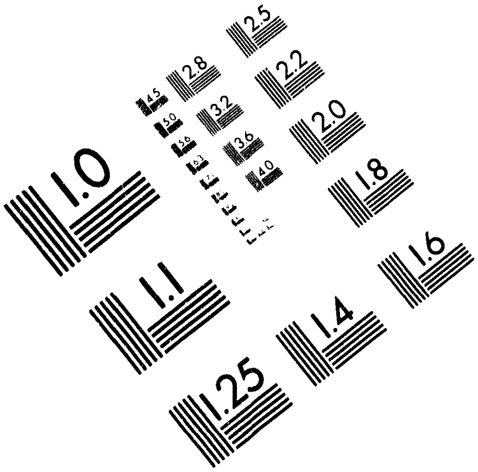
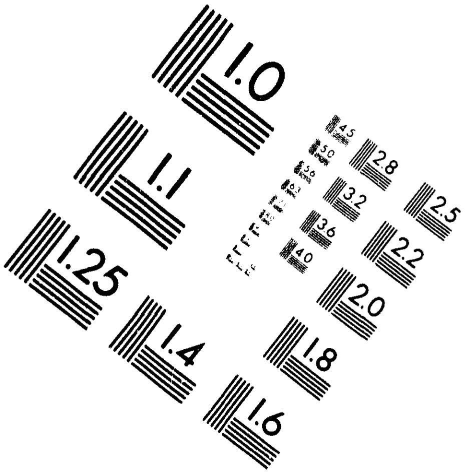


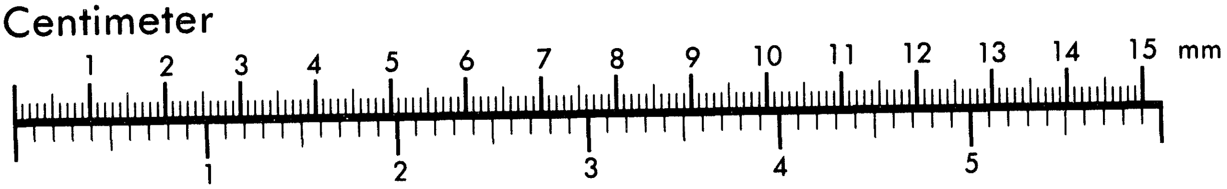
AIIM

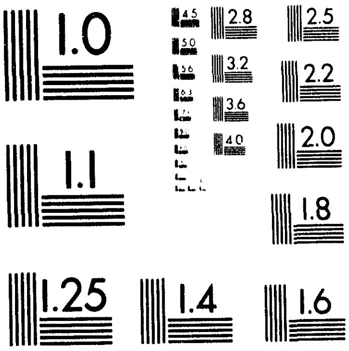
Association for Information and Image Management

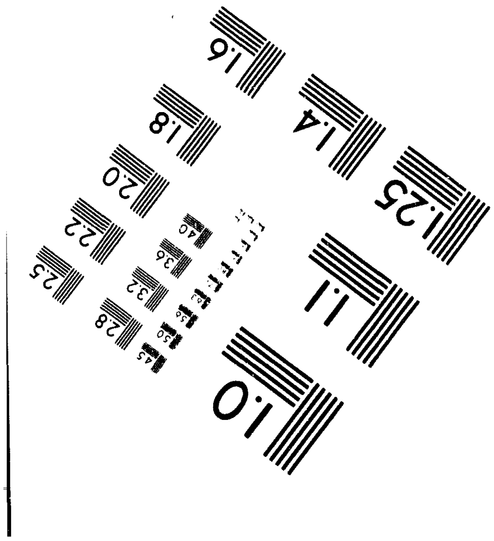
1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

301/587-8202

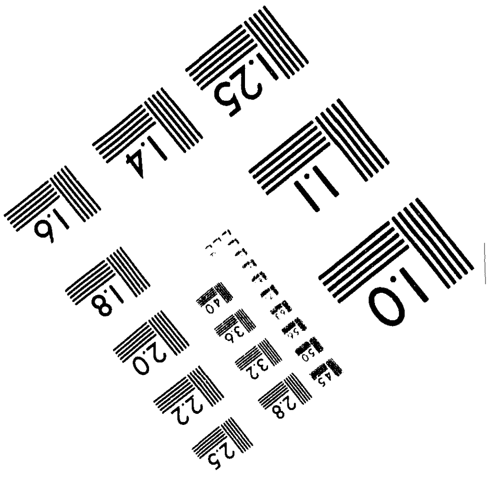








MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.



1 of 1

2

Conf-940873--4

PNL-SA-24743

INFORMATION SYSTEM ARCHITECTURE TO SUPPORT TRANSPARENT ACCESS TO DISTRIBUTED, HETEROGENEOUS DATA SOURCES

J. C. Brown

August 1994

Presented at the
11th Office Information Technology Conference
August 23-25, 1994
Chicago, Illinois

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

Pacific Northwest Laboratory
Richland, Washington 99352

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

8TD

Information System Architecture to Support Transparent Access to Distributed, Heterogeneous Data Sources

Jim Brown
Pacific Northwest Laboratory*
Battelle Boulevard
Richland, WA. 993352
(509)375-3626
jc_brown@pnl.gov

ABSTRACT

Quality situation assessment and decision making require access to multiple sources of data and information. The question facing today's analyst is not so much "Is the data I need to do my work available?" as "How can I get to the data I need to make an informed decision?" Insufficient accessibility to data exists for many large corporations and Government agencies. By utilizing current advances in computer technology, today's situation analyst's have a wealth of information at their disposal.

There are many potential solutions to the information accessibility problem using today's technology. The United States Department of Energy (US-DOE) faced this problem when dealing with one class of problem in the United States. The result of their efforts has been the creation of the Tank Waste Information Network System -- TWINS.

The TWINS solution combines many technologies to address problems in several areas such as User Interfaces, Transparent Access to Multiple Data Sources, and Integrated Data Access. Data related to the complex is currently distributed throughout several US-DOE installations. Over time, each installation has adopted their own set of standards as related to information management. Heterogeneous hardware and software platforms exist both across the complex and within a single installation. Standards for information management vary between US-DOE mission areas within installations. These factors contribute to the complexity of accessing information in a manner that enhances the performance and decision making process of the analysts.

This paper presents one approach taken by the DOE to resolve the problem of distributed, heterogeneous, multi-media information management for the HLW Tank complex. The information system architecture developed for the DOE by the TWINS effort is one that is adaptable to other problem domains and uses.

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

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

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 radio-nuclides, and 5) are 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, and 6) poor data traceability.

To address these problematic conditions, the Pacific Northwest Laboratory developed the Tank Waste Information Network System (TWINS). The system architecture developed for TWINS is applicable to other information systems, on a local, national, or international in scope.

2. BACKGROUND

A primary issue facing many industries and several government agencies involved with environmental issues is their efforts to come to grips with the information management. Management, scientists and engineers require timely access to information from multiple data sources to make informed national 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.

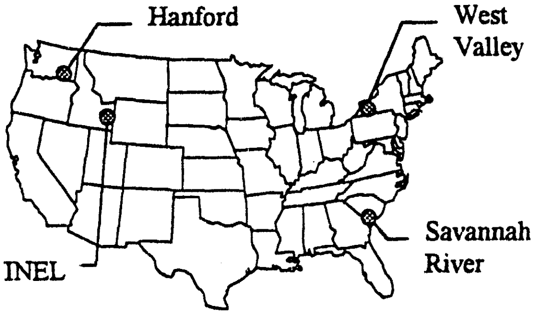


Figure 1 – Waste Storage Installations

To address one area of US-DOE's information management problems, Battelle and the Pacific Northwest Laboratory (PNL) have created an access mechanism to a network of information systems that provide 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 TWINS network centers on a concept of relating data to a geographical reference point or object of interest. The TWINS 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 information system described in this paper, TWINS, targets one of the US-DOE's information management problems, it is important to realize this just a special case of a general problem -- that of managing, identifying, accessing, and presenting many types of information across the DOE complex. Also, these issues addressed by TWINS 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 that will establish a baseline assessment of 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-systems' data at one time. The users will often look at information from many eco-systems at one time to help determine correlation's between them. For example, a user may be concerned about how forestry harvests within 1000 meters of a wetland effects 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.

Because of public concern over the environmental impact of a particular installation, a group of constituent minded Senator's request an investigation to look into the situation and to report back to them. Clearly, it is imperative that any information reported back must be completely defensible under public scrutiny. This request would a series of events to take place. These events start at the level of a public or Congressional inquiry into the situation 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 correct potential problems. Continuous monitoring will be performed after any corrective actions to ensure satisfactory resolution of the situation.

2.2. Information Needs By Process Step

In each step of this problem scenario, multiple types and volumes of data and information products are generated and/or collected. These data need to be accessible by others to enhance 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. The ability to capture, manage, coordinate, and present these types of data for many different data categories is critical. Timely presentation of data to users is the principle focus of the TWINS effort. With access to multiple types and levels of information, decisions are made confidently based on a more complete understanding of the situation. Some of the data likely to be produced at each step are listed in Table 1.

Process Step	**Potential Information Products**
Public or Congressional inquiry	Report or letter to DOE
Agency guidance or directives	Directive to applicable agency Field Offices
Field Offices	Data Quality Objectives 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 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 Qualified data and reports
Situation assessment	Report on site situation
Corrective alternatives assessment	Corrective alternative selection report

Table 1 -- Information Products For Process Steps

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 (Cont.)

3. Issues To Be Addressed

Issues related to complex information management systems, like TWINS, may be categorized in many ways. The most prevalent are 1) Political / Cultural and 2) Technical.

Political / Cultural issues are often time 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 some other 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.

Along with these issues are those of a more technical nature. These issues and challenges 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 Platforms (Hardware and Software)
- Client / Server vs. Centralized vs. Replicated Systems
- Communications
  - ◊ Local Area Network
  - ◊ Wide Area Network
  - ◊ Dial-up Modem

4. The TWINS Architecture

In its most basic form, an information system for 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. This simplistic view is represented in Figure 2.

Each component of this architecture in the TWINS environment has several more layers of complexity designed and built into them. These more complex layers have been implemented to achieve the functionality required by scientists, researchers, management and other users of the system.

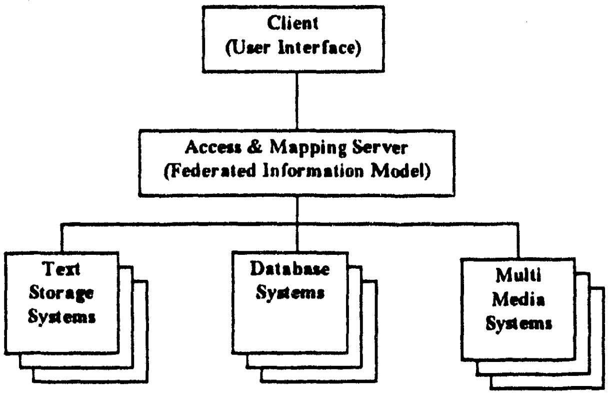


Figure 2 – Basic Three Component Architecture

4.1 Details Of The TWINS Architecture

The simple 3 layer system diagram shown in Figure 2 above can be expanded to show some of the more detailed components. Each of these components is directly related to the capabilities described in Table 2 above to show how these technical components tie together into a complete system.

4.1.1 Client User Interface

Using current terminology present in today's computer industry, the TWINS interface is described as a Client / Server / Server system. The Client is the software component that resides on the user' workstation. The Client is composed of several pieces of software as shown in Figure 3. The components of the Client include 1) the Client application itself, 2) software libraries for the specific database(s) and storage mechanisms, and 3) communications software that accommodates Local Area Network (LAN), and Wide Area Network (WAN) connectivity as well as access via dial-up modems.

The users interface developed for TWINS is a fully multi-media capable interface. To gain a high level overall understanding of a situation, scientists and engineers often require access to multiple types of information. It is important that the user interface enable the user to 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

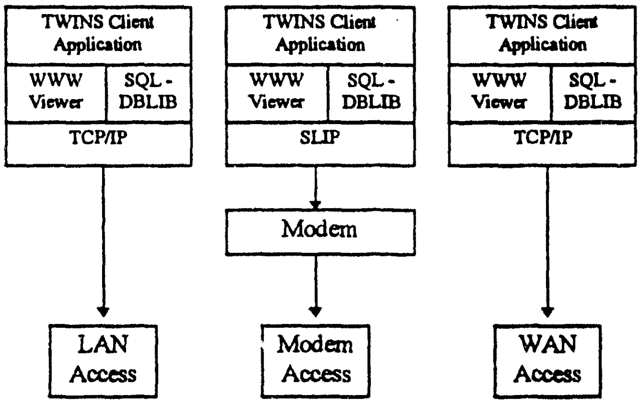


Figure 3 – TWINS Client Software Configuration

information.

It is also important that the user interface be as intuitive as possible. The TWINS 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 what TWINS calls the Access and Mapping Server as shown in Figure 4. This component handles the user access and query control aspects of TWINS. 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 that attempts 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. To accomplish this, TWINS uses a 'Federated' (or Global) information model to maintain current mappings of individual databases and storage systems into a global information model.

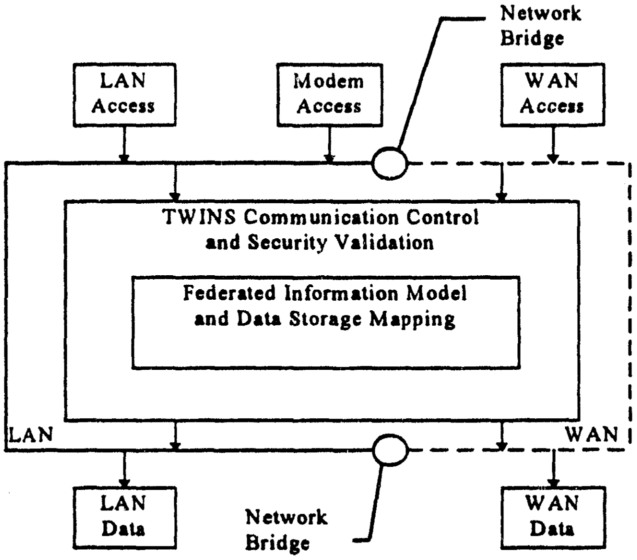

Network Bridge
LAN Access
Modem Access
WAN Access
TWINS Communication Control and Security Validation
Federated Information Model and Data Storage Mapping
LAN
WAN
LAN Data
Network Bridge
WAN Data

Figure 4 -- 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 5 shows briefly how information from multiple systems maps into one Federated (global) information model. The information structure of Databases 1 and 2 are integrated into one global information model.

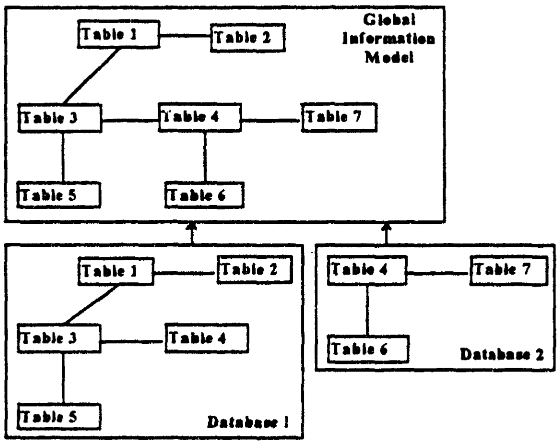


Figure 5 -- Global Information Model Mapping

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.

4.1.3 Information Storage Servers

The final layer in the TWINS architecture is the actual data / information storage level. Storage systems that TWINS accesses reside in multiple locations across the country. These storage facilities are implemented using three primary technologies: 1) Relational Database Management System (RDBMS), 2) File Management Systems (FMS), and 3) Wide Area Iı.ıormation 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 6 shows a typical distribution of data storage for TWINS.

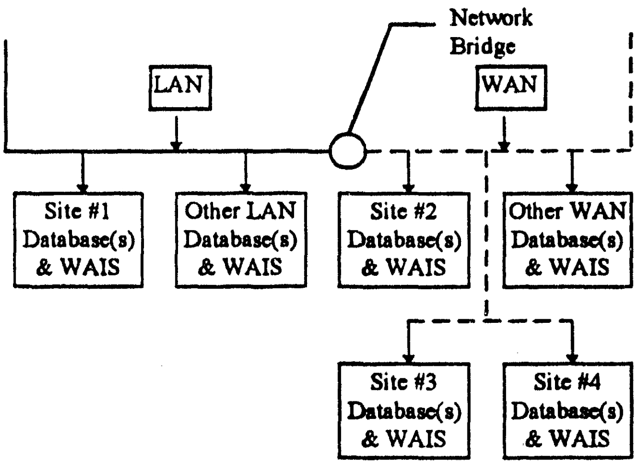

Network Bridge
LAN
WAN
Site #1 Database(s) & WAIS
Other LAN Database(s) & WAIS
Site #2 Database(s) & WAIS
Other WAN Database(s) & WAIS
Site #3 Database(s) & WAIS
Site #4 Database(s) & WAIS

Figure 6 – Typical Data Storage Configuration

The three primary layers of TWINS 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 integral parts of TWINS. 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.

The communications between the Access and Mapping Server and the Data Storage Servers are supported through LAN's and WAN's. Figure 7 shows the overall architecture of TWINS.

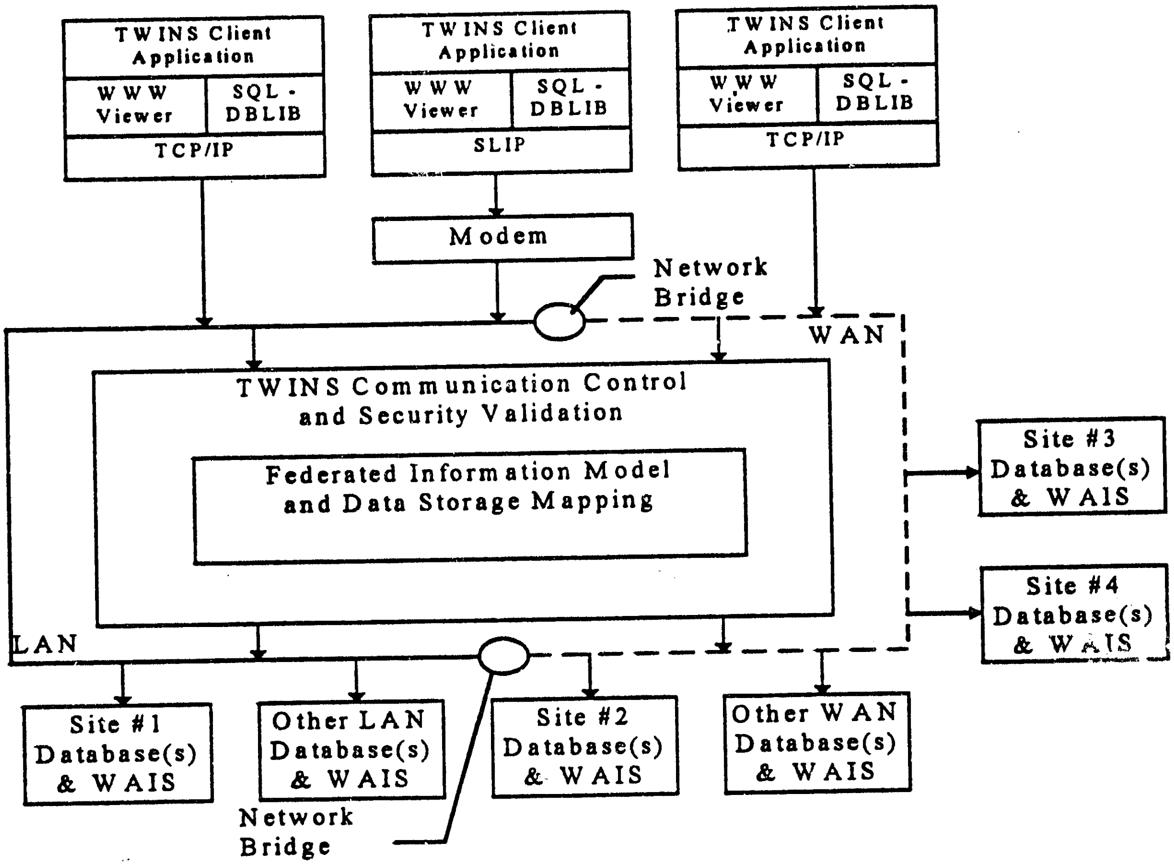

TWINS Client Application
WWW Viewer
SQL - DBLIB
TCP/IP
TWINS Client Application
WWW Viewer
SQL - DBLIB
SLIP
Modem
TWINS Client Application
WWW Viewer
SQL - DBLIB
TCP/IP
Network Bridge
WAN
TWINS Communication Control and Security Validation
Federated Information Model and Data Storage Mapping
Site #3 Database(s) & WAIS
Site #4 Database(s) & WAIS
LAN
Site #1 Database(s) & WAIS
Other LAN Database(s) & WAIS
Site #2 Database(s) & WAIS
Other WAN Database(s) & WAIS
Network Bridge

Figure 7 -- TWINS Technical System Architecture

4.2 Technology Used In TWINS

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 TWINS. Table 2 describes briefly how these features integrate into the TWINS environment.

Information System Capability	**How The Capability Is Used In The TWINS Environment**
Distributed databases	TWINS accesses information systems that are widely distributed across the DOE complex.
Heterogeneous computing environments	Each site may have different computing environments with regard to both hardware and software.
Client / Server applications (C/S)	The TWINS application is a full Client / Server application. 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	TWINS makes use of networking in two stages. Users can 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 TWINS user interface is a full Windows based graphical user interface. The users are not expect to ever 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 TWINS 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 them. Information in TWINS 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 are 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. .
Wide Area Information Server (WAIS)	The general reference material type of information available via TWINS is stored and retrieved through the use of WAIS systems. This gives TWINS 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 TWINS Environment

5. Benefits Of TWINS And The TWINS Architecture

TWINS enhances the ability of researchers, scientists, and management to better understand issues related to the US-DOE's environmental restoration and waste management programs. Scientists and researchers from around the US-DOE complex require up to date information to make comparisons between similar installations -- those with like or similar processes. 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 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 needs to maintain control over the system(s) that manage their data. Because of this, TWINS has adopted the distributed information system approach of data management. This approach allows TWINS 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 complex
- Ability to use multiple media types to better understand situations by merging of tabular data with other data media types.

6. Conclusions

By implementing TWINS, the US-DOE and stake-holders have benefited through:

- Better communications between sites and DOE-HQ;
- 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 TWINS effort has concentrated on the US-DOE issues, 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 TWINS architecture can be applied to other databases as they are developed and placed on-line

The distributed system architecture of TWINS is also scaleable. The basic system can be applied to a single site or an entire US-DOE or industrial complex on a localized or world-wide level. TWINS gives a user the ability to pick a specific location and access pertinent relative data. The 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 accessible to the correct users will enable management, scientists, researchers, and others, to better decide how to manage complex, information intensive situations.

DATE
FILMED

10/17/94

END