem, M.A. Gennert, and M.O. Ward, “A DBMS Architecture for Global Change Research,” Proc. ISY Conf. on Earth and Space Science, Pasadena, CA, pp. 186–194, Feb. 1992.

25. R. Hull and R. King, “Semantic Database Modeling: Survey, Applications, and Research Issues,” ACM Computing Surveys, Vol. 19, No. 3, pp. 201–260, 1987.

26. Y. Ioannidis et. al., “Desktop Database Management,” IEEE Data Engineering Bulletin, Vol. 16, No. 1, pp. 19–23, Mar. 1993.

27. M.E. Jakubauskas, “Utilizing a geographic information system for vegetation change detection,” GIS/LIS, pp. 56–64, Apr. 1989.

28. S.M. Jampoler and B.N. Haack, “Use of GIS to identify land use change: Kathmandu, Nepal,” GIS/LIS, pp. 75–84, Apr. 1989.

29. J. Rasure, D. Argiro, T. Sauer, and C. Williams, “A Visual Language and Software Development Environment for Image Processing,” Int. J. Imaging Systems and Technology, Vol. 2, pp. 183–199, 1990.

30. R. Leemans, “Possible changes in natural vegetation patterns due to global warming,” Global Change Database Project-Pilot Project for Africa, Workbook, J.R. Eastman, J. Kineman, R. Dodson, M. Livingston, N. Azimi, ICSU-IGBP and Panel on World Data Centers, 1990.

31. L.E. McKenzie, Jr. and R. Snodgrass, “Evaluation of Relational Algebras Incorporating the Time Dimension in Databases,” ACM Computing Surveys, Vol. 23, No. 4, pp. 501-543, Dec. 1991.

32. B.C. Ooi., “Efficient Query processing in Geographic Information Systems,” In Lecture Note in Computer Science, Spring-Verlag, 1990.

33. J. Peckham and F. Maryanski, “Semantic Data Models,” ACM Computing Surveys, Vol. 20, No. 3, pp. 153–189.

34. K. Qiu, N.I. Hachem, M.O. Ward, and M.A. Gennert, “Providing Temporal Support in Data Base Management Systems for Global Change Research,” Proc. 6th Int. Working Conf. Statistical and Scientific Database Management, ETH, Switzerland, June 1992.

35. J.A. Richards, “Multispectral Transformations of Image Data,” Ch. 6 in Remote Sensing Digital Image Analysis, Springer-Verlag, pp. 127–145, 1986.

36. A. Segev and A. Shoshani, “Logical Modeling of Temporal Data, ” Proc. ACM SIGMOD Conf., pp. 454–466, 1987.

37. M. Shapiro, et. al., “GRASS 4.0 Programmer’s Manual (Draft),” U.S. Army Construction Engineering Research Laboratory, Apr. 1992.

38. L.T. Steyaert, “Investigating the use of geographic information systems technology in the computer workstation environment for global change research,” Proc. 1989 ASPRS/ACSM Annual Convention, Baltimore, pp. 46–54, Apr. 1989.

39. H. Sioli, “Introduction to the Symposium: Amazonia – Deforestation and Possible Effects,” Forest Ecology and Management, 38, pp. 123–132, 1991.

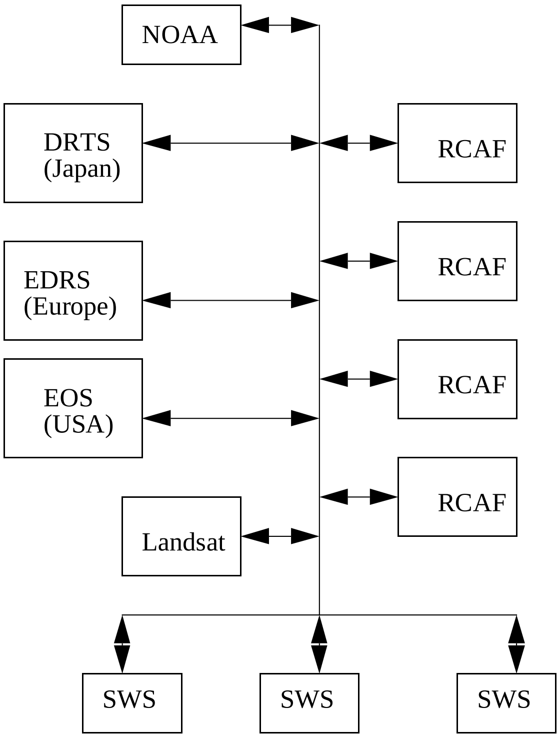
40. R. Snodgrass, I. Ahn, “Temporal Databases,” IEEE Computer, Vol. 19, No. 9, pp. 35–42, Sept. 1986.

41. M. Stonebraker and G. Kemnitz, “The POSTGRES Next-Generation Database Management System,” Comm. of the ACM, Vol. 34, No. 10, Oct. 1991, pp. 78–83.

42. J. Westervelt, “Introduction to GRASS 4,” U.S. Army CERL, July 1991.

43. Y. Zhang, M.O. Ward, N.I. Hachem, and M.A. Gennert, “A Visual Programming Environment for Supporting Scientific Data Analysis,” Proc. Int. Workshop on Visual Prog. Languages, to appear Aug. 1993 (full version available as WPI-CS-TR 93-01, Mar. 1993).

44. Y. Zhou, M.A. Gennert, M.O. Ward, and N.I. Hachem, “Requirements of a Database Management System for Global Change Studies,” Proc. ASPRS/ACSM ’92—Mapping and Monitoring Global Change, Vol. 1, pp. 186–194, Baltimore, MD, Aug. 1992.



SWS: Scientific Workstation
RCAF: Remote Computing and Analysis Facility
EOS: Earth Observation System
EDRS: European Data Relay Satellite
DRTS: Data Relay and Tracking Satellite
NOAA: National Oceanographic and
Atmospheric Administration

Figure 2: A Distributed Scientific Data Analysis Architecture

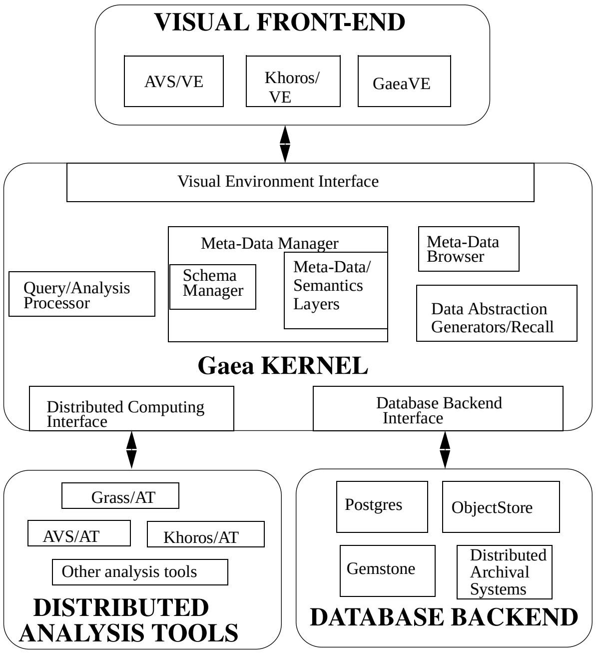


Figure 3: A Visionary Look into the Architecture of Gaea

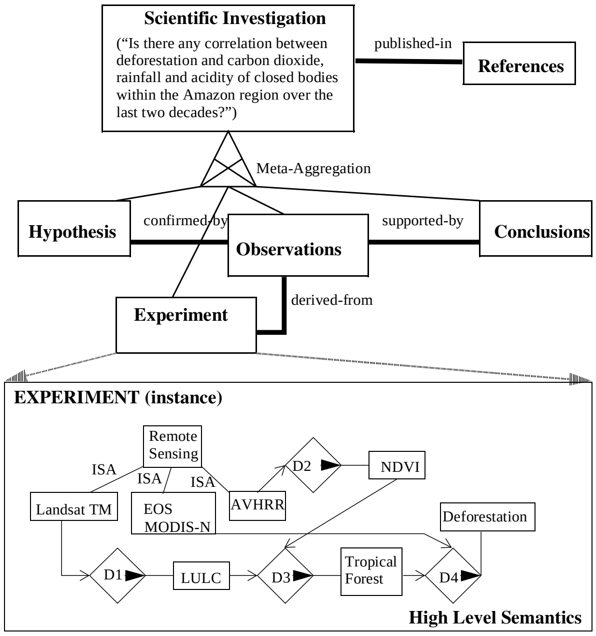


Figure 4: Generic Semantic Model for Scientific Investigations

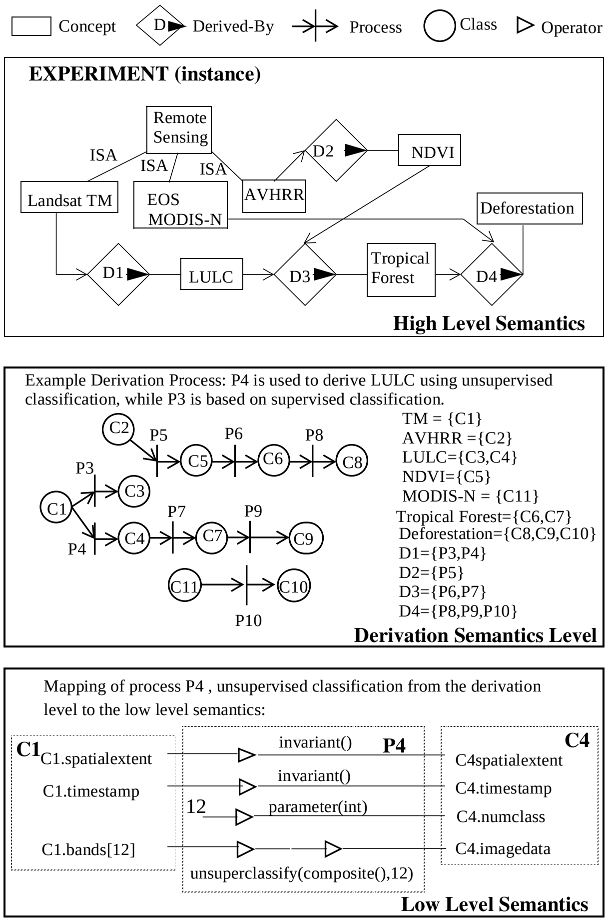
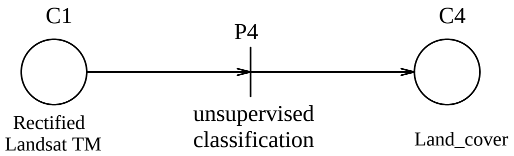


Figure 5: Derivation Management Layers




```
DEFINE PROCESS     P4
OUTPUT    C4
ARGUMENT (  bands  SETOF   C1  )
TEMPLATE {
  ASSERTIONS:
        card ( bands ) = 3;        // need three bands
        common ( bands.timestamp );
        common ( bands.spatialextent );
  MAPPINGS:
        C4.spatialextent = ANYOF bands.spatialextent;
        C4.timestamp = ANYOF bands.timestamp;
        C4.numclass = 12;
        C4.data = unsuperclassify ( composite ( bands ),  12 );
}
```

Figure 6: Derivation Process for Unsupervised Classification

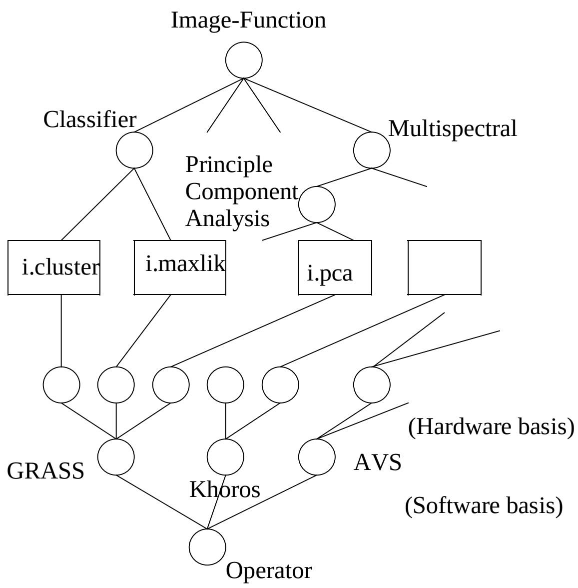


Figure 7: Operator taxonomies