Conf-9505236--1

PNL-SA-26111

ENTERPRISE WIDE TRANSPARENT INFORMATION ACCESS

J. Brown

May 1995

Presented at the
1995 Sybase Users Group Conference
May 15-18, 1995
Dallas, Texas

Prepared for
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Enterprise Wide Transparent Information Access

Jim Brown
Pacific Northwest Laboratory*
Battelle Boulevard
Richland, WA. 99352
(509)375-3626
jc_brown@pnl.gov

ABSTRACT

The information management needs of the Department of Energy (DOE) represents a fertile domain for the development of highly sophisticated yet intuitive enterprise-wide computing solutions. These solutions must support business operations, research agendas, technology development efforts, decision support, and other application areas with a user base ranging from technical staff to the highest levels of management.

One area of primary interest is in the Environmental Restoration and Waste Management Branch of DOE. In this arena, the issue of tracking and managing nuclear waste related to the long legacy of prior defense production and research programs is one of high visibility and great concern. The Tank Waste Information Network System (TWINS) application has been created by the Pacific Northwest Laboratory (PNL) for the DOE to assist in managing and accessing the information related to this mission.

The TWINS solution addresses many of the technical issues faced by other efforts to provide integrated information access to a wide variety of stakeholders. TWINS provides secure transparent access to distributed heterogeneous multi-media information sources from around the DOE complex. The users interact with the information through a consistent user interface that presents the desired data in a common format regardless of the structure of the source information.

The solutions developed by the TWINS project represent an integration of several technologies and products that can be applied to other mission areas within DOE and other government agencies. These solutions are now being applied to public and private sector problem domains as well. The successful integration and inter-operation of both commercial and custom modules into a flexible and extensible information architecture will help ensure that new problems facing DOE and other clients can be addressed more rapidly in the future by re-use of existing tools and techniques proven viable through the TWINS efforts.

Keywords: Environmental Data, Distributed, Heterogeneous, Graphical User Interface, Meta-data, Multi-media.

1. INTRODUCTION

The United States Department of Energy (US-DOE) is responsible for multiple installations distributed throughout the United States. Many of these installations contain waste products that, when released, would negatively impact the environment. Some of the waste sites are: 1) temporary, 2) unsafe, 3) leaking, 4) filled with an undetermined amount and type of radionuclides, and 5) in need of immediate attention. To address these situations effectively, scientists and engineers require timely access to information from multiple databases and data storage systems across the DOE complex.

Common problems facing the US-DOE in this and other areas are similar to those facing most large distributed information intensive institutions. Some of these problems include: 1) excessive data reporting time, 2) multiple data types, 3) disparate report formats, 4) high cost of data, 5) unknown or inconsistent data quality indicators, 6) poor data traceability, and 7) uncontrolled and unsynchronized data replication and duplication.

To address these problematic conditions, the Pacific Northwest Laboratory (PNL) developed the Tank Waste Information Network System (TWINS). The system architecture developed for TWINS is applicable to other

* Pacific Northwest Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC06-76RLO #1830.

information systems that are local, national, or international in scope.

2. BACKGROUND

Impacts of environmental restoration activities, regulatory compliance, and pollution prevention are just a few of the issues facing many commercial, industrial and government entities today. Involvement with environmental issues and their efforts to come to grips with information management needs have many offices scrambling. Management, operators, scientists and engineers require timely access to information from multiple data sources to make informed national and business policy decisions. In the case of the US-DOE, data currently exists in many forms at locations distributed across the DOE complex as shown in Figure 1. Industrial and commercial operations also exist where their information resources are often distributed across wide geographic regions. These systems are also often implemented on heterogeneous hardware, software and network platforms.

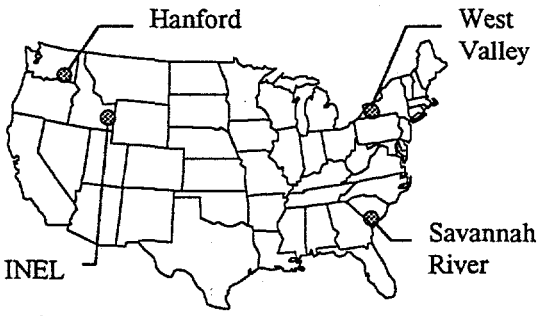

Figure 1 -- Waste Storage Installations

To address one area of US-DOE's information management problems, Battelle Memorial Institute (BMI) and PNL have created an information architecture and access mechanism to a network of information systems that provides immediate access to site data. The system resulting from this task is the Tank Waste Information Network System - TWINS.

TWINS is a computerized geographically-oriented system designed for selecting and accessing information from databases and other sources on various platforms across the US-DOE complex. TWINS provides a standardized database format with an integrated graphical user interface allowing quick, easy, and intuitive access to data. The application centers on a concept of relating data to a geographical reference point or object of interest. The user interface allows the user to select areas or objects of interest ranging from entire installations to individual sites. The user then requests data related to the object(s) from a wide variety of data sources.

While the actual information system described in this paper targets one of the US-DOE's information management problems, it is important to realize this is a special case of a general problem -- that of consistently and effectively managing, identifying, accessing, and presenting many types of complex-wide information without the need for users to know exact details of how to access the data. Also, the issues addressed here go beyond the DOE. Other government, commercial, and industrial installations across the country have similar problems related to a wide range of information types.

For example, the United States Environmental Protection Agency (US-EPA) has established a program to baseline the different ecological systems across the United States. While the application is different, the functionality and system capabilities required are very similar. Many forms of information are distributed across the country. Users of this information will not necessarily look at only one eco-system's data at a time. They will often look at information from many eco-systems at one time to help determine correlations between them. For example, a user may be concerned about how forests harvested within 1000 meters of a wetland affects aquatic and wildlife habitat. To answer this question, the user will need to access forestry and wetland information at the same time

2.1. Problem Scenario

Perhaps the best way to understand how environmental information management can be addressed by TWINS is to walk through the steps in a hypothetical problem case.

Public concern over the environmental impact of a particular installation may prompt a group of constituent minded senators to request an investigation into the situation. Clearly, it is imperative that any information reported back must be completely defensible under public scrutiny. This request would cause a series of events to take place. These events begin at the level of a public or Congressional inquiry into the situation and continue all the way down to specific sampling and data collection activities. After the collection and assimilation of information, decisions are made and activities carried out to identify potential problems. Continuous monitoring

will be performed after any corrective actions to ensure satisfactory resolution of the situation.

Public / Congressional inquiry ==>
==> Agency guidance / directive
==> Field offices
==> Site operators
==> Field data collection
==> Analysis of data
==> Aggregation of data
==> Situation assessment
==> Corrective alternatives assessment
==> Corrective action
==> Response to inquiry
==> Compliance monitoring

2.2. Information Needs By Process Step

Multiple types and volumes of data and information products are generated and/or collected at each step of this process. Some of the likely data produced for each step are listed in Table 1.

Process Step	Potential Information Products
Public/Congressional inquiry	Report or letter to DOE
Agency guidance / directive	Directive to applicable agency Field Offices
Field Offices	Data Quality Objectives (DQO's) Sample and analysis plan
Site Operators	Site identification and general data for the site Preliminary characterization estimates based on historic data
Field data collection	Multiple data types from the field including samples, observations, photographs, videos, audio notes, field note books, etc.
Analysis of data	Validated and verified data points Characterization estimate refinement
Aggregation of data	Data Quality Assessment's (DQA's) Qualified data and reports
Situation assessment	Report on site situation
Corrective alternatives assessment	Corrective alternative selection report
Corrective action	Report on corrective alternative outcome
Response to inquiry	Report on resolved situation
Compliance monitoring	Periodic report on site status

Table 1 – Information Products For Process Steps

These data need to be accessible by others who will use the data to draw their own understanding of the situation. This data may be of several different types including: 1) tabular data, 2) textual data such as reports and documents, 3) photographs, 4) videos, 5) graphs and information visualization products, 6) drawings, and even 7) audio recordings. As well as the 'hard data' associated with the process, other data are generated such as schedules, cost information, risk projections, etc. Being able to capture, manage, coordinate, and present these types of data for many different data categories is critical. Presenting data to users in a timely manner (as required) is the principle focus of the TWINS effort. With access to multiple types and levels of information, decisions can confidently be made based on a more complete understanding of the situation.

3. ISSUES TO BE ADDRESSED

Issues related to providing transparent access to complex information management systems can be categorized in many ways. The most prevalent are: 1) political / cultural and 2) technical.

Political / cultural issues are often harder to overcome than technical issues. These need to be addressed adequately before any developed system can be termed a success. A few examples of political / cultural issues include (but are not limited to):

- Not-Invented-Here:
  The solution was designed and implemented by another organization; therefore, it cannot be of any use to us.
- Procedures and Policies:
  Policies and procedures may be obsolete or in need of review.
- Lack of Automated Systems:
  The institution may not have or be able to afford proper hardware / software infrastructure.
- Insufficient Resources:
  Staff, system, and monetary resources are often not sufficient for organizations to be able to implement the necessary applications.
- Security:
  Adequate protection of both the system(s) and the information may not be provided.
- Expense of Upgrading and Connecting:
  The cost of implementing an automated solution may be prohibitive.
- Information Sharing:

Independent data owners may not be willing to share their information with others.

- Funding Across Independent Groups:
  Funding becomes a larger issue when there are multiple groups / organizations involved.

Issues and challenges of a more technical nature relate directly to the information system architecture and infrastructure. The components of the system can be broken down into three lower level architectures. These include:

Data Architecture

Within the realm of the data architecture, choices can be made among:

- Distributed Databases
- Central Repository
- Replicated Systems

Process Architecture

The process architecture is a logical mapping of all applications and processes that occur within the system infrastructure. Choices here include:

- Fixed Processes For All Sites
- Flexible Processes For All Sites

Technology Architecture

The last component architecture is the technology architecture. Within the technology architecture, decisions need to be made regarding:

- Heterogeneous Hardware/Software Platforms
- Client/Server vs. Centralized
- Communications
  - ◊ Local Area Network
  - ◊ Wide Area Network
  - ◊ Dial-up Modem

4. ASPECTS OF A TRANSPARENT ARCHITECTURE

The complexities of today's information management environments require equally sophisticated architectures to be effective. Some of the characteristics a state-of-the-art (SOTA) architecture must possess include flexibility, expandability, modularity of function, integrated multi-level security, multi-network / application interoperability, user friendly interfaces, linkages between multiple applications, etc. Only a few years ago, it was not possible to deliver these features because of technology limitations. Today's technology and development tools have advanced to the point where these features (and more) are easily developed.

In its most basic form, an information system for transparently managing and tracking data must have three primary components. These include: 1) an interface for users to interact with the system, 2) a global information model that acts as a road map to data storage environments, and 3) one or more storage systems for maintaining the data of interest (Figure 2).

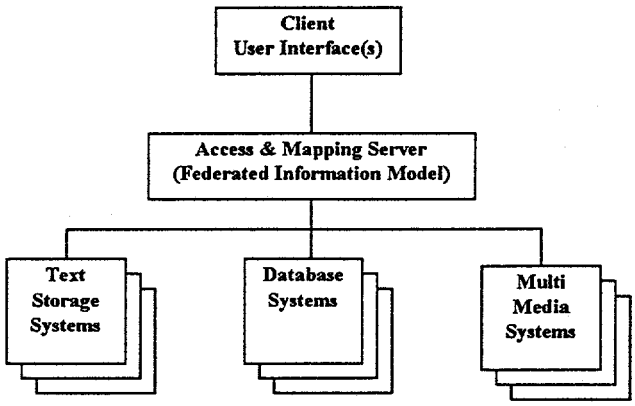

Figure 2 – Basic Three Component Architecture

Multiple business functions within an organization may have the need to evolve independently of others. However, the need and desire to share information across business area boundaries is increasing daily. Even though an area may develop its own systems, coordination of these efforts can enable information sharing, expandability and flexibility thus preparing an organization for future growth. Expanding on the diagram in Figure 2, one can wrap the basic architecture to represent a single business area (Figure 3).

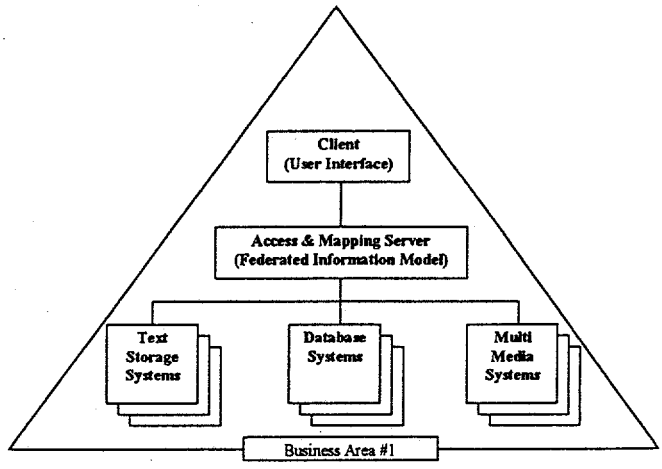

Figure 3 – Business Area Architecture

By replicating the basic architecture and horizontally linking middleware components, business areas can now access / share information with other areas (Figure 4).

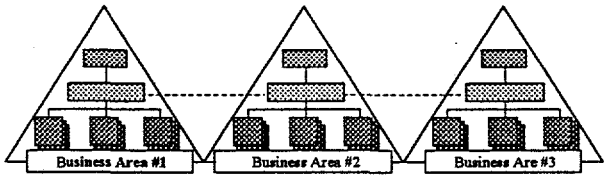

Figure 4 -- Multiple Business Areas

This same expandability paradigm can be applied vertically as well as horizontally to capture the higher level operations and functions of a business entity (Figure 5).

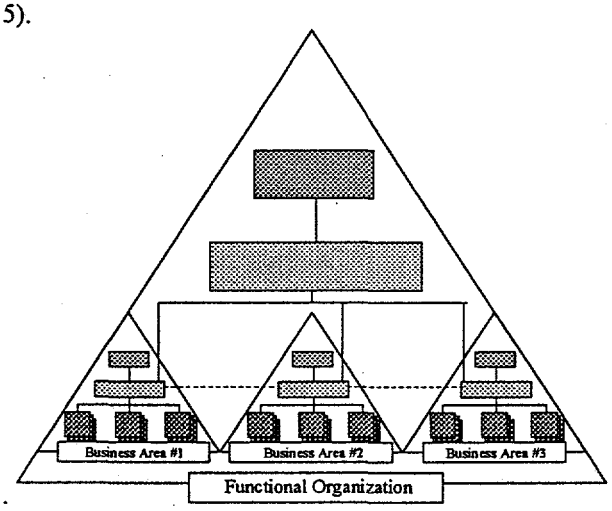

Figure 5 -- Business Area Aggregation

Each component of this architecture has several more layers of complexity designed and built into them to achieve the functionality required by scientists, researchers, management and other users of the system.

4.1 Finer Details of the Architecture

The simple three layer system diagram shown in Figure 2 above can be expanded to show some of the more detailed components. These components are directly related to the capabilities described in Table 2.

4.1.1 Client User Interface

Using terminology present in today's computer industry, the architecture is best described as a three-tier Client/Server architecture. The Client is the software component that resides on the user's workstation. It is composed of several pieces of software as shown in Figure 6. The components of the Client include: the Client application itself, software libraries for the specific database(s) and storage mechanisms, and communications software that accommodates Local Area Network (LAN), and Wide Area Network (WAN) connectivity as well as access via dial-up modems.

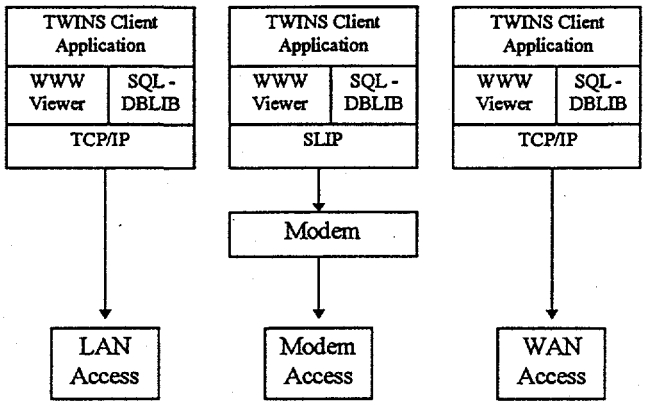

Figure 6 -- TWINS Client Software Configuration

Because of the different types of data in use today, the user interface must be fully multi-media capable. To gain a high-level overall understanding of a situation, users often require access to multiple types of information at one time. It is important that the interface enable the user to transparently access and present these information types simultaneously. This ability helps show relationships between data types and helps the users draw better conclusions based on a more thorough presentation of information.

It is also important for the interface to be as intuitive as possible. The interface is developed around the paradigm of attaching information to a spatially oriented object. This allows the user to select the object(s) of interest and then request information about the selections. The types of information available via the interface are identified in the menu system of the interface. In this manner, the user can select many different categories and types of information about a specific selection. The access to the information is transparent to the user. By that, we mean that the interface forms a general request for information and passes it to the Access and Mapping Server for processing. The user is not required to know where the data is stored. They query the global information model, not the actual data storage systems.

4.1.2 Access and Mapping Server

The first level of Server software is the Access and Mapping Server as shown in Figure 7. This component handles the user access and query control aspects of the architecture. Because of network security reasons, each user is required to have proper authorization to access the system. The access and mapping server validates each user attempting to access the data.

The mapping aspect of this server is the heart of the TWINS' ability to access multiple types of information from distributed, heterogeneous systems. It is also what makes transparent access possible. To accomplish this, it uses a 'Federated' (global) information model to maintain current mappings of individual databases and storage systems into a global information model.

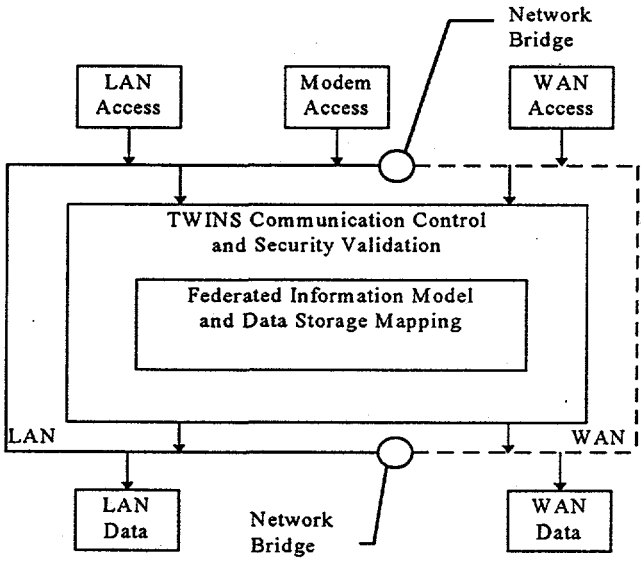

Figure 7 – Access and Mapping Server Configuration

The individual databases to be accessed are first modeled independently. These independent models are then integrated into one global information model. Other multi-media information types, such as photographs, videos, textual documentation, etc., can be treated as file systems or stored directly in the database management system. The meta-data about the multi-media information is captured and can be stored in two places depending on the nature of the access requirements. If it is sufficient to have access to this meta-data without displaying the actual photograph, video, etc., the meta-data is better treated as standard data and stored in the database. If it is always necessary to display the actual multi-media information, the meta-data can be stored in the same system. In the TWINS application, the meta-data is stored in a database. A sub-set of the meta-data is also stored with the multi-media information in a Wide Area Information Server (WAIS). Figure 8 shows briefly how information from multiple systems map into one Federated (global) information model. The information structure of Databases 1 and 2 are integrated into one global information model.

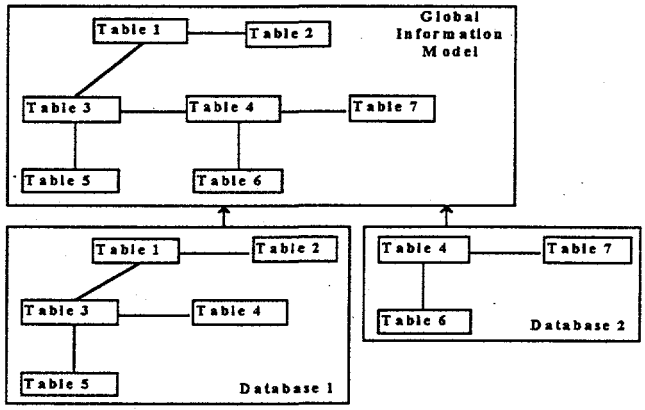

Figure 8 – Global Information Model Mapping

The global model is what the users access when they retrieve data. The global model also contains the mappings into the remote data storage systems. By allowing the users to access information through the global information model, the users are not required to have a thorough understanding of the underlying databases. This allows the users transparent access to information.

Also included in the middleware component of the architecture is all the meta-data required to drive the user interface. This enables the system to add mappings to additional categories of data in the background without the need to deliver an entirely new user interface. Since the interface is totally data driven, it modifies its presentation of options and data availability based on information stored in the middleware data catalogs.

4.1.3 Information Storage Servers

The final layer in the architecture is the actual data / information storage level. Accessible storage systems actually reside in multiple locations across the country. These storage facilities are implemented using three primary technologies: Relational Database Management

Systems (RDBMS), File Management Systems (FMS), and Wide Area Information Servers (WAIS). The types of information contained in these storage systems, as well as the information structures, are integrated into the global model in the access and mapping server layer. Figure 9 shows a typical distribution of data storage.

The three primary layers of the architecture can be brought together into one diagram showing the overall architecture of the system. To do this, the communications aspects must also be included. Two areas of communication are critical parts. The first is the communication between the Client and the Access and Mapping Server. This communication is supported through the use of Local Area Networks (LAN), Wide Area Networks (WAN - InterNet), and through the use of modems. The Transmission Control Protocol / Internet Protocol (TCP/IP) is utilized as the communications protocol on the LAN and WAN. Serial Line Internet Protocol (SLIP) is used for modem communications.

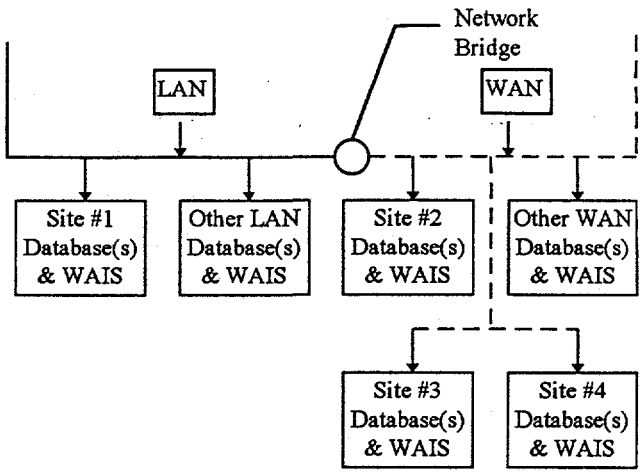

Figure 9 – Typical Data Storage Configuration

The communications between the Access and Mapping Server and the Data Storage Servers are supported through LAN's and WAN's. Figure 10 shows the resulting overall architecture.

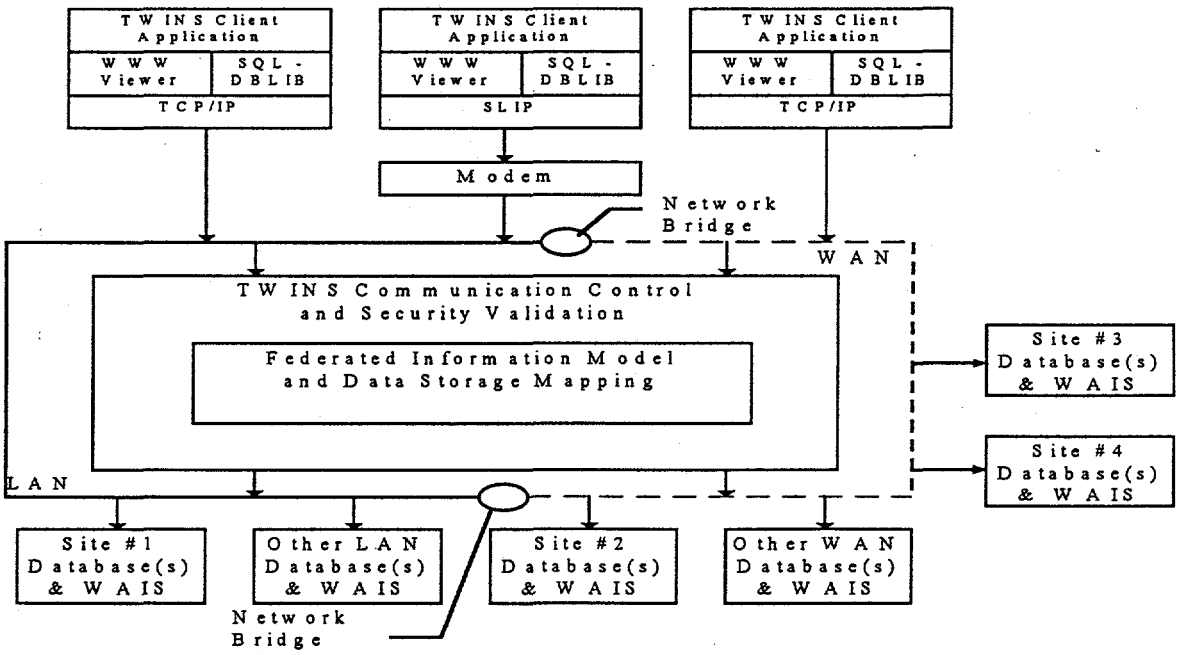

Figure 10 – Overall Technical System Architecture

4.2 Technologies Used

Throughout the computer industry today, there are many terms used to describe features of information systems. Many of these terms describe capabilities that complex information intensive enterprises require to ensure that adequate and appropriate information is available to the correct decision makers in a timely manner. Some of the information system features in today's computer industry include:

- Distributed databases
- Heterogeneous computing environments
- Client / Server applications (C/S)
- Multi-level cross-platform networking
- Graphical User Interface (GUI)
- Multi-media information display
- Geographic Information Systems (GIS)
- Data fusion
- Wide Area Information Servers (WAIS)

Each of these phrases relates to a very important part of overall architecture. Table 2 describes briefly how these features integrate into the environment.

Information System Capability	**How the Capability is Used in the PNL Architecture**
Distributed databases	Information systems are widely distributed across the nation.
Heterogeneous computing environments	Each site may have different computing environments with regard to both hardware and software.
Client / Server applications (C/S)	Applications are full Client / Server applications. The *client* component (the user interface) resides on the end users' desk top. The access and data *servers* are installed on larger systems that actually manage access and store data.
Multi-level cross-platform networking	The architecture makes use of networking in two stages. Users connect to the access servers via modems, local and/or wide area networks. The access server and global information model access the sites' databases via local and/or wide area networks.
Graphical User Interface (GUI)	The primary user interface is a Windows based graphical user interface. The users are not expected to see a command line for accessing information. Instead, they use a pointing device to select the objects of interest and to enter selection criteria.
Multi-media information display	Often times, to gain a better perspective for interpretation of information, other media types, such as photographs, videos, textual documents, etc., provide valuable augmentation to standard database information. The user interface supports the presentation of these media types in conjunction with standard tabular database information.
Geographic Information Systems (GIS)	Most information of interest to the complex has a spatial component to it. Information is linked to the actual location of the objects of interest. It is also important to be able to visualize data with respect to physical locations.
Data fusion	The integrated viewing and analysis of different information types is critical to gaining a better understanding of what the situation actually is. This integration is referred to as 'Data Fusion.' To form a complete picture of a particular situation, it is necessary to be able to fuse tabular data, photographs, videos, etc. .
World Wide Web (WWW) and Wide Area Information Server (WAIS)	The general reference material type of information is stored and retrieved through the use of WWW and WAIS systems. This gives the architecture the flexibility of adding information in different formats and still being able to present that information through one interface

Table 2 -- Information System Components and the PNL Architecture

5. RECOGNIZED BENEFITS OF A TRANSPARENT INFORMATION MANAGEMENT ARCHITECTURE

Applications built using the architectural concepts presented here enhance the ability of researchers, scientists, and management to better understand complex information intensive issues. Users from around the country require up-to-date information to make informed comparisons and decisions. To accommodate having up-to-date data access, each data generating site must have their information in an electronic form that is accessible through a wide area network. Maintaining ownership of data is a critical aspect to getting each site to buy into the concept of a national network of information systems. Having all data housed on one centralized system reduces the comfort level of the data owners because they no longer have direct control over the data for which they are responsible. Each site must maintain control over the system(s) that manage their data. Because of this, PNL has adopted the distributed information system approach of data management. This approach allows PNL-developed systems to provide the following features:

- Site ownership of data
- Builds on systems that are already in place
- Consistent data definition across the complex for new development
- Consistent data presentation of data regardless of the underlying site specific information structure
- A uniform access mechanism to site specific data
- Scalability, built for growth and expansion
- Distributes processing across multiple platforms across the complex
- Ability to use multiple media types to better understand situations by merging tabular data with other data media types.

By implementing systems using the architectural concept identified in this paper, the US-DOE and stake-holders have benefited through:

- Better communications between sites,
- More efficient and/or cost effective assimilation and reporting of data,
- Making the data available to the correct users in a timely manner,
- Better understanding of data resulting in better (more informed) decisions,
- Elimination of data duplication,
- Single point data validation; and
- Site ownership of data and data quality.

While the specific application exemplified in this paper is focused on a specific DOE issue, it is not limited to this purpose. The information model and system architecture of TWINS have been purposefully designed to be general, expandable, and flexible. Other types of installations can be added to the information network. Also, the architecture can be applied to other databases as they are developed and placed on-line

The distributed system architecture is also scalable. The basic system can be applied to a single site or an entire government, commercial, or industrial complex on a localized or world-wide level. Data may be in a structured database or in a less structured file management system. As long as the information in the system is associated with a geographical reference point, the user can access data in expanding queries and relate all significant data to the subject reference point. Ever expanding capabilities of computer access allow the linking of multiple information sources permitting managers to use information from seemingly unrelated files. Linking other data, such as budgeting and scheduling, to scientific data will allow managers the added flexibility of incorporating cost analysis into overall project strategy.

Linking information in this manner and making it transparently accessible to the correct users at the right time (as needed) will enable management, scientists, researchers, and others, to better decide how to manage complex, information intensive situations.