Integration of Distributed and Heterogeneous Information for Public-Private Policy Analyses

David Su-Kai Cheng

Working Paper CISL# 2004-06

June 2004

Composite Information Systems Laboratory (CISL)
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02142

Integration of Distributed and Heterogeneous Information for Public-Private Policy Analyses

by

David Su-Kai Cheng

Submitted to the Engineering Systems Division
and the Department of Electrical Engineering and Computer Science

May 13, 2004

in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Technology and Policy
and Master of Science in Electrical Engineering and Computer Science

ABSTRACT

Databases and analysis tools currently being used to study carbon dioxide capture and storage (CCS) options are managed by diverse organizations and are heterogeneous in format. Tools to study the various components of a CCS system have been developed in several fields including chemistry, geology, and economics. Data being used to run analyses are being obtained from an equally diverse set of organizations, from data collected for environmental assessments to data on oil and gas exploration. These variations in tools and data cause complications in systems-level analyses, resulting in additional effort expended in data collection and opportunities for human error.

A geographic information system has been implemented to automate and support robust studies of both component and system options. Context management and information integration techniques have been designed into the system. The system improves the availability and quality of information by automatically managing the distributed and heterogeneous data sources. The resulting information is being used to advance research and development of CCS systems through efforts such as the NETL sponsored Regional Carbon Sequestration Partnerships. This paper will present an overview of the system and initial results of its application to CCS-related data.

Thesis Supervisors:
Howard J. Herzog
Title: Principal Research Engineer, Laboratory for Energy and the Environment
Stuart E. Madnick
Title: John Norris Maguire Professor of Information Technology
and Professor of Engineering Systems

Table of Contents

Chapter 1: Introduction

Section 1.1: Motivation

An increasing amount of complex and diverse data from distributed sources are being used in analyses of carbon dioxide (CO_2) capture and storage (CCS) systems (systems in which CO_2 are captured from sources, redirected, and stored in non-atmospheric sinks in order to reduce the levels of CO_2 in the atmosphere). In order to make these data coherent and understandable to decision makers, we will incorporate information technologies such as context mediation and information integration into a information system. The resulting Distributed Information Management System (DIMS) will be used to inform future policy decisions.

Government, industry, academic, and non-government organizations are collaborating in the research, with each group specializing in the development of analyses and data sources that relate to different aspects of CCS. They are utilizing tools from many fields of research including chemistry, geology and economics, as well as developing new analysis techniques for understanding CCS costs and project options. The data is likewise being collected from many different origins ranging from geologic exploration to environmental regulation.

Available tools for CCS analysis are component-based, specific for a particular piece of the CCS framework. System-level connections and considerations are left as work for human analysts. Current efforts in CCS are working to combine these components into more complete analyses, but these tools and the associated data are disperse in physical location, managed by different groups, and diverse in context. For example, characterizations of geologic reservoirs, emissions, and geography are necessary for the full analyses, but are administered by different groups. General geologic information is maintained by the US Geological Survey while specific geology of reservoirs is maintained by the Department of Energy. Emissions information is collected by the Environmental Protection Agency (EPA). Additionally, various individual research groups provide specialized data that supplements these basic data.

Unfortunately, these databases are not available in a consistent format and must be gathered together to provide the required information. The process of gathering and coordinating data has been done manually by a number of groups for their individual research but not for general use. Because this process is both time-consuming and error-prone, it is important to develop and distribute automated mechanisms to gather relevant data from diverse data sources.

The work in this thesis considers the application of context management and information integration technologies to the data available for CCS analysis. Through development of a specific Geographic Information System (GIS), we are able to explore the best methods for integration in this field. The improved data and the integration methods can be applied to other projects to improve the consistency, usability, and quality of CCS-relevant information.

Section 1.2: Carbon Dioxide Capture and Storage

The use of fossil fuels in human activities such as electricity generation, industrial processes, transportation and residential heating generates CO_2. Research suggests that this anthropogenic CO_2 may cause global climate change, driving changes in weather patterns, the sea level, agricultural compatibility, and oceanic acidity [Webster et al, 2002]. Due to these concerns about global climate change, industrial and governmental organizations are considering various strategies to reduce anthropogenic CO_2 emissions.

While most of the public has heard about reducing CO_2 emissions through measures such as improved efficiency and use of hydrogen fuel cells, most have not heard about the option of CCS [Curry, 2003]. CCS refers to technologies that capture CO_2 and redirect it to non-atmospheric storage reservoirs, called sinks. The major components of CCS are capture from sources, transport to sink, and storage in sinks. Capture includes removing CO_2 from emission streams, purifying it into a sufficiently high concentration, and compressing it for transport. Transportation of CO_2 may be done through pipelines, or shipping of refrigerated containers. Storage includes preparing a site, injecting CO_2, and monitoring storage integrity. Additional steps in storage may include managing the long term integrity of the sinks.

Figure 1.1 shows the breakdown of CO_2 emissions in the U.S. from electric power, transportation, industry, commercial, and residential sectors. Research focus in capture is on the electric power and industrial sectors. These sectors contribute 57 percent of emissions and offer cost-effective targets for CCS technologies. Technologies exist to capture CO_2 from the large and stationary facilities represented in the fleets of these sectors. Costs are significantly larger for capture technologies dealing with smaller or mobile sources.

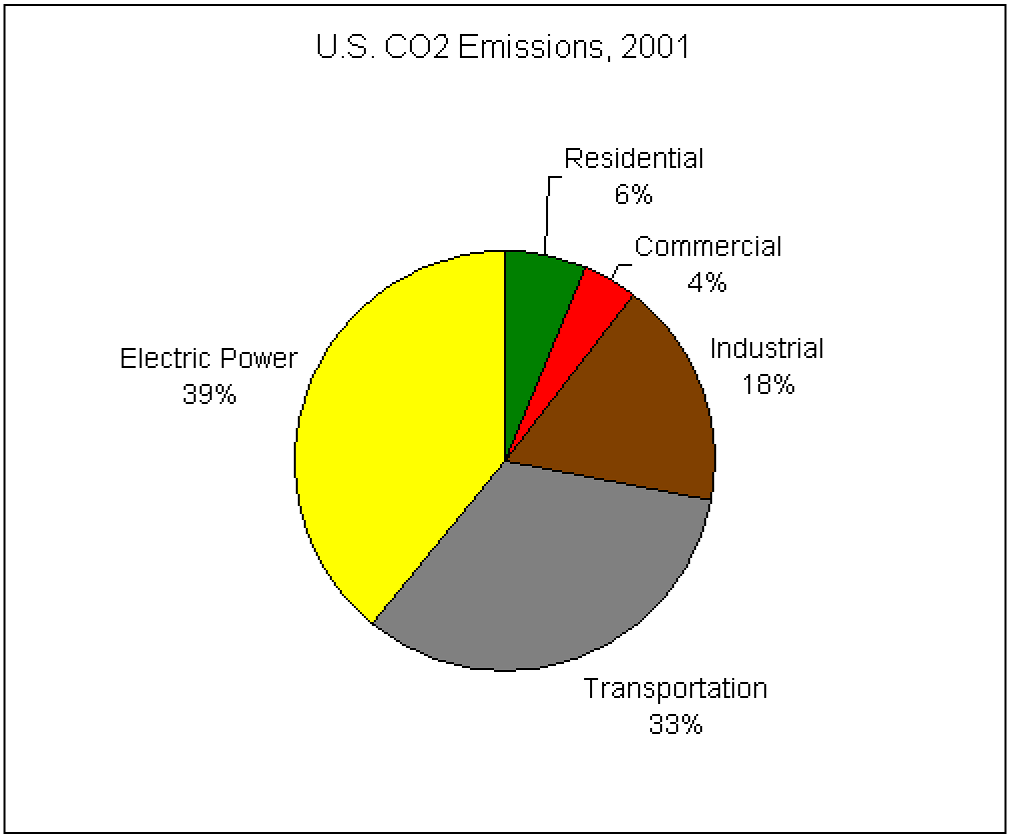


Figure 1.1: Anthropogenic CO_2 Emissions, 2001 [Energy Information Agency, 2002]

Researchers are also considering different types of sinks for CCS. Considerations for sink selection include the potential storage volume, the ease of injection, and the expected duration of storage. Herzog and Golomb [2004] suggest that, while actual storage volume is uncertain, volume estimates are orders of magnitude larger than the current emission rate. Figure 1.2 shows a logarithmic graph of these estimates, in giga-tons carbon, for storage in the ocean, saline aquifers, depleted oil and gas reservoirs, coal seams, and terrestrial sinks. To provide a reference, the current worldwide carbon emissions are estimated at seven giga-tons carbon per year.

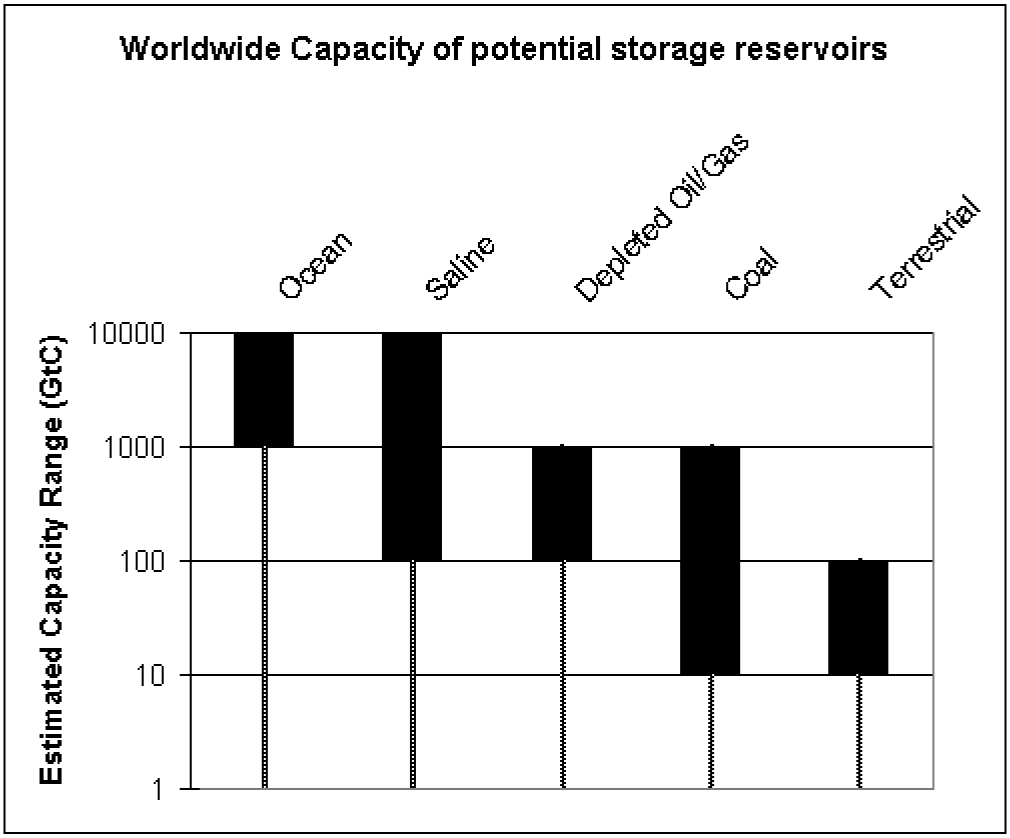


Figure 1.2: Estimated Storage Volumes [Herzog and Golomb, 2004]

Section 1.3: Project Description

The need to understand strategy and policy options for CCS requires a systems analysis approach taking all of the factors from source to sink into account. Technical analysis tools include calculations of coal volumes [Brennan and Burruss, 2003], brines aquifer adsorption of CO_2 [Maroto-Valer et al, 2003], fluid flows for injection [Bock et al, 2002], and reservoir sealing characteristics [Grigg et al, 2003] [Freidmann and Nummedal, 2003] [White, 2003]. Beyond the technical studies are considerations of the economic viability [Dooley et al, 2002] of a project. Organizations including Pacific Northwest National Laboratory (PNNL), Mid-continent Interactive Digital Carbon Atlas and Relational dataBase (MIDCARB), Ecofys, and Massachusetts Institute of Technology (MIT) are working to combine these individual tools into systems-level CCS analysis projects.

The Carbon Capture and Sequestration Technologies Program (CCSTP) at MIT's Laboratory for Energy and the Environment conducts research into technologies to capture, utilize, and store CO_2 from large stationary sources. A major research focus is the development of a Geographic Information System (GIS) that is used as a basis for analysis tools. These analyses address the complex systems approach to CCS. An integral part of this GIS is the Distributed Information Management System (DIMS). DIMS incorporates context mediation and information integration technologies into the GIS in order to manage the issues relating to utilization of multiple heterogeneous sources in a single complex analysis. DIMS handles database connectivity, format and context mediation, and information integration from multiple sources in order to provide a more unified view on the available data. This allows users and developers of GIS tools to focus on the analyses instead of the complexities of data management.

This thesis provides an overview of the GIS development and detailed discussion of the DIMS technology and implementation. Chapter 2 provides background information on current GIS systems used in CCS, information on data integration, and the issues of distributed data. Chapter 3 gives an overview of the CCSTP GIS, working from the user interface down to data sources. Chapter 4 explains the current implementation of DIMS. Chapter 5 discusses specific designs that would improve the performance and scalability of DIMS. Chapter 6 highlights the utilization and policy implications of DIMS. Chapter 7 states the conclusions of the thesis.

Chapter 2: Background

Section 2.1: Current GIS efforts

Other CCS GIS systems have helped to describe benefits and limitations of current efforts. From a set of projects that have been developed worldwide [Gale, 2002], this section discusses the Pacific Northwest National Laboratory's (PNNL) GIS, the Mid-continent Interactive Digital Carbon Atlas and Relational dataBase (MIDCARB), and Ecofys' decision support system (GESTCO).

2.1.1 PNNL GIS

Researchers at PNNL chose to develop a GIS for CCS analysis in order to "visually display spatial relationships and perform queries and screening analyses with ease" [Dahowski et al, 2001]. They have developed a database with CO_2 sources, pipelines, and potential sinks, and used the database and GIS to develop in-house capture and storage screening analyses.

Public access to the database associated with GIS is not allowed, but PNNL has described the contents. The CO_2 sources in the database include large power plants and anthropogenic sources that serve enhanced oil recovery (EOR) projects. Transport data on major CO_2 distribution pipelines is also included. The database also includes potential sinks such as EOR projects, enhanced coal bed methane (ECBM) projects, coal basins, brine aquifers and CO_2 domes.

Based on this GIS, PNNL provides analyses and suggestions. While most of the analyses are propriety, Dahowski and Dooley [2002] have presented one recent analysis to the public. This analysis "examines the existing stock of fossil-fired power" plants "that have a minimum of a decade's worth of productive life" and the "relationships between plant type, location, emissions, and vintage" to consider the economics of plant retrofit and sink storage.

Because the GIS and database are proprietary, it is more difficult to study the processes leading to the final published results. Work for this thesis suggests that value can be gained through information transparency. Information transparency allows stakeholders to view the data and methods supporting analyses and provide input.

2.1.2 MIDCARB

The MIDCARB database project is under development by a consortium of five state geological surveys. This early collaboration toward CCS analysis includes the state geological surveys of Illinois, Indiana, Kansas, Kentucky, and Ohio, and is led by the Kansas Geological Survey (KGS). The consortium was formed due to locale as well as technical capabilities. The stated goals of the project are characterizing major CO_2 sources and storage sites, developing databases, and supplying the data to the public [Carr et al, 2002].

Each of the member geological surveys' function is to record the geology in their state for analyses and historical knowledge. Personnel specialized in various geological sciences and knowledgeable in the regional specifics work to improve the data integrity.

Importantly, each of the surveys already had developed electronic databases on the relevant data prior to the start of the project. Because these databases were in place, the effort in aggregating the databases into a single portal was much smaller than the effort needed to collect the data from unconsolidated sources, such as regular computer files or even paper files.

In order to provide the data from the geological surveys to the public, MIDCARB has chosen to develop a World Wide Web (WWW, web) interface (http://www.midcarb.org). Through this portal the public can view the maps and data that MIDCARB publishes on its web site. The data displayed to the user remains stored on the source survey's database, only the requested data is transferred through the portal to the user.

Figure 2.1 shows the data pathways in using the portal. The portal receives requests from a web browser either through the web interface (ArcIMS) or as a server request (ColdFusion). Each of the state geological surveys provides an accessible database which has program (ArcSDE) that handles the data transmission. ArcIMS displays maps from the data collected through the ArcSDE interfaces. Both ArcIMS and ArcSDE are complementary parts of the off the shelf GIS package used by MIDCARB. ColdFusion is a web server programming language that has been used to develop more calculation intensive programs. ColdFusion programs can provide data reports by first querying the state databases directly to retrieve data, then running programs on the server on the data, and then formatting the results for the user in the report.

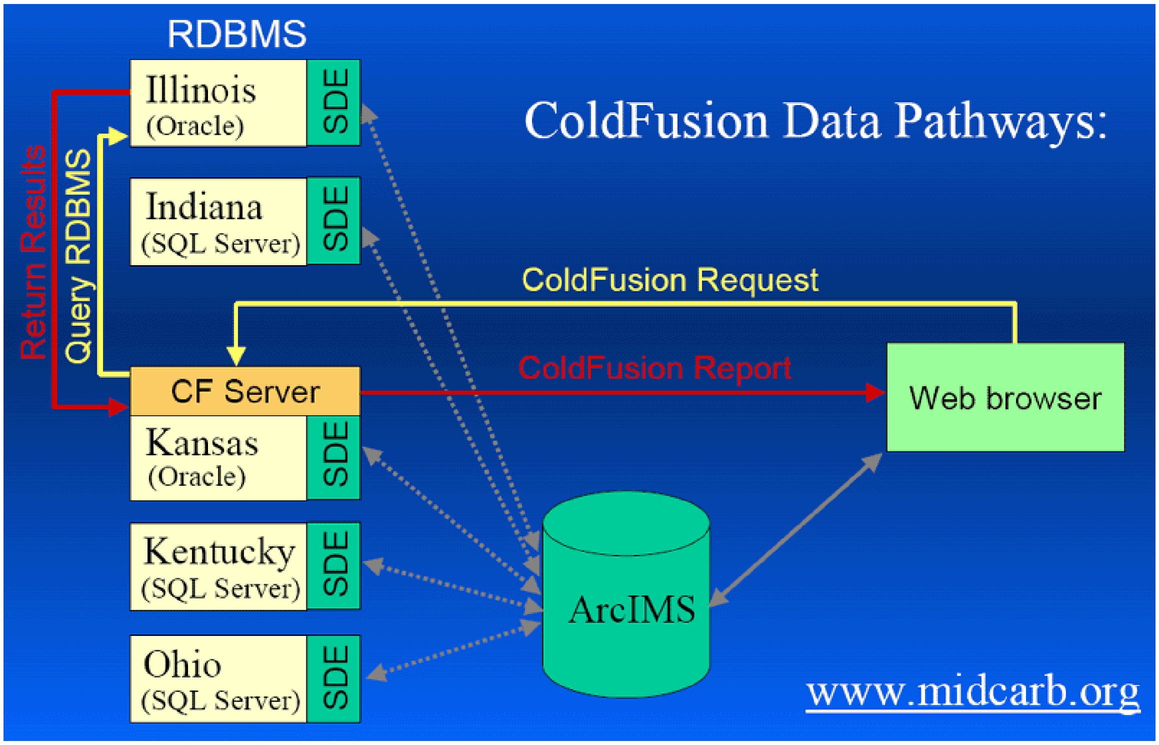


Figure 2.1: MIDCARB System Diagram [White et al, 2003]

DIMS has learned a great deal from interactions with the group developing MIDCARB. MIDCARB has built coalitions of data providers and offers data from many distributed sources through their portal. KGS presents the highest quality data to the user without overwriting data sources. While DIMS is not concerned with the collection of data itself, work with MIDCARB has helped to build the linkages with data providers. DIMS manages data in a similar fashion by presenting data to the user as an integrated whole while maintaining the source information. This additional information can be retrieved by users to better understand the data management process.

In addition to the aggregation and portal aspects developed in MIDCARB, DIMS explores data integration. Data integration takes the information from the distributed sources and consolidates data representing the same real-world entities. This provides a more coherent picture of the information required in analyses.

2.3.3 Ecofys GESTCO-DSS

Ecofys is a European company focused on sustainable energy solutions. Through a project funded by the European Union, Ecofys is developing a decision support system as part of the European Potential for Geological Storage of Carbon Dioxide from Fossil Fuel Combustion (GESTCO) Program. The primary goal of GESTCO is the development of tools to estimate carbon storage costs in Europe and worldwide, but a significant part of the project is the collection of data on worldwide sources of emissions.

GESTCO contains general data on sources and sinks worldwide, and factors that modify transport cost. Data was selected from a wide variety of sources [Hendriks et al, 2002a] with a focus on creating a low resolution, worldwide coverage of information. In some cases, relevant data on emissions was estimated using available data such as production quantities and assumptions on emissions that would arise from the production. The additional data used in estimating transportation costs include the location of existing pipelines, land-type, and terrain.

Ecofys divides their analysis tool into four calculation modules: separation, transportation, storage, and cost-engineering. In their design document, three of the modules are described as follows [Floris and Wildenborg, 2000]:

- Separation:
    - Calculate extra cost for CO_2 separation
    - Measure reduction of CO_2
- Transportation
    - Calculate optimal pipeline diameter
    - Model transportation paths including existing infrastructure
- Storage
    - Estimating volumetrics of subsurface storage
    - Compression requirements for injection
    - Costs based on exploration risks
    - Measurement of possible extra hydrocarbon production

GESTCO is an add-on to ESRI's ArcGIS 8 software package. In order to utilize the GESTCO system to perform analyses, users must have the ArcGIS software package installed on their computer as well as the GESTCO add-on and have connections to databases which are supported by GESTCO.

Figure 2.2 is a screen capture from the GESTCO system. This shows the user interface for the source-sink transport routing. The circles represent emission sources, triangles and squares represent gas and oil sinks respectively. Lines represent current pipelines and rivers. When a source and sink are selected by the user, the DSS checks if the sink has enough capacity for the selected source's emissions and then calculates the capture, transport, and storage costs [Egberts et al, 2003]. The highlighted line in the figure shows the calculated least-cost route.

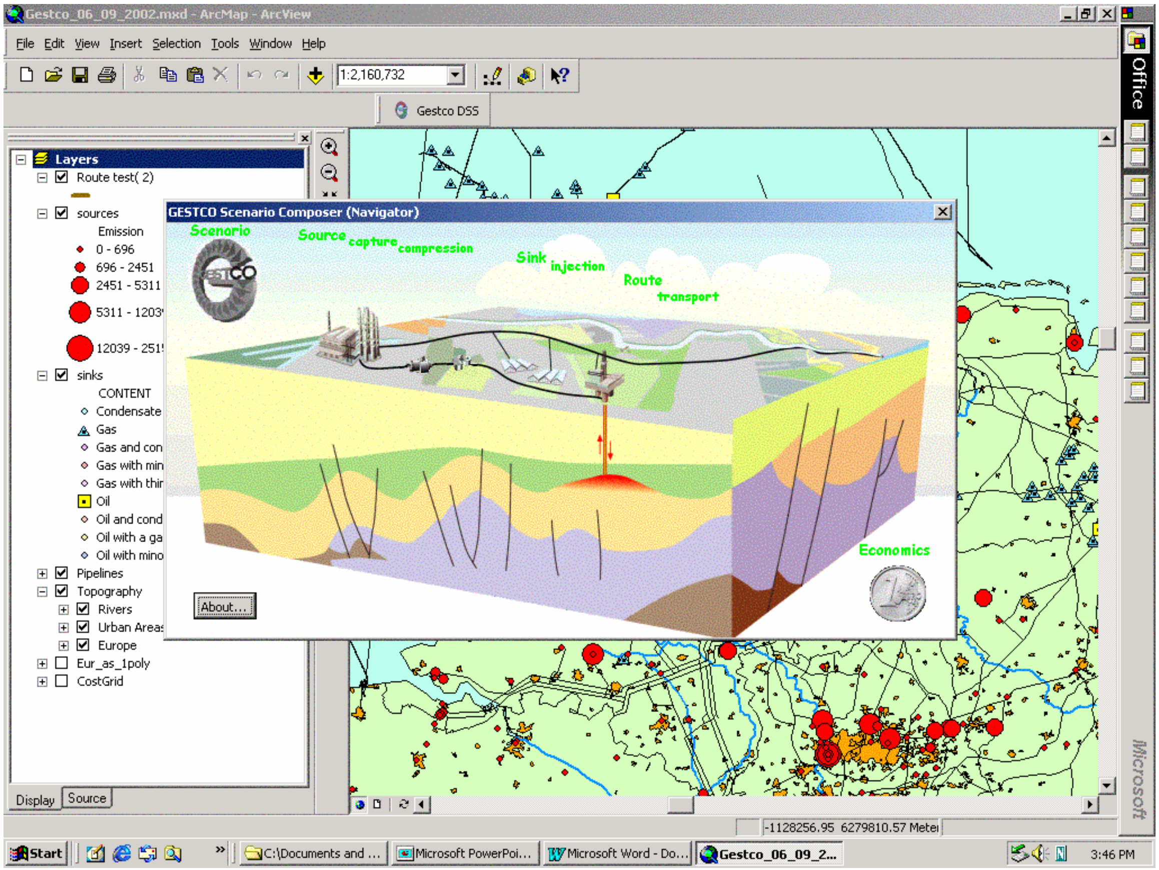


Figure 2.2: CO_2 Routing Analysis [Hendricks et al, 2002b]

One component of the GESTCO database is a cost surface. This surface associates map locations to a cost for building pipeline over that land. It is derived from the data on existing pipeline locations and features such as rivers and populated areas. Using the least cost path algorithm of ArcGIS on the cost surface provided by the GESTCO database, GESTCO determines a least-cost transport path from the source to sink.

The data used in the system are gathered together but remain a static snapshot of information unless the operators manually update the databases. The surface described above is one example of the static data. Each path is generated as a single component on the current infrastructure. When the DSS calculates a new path, it can not be incorporated back into the cost surface for future calculations.

GESTCO is primarily focused on providing a tool for global storage cost analysis. Ecofys has gathered a database of world-wide information and worked on analysis tools for source-sink matching and storage economics. DIMS can draw on the requirements set by GESTCO to better understand the data needs for analyses that are relevant to the global CCS community. However, DIMS is focused on distributed data management issues from sources within the U.S. and on providing a means of integrating these data for improved analysis and end-use. DIMS is also designed to manage dynamically updated data. For example, it would be possible to incorporate transport routes developed in a routing analysis back into subsequent analyses in order to build a transport network.

Section 2.2: Context Interchange (COIN)

Integration systems have also played an important role in the development of this system. Systems such as the Context Interchange System (COIN) have been studied to help frame the technical needs of data integration.

COIN is exploring the concept of "logical connectivity ... to support the acquisition, organization, and effective intelligent usage of distributed context knowledge" [Madnick, 1999]. It is a collection of programs that mediate information queries by accepting the queries in the user's context, and deconstruct the query into the relevant sub-queries to the underlying data sources.

COIN is being used in the Laboratory for Information Globalization and Harmonization Technologies and Studies (LIGHTS) project. The goal of LIGHTS is to understand the inter-relations between utilization of information technology and the realities of political international relations. The work is focused on using distributed information integration as it applies to complex global issues such as conflict and emergent risks, threats, and uncertainties [Choucri et al, 2003].

LIGHTS is directed towards understanding policy applications of information systems and in standardizing and warehousing information for use in this application. The project will use the data and the technology in the COIN system to assist in understanding the policy issues relating to world conflict and in developing policy analyses and interpretations.

COIN has developed as an integration system with a focus on financial analyses and, in its application with LIGHTS, to help standardize and warehouse information relating to world conflict policy. It parses a user query into separate sub-queries for separate underlying databases. The DIMS system leverages the knowledge and research from the COIN system to apply integration techniques to the field of CCS because of the context differences in the current information systems available for use in CCS analysis.

Chapter 3: GIS Design and Implementation

This chapter describes the design of CCSTP's Geographic Information System (GIS). A system overview is followed by specifics of each system layer. Each section highlights the general goals and important design points, then discusses the reasoning and details of implementation.

Section 3.1: System Overview

3.1.1 System Goals

- Reproducibility: Other CCS groups, especially any RCSPs that still need to develop their GIS system, could benefit from reproducibility of the CCSTP GIS. Design and implement with simplicity and interoperability in mind.
- Extensibility: Because CCS is still a growing and changing field, extensibility will allow incorporation of unexpected tools and techniques. Components of the system are designed, implemented, and used as distinct modules. Each module offers an external interface so that other modules need not know the internal implementation details, allowing modules to be developed separately, upgraded individually, and extended as needed.
- Maintainability: Clear documentation of the development process and the reasons for implementation choices will enable future maintenance of the system. This GIS project is likely to be long-lived, therefore time spent in support of maintainability will have future benefits for developers and users.

3.1.2 System Design

Figure 3.1 is a diagram of the layers in the GIS as well as the control and data flows. Thin dotted lines represent control signals and double lines represent data flows. Layers of the system are labeled, representing distinct modules. The local system is enclosed in the solid box.

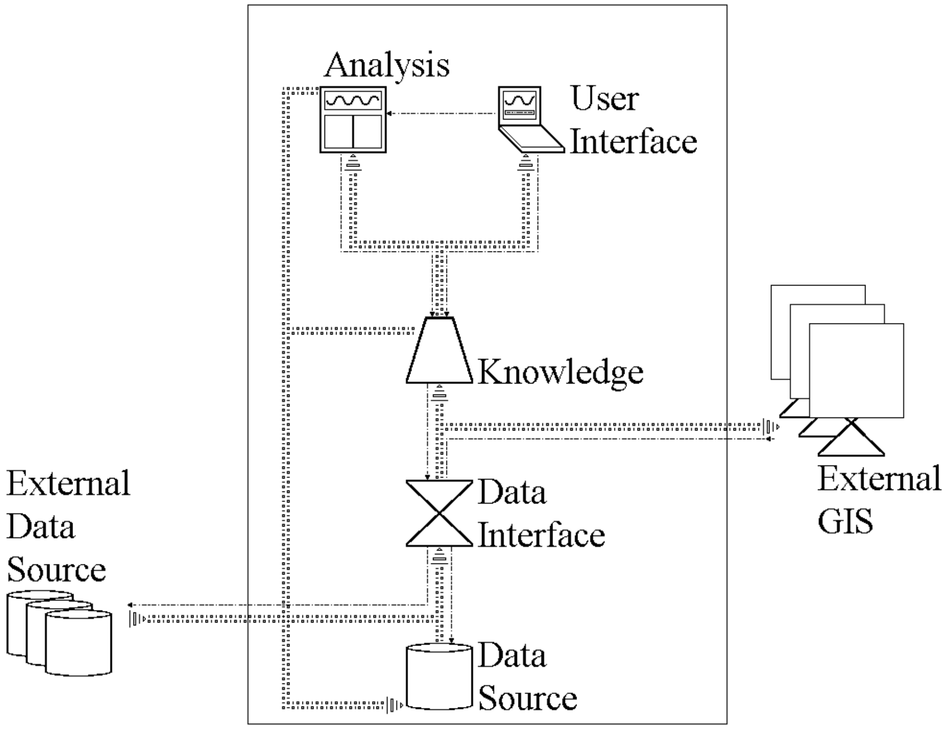


Figure 3.1: DIMS System Schematic

The layers of the system are as follows:

- The User Interface Layer (UIL) provides access to the system in a human usable format. Based on this interaction, the UIL sends control signals to the Analysis Layer (AL) to initiate analyses and to the Knowledge Layer (KL) to request information.
- The AL runs analyses and models. The data required for the analyses are retrieved from the KL, and results are stored into the Data Source Layer (DSL).
- The KL integrates data into coherent sets of information based on user requirements and available data. The KL receives requests from the UIL or AL, then collects data from the Data Interface Layer (DIL) and integrates the data that is relevant to the request. These results can be stored in the DSL for future use.
- The DIL provides the connectivity to external and internal data sources and mediates differences in source context such as unit of measurement. The DIL provides data to the KL and to external GIS requests.
- The DSL represents all the databases and sources of data available to the GIS. It provides data based on queries from the DIL.

Although each layer could be developed and hosted on different servers, a local system has been defined as the set of system components which utilize the same DIL and KL. This is represented in Figure 3.1 by a box surrounding local system components. As depicted in the figure, external data sources provide input to the system but are not used to store system data. Likewise, external GIS are allowed to access the data of the system through the DIL, but are not given access to store data in the local data source.

3.1.3 System Implementation

The CCSTP GIS is implemented on two consumer grade desktop PCs. Each of these machines is running primarily standard software with some new programs to support the DIMS layers. Details of the system configuration, hardware and software, are listed in Appendix C.

Oracle and ESRI products - Oracle database, Oracle application server, ESRI ArcGIS, and ESRI ArcIMS - have been chosen to support the reproducibility of the system. These software products are heavily used in the CCS community and at MIT. Therefore, other groups will be able to reproduce techniques developed in DIMS in their own systems.

The software code developed to incorporate information integration technologies are written in the Oracle database in the PL/SQL and Java programming languages. Some analysis programs have also been written for ArcGIS in Visual Basic, the scripting language used by that software package.

DIMS is currently being deployed as a production system in order to meet the demands of use by the public, RCSP members, and NATCARB users. For this, new server hardware has been acquired and is being prepared with the software discussed above. This will allow us to test the scalability and stability of the DIMS methodology.

Section 3.2: User Interface Layer

3.2.1 User Interface Goals and Design

- Goal: Provide an interactive environment that can accommodate all of the potential users that include technical analysts, policy decision-makers, and the public.
- Goal: Provide a non-intrusive interface by having minimal software and processor requirements.
- Design: Handle user input to trigger queries to the KL and commands to the AL.
- Design: Display map and information screens that are graphical and straightforward.

3.2.2 User Interface Implementation

In order to fulfill the goals of an accommodating and non-intrusive user interface, we have implemented an internet website. The website is developed in the Apache web server within Oracle 9i Application Server. These software packages form the basic web server, onto which the ESRI ArcIMS software is added to provide graphical interfaces and map displays. These off-the-shelf products allow us to quickly prototype a User Interface. Additionally, the requirements on the user side are minimal because all of the computation is done on the server side. A user of the GIS only needs a web browser and internet connection.

ArcIMS display maps and data that are stored in a database or on the server's file system. Specific scripts (ArcSDE) are required by ArcIMS and have been added to CCSTPs database to handle control and data connections between the KL and ArcIMS.
The interface to send controls to the AL has not yet been implemented. These will be handled by programs in the web server that will take user commands and run the appropriate analysis program.

Section 3.3: Analysis Layer

3.3.1 Analysis Goals and Design

- Goal: Enable tools and models to analyze major components of a CCS system: capture feasibility, transport routing, sink selection, and cost estimates for each of the components.
- Goal: Allow extensibility of analysis modules in support of system design goal.
- Design: Capture feasibility requires measurement or estimation of emission quantity and concentration, and requirements of capture technologies.
- Design: Transport calculations require the location of a source and sink pair, and factors that modify costs such as terrain, right of ways, and transport options.
- Design: Sink selection requires information on the reservoir characteristics: depth, thickness, permeability, porosity, pressure, and temperature.
- Design: Analysis tools should communicate with the database but be implemented in any programming tool appropriate.

3.3.2 Analysis Implementation

Analyses are developed at CCSTP and in other groups using a variety of programming languages and tools. The CCSTP GIS can incorporate these tools into the system as long as the results are stored into the database for display and further analysis.

For example, the sink injection model developed by CCSTP runs in ArcGIS. The model takes the reservoir characteristics as input in the form of two-dimensional grids, stored in ArcGIS raster files. It then runs the injection costing algorithm [Heddle, 2003] that was developed at MIT on the grids to calculate the estimated cost of drilling wells and a per-ton cost for injection. Other analysis programs have been written for the Oracle database, including calculations of emissions from sectors in each of the regions of the U.S.

Section 3.4: Knowledge Layer

3.4.1 Knowledge Goals and Design

- Goal: Integrate data from different sources into collections of information that relate to the same real-world entity.
- Goal: Provide a single information interface for users and analyses, and supply integrated information to users in an understandable way

- Design: Integration is performed by building knowledge objects that represent the available information on the entity.
- Design: The knowledge objects will retain information on the data interfaces used in integration to enable tracing of data flows and data quality.
- Design: Correlate naming, labeling, and primary key conventions from the different data sources to locate related information.

3.4.2 Knowledge Implementation

This layer is one of the primary focuses of the research in DIMS. The integration concepts and description of module implementation for this layer are covered in Chapter 4: DIMS Implementation.

Section 3.5: Data Interface Layer

3.5.1 Data Interface Goals and Design

- Goal: Mediate context differences between sources and the local system through data conversion and translation.
- Design: Each type of data source that DIMS will utilize will have an associated DIL module tasked with interpreting the data.
- Design: Translations between source and local context are centralized to allow reuse and avoid errors.

3.5.2 Data Interface Implementation

This layer is one of the primary focuses of the research in DIMS. Definitions of context issues as well as implementation details of the Data Interface modules are described in Chapter 4: DIMS Implementation.

Section 3.6: Data Source Layer

3.6.1 Data Source Goals

- Goal: Build collaborations with data collectors.
- Goal: Understand current state of data sources available to research in CCS.
- Goal: Coordinate with data collectors to provide source data with improved information quality.
- Design: Support the goal of reproducibility with efficient, simple and interoperable database.

3.6.2 Local Data Source Implementation

The local Data Source is used to maintain process information needed by layers and to store the results of integration and analyses. It is implemented in an Oracle 9i database with additional ArcSDE scripts that are used by the ArcGIS analysis tool and ArcIMS web mapping software to interface with the database.

3.6.3 External Data Sources

CCSTP conducted a study of data sources relating to the field of carbon dioxide capture and storage (CCS) focusing on data that are national in coverage, detailed in characterization, current, updated, and publicly available. Specific information about each of these data sources is available in Appendix D.

Large point-sources emitters of CO_2 include power plants and industrial facilities. Data sources that have been evaluated are:

- eGRID: An EPA database on electricity generation plants in the U.S. The database includes several important characteristics on boilers and power plants, including the location and ownership of the plants as well as the production capacity, fuel used, and emissions of criteria pollutants and CO_2. Figure 3.2 lists a selection of the 142 fields in the 2000 release of eGRID.

Field	Name	Description
2	PSTATABB	State Abbreviation
3	PNAME	Plant Name
20	CNTYNAME	Plant county name
21	LAT	Plant latitude
22	LON	Plant longitude
30	NAMEPCAP	Plant generator capacity (MW)
39	PLNGENAN	Plant 1998 annual net generation (MWh)
42	PLNOXOZ	Plant 1998 ozone season NOx emissions (tons)
43	PLSO2AN	Plant 1998 annual SO2 emissions (tons)
44	PLCO2AN	Plant 1998 annual CO2 emissions (tons)
45	PLHGAN	Plant 1998 annual mercury emissions (lbs)
58	PLGENACL	Plant 1998 annual coal net generation (MWh)
59	PLGENAOL	Plant 1998 annual oil net generation (MWh)
60	PLGENAGS	Plant 1998 annual gas net generation (MWh)
61	PLGENANC	Plant 1998 annual nuclear net generation (MWh)
62	PLGENAHY	Plant 1998 annual hydro net generation (MWh)
63	PLGENABM	Plant 1998 annual biomass/wood net generation (MWh)
64	PLGENAWI	Plant 1998 annual wind net generation (MWh)
65	PLGENASO	Plant 1998 annual solar net generation (MWh)
66	PLGENAGT	Plant 1998 annual geothermal net generation (MWh)
67	PLGENAOF	Plant 1998 annual other fossil (tires, batteries, chemicals, etc.) net generation (MWh)
68	PLGENASW	Plant 1998 annual solid waste net generation (MWh)
85	OWNRNM01	Plant 2000 owner name (first)
86	OWNRUC01	Plant 2000 owner code (first)
87	OWNRPR01	Plant 2000 owner percent (first)

Figure 3.2: Selected columns, eGRID [U.S. EPA, 2001]

- MIDCARB: An aggregate database of five state surveys that contains information on various emissions sources in electricity generation and other industries such as ammonia or concrete manufacturing. The web interface uses ArcIMS to display maps of the facility location and industry.

- GESTCO: A collection of data from several journals and databases on sources of CO_2 worldwide. The database contains estimates of CO_2 emissions, either as a reported figure or as an estimate based on production from the facility.

The following data sets on sink characteristics have also been evaluated:

- GASIS: A NETL/DOE database of data on gas reservoirs. Data was consolidated from several previous regional atlases of gas data. The database contains fields for reservoir properties such as depth, porosity, permeability, and temperature, but the data is not complete for many fields. Figure 3.3 shows the completion percentage for the data fields relevant to CCS analysis. The figure shows that, for example, geographic location is supplied for only 14 percent of the entries.

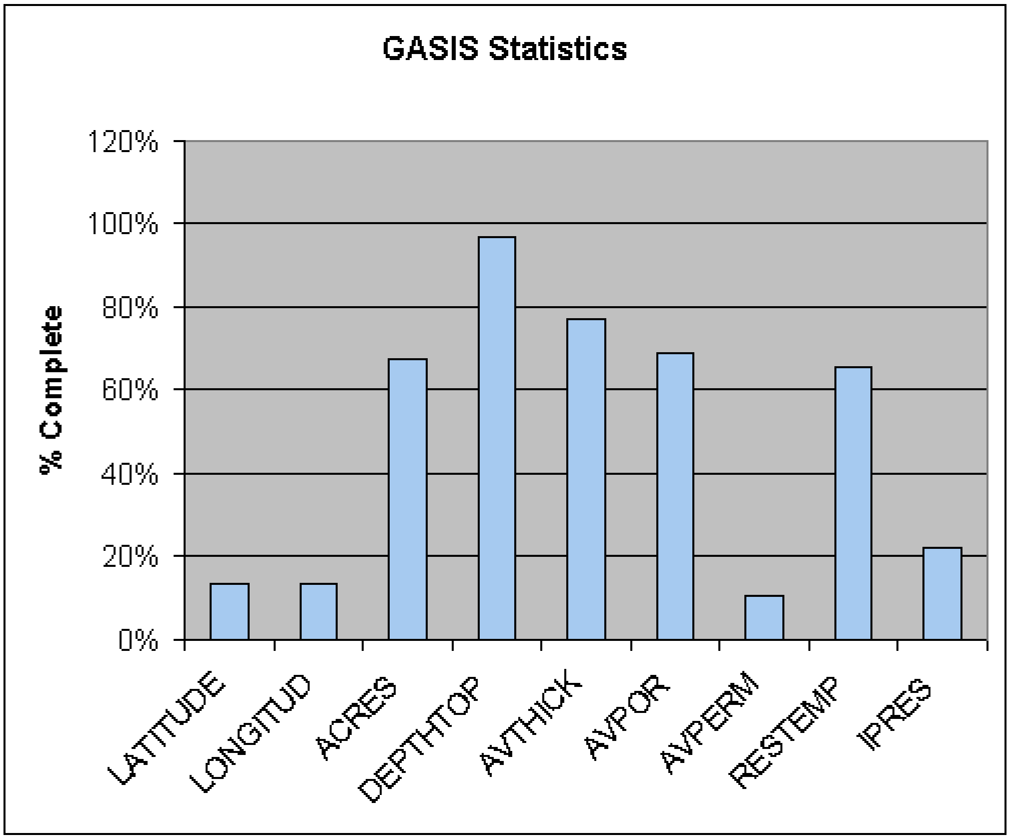


Figure 3.3: Gasis Statistics, percent complete

- UT-BEG's Brine DB: A set of GIS shapefiles and rasters that have been developed at the University of Texas, Bureau of Economic Geology. These GIS files cover 21 brine aquifer formations and include 16 characteristics such as

depth, porosity, permeability. The database was developed specifically to characterize brine aquifers with high storage potential.
- TORIS: The Total Oil Recovery Information System (TORIS) is a database on Oil Reservoir properties which is maintained by the National Petroleum Technology Office of DOE. It includes reservoir characteristics such as porosity and permeability. The publicly available database is the version produced in 1984, and does not contain current information on production or new oil exploration.
- COALQual: The Coal Quality (COALQual) is a database produced by the U.S. Geological Survey (USGS). It contains a set of shapefiles that define regions of coal that are either considered mineable or unmineable by the USGS.

Other types of information are relevant when analyzing CCS options, including physical terrain, political boundaries, population centers, demographic information, and regulatory information.

- ETOPO5: This is a set of data on the average elevation for the land in a five minute latitude by five minute longitude area. From this database, we have extracted the topography of the U.S.
- USGS Boundaries: The state and county boundaries. Boundaries of urban and metropolitan areas.
- GNIS: The Geographic Names Information System is a product of the USGS that provides the name and location of all federally recognized locations.
- Census Population: Data on the 2000 Census data by county.
- USGS Hydrography: Polygon and line water features of the U.S. intended for regional or national display. Includes lakes, reservoirs, rivers, shoreline, and other waterways.

3.6.3 Data Source Findings

We have focused our efforts in collecting the data required for developing our CCS analyses. However, in locating and evaluating data sources, a number of issues have become apparent with the data. These issues are consistent with the fact that data relating to CCS has not been specifically collected in the past, so we must use data from other available sources.

One issue is that the relevant data are dispersed among many databases and organizations. This creates difficulties because the appropriate databases must first be discovered, and there is no list of the best databases of CCS data. Further, even after the sources have been located, the ease of access to the data varies. Although all of the databases described above are public domain, some provide their data as publicly accessible computer files that are distributed on the internet, some are only available on CD by sending a request to the source, while others require the request of a username and password on the source database for access.

A related issue is that these data are offered in heterogeneous formats, both in the type of file used and the manner of data representation. This arises from the previous issue of disperse source organizations. Since the sources were used in a variety of different ways, they evolved to meet different needs. The files types seen among the sources of interest range from delimited text files to complex Excel spreadsheets and Access databases. Data that is provided within these files are represented in a variety of different measurement units.

The last issue that we are considering is the way to combine the sources together. This requires first determining the correlations between databases in order to merge data from databases together. It also requires selecting the best database entries when multiple databases have duplicate entries.

Chapter 4: DIMS Implementation

The Distributed Information Management System (DIMS) provides a consistent means of accessing distributed data sources, manages the contextual differences found in the heterogeneous data, and integrates these data into coherent collections of data. In terms of the system architecture, DIMS consists of the Data Interface Layer (DIL) and the Knowledge Layer (KL).

This chapter discusses the current implementation of DIMS by reiterating the goals of the DIMS layers, covering the process requirements of each layer, and showing the execution of the process by example. The sections cover an overview of the organization of DIMS, the DIL implementation, and the KL implementation.

Section 4.1: General Organization

DIMS provides users with location transparency and transaction transparency [Stonebraker and Hellerstein, 1998]. This way, multiple databases (locations) and queries (transactions) can be used in the system without changing the user's perception of the data. This is done through the DIL and KL.

The DIL provides access to the data sources and manages the context differences. The layer is implemented through tables and scripts that are used to connect to source data, document the meta-data, manage context translation formulas, and present the data.
The KL integrates information into specific topics. It is implemented through tables and scripts that define which tables and columns are related and how to bring the columns together into an integrated whole.

The tables and scripts that are produced in this implementation are grouped together under database user names to keep the database organized and understandable. DIL tables are grouped according to the data provider name. KL tables are grouped according to the name of the user of the table. The SQL code below creates the EPA user and grants it the right to connect to the database and store some data in the database to temporarily cache data and store the results of analyses. This user will be the owner of DIL tables that access EPA data and KL tables that are used by the EPA.

```
create user epa
  identified by
  default tablespace gis0
  temporary tablespace temp
  quota 102400 K on gis0
  quota 10240 K on temp
  account unlock;

grant connect to epa;
```

Data warehouses are traditionally considered to be a way to maintain a historical store of information, to provide transformations of the data for use in business analyses [Gupta, 1997], and to generate reports from that data in the context of the historical time [Greenfield, 1995].

Though there are some similarities between this implementation of DIMS and data warehouses, DIMS is primarily focused on providing access to data as opposed to storing the data. Some of the data that is used by DIMS is non-volatile, historical data, but they are not used to make analyses of how things were, but analyze how CCS systems can be in the future. These data are not saved locally for the purposes of warehousing. Instead, DIMS provides a pathway connecting the analysis tools and users to the data providers and warehouses which also ensures that the tools and users can understand the data.

Section 4.2: Data Interface

The basic steps that are required to create interfaces are as follows. First, a connection is made to the source data in order to access the raw data. Second, the context of the source data is determined and conversion functions are defined. These functions will be used to translate the source data into a common context used in DIMS. After this, the DIL table is created in the database.

4.2.1: Connections

The connection is the basic link used to retrieve raw data from a source data. Depending on the nature of the data source, the process used to create a connection differs. The different connections made in the DIMS system are to computer files and to remote databases.

Files

Computer files have been obtained for the system in a variety of file formats including: Microsoft Excel, Microsoft Access, FoxPro, and delimited txt. These files are not readily usable by the Oracle database used in DIMS. Therefore, files are first exported into a comma separated value (CSV) file format and then linked to the database.

Excel files, for example, are opened in the Excel program. After the .xls file is loaded, the data is exported to a CSV file by selecting Save As under the File menu and choosing the CSV (Comma delimited)(*.csv) option from the Save as type box. Other file formats can be exported by using similar facilities in the software program that is appropriate to the file.

In order to keep these files organized, they are grouped into directories named for the data provider. In our server, this directory is created under the /u01/rawdata directory. Therefore all CSV files for the EPA are stored in the /u01/rawdata/epa directory. The directory also has to be defined in the database using the create directory command. The following script shows how this is done for the EPA user.

```
create directory d_epa
  as '/u01/rawdata/epa';

grant read on directory d_epa to epa;
grant write on directory d_epa to epa;
```

An external table is used to connect to the CSV file. The columns of this table mirror the source file, named identically and in the same order, to simplify the script-making process. It is loaded using the organization external command, which makes the table retrieve the data from the CSV file. DIMS dynamically accesses and mediates the data. This reduces the storage requirement by accessing data from the data file, but also increases the access time to the data because it is not stored as efficiently as can be done in the database. In the example script below, the basic oracle import tool (type oracle_loader) is used. It is configured to find data entries on each line of the file and use the comma character as the separator between columns unless it is enclosed by quotes.

```
create table epa.egrdplnt
  (
  seqplt98 varchar(255),
  pstatabb varchar(255),
  pname varchar(255),
  orispl varchar(255),
  pltype varchar(255),
  ...
  )
  organization external
    (
      type oracle_loader
      default directory d_epa
      access parameters
      (
        records
          delimited by newline
          badfile d_epa:'egrid98_egrdplnt%a.bad'
          logfile d_epa:'egrid98_egrdplnt%a.log'
        fields
          terminated by ','
          optionally enclosed by '"'
          missing field values are null
      )
      location ('egrid98_egrdplnt.csv')
    )
  reject limit 200;
```

By loading the data from the file instead of actually duplicating the data in the database, DIMS acts as a dynamic access and mediation system as opposed to a data storage system. In this implementation, most decisions attempt to lean to the dynamic access side. The tradeoffs in this decision are between reducing the local storage requirement by accessing data from the data file, and improving access time by caching the data in the database.

Remote Databases

Other data providers have allowed direct access to their databases. Connections to remote databases can be created in the DIMS database by registering the remote database in the local names registry and running simple SQL commands to connect to the remote database.

First, the service name of the remote server is registered in the local system. This is done by adding an entry to the the local system's name registry. The entry describes the network protocol, the host name, port number and the database name (SID) used by the remote server. The entry is added to a registry file (tnsnames.ora) found in the $ORACLE_HOME/network/admin directory. The entry for the Kansas Geological Survey's (KGS) database is shown below.

```
abyss =
  (DESCRIPTION =
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = abyss.kgs.ku.edu)(PORT = 1521))
    )
    (CONNECT_DATA = (SID = abyss))
  )
```

Next, a database link is created in the local database. This SQL command is used to store the user name and password that is used to connect to the remote database and associate it to a name that can be used in the local database. The following code creates a link to the service defined above (abyss).

```
create database link abyss.kgs.ku.edu
  connect to MIT_GIS identified by
  using 'abyss';
```

The final step in establishing connections to remote databases is to set up synonym tables. creating a synonym stores an association between a local database table name and a remote database table. While they can be used as local tables by the user, the data is actually accessed from the remote database. The code below generates synonyms for tables in the KGS database that relate to Kansas power plants and emissions from those plants.

```
create synonym kgs.ds_facilities
  for midcarb.facilities@abyss.kgs.ku.edu;

create synonym kgs.ds_facilities_emissions
  for midcarb.facilities_emissions@abyss.kgs.ku.edu;
```

4.2.2: Context and Conversion

The next step in creating the data interface is to determine the context of the source, and how that context can be related to the DIMS context. Context refers to the the set of assumptions about how data is represented in the system and how it should be interpreted when retrieved from the system [Madnick, 1999]. This includes the measurement units, geographic projection, and precision used in collecting and storing the information. This context information can be reflected in meta-data files, but is often incompletely characterized because the data provider believes that certain assumptions are "obvious".

The DIL uses three database tables to document the metadata from data sources as they are entered into the system. The tables are the context descriptor table, the context matching table, and the context conversion table. When new data sources are added to the system, new metadata from the source is added. When a data interface is generated for a particular data source, these tables are referenced to determine the conversion methods appropriate for each column of source data.

The descriptor table is a list of the different measurement units that are used by data sources in the system. This table provides a centralized repository of units that are handled in the system. When new sources are added to the system, only previously undefined units of measurements have to be added to the table. Each entry of the table consists of an ID, unit label, and description. The ID is a number that is unique to an entry in the table. The ID is used to reference the entry from other tables. The label is a short text version of the measurement unit. It can be used for purposes of display in the User Interface Layer. The description field of an entry is an informative and descriptive text about the type of unit. This can be used to explain special cases or codes that are used. Figure 4.1(a) shows the first several entries of the descriptor table.

The matching table is used to assign measurement units to the columns of source data. For each column of source data that is used by the DIL, an entry is added to the matching table. This entry identifies the column, and relates it to the descriptor. A column is identified by the name of the table owner, the table name, and the column name. It is related to the descriptor using the ID number that matches the column's context. Figure 4.1(b) shows the entries of the matching table that correspond to columns of the eGRID database used by the DIL.

The context conversion table stores the functions that are used to convert data between different contexts. The columns in this table are the source ID, the destination ID, a description, and a conversion method. The source and destination IDs indicate the units for the input and output respectively, and reference the ID column of the context descriptor table. The description is a text field that describes the method used in the conversion. The function column holds the actual conversion function. The function is represented as a PL/SQL code fragment, which can be retrieved when the DIL table is created. Figure 4.1(c) shows the conversion functions that are needed to handle the databases of emissions sources. In this table, the "$1" in the text of the function column is used to represent the input variable of the function.

ID	Label	Description
1		Undefined Text
2	Degrees	Geographic Degrees Latitude
3	Degrees	Geographic Degrees Longitude
4	Tonnes CO2	Metric tons of CO2
5	Degrees	Degrees North
6	Degrees	Degrees West
7	State	Full State Name
8	State	State Postal Abbreviation
9	MW	Mega-watts
10	Tons CO2	Short tons CO2. 'N/A' = unknown
11	Gg CO2	Gigagrams CO2.
12	Gg CO2	Gigagrams*(short tons)/(metric tons) CO2. GESTCO incorrectly assumes eGRID data is in metric tons

(a) Context Descriptor Table (dims.di_context_descriptor)

Owner	Table_name	Column_name	Descriptor_ID
EPA	DS_EGRID	PNAME	1
EPA	DS_EGRID	PSTATABB	8
EPA	DS_EGRID	LAT	5
EPA	DS_EGRID	LON	6
EPA	DS_EGRID	NAMEPCAP	9
EPA	DS_EGRID	PLNGENAN	9
EPA	DS_EGRID	PLCO2AN	10
EPA	DS_EGRID	PLGENACL	9
EPA	DS_EGRID	PLNGENAOL	9
EPA	DS_EGRID	PLNGENAGS	9

(b) Context Matching table (dims.di_context_matching)

Source Context	Destination Context	Description	Conversion Method
5	2	Degrees Latitude North to Degrees Latitude	$1
6	3	Degrees Longitude West to Degrees Longitude	-($1)
10	4	Convert short tons to metric tons. Filter out 'N/A'	0.9072 * decode($1, 'N/A', NULL, $1)
11	4	Convert gigagrams to metric tons	1000 * $1
12	4	Convert gigagrams*(short tons)/(metric tons) to metric tons	907.1847 * $1

(c) Context Conversion Table (dims.di_context_conversion)
Figure 4.1: Fragments of the DIMS metadata tables.

4.2.3: Generating the DIL table

The final step in the process is to actually create the interface through a SQL script. The DIL table is implemented through the creation of a database view. The view encapsulates the source connection and meta-data information that have been produced in the previous steps into a single table for access by users of the DIL.

For example, the view that has been created for the eGRID data interface accesses data from the connector table described above (epa.ds_egrdplnt) and converts the source columns into the DIMS context. In this instance, the context of three source columns need to be converted: the latitude (LAT), longitude (LON), and annual CO_2 (PLCO2AN). The SQL code that has been written based on the meta-data is shown below, with a number of the unconverted lines removed for brevity.

```
create or replace view
  epa.di_egrid
as
select
  PNAME,
  PSTATABB,
  to_number(LAT),
  - to_number(LON),
  to_number(NAMEPCAP),
  to_number(PLNGENAN),
  0.9072 * to_number(decode(PLCO2AN, 'N/A', NULL, PLCO2AN)),
  ...
```

```
from epa.ds_egrid
;
```

For the latitude (LAT), the metadata tables show that the source context is in degrees north (ID 5). The desired DIMS context for the resulting view is geographic degrees latitude (ID 2). The function for this conversion is simply the identity function, because these two contexts are numerically equivalent. However, the column in the resulting data interface will be labeled as geographic degrees latitude, allowing it to be compared with other latitude data.

Longitude (LON) is defined in degrees west (ID 6), and is converted into geographic degrees longitude (ID 3). This conversion requires an inversion of sign.

Annual CO_2 emissions (PLCO2AN) is defined in short tons, with the text "N/A" representing an unknown number (ID 10). In order to convert from this context to the local context of metric tons (ID 4) the conversion function first decodes the source column in order to convert the text into a NULL before performing the arithmetic conversion. DIMS uses the NULL value because the database can store and calculate numeric data with NULL values, but not with text.

Figure 4.2(a) and 4.2(b) show a selection of data from the connector table and from the resulting data interface view. The columns that have been converted due to context differences are highlighted in figure 4.2(b) using bold-italic face.

PNAME	PSTATABB	LAT	LON	NAMEPCAP	PLNGENAN	PLCO2AN	PLGENACL	PLGENAOL	PLGENAGS
J R WOOD INCORPORATED	CA	37.1871	120.6414	1.05	34	'N/A'	0	0	34
J S EASTWOOD	CA	36.7465	119.6395	199.8	374988	0	0	0	0
JACKSON VALLEY ENERGY L P	CA	38.4656	120.5493	18.5	103754.7	379642.95	76261.8	0	57.3

(a) eGRID Data Source (epa.ds_egrid)

PNAME	PSTATABB	***LAT***	***LON***	NAMEPCAP	PLNGENAN	***PLCO2AN***	PLGENACL	PLGENAOL	PLGENAGS
J R WOOD INCORPORATED	CA	***37.1871***	***-120.6414***	1.05	34		0	0	34
J S EASTWOOD	CA	***36.7465***	***-119.6395***	199.8	374988	***0***	0	0	0
JACKSON VALLEY ENERGY L P	CA	***38.4656***	***-120.5493***	18.5	103754.7	***344412.08***	76261.8	0	57.3

(b) Data Interface Table (epa.di_egrid)

Figure 4.2: Selections from eGRID Tables

The context mediation step performed with this data interface resulted in the conversion of the LAT, LON, and PLCO2AN columns. With LAT, the numeric value has not been changed, but the resulting column can be marked as being in the DIMS context. With LON, the value has been negated. With PLCO2AN, the text values representing unavailable entries ('N/A') have been replaced with NULL values and the numeric values have been converted from short tons to metric tons.

Section 4.3: Knowledge Layer

In order to develop a knowledge layer table, a central topic for the information is first identified, then the data interfaces and columns that supply information on the topic are selected, and finally the conflicts that arise between interfaces are resolved. The knowledge topic can be specific (i.e. high emission power plants in Kansas) or general (i.e. carbon sources in the U.S.) in nature. The selection of topic will help determine which interfaces and columns are appropriate for use in the resulting table. In some cases, multiple sources provide the same type of data on a topic. In these instances, decisions are made as to how to integrate the multiple sources into the final table.

Section 4.3.1: Topic Identification

The topic of a knowledge table simply defines a set of information needs to be addressed with the available data. It clearly states the expected utilization of the information and the data attributes that are desired.

The topic of U.S. power plants is used as an example to show the steps required to implement a knowledge table. This table is intended to be used to estimate CO_2 concentrations in power plant emissions streams and the total emitted CO_2. Additional data that is required is the basic plant identification information so that the emissions information can be connected to a specific plant. The following information sets are needed:

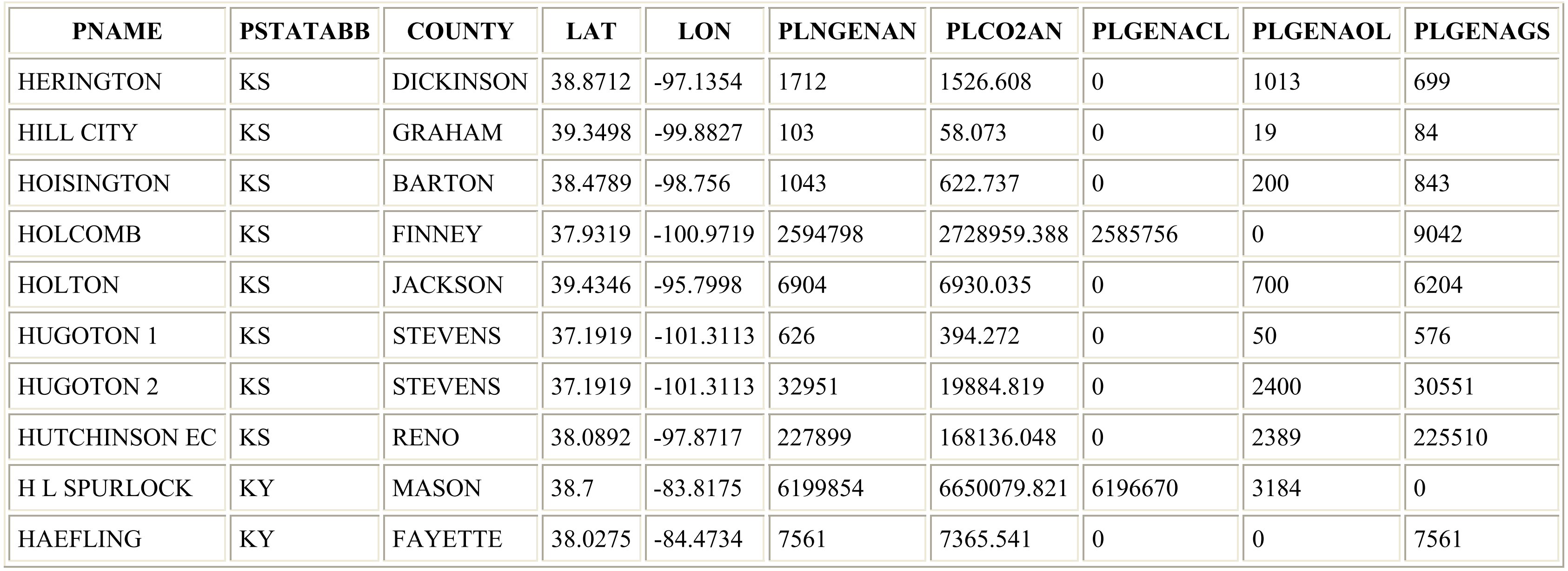
- Plant description
    - Plant name
    - Ownership information
- Location
    - Political Location: State and county
    - Geographic Location: Latitude and longitude
- Generation Information
    - Primary Fuel / fuel mix
    - Electricity production
- Emissions
    - Quantity of CO_2 emitted

Section 4.3.2: Interface Selection

In order to gather the data for the power plant knowledge table, source interfaces are chosen that contain data that is relevant to the topic. eGRID, MIDCARB, and GESTCO are the three source databases that contain data on power plants and emissions. Each of these databases contains a subset of the necessary data, and each has a different data focus.

eGRID

eGRID contains some data for each of the characteristics listed, but the database is focused collecting data on emissions from plants in order to ensure compliance with emissions regulations. Because the focus is on total emissions, other characteristics such as the location are not as important to the EPA and are therefore not carefully checked for accuracy. Figure 4.3 shows a selection of the data provided by the eGRID data interface that is described above. This selection shows the converted values of entries for several power plants in Kansas (KS) and Kentucky (KY).

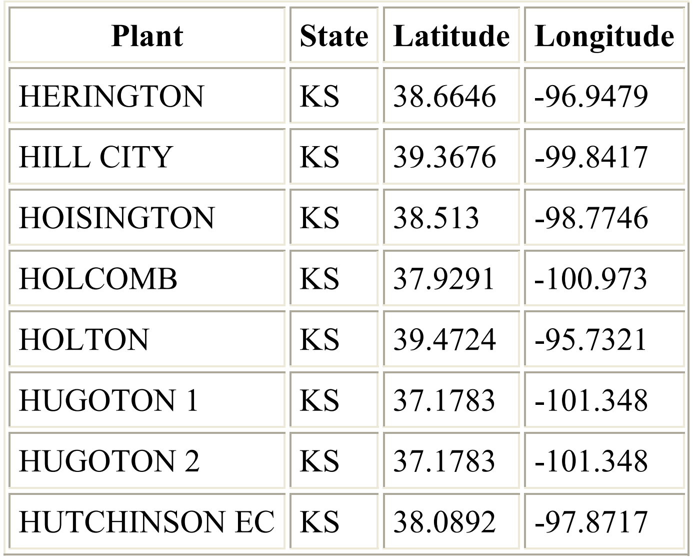


PNAME	PSTATABB	COUNTY	LAT	LON	PLNGENAN	PLCO2AN	PLGENACL	PLGENAOL	PLGENAGS
HERINGTON	KS	DICKINSON	38.8712	-97.1354	1712	1526.608	0	1013	699
HILL CITY	KS	GRAHAM	39.3498	-99.8827	103	58.073	0	19	84
HOISINGTON	KS	BARTON	38.4789	-98.756	1043	622.737	0	200	843
HOLCOMB	KS	FINNEY	37.9319	-100.9719	2594798	2728959.388	2585756	0	9042
HOLTON	KS	JACKSON	39.4346	-95.7998	6904	6930.035	0	700	6204
HUGOTON 1	KS	STEVENS	37.1919	-101.3113	626	394.272	0	50	576
HUGOTON 2	KS	STEVENS	37.1919	-101.3113	32951	19884.819	0	2400	30551
HUTCHINSON EC	KS	RENO	38.0892	-97.8717	227899	168136.048	0	2389	225510
H L SPURLOCK	KY	MASON	38.7	-83.8175	6199854	6650079.821	6196670	3184	0
HAEFLING	KY	FAYETTE	38.0275	-84.4734	7561	7365.541	0	0	7561

Figure 4.3: Selection of eGRID Data Interface (usgs.di_egrid)

MIDCARB

The MIDCARB data is interested in providing precise data for each of the power plants that are located in the MIDCARB states. In Kansas, the Kansas Geological Survey (KGS) has worked to locate each of the power plants. Using the eGRID data as a starting point, KGS looked at overhead photos and street maps to determine the geographic coordinates of each of the 89 power plants in Kansas. Of these plants, they were able to update 78 (88%) using digital orthophotos and street maps. They updated the location of five (6%) of the power plants by approximating the location relative to another plant. After verifying the coordinates using the orthophotos and maps, they found that only six (7%) of the geographic locations in eGRID were correct. Figure 4.4 shows a selection of the updated locations that are available from the MIDCARB database.



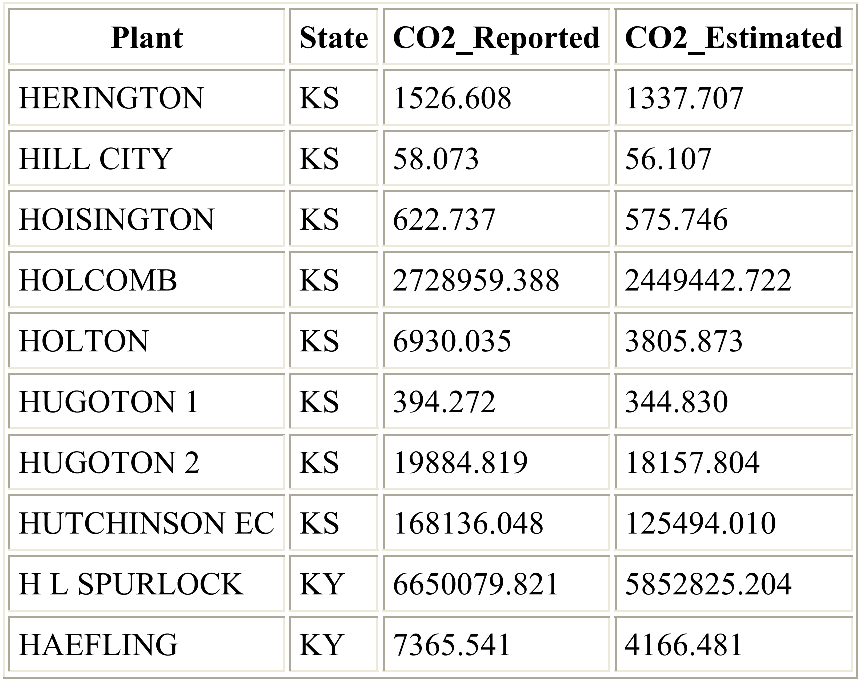
Plant	State	Latitude	Longitude
HERINGTON	KS	38.6646	-96.9479
HILL CITY	KS	39.3676	-99.8417
HOISINGTON	KS	38.513	-98.7746
HOLCOMB	KS	37.9291	-100.973
HOLTON	KS	39.4724	-95.7321
HUGOTON 1	KS	37.1783	-101.348
HUGOTON 2	KS	37.1783	-101.348
HUTCHINSON EC	KS	38.0892	-97.8717

Figure 4.4: Selection of MIDCARB facilities data interface (kgs.di_facilities)

GESTCO

The GESTCO database focuses on gathering data for as many sources of worldwide CO_2 as possible. Ecofys, the producers of the GESTCO database, use these data for their analyses. Because many of the emissions sources do not report their CO_2 emissions, Ecofys has included estimated emission for each of the sources using the standard IPCC method of estimating CO_2 emissions, based on the type of input fuel and the total power produced [Hendriks et al, 2002a].

For power plant data in the US, the database includes the plant name, state location and CO_2 as reported in the eGRID 2000 database. In addition, the GESTCO database includes the estimated CO_2 emissions, which can be used as a comparison to the reported emissions provided by eGRID. Figure 4.5 shows a selection of the GESTCO data that corresponds to the data for power plants shown in Figure 4.3.



Plant	State	CO2_Reported	CO2_Estimated
HERINGTON	KS	1526.608	1337.707
HILL CITY	KS	58.073	56.107
HOISINGTON	KS	622.737	575.746
HOLCOMB	KS	2728959.388	2449442.722
HOLTON	KS	6930.035	3805.873
HUGOTON 1	KS	394.272	344.830
HUGOTON 2	KS	19884.819	18157.804
HUTCHINSON EC	KS	168136.048	125494.010
H L SPURLOCK	KY	6650079.821	5852825.204
HAEFLING	KY	7365.541	4166.481

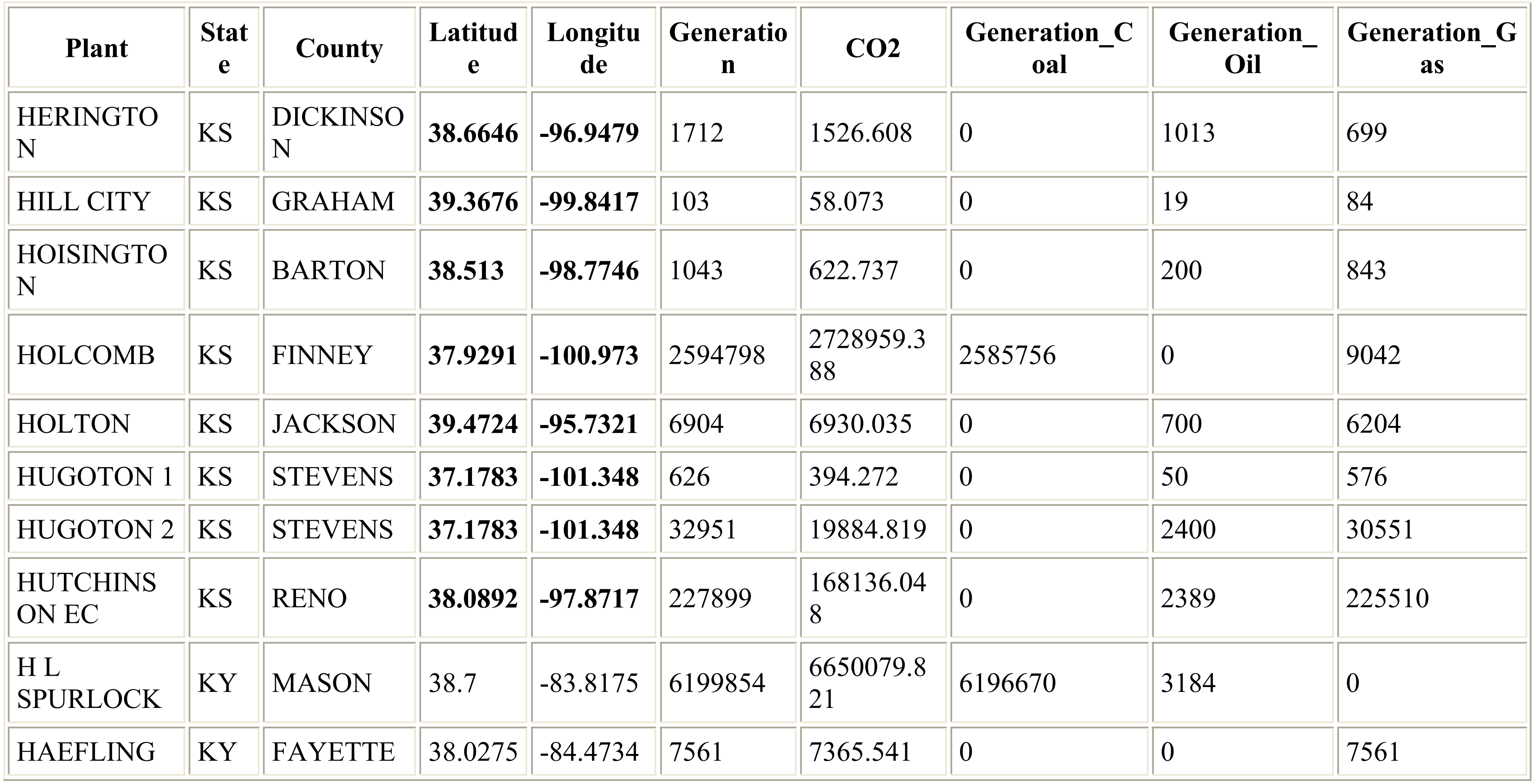
Figure 4.5: Selection of GESTCO data interface (ecofys.di_tblindustries)

Section 4.3.3: Integration and Conflict Resolution

Common naming schemes or identification fields can be used to correlate two data sources in some instances, but more complex linkages between multiple sources are required in many cases. These more complex linkages may use several columns as an aggregate key, or may span multiple tables. In each of these cases, the goal is to determine which rows of data in different databases are being used to represent the same entity.

The data is correlated by determining a set of data fields that uniquely define a power plant in each of the sources and each field that represents the same power plant attribute. Each data source is given a subjective quality rating by the user of the integration, which is based on the accuracy and percieved utility of the data. For the example, the CCSTP research group believes that the MIDCARB data is more accurate based on the extra effort made to check plant location, and that the CO_2 estimates of the GESTCO database are less accurate because they do not account for many variabilities in power production that alter the emissions rate. After these quality ratings are determined, data for each attribute is retrieved from the source with the highest quality rating for use in the KL object.

The resulting integration draws primarily from the eGRID database, but uses the higher quality coordinates available in the MIDCARB database. In this instance, the CO_2 emissions estimates from GESTCO were found to be of lower quality than the eGRID data, and so they were not used in place of the eGRID emissions data. Figure 4.6 shows the resulting table with integrated data. The latitude and longitude values that have been selected from the MIDCARB database for power plants in Kansas (KS) are highlighted in bold-italics.



Plant	**State**	**County**	**Latitude**	**Longitude**	**Generation**	**CO2**	**Generation_Coal**	**Generation_Oil**	**Generation_Gas**
HERINGTON	KS	DICKINSON	**38.6646**	**-96.9479**	1712	1526.608	0	1013	699
HILL CITY	KS	GRAHAM	**39.3676**	**-99.8417**	103	58.073	0	19	84
HOISINGTON	KS	BARTON	**38.513**	**-98.7746**	1043	622.737	0	200	843
HOLCOMB	KS	FINNEY	**37.9291**	**-100.973**	2594798	2728959.388	2585756	0	9042
HOLTON	KS	JACKSON	**39.4724**	**-95.7321**	6904	6930.035	0	700	6204
HUGOTON 1	KS	STEVENS	**37.1783**	**-101.348**	626	394.272	0	50	576
HUGOTON 2	KS	STEVENS	**37.1783**	**-101.348**	32951	19884.819	0	2400	30551
HUTCHINSON EC	KS	RENO	**38.0892**	**-97.8717**	227899	168136.048	0	2389	225510
H L SPURLOCK	KY	MASON	38.7	-83.8175	6199854	6650079.821	6196670	3184	0
HAEFLING	KY	FAYETTE	38.0275	-84.4734	7561	7365.541	0	0	7561

Figure 4.6: Selection of Power Plant Knowledge view (ccstp.k_power_plant)

Section 4.4: Implementation Comments

This version of DIMS has been implemented entirely within an Oracle database using straightforward and relatively simple processes and code. It provides the context mediation and integration results that were expected in the design through the views that described above. However, It still remains a work in progress with many potential areas of improvement. Some of the ways to improve the system are discussed in Chapter 5: Scalable Integration Designs.

Chapter 5: Scalable Integration Designs

The implementation of DIMS presented in this thesis has been driven by the reality of distributed information available for CCS analyses and the needs of the Carbon Capture and Sequestration Technologies Program's (CCSTP) GIS. It has proven sufficient and usable for the needs of the project, but it requires a significant amount of manual design and programming. Future uses of the system are likely to require processing of much more data from a greater diversity of sources. With this in mind, this chapter discusses designs that will improve the scalability of the system.

Section 5.1: COIN

In order to reduce the amount of manual work that has to be done within the DIMS system, the system can be modified to utilize existing middleware such as the Context Interchange (COIN) System. These modifications would effectively replace the Data Interface Layer (DIL) implementation with COIN, localizing changes to this Layer and modules that access the DIL. This would improve the scalability of the system because the COIN "mediation service requires only ... a logical specification of how data are interpreted ... and how conflicts ... should be resolved ... but not what conflicts are present" [Goh et al, 1999].

In order to make the modification, we define specification and conflict resolution procedures for each data source in the COIN standard and then utilize the COIN interface to access the data sources instead of the DIL. The specification files are similar in nature to the meta-data tables that are created in the DIL. They clearly define the units of each data column and the way that one unit is converted into another.

The specification file could be generated directly from the meta-data tables of DIMS for the current data sources and written separately for new data sources that are added to the system. To extract the information from the meta-data tables, first select entries from the context matching table, grouping by data source. Then, for each entry of a group, create a specification file and write out the column name and context name to file in the COIN specification format.

The conflict resolution information defines the method for reconciling differences in contexts between the source data and the user (receiver) of the data. These methods are derived from the context translation table of DIMS. The following pseudocode lays out the implementation of the conversion from the context translation table to the COIN specification.

```
create_coin_resolution ( file resolution_file )
 for each entry e_cct in context conversion table
  write to resolution_file:
   cdt1.column_name, cdt2.column_name, cct.function
  from context descriptor table cdt1,
   context descriptor table cdt2,
   context conversion table cct
  where cct.source == cdt1.id
   and cct.destination == cdt2.id
```

In the example eGRID context from the previous chapter, this would result in three scaling functions being added to the conflict resolution file. The eGRID longitude would scale by 1, latitude by -1, and CO_2 by 0.9072. The other columns of the database do not conflict with the receiver context, and therefore would not need resolution.

By creating a coin specification for the data sources, DIMS will be able to use the COIN context mediation service to automatically detect and resolve context differences between source data and the context of the system. DIMS would then provide access to the source data tables and the specification for use by other system components and external GIS. These users of the data would produce their own local context definition and necessary conflict resolution procedures. However, they would not have to explicitly manage the source data context, as this would be automatically managed through use of the COIN system.

Section 5.2: Information Quality for Integration

The current implementation of DIMS requires manual development of the collection and unification methods. While the implemented integrations are straightforward and reusable, it is possible to automate the process using the notion of Information Quality (IQ), which is a quantitative representation of the value of the information users.

In order to develop IQ in CCS and use it to support integration, it is important to consider a few steps of the IQ process. First, the important metrics must be determined. Second, strategies to implement integration with IQ that dynamic and functional methods for unifying contradictory data are designed. These methods include selecting values for entities from the highest quality source and taking an average of values weighted by the sources quality score.

Frequently cited goals of information quality listed in IQ literature are accuracy, timeliness, completeness, consistency, usability, reliability, and believability [Wand and Wang, 1996] [Strong et al, 1997] [Kovac et al, 1997] [Shankaranarayan et al, 2003]. Though there are multiple contextual and semantic variations in the terms, the metrics that are most relevant to the field of CCS are the following:

- Accuracy: consistency of data with the true value. We assume that measured values, such as emissions data reported on emissions forms, are accurate.

Calculated values, such as emissions estimates based on production levels, are less accurate.

- Precision: resolution of the information. Specific geographic coordinates of a power plant are more precise than the county in which the power plant runs. In the gas reservoir integration example, the precision of the information is increased by providing geographic coordinates, but the accuracy is reduced.
- Timeliness: nearness of the data collection or delivery time to the requirements of the data analysis and use. Data on geologic reservoir characteristics may still be timely even after decades, but data on emissions from a source should be taken during the time period being analyzed.
- Completeness: ratio of delivered data to the possible or expected data.
- Reliability: subjective expectation that a source provides information per request.
- Believability: subjective belief that delivered data represents reality.
- Consistency: measure of variance between data delivered by different data sources. Measures of consistency can be used to highlight characteristics that vary so that the underlying causes can be explored.

Kovac [1997] suggests a method of determining information quality by assigned a score of timeliness, reliability, and accuracy to data, then taking the sum of the average score as the quality ranking. Shankaranarayan [2003] incorporates the users perceived value of the quality goals by calculating quality as a weighted sum of timeliness, accuracy, and completeness, weighing each factor with user defined relevance modifiers.

Because IQ is itself dependant on context, the current DIMS framework can be used to develop and utilize it. Each data source would provide measurements of quality of the data, which would be converted by the system when the data is interfaces. The KL would then integrate sources based on these quality metrics.

The following is a simple example of how integration could be managed more automatically, based on the power plant integration discussed above and the quality ranking scheme described by Shankaranarayan. For this example, assume that each source has an equivalent quality ranking prior to use in DIMS. In our case, this is because the sources do not provide quality metrics. Because of this, relevance modifiers must be determined in order to have some means of selecting the highest quality values. These relevance modifiers are assigned in the range of zero to one based on the user's belief in the quality of a source. Figure 5.1 shows an example of the quality metrics that. These numbers were chosen to show the relative quality that we have determined by considering how each of the sources generated their data. The figure indicates that while eGRID is believed to be the most accurate and precise source of plant name and emission data, the KGS data is a more accurate source of latitude and longitude data.

Organization	Table	Column	Accuracy	Precision	Destination	Destination Column
EPA	di_egrid	pname	1	1	k_power_plant	plant
EPA	di_egrid	lat	.75	.9	k_power_plant	latitude
EPA	di_egrid	lon	.75	.9	k_power_plant	longitude
EPA	di_egrid	plco2an	.8	.7	k_power_plant	co2
KGS	di_facilities	latitude	.9	.9	k_power_plant	latitude
KGS	di_facilities	longitude	.9	.9	k_power_plant	longitude
ECOFYS	di_tblindustries	co2_reported	.6	.7	k_power_plant	co2
ECOFYS	di_tblindustries	co2_estimated	.5	.7	k_power_plant	co2

Figure 5.1: Sample Quality Table

Using these metrics, a simple integration function would simply chooses the highest quality value from among the possible sources. Below is the pseudocode for the integration function, this scans the quality table for all entries that correspond to a destination table, and creates a view that uses the source column of the highest quality.

```
function dims_integrate( string a_dest )

  string sql_code;

  # Initialize code fragment
  sql_code = "create or replace view " + a_dest +
    " as " + newline + "select " + newline;

  # Select columns
  select entries in quality table
    where a_dest == qt.destination
    group by destination_column
    (
      for each group g
      (
        select entry e with the highest (e.accuracy + e.precision)

        sql_code += e.organization + "." +
          e.table + "." + e.column +
          " as " + e.destination_column +
          ", " + newline;
      )
    )

  remove trailing ", " + newline;

  # Select tables
  sql_code += "from " + newline;

  select unique (organization, table) from quality table
    sql_code += organization + "." + table + ", ";

  remove trailing ", ";

  # Create view
  run sql_code;
```

This allows for a more scalable and flexible design by easing the management needs of individual knowledge tables. New knowledge tables could be generated by adding in a new destination and the rows indicating which sources are usable into the quality table.

New data sources could be inserted into the quality system by adding an entry for an existing knowledge table. Running the code above after either of these changes would automatically incorporate the new information.

Chapter 6: DIMS Applications and Implications

The Distributed Information Management System (DIMS) has been developed to address data issues found in current carbon dioxide capture and storage (CCS) data. This implementation incorporates information management techniques of context management and information integration to bring together data from diverse and distributed data sources and provide a means of querying the information from a single interface. This chapter discusses how the DIMS technology can be applied for work in CCS development and policy-making. First, the stakeholders of DIMS information are identified. Next, primary applications and benefits are considered. Finally, DIMS is related to current government initiatives regarding CCS data.

Section 6.1: Who are the stakeholders

There are a variety of organizations within the government, industry, and public that are interested in CCS information. These groups include NETL, EPA, CO_2 emitters, transport services, storage facilities, and non-government organizations (NGO). Each of these groups has goals and interests in the CCS information and in DIMS and similar information management technologies that can be used to improve the information's quality and accessibility.

NETL is supporting efforts to generate and collect CCS relevant data. The primary goal of these efforts is to develop the projects and policy recommendations that will lead to the reduction the nation's carbon intensity. Because it is supporting many different efforts, NETL's interests would be to aggregate the resulting data to compare results across different projects and to produce a complete national database.

The EPA's primary goal is to keep track of emissions regulations and the levels emissions that are entering the environment. As such, it has been a primary source of information on CO_2 emissions data. However, because the focus is on emissions, it has not verified the spatial data associated with emitters and is interested in accessing integrated CCS data in order to improve the accuracy of their own data.

CO_2 emitters are preparing for future possibilities carbon reduction policy. In order to make strategic decisions on how to meet the regulations, these companies want to understand the different CCS options that are available to their facilities. In particular, these companies would be able to use integrated CCS information to analyze the costs associated with using different transport and sink options.

Similarly, the other organizations that would be involved in the deployment of a CCS project are interested in the quality of CCS information. These organizations include transport services and CO_2 sink providers. These companies would use the data to understand how and where the markets for CO_2 storage are likely to evolve. Non-governmental organizations (NGO) are also interested in the quality of CCS information. At a national level, NGOs would like to use the information to help compare the environmental benefits to the costs of CCS options. Based on this information, they would be able to take informed action to help motivate policy direction that they believe is appropriate for the public interest. At a local level, public interest groups are interested in the integrated information in order to promote the outreach and education that can accompany the development of projects. This sort of public interest has been shown to be an integral factor in the support [Heinrich et al, 2003] or suppression [de Figueirido, 2003] of projects.

Section 6.2: CCS Project Identification

The underlying theme of many of the stakeholders is the understanding of how CCS projects will be identified and developed. This identification includes understanding the necessary analyses of CCS components, how the components interact spatially, and how projects can be selected for near term and long term development. For each of these tasks, the quality of information available and the ability to integrate data from a variety of sources are primary components.

6.2.1: CCS Component Characterization

General consensus among CCS researchers states that the major types of components in a CCS system are the CO_2 emissions source, the CO_2 sink, and the transport system that takes the CO_2 from source to sink. However, the process of determining the essential data characteristics that are needed for analysis is still underway.

For the sources, the quantity emissions of power plants and the high purity emissions of industrial facilities are believed to offer the most economical capture opportunities. At the level of project identification, analysis of these sources should produce the costs for capture and the amount of CO_2 that needs to be storage. In order to accurately model these costs, specific facility characteristics are required. Some of the major characteristics that will affect the capture costs are the type of plant, the current production technology, the primary fuel used, the plant size, and surrounding land use. Each of these characteristics will change the types of retrofit and capture technologies available.

For sinks, current projects are limited to injection into deep saline aquifers for storage, and injection into oil reservoirs for enhanced oil recovery (EOR). However, there is interest in migrating the techniques developed for aquifer storage and EOR to other sinks such as depleted oil and gas and coal beds. Analyses of sinks should include measurements of storage capacity, injection costs, and storage duration. Analyses of geologic sinks require characterization of the sink's physical properties: porosity, permeability, pressure, temperature, depth, thickness, and seal type.

In terms of transport, current options include using available CO_2 transport facilities such as trucking, freight, and built pipelines or building new pipeline infrastructure. The costs associated with using the current infrastructure will depend on the characteristics such as the accessibility of the source and sink to the available pathways: roads, rivers, and pipes, and the quantity of CO_2 being transported. In order to calculate the costs of building new pipeline infrastructure, the topography of the land, the land use, and the barriers to construction are needed. For low quantities of CO_2 that may occur during small scale projects, the flexibility and low capital cost of trucking may prevail where new pipelines would be more suited to large flows of CO_2 in a large and long term project.

6.2.1: Spatial Relationship

Aside from characterizing individual components, identification of projects must consider how the components are spatially related. This relationship helps to describe the costs that are associated with a projects transport component.

On a national level, the spatial relationship among the CCS components can be used to determine the storage potential of various regions. High potential regions will contain sources and sinks with low capture and injection costs that are in close proximity to each other and that do not have major transport obstacles between them. On the other hand, if one region contains sources that are suitable for a distant region's sinks, the transport costs between them may make it prohibitively difficult to connect them.

At the local level, detailed project analyses can be made using the spatial relationships. This includes matching of sources and sinks and selection of transportation options. The matching process determines which of the available sinks can store the emissions from each of the sources. After matches are made, the different transport options can be considered by calculating the available paths and costs.

Initially, matching can use the characterization of sink capacity and injectivity paired with the characterization of a source's emission rate to determine if the sink is capable of storing the source's CO_2. A more complex approach would also measure how matching a source and sink would affect the costs of connecting other sources and sinks. This could include effects such as providing a more efficient route from a source to a sink by combining the flows from multiple sources, or by reducing the available capacity of a sink and thereby forcing sources to be matched to more expensive sink options.

Likewise, the transport structure can develop simply between one source and sink pair or in a setting with multiple sources and sinks. In each case, the goal is to either find a path using the existing transport facilities or a path through the terrain that can be used to build new pipelines.

Section 6.3: Marginal Abatement Curves

After these characterizations and spatial relationships are established, analysis tools can be employed to help select projects for near-term pilots and forecast the costs and benefits for long-term CCS development. Tools such as CO_2 abatement curves enable the dissemination of the cost information critical for making these decisions.

A marginal abatement curve shows the relationship between the amount of CO_2 that can be stored and the cost of storing the last unit of CO_2. It can be generated with more ease and less precise data than some other analysis techniques, but still delivers results that are useful to policy development. This type of curve will be most important in early stages of

CCS analysis while data is still uncertain and project are still being considered in a general fashion.

This curve is produced by first calculating the cost and amount of CO_2 stored for individual projects, then combining them together in order of increasing unit cost. Consider that projects P_0 to P_n have respective costs and storage quantities of C_0 to C_n and Q_0 to Q_n. The costs can be calculated by adding up the component costs associated with a particular CCS project. The storage quantity can be calculated by adding up the amount of CO_2 captured from each source in the project. After these numbers are found, the unit cost of storage for each project P_i is simply Q_i/C_i. As an example, Figure 6.1 shows what a marginal abatement curve would look like with six projects ranging in size and cost.

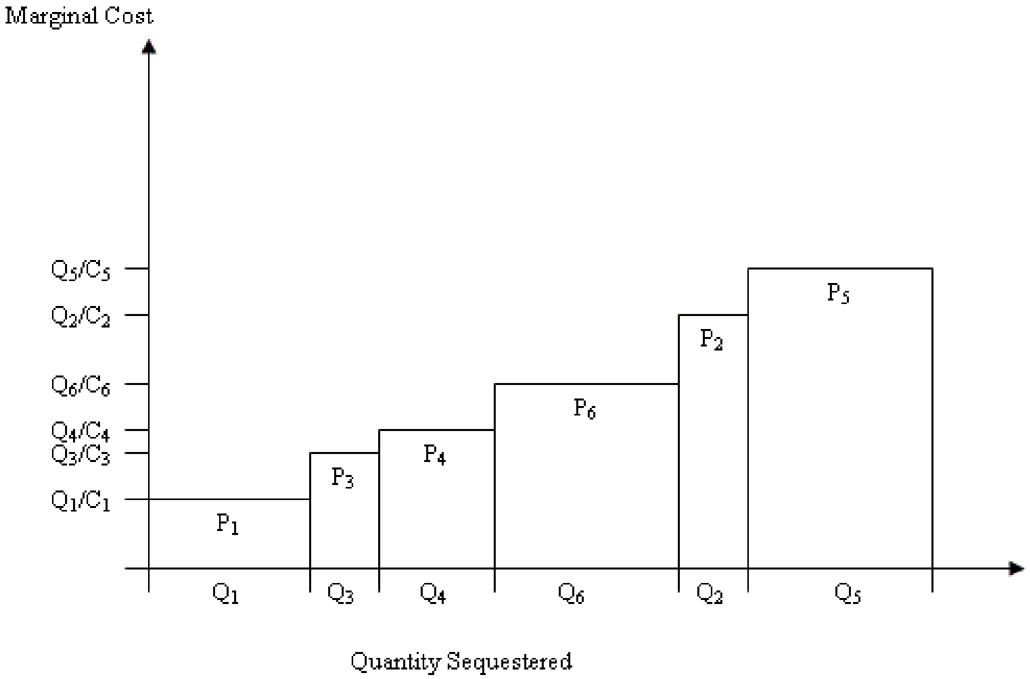


Figure 6.1: Example marginal abatement curve

In the application to CCS analyses, it is unlikely that the cost of every project will be determined in order to generate an abatement curve. Instead, case studies in a variety of potential projects would be used to understand the trends of costs for different types of projects. These cost trends could then be used to extrapolate the costs of projects with similar attributes and build an abatement curve.

Marginal abatement curves are an important factor in the making CCS policy decision. They enable decision makers to compare the cost for abatement through CCS verses other carbon management options. The curves also allow policy makers to estimate the costs that will be induced due to potential policies that are intended to induce CO_2 reductions.

Section 6.4: Current Initiatives

Two current initiatives are advancing the work done in characterizing major components, developing the understanding of spatial relationships, and performing cost analyses for CCS. The Regional Carbon Sequestration Partnerships (RCSP) are focusing on efforts to build the basic knowledge of CCS. The National Carbon Sequestration Atlas (NATCARB) is developing the information management strategies that can be used to bring the data together. The results of these projects will be used by NETL to develop policies and practices for carbon management in the future.

Database and GIS tools, as well as the techniques of context mediation and information integration studied in this thesis will be essential to manage the large amounts of information are being gathered and generated in these project. The use of or parallel between DIMS and these projects is discussed after the project description.

6.4.1: Regional Carbon Sequestration Partnerships

Regional Carbon Sequestration Partnerships (RCSP) are collaborative efforts between government, industry, academia, and non-profit organizations that are focused on studying the options for CCS in a specific region of the U.S. NETL has selected seven RCSP (RCSP) from proposals around the nation to work on the studies. The RCSPs are tasked with up to three phases of work, with the continuation of projects depending on the results of previous phases. The first phase is currently underway and consists primarily of information gathering and analysis of CCS alternatives within each region.

The second phase is the deployment of a field study through a small test project that is designed in Phase 1. The third phase is a larger scale deployment of CCS technologies into the region to affect significant reduction in the region's carbon intensity.

Each RCSP defined its own region by considering similarities in geographic properties and CO_2 emissions characteristics. This allows the partnership to concentrate effort in understanding the region's CCS potential. Figure 6.2 shows the extents of each of the RCSPs. The states which are associated with each partnership is filled with a color representing the partnership, with a few states showing two colors because they are considered in two partnerships. A number of states are not considered in any of the original partnerships but may be brought into a partnership as the project continues. The figure also marks the location of each partnership's lead organization with a star. The seven partnerships are:

1. Midwest Regional Carbon Sequestration Partnership
2. Midwest Geological Sequestration Consortium
3. Southeast Regional Carbon Sequestration Partnership
4. Southwest Regional Partnership for Carbon Sequestration
5. West Coast Regional Carbon Sequestration Partnership
6. Northern Rockies and Great Plains Regional Carbon Sequestration Partnership
7. Plains CO_2 Reduction Partnership

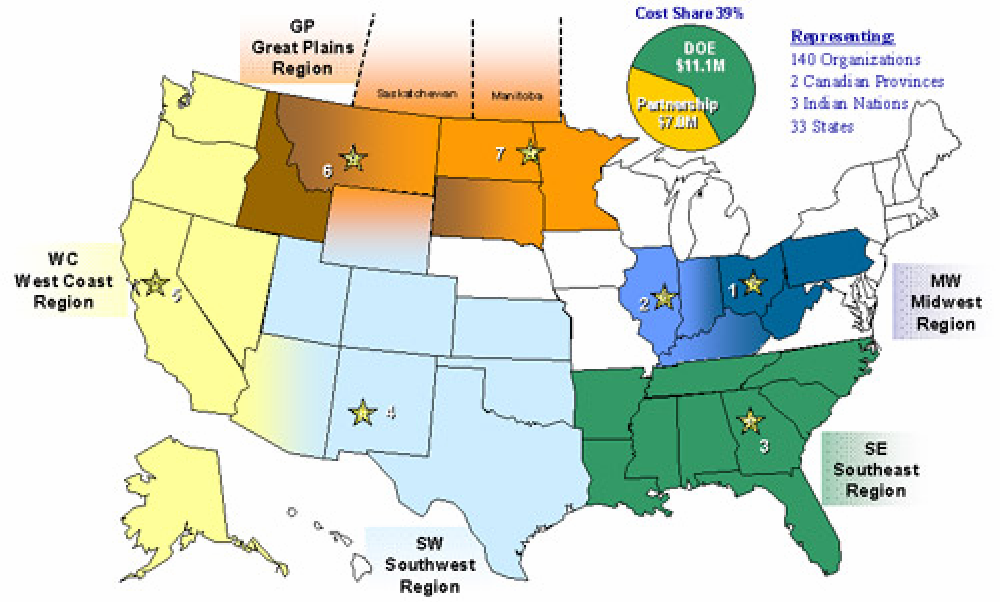


Figure 6.2: Map of Regional Partnerships [NETL, 2004]

Each of these partnerships is current working on Phase 1 of the project. This is an 18 month intensive study into the regions CCS potential. The Phase 1 goals include [NETL, 2003] [U.S. Newswire, 2003]:

- Characterization of regional CCS options
  - Options and opportunities for CO_2 capture and storage
  - CO_2 transport options
  - Regulatory permitting
  - Communications and outreach
  - Public acceptance
  - Monitoring and verification requirements
  - Environmental efficacy of sequestration
- Identification of the most promising options
  - Development of tools and analyses
- Preparation of plans
  - Cost-effective CCS systems
  - CCS systems suitable for pilot projects

In order to allow the partnerships to produce analyses that are most appropriate for their region, they are essentially given free reign on how to meet the goals of the phase. One of the few common requirements is that each partnership builds a Geographic Information System (GIS) for their region. The GIS will act as a central repository for data and results of the research, and will be provided as one of the deliverables to NETL at the conclusion of the phase. Technical data on sources and sinks are being culled from previously developed databases and produced through the partnership members with direct contacts to companies with more accurate data on the characteristics of the sources and sinks. The GIS will also be used during the phase as a communication channel between different groups of the partnership. Preliminary analysis results that are developed are stored in the

GIS and can then be accessed by the other members for consideration in developing public outreach plans and for incorporation into the final studies.

The results of data gathering and analyses will be brought together in a test project plan for each region. These plans will describe the region's CCS options and project possibilities and provide detailed for implementing some of the most promising CCS options as a Phase 2 project. From the results of each of the RCSPs, NETL and policy-makers will be provided with a set of the most relevant information for CCS and several CCS systems to consider.

The CCSTP GIS is being utilized directly by two of the RCSPs: the West Coast Regional Carbon Sequestration Partnership (WCRCSP) and the Southeast Regional Carbon Sequestration Partnership (SERCSP). The Analysis Layer of the GIS is being utilized by the WCRCSP to centralize all of their analysis tools. These tools are benefiting from the ability to use diverse data sources through the single interface provided by the Knowledge Layer. The local Data Source Layer is being used by the SERCSP to manage their developing data sets and analysis results. The GIS will serve these data through the Data Interface Layer in order to provide the benefits of context management to the SERCSP members.

6.4.2: National Carbon Sequestration Atlas

The National Carbon Sequestration (NATCARB) Atlas is a project intended to gather geologic and geographic data on the many components of CCS into a single accessible location. It is specifically developed as a portal with minimal storage of actual data. Instead, the data and tools that are available through the portal are stored on separate and distributed servers across the country. The goals of NATCARB are as follows [Bartley et al, 2004]:

- Provide an intelligent portal to users
    - Access to national data on carbon sequestration
    - Access to distributed tools
- Query data and tools from federation of distributed servers
- Develop partnerships
    - Synergy and communication in the carbon sequestration community

An intelligent portal is defined as one which is able to process a user's request for data in a specific geographic location or a specific analysis tool, automatically determine where the data or tool is stored among the distributed servers, and create individual requests to the distributed servers to retrieve the desired information or tool. This portal is a continuation of the work done with Mid-continent Interactive Digital Carbon Atlas and Relational dataBase (MIDCARB), and will leverage existing technology and expertise, as well as the difficulties uncovered in the previous work.

Initially, NATCARB will be developed in conjunction with the RCSPs, allowing NATCARB to quickly gather data linkages and allowing the RCSPs to provide an aggregated national view on their data. The process used in MIDCARB to aggregate data from the five states is being extended to include data from the seven RCSPs. However, there are some significant changes to the process since MIDCARB's previous mechanism often encountered significant delays and bottlenecks while gathering data.

Two of the major issues uncovered during the MIDCARB project are the difficulty in managing many layers of information and performance problems with the architecture of the system. The portal accessed 125 different layers of information, each representing a view on a database table that was stored in one of five databases. Each of these layers required manual management of the metadata and configuration within the portal in order to properly access the data. In addition, all of the raw data was requested from the distributed servers on each user request. The data was used to generate the maps, then discarded. This architecture caused a great deal of network traffic for information that was never used.

The number of layers serviced caused management problems because metadata was managed manually and centrally. Each time a new layer of data was added to the MIDCARB system, the state that generated the layer contacted the portal administrator and requested that the inclusion of the new layer. The portal administrator was then required to modify parameters in the portal before this new data was usable. This technique is not scalable to the larger number of sources and layers at the national level that NATCARB intends to service. In order to alleviate this manual bottleneck, NATCARB builds a repository of metadata to contain the necessary information needed to connect to each distributed database and the detailed metadata about the data layers available in the database. Instead of centrally and manually managing this metadata, it is populated and managed by the administrators of the distributed servers. These remote administrators use an internet webpage connected to the NATCARB portal to enter the connection information for their own remote server. After this initial connection is made, the NATCARB portal automatically queries the distributed servers in order to discover all of the available layers. The remote administrator can then manage these layers remotely, indicating which layers the portal should allow users to view and/or query.

Performance problems in the MIDCARB portal were primarily caused by network constraints. In that portal, all of the raw data of a layer was copied to the MIDCARB server and then was processed into an image for the portal and finally published to the website. Because of the configuration of the server, this process occurred for each request, ensuring that the most current data was being used in the portal, but also incurring large penalties to the amount of data that was being requested. This made the network bandwidth and delays between the portal and other state servers a major factor in the responsiveness of the system, with responsiveness to user queries affected dramatically by the number of users and frequency of queries. NATCARB intends to reduce the effect of network speed to the responsiveness of the system by initially requesting a much smaller amount of data from RCSP servers. Instead of the actual data, the portal will request the mappable images of the regional data from the remote servers.

The remote servers will generate the image and send it back to the portal. The portal will then collect all of the regional layers and generate a national layer of background data and then combine all of these into the final image that is delivered from the portal. Actual data is only requested from the remote servers when a user specifically queries information from a layer. This technique reduces both the quantity of data transmitted between servers as well as the amount of processing required at the portal.

This national database will be beneficial to the policy process because broad analyses can be performed on national level. In this early stage of development, NATCARB is providing a way to see the results of the RCSPs in a side-by-side comparison, benefitting NETL when studying the Phase 1 results of the RCSPs. It also is bringing together the major regional data providers so that the data issues can be discussed and jointly addressed.

Both the research done for this thesis and the work done in NATCARB focus on improving the accessibility to CCS data through automating some data processing. The Data Interface Layer of DIMS is similar in nature to the meta-data registry provided by NATCARB. Through these mechanisms, each project is able to aggregate data that is coming from many different sources.

However, there are also differences between the two systems that are summarized below:

	NATCARB	DIMS
Focus	Provide accessible and easy to use portal to view data and tools. Allows users to quickly see data.	Provide access to integrated data. Allows users to utilize the data for computational and programatic analyses
Metadata	Stored in database registry. Input and updated by data providers. Used to store remote server access parameters.	Stored in database registry. Input and updated by DIMS users. Used to store remote database access parameters and context information for context mediation
Data connections	Maintained through ArcIMS protocols. Portal server connects to distributed ArcIMS servers	Uses a variety of methods to access data. DIMS server connects to distributed databases and files
Data transfer	First transfers image from remote server. Portal only queries data upon user request for further information.	Transfers data from specific columns that are accessed in DIL and KL tables

NATCARB may benefit from DIMS or similar context management and integration frameworks to manage increases in size and complexity of its network of data sources. For example, integration procedures will be important if remote servers are added to the NATCARB system that are not working as collaboratively as the RCSPS. These new servers may contain repeated or conflicting information. Instead of duplicating this information in the user's view, it would be most appropriate to integrate the two sources together, as has been done in DIMS.

Chapter 7: Conclusion

The Distributed Information Management System (DIMS) is an implementation of novel information management technologies in the area of carbon dioxide capture and storage (CCS) research and analysis. These technologies are used to mediate the context differences between data sources and to integrate the databases together. The need for context mediation arises because the currently available data that is being culled from a variety of sources that were not originally intended for CCS use. The need for integration brings together data that is required for analyses from the sources that, individually, only supply part of the necessary data.

DIMS is already being utilized in projects that are supporting CCS analyses and the development of carbon management policies. The improved quality and completeness of data in the DIMS system demonstrates the benefits of using the information management technologies.

The development and implementation of DIMS has uncovered a number of issues with CCS information. These issues are the basis for the following recommendations to the CCS community:

- Manage current data using integration technologies: The current data was not collected for the purpose of CCS analyses. However, integration can be used to maximize the usability and value of the data by allowing the users of data to define information topics and then retrieve only the relevant data.
- Support the collection of new CCS data: The current data sources can undervalue or ignore CCS factors. Initiatives to collect and improve the data with specific focus on CCS requirements will enable the analyses to be more accurate.
- Encourage development of information quality: Building quality throughout the data development process improves the information because specific knowledge and local expertise can be applied to the information. Important Information Quality (IQ) metrics to consider in the area of CCS are accuracy, precision, timeliness, completeness, reliability, believability, and consistency.

Acting on any of these recommendations will improve the state of CCS information that is being used to develop analyses and policies. By improving the available information, DIMS and related GIS systems will provide several benefits to the research and analysis in CCS and the future policy development. In particular, DIMS will be beneficial in the following policy applications:

- Managing data sources to provide consistent access: DIMS provides decision-makers the ability to access data in the context that is most comfortable for them. This reduces the amount of confusion that the decision-makers will encounter when considering data.
- Integration of data for system-level analyses: The ability of DIMS to bring together data from a variety of sources enables the Carbon Capture and Sequestration Technologies Program's (CCSTP) GIS to develop system's

analyses. These analyses will aid in the understanding of interactions between CCS components and provide a more complete picture for decision-makers.

- Improving public awareness and education of CCS: Public knowledge in CCS is currently limited, but is a primary factor in the motivating expansion of CCS. Integration systems can help in the education process by highlighting relevant information, correlations, and contradictions. These can then be delivered in a context that is consistent with the viewpoint of the public.

References

J. Bartley, T. Carr, D. Cheng, et al. "Creating a distributed national database for carbon sequestration". Presented at ESRI Petroleum Users Group Conference. February 2004.
H.J. Herzog, D. Golomb. "Carbon Capture and Storage from Fossil Fuel Use". Encyclopedia of Energy. To be published 2004.

Energy Information Agency. "Emissions of Greenhouse Gases in the United States". EIA Report, #EIA/DOE-0573(2001). Released December, 2002. http://www.eia.doe.gov/oiaf/1605/gg02rpt/carbon.html

G. Heddle, H. Herzog and M. Klett. "The Economics of CO2 Storage". MIT LFEE 2003-003 RP. August 2003.

P.J.P. Egberts, J.F. Keppel, A.F.B. Wildenborg, et al. "A Decision Support System for Underground CO_2 Sequestration". Greenhouse Gas Control Technologies (GHGT6) Proceedings. October 2002.

J.J. Dooley, J.A. Edmonds, R.T. Dahowski, et al. "Modeling Carbon Capture and Storage Technologies in Energy and Economic Models". IPCC Workshop on Carbon Dioxide Capture and Storage Proceedings. 2002.

J. Gale. "Overview of CO_2 emission sources, potential, transport, and geographical distrobution of storage possibilities. IPCC Workshop on Carbon Dioxide Capture and Storage Proceedings. 2002.

P. Freund, J. Davison. "General overview of costs". IPCC Workshop on Carbon Dioxide Capture and Storage Proceedings. 2002.

B. Bock, R. Rhudy, H. Herzog. "Economic Evaluation of CO_2 Storage and Sink Enhancement Options: Interim Final Technical Report". Tennessee Valley Authority Public Power Institute. December 2002.

M. Webster, C. Forest, J. Reilly, et al. "Uncertainty Analysis of Climate Change and Policy Response". MIT Joint Program on the Science and Policy of Global Change. Report No. 95. December 2002

S.T. Brennan, R.C. Burruss. "Specific Sequestration Volumes: A Useful Tool for CO2 Storage Capacity Assessment". Second Annual Conference on Carbon Sequestration Proceedings. May 2003.

C.O. Karacan, P.M. Halleck, A.S. Grader, et al. "Kinetics of the Physical Changes and Gas Storage Capacity Induced by Carbon Dioxide Sequestration in Coal". Second Annual Conference on Carbon Sequestration Proceedings. May 2003.

M.M Maroto-Valer, M.L. Druckenmiller, J.M. Andresen. "In-Situ Study of Carbon Dioxide Sequestration in Saline Brine Formations". Second Annual Conference on Carbon Sequestration Proceeding. May 2003.

R.B. Grigg, B.J. McPherson, R.K. Svec. "Laboratory and Model Tests at Reservoir Conditions for CO2-Brine-Carbonate Rock Systems Interactions". Second Annual Conference on Carbon Sequestration. May 2003.

S.J. Freidmann, D. Nummedal. "Reassessing the Geological Risks of Seal Failure for Saline Aquifers and EOR Projects". Second Annual Conference on Carbon Sequestration. May 2003.

C.M. White. "An Initial Set of Working Hypotheses Concerning Some Chemical, Physical, and Thermodynamic Phenomena That Occur when CO2 is Injected into a Coalbed". Second Annual Conference on Carbon Sequestration. May 2003.

R. Kovac, Y.W. Lee, L.L. Pipino. "Total Data Quality Management: The Case of IRI". Proceedings of the 1997 Conference on Information Quality, October 1997, pp. 63-79. http://web.mit.edu/tdqm/www/tdqmpub/IRITDQMCaseOct97.pdf

W.Y. Chung, C. Fisher, R. Wang. "What Skills Matter in Data Quality?". Proceedings of the Seventh International Conference on Information Quality, November 2002, pp. 331-342.
http://web.mit.edu/tdqm/www/tdqmpub/WSMDQ-ICIQNov02.pdf

J.D. Funk, Y.W. Lee, R.Y. Wang. "Institutionalizing Information Quality Practice: The S. C. Johnson Wax Case," Proceedings of the 1998 Conference on Information Quality, October 1998. pp. 1-17.
http://web.mit.edu/tdqm/www/tdqmpub/SCJTDQMCaseOct98.pdf

D.M. Strong, Y.W. Lee, R.Y. Wang. "10 Potholes in the Road to Information Quality". IEEE Computer, Vol. 30, No. 8. August 1997. pp. 38-46

S.E. Madnick. "Metadata Jones and the Tower of Babel: The Challenge of Large-Scale Semantic Heterogeneity". MIT Sloan School of Management Working Paper, WP#4069. 1999.

M. Hansen, S. Madnick, M. Siegel. "Data Integration using Web Services". MIT Sloan School of Management Working Paper, WP 4406-02. May 2002
H. Zhu, S.E. Madnick, M.D. Siegel. "Global Comparison Aggregation Services. MIT Engineering Systems Division Working Paper Series, ESD-WP-2002-08. December 2002.

S.E. Madnick. "The Misguided Silver Bullet: What XML Will and Will Not Do to Help Information Integration". MIT Sloan School of Management, WP 4185-11. October 2001.

G. Shankaranarayan, M. Ziad, R.Y. Wang. "Managing Data Quality in Dynamic Decision Environments: An Information Product Approach. Journal of Data Managment, Forthcoming. 2003.

Y. Wand, R.Y. Wang. "Anchoring Data Quality Dimensions in Ontological Foundations". Communications of the ACM, Vol. 39, No. 11. November 1996. pp. 86-95.

A. Firat, S. Madnick, B. Grosof. "Knowledge Integration to Overcome Ontological Heterogeneity: Challenges from Financial Information Systems". Internation Conference on Information Systems Proceedings. December 2002.

"Users Manual: Emissions & Generation Resources Integrated Database". U.S. Environmental Protection Agency, Office of Atmospheric Programs. September 2001.

S.W. White, T.R. Carr, J.A. Drahovzal, et al. "An Update on the Midcontinent Interactive Digital Carbon Atlas and Relational dataBase (MIDCARB) and its Future". Second Annual Conference on Carbon Sequestration. May 2003.
http://www.midcarb.org/Documents/NETL-May-2003.pdf

T.R. Carr, J.D. Bartley, K.A. Nelson, et al. "The MIDCARB Carbon Sequestration Project: Midcontinent Interactive Digital Carbon Atlas and Relational dataBase". GSA Annual Meeting. October 2002.
http://www.kgs.ku.edu/PRS/publication/2002/ofr2002-45/GSA2002.pdf

F. Floris, T. Wildenborg. "GESTCO-DSS: Software Requirements Specification (Draft)". Netherlands Institute of Applied Geoscience TNO Report. October 2000.

R.T. Dahowski, J.J. Dooley. "Carbon Management Strategies For Existing U.S. Generation Capacity: A Vintage-Based Approach". Greenhouse Gas Control Technologies (GHGT6) Proceedings. October 2002.

R.T. Dahowski, J.J. Dooley. "A Vintage Based Approach for Assessing Carbon Sequestration Options for U.S. Power Plants". Second Annual Conference on Carbon Sequestration Proceedings. May 2003.

C. Hendriks, A.S. van der Waart, C. Byrman, et al. "Building the Cost Curve for CO2 Storage: Sources of CO2". IEA Greenhouse Gas R&D Programme, Final Report: M70012. July 2002.

C. Hendriks, A.S. van der Waart, C. Byrman. "A Decision Support System for Underground CO_2 Storage". Greenhouse Gas Control Technologies (GHGT6) Proceedings. October 2002.

D. Cheng, T. Curry, A. Smith. "Analysis of Carbon Management GIS Data (draft)". Carbon Capture and Sequestration Program Working Paper. September 2003.

R. Dahowski, J. Dooley, D. Brown, et al. "Understanding Carbon Sequestration Options in the United States: Capabilities of a Carbon Management Geographic Information System". Battelle/PNNL. 2001.

"Regional Carbon Sequestration Partnerships". National Energy Technology Laboratoy. January 2004.
http://www.netl.doe.gov/coalpower/sequestration/partnerships/

"Energy Secretary Abraham Creates Regional Partnerships to Develop Carbon-Sequestration Options; Initiative to Address Options". U.S. Newswire. September 2003
http://releases.usnewswire.com/GetRelease.asp?id=121-09022003

H. Zhu, S.E. Madnick, M.D. Siegel. "The Interplay of Web Aggregation and Regulation". MIT Engineering Systems Division Working Paper Series, ESD-WP-2002-07. November 2002.

G.W Bush. "President Announces Clear Skies & Global Climate Change Initiatives". February 2002.

http://www.whitehouse.gov/news/releases/2002/02/20020214-5.html
"Fact Sheet: President Bush Announces Clear Skies & Global Climate Change Initiatives". February 2002.
http://www.whitehouse.gov/news/releases/2002/02/20020214.html

"Global Climate Change Policy Book". February 2002.
http://www.whitehouse.gov/news/releases/2002/02/climatechange.html

M. Webster, C. Forest, J. Reilly, et al. "Uncertainty Analysis of Climate Change and Policy Response". December 2002.
http://web.mit.edu/globalchange/www/MITJPSPGC_Rpt95.pdf

S. Ernst. "Bill takes aim at greenhouse gas emissions". Puget Sound Business Journal. March 3, 2003.
http://seattle.bizjournals.com/seattle/stories/2003/03/03/story4.html

J. Lieberman. "Climate Stewardship Act of 2003". Bill Number: S139. 108th Session of the U.S. Congress. January 2003.

"California Governor Signs Nation's First Law To FIght Global Warming With Forest Conservation". The Pacific Forest Trust. September 9, 2002.
http://www.pacificforest.org/news/sb812.html

"Voluntary Reporting of Greenhouse Gases 2001 Summary". Energy Information Administration. Report #DOE/EIA-0608. February 19, 2003

N. Choucri, S. Madnick, M. Siegel. "LIGHTS: Laboratory for Information Globalization and Harmonization Technologies and Studies". MIT Working Paper, CISL #2003-08. February 2003

N. Choucri, S. Madnick, M. Siegel. "Laboratory for Information Globalization and Harmonization Technologies: A New Research Initiative". MIT Sloan School of Management Working Paper, WP 4350-01. December 2001.

R. Wang, T. Allen, W. Harris, et al. "An Information Product Approach for Total Information Awareness". IEEE 2003. 2003

D. Caterinicchia. "DARPA builds open-source rankings". Federal Computer Week. March 2002.

D. Caterinicchia. "Data mining aims at national security". Federal Computer Week. March 2002.

B. Perens. "Why Security-Through-Obscurity Won't Work". Slashdot Feature. July 1998. http://slashdot.org/features/980720/0819202.shtml

"Copyright Basics (Circular 1)". U.S. Copyright Office. 2003. http://www.copyright.gov/circs/circ1.html

"MOHOMINE LAUNCHES UNSTRUCTURED DATA MANAGEMENT SOFTWARE AS OEM PRODUCT FOR ENTERPRISE APPLICATION VENDORS, GOVERNMENT". Mohomine Press Release. December 2001. http://mohomine.com/news/companynews20011205.asp

"Knowledge Management Strategic Investments". In-Q-Tel. 2003. http://www.in-q-tel.com/tech/km.html

D.S. Cheng. "Balanced Location Information Policies: A Stakeholder Analysis Based on Increased User Management of Location Information". MIT 6.805 Paper. May 2002.
M. de Figueiredo. "The Hawaii Carbon Dioxide Ocean Sequestration Field Experiment: A Case Study in Public Perceptions and Institutional Effectiveness". MIT Theses Collection. March 2003.

M.A. de Figueiredo, D.M. Reiner, H.J. Herzog. "Towards a Long-Term Liability Framework for Geologic Carbon Sequestration". Second Annual Conference on Carbon Sequestration Proceedings. May 2003.

J.J. Heinrich, H. J. Herzog, and D.M. Reiner, "Environmental Assessment of Geologic Storage of CO2, MIT LFEE 2003-002 RP, December (2003).

J. Gray. "Distributed Computing Economics". Microsoft Research Technical Report #MSR-TR-2003-24. March 2003.

F. Chen, B.D. Ripley. "Statistical Computing and Databases: Distributed Computing Near the Data". Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC2003). March 2003.

T. Curry. "Public Survey of Opinions on Carbon Capture and Storage: Preliminary Results". Proceedings of Alliance for Global Sustainability (AGS) Technical Meeting. November 2003.

H. Zhu. "Temporal Contexts". COIN Presentation. November 2003.
C. Ding, H. Zha, X. He, et al. "Link Analysis: Hubs and Authorities on the World Wide Web". LBNL Tech Report 47847. May 2001.

S.D. Hovorka, M.H. Holtz, P. Knox, et al. "Technical Summary: Optimal Geological Environments for Carbon Dioxide Disposal in Brine Formations (Saline Aquifers) in the United States". University of Texas, Bureau of Economic Geology. 2002.

D. Cheng, T. Curry, A. Smith, et al. "Analysis of Carbon Management Data". Second Annual Conference on Carbon Sequestration Proceedings. May 2003.

Chapter 9: Appendices

Appendix A: List of Acronyms

AL	Analysis Layer
CCS	Carbon Capture and Sequestration
CCSP	Carbon Capture and Sequestration Technologies Program
COALQual	Coal Quality Database
COIN	COntext INterchange system
CSV	Comma Separated Values
DIL	Data Interface Layer
DIMS	Distributed Information Management System
DOE	Department of Energy
DSL	Data Source Layer
ECBM	Enhanced Coal Bed Methane
EOR	Enhanced Oil Recovery
EPA	Environmental Protection Agency
ESRI	Environmental Systems Research Institute, Inc.
Gg	Giga-grams
GESTCO	European Potential for Geological Storage of Carbon Dioxide from Fossil Fuel Combustion
GIS	Geographic Information System
IQ	Information Quality
KGS	Kansas Geological Survey
KL	Knowledge Layer
LIGHTS	Laboratory for Information Globalization and Harmonization Technologies and Studies
MB	Mega-Bytes
MIDCARB	Mid-continent Interactive Digital Carbon Atlas and Relational dataBase
MIT	Massachussetts Institute of Technology
NATCARB	NATional CARBon sequestration atlas

NETL	National Energy Technology Laboratory
PNNL	Pacific Northwest National Laboratory
RCSP	Regional Carbon Sequestration Partnership
TORIS	Total Oil Recovery Information System
UIL	User Interface Layer
USGS	United States Geological Survey
WWW	World Wide Web

Appendix B: System Design Supplement

9.B.1 User Interface Layer

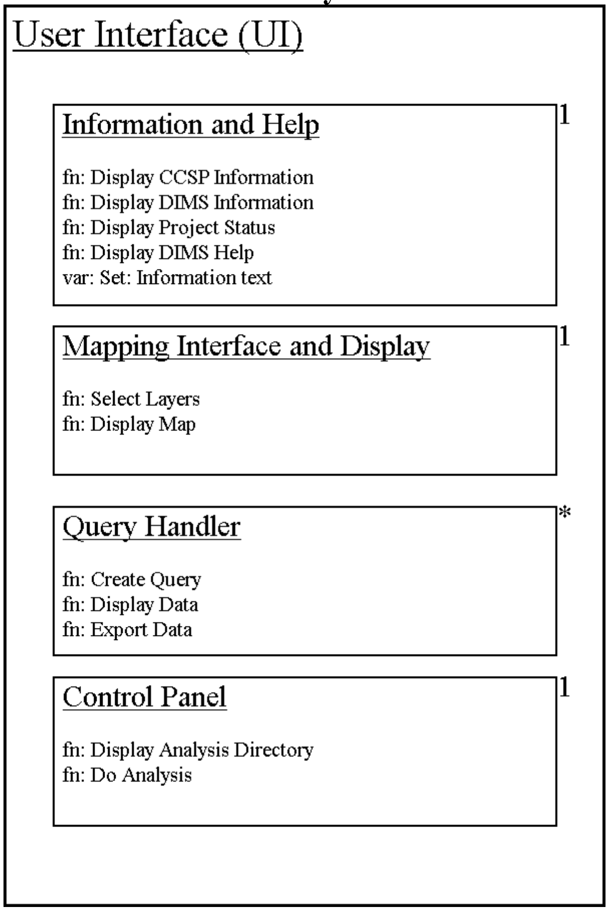


Figure 9.B.1: User Interface Layer Diagram

Modules in UI include:

- Informational and help screens: A module that will provide introductory information and context for the project and provide help in using the system.
  - Display CCSP Information: Displays information on CCSP to the user
  - Display DIMS Information: Displays information on DIMS to the user
  - Display DIMS Project status: Displays the status of the DIMS project and milestones to the user
  - Display DIMS Help: Displays help screens to assist user in working with the system.
- Spatial mapping interface and display: A module that will retrieve information from K that pertains to the user request and display it to the user in an easily understandable form.
  - Select Layers: Selects the layers to be displayed on the display
  - Display Map: Displays a requested graphical map
- Query handler: Modules that the user can interact with to retrieve subsets of data specific to a particular question, and display the results.

    - Create Query: Assists in creation of a information query
    - Display Data: Displays a set of query results that the user requests
    - Export Data: Exports data into another format for the user
- Control panel: A module that allows the user to interact with the system and request new analyses. This module will translate the user commands into control signals for A.
    - Display Analysis Directory: Displays the set of available analysis tools
    - Do Analysis: Requests that the DIMS system performs an analysis

9.B.2 Analysis Layer

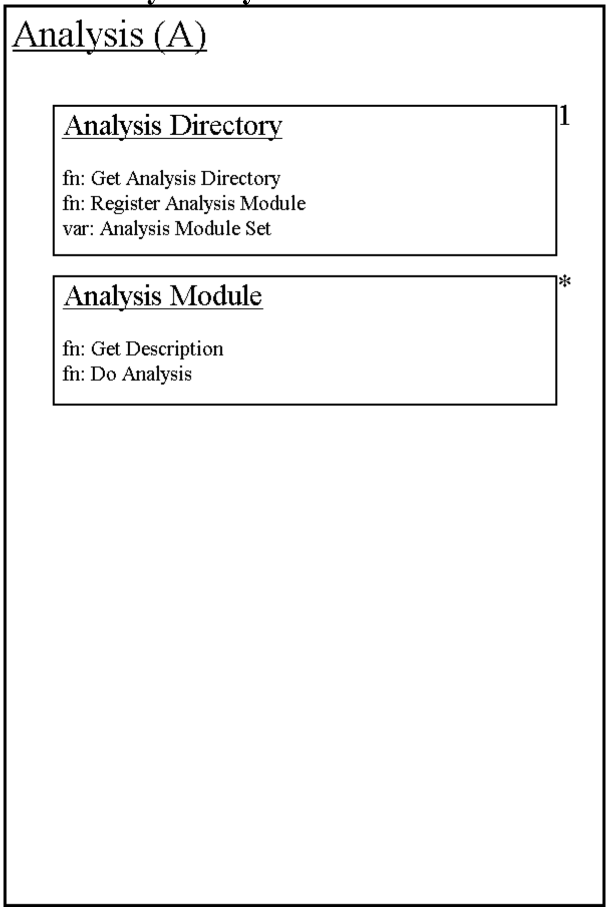


Figure 9.B.2: Analysis Layer Diagram

Modules in A include:

- Analysis Directory: A module that can be used to discover the analysis modules that are available in the system. The directory will describe each analysis module, and its methods of invocation.
    - Get Analysis Directory: Returns the set of available analysis modules
    - Register Analysis Module: Adds an analysis module to the set of available modules
- Analysis Module: Modules that perform computational analyses. These are the workhorses of the system. Each analysis module can be developed to perform a different type of analysis.
    - Get Description: Gets the description of the analysis module

    - Do Analysis: Performs an analysis based on specified parameters

9.B.3 Knowledge Layer

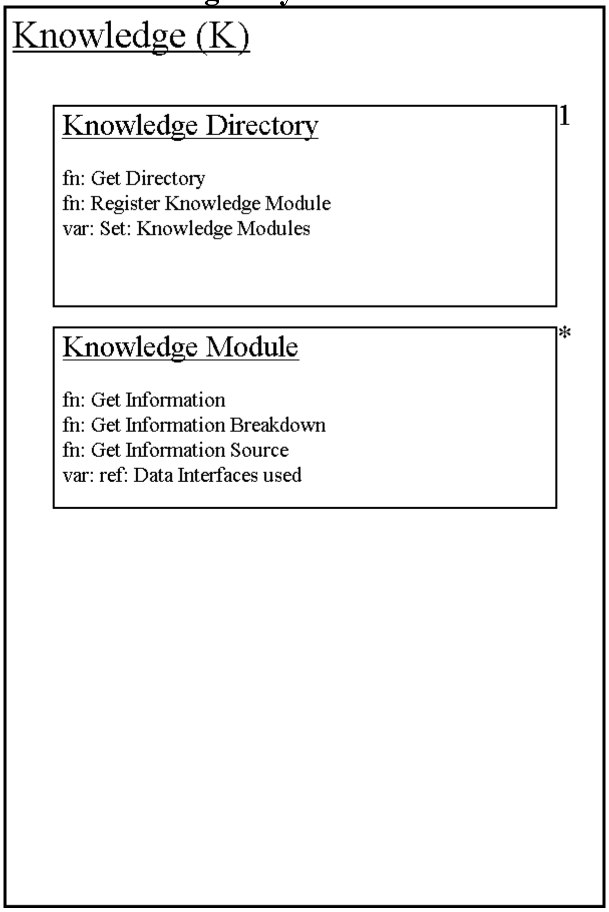

Knowledge (K)
Knowledge Directory
fn: Get Directory
fn: Register Knowledge Module
var: Set: Knowledge Modules
1
Knowledge Module
fn: Get Information
fn: Get Information Breakdown
fn: Get Information Source
var: ref: Data Interfaces used
*

Figure 9.B.3: Knowledge Layer Diagram

Modules in K include:

- Knowledge Directory: A module that can be used to discover the different knowledge and integration
    - Get Directory: Returns the set of available knowledge modules
    - Register Knowledge Module: Adds a new knowledge module to the set of available modules in the directory
- Knowledge Module: Modules that perform the task of gathering and integrating data from different data interfaces. These modules will be programmed with rules that define how various data can be integrated, so that the rules can be applied dynamically to new and updated data.
    - Get Information: Returns the integrated information that has been requested
    - Get Information Breakdown: Returns specifics on the data interfaces used in integrated information
    - Get Information Source: Returns specifics on the sources used in the integrated information

9.B.4 Data Interface Layer

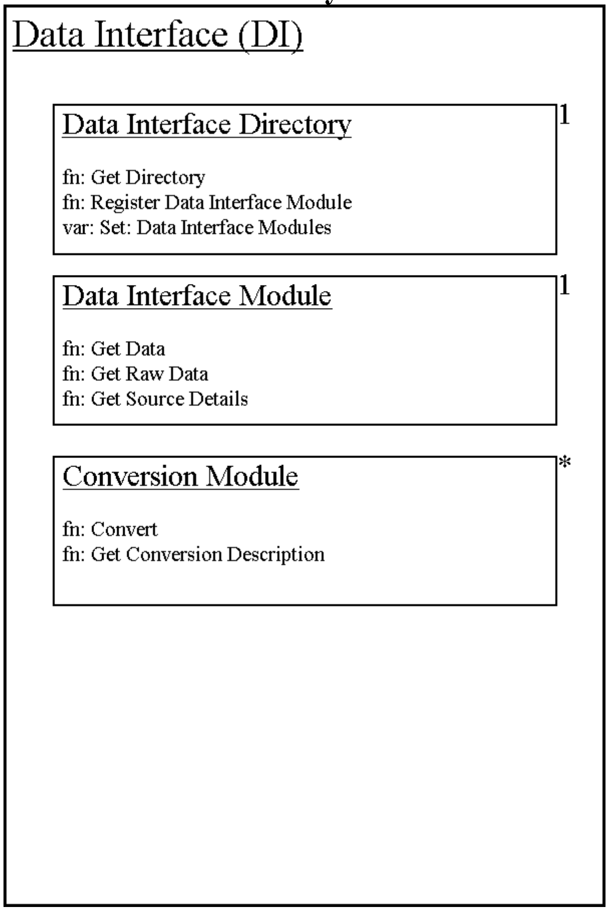
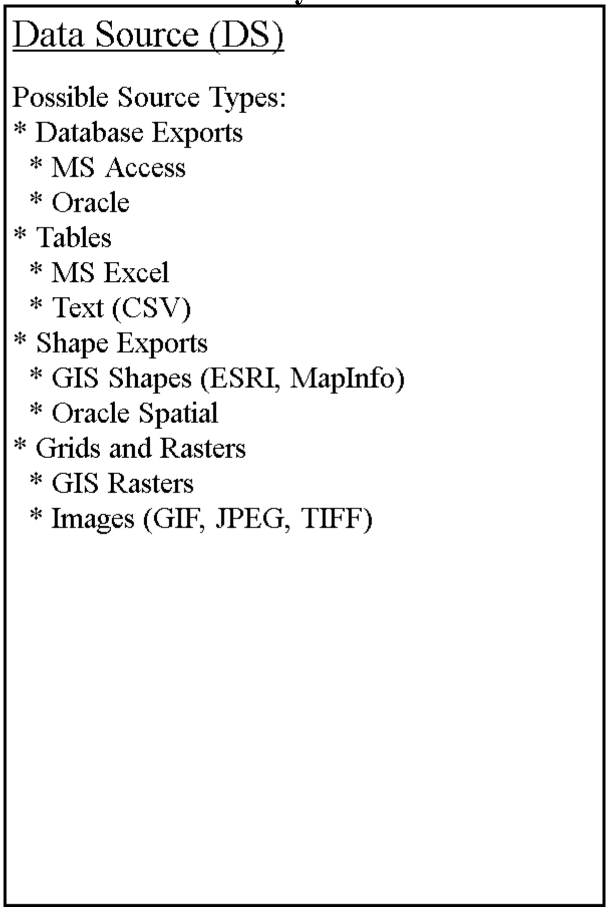


Figure 9.B.4: Data Interface Layer Diagram

Modules in DI include:

- Data Interface Directory: A module that can be used to discover the different data interfaces (i.e. data source) that are available
  - Get Directory: Returns the set of data interfaces available
  - Register Data Interface Module: Adds a new data interface module to the set of available modules
- Data Interface Module: Modules that provide the informational interface to various data sources
  - Get data: Returns the data in the local context
  - Get raw data: Returns the data as delivered by the data source
  - Get source details: Returns information about the source of the data
- Conversion Module: Modules that assist in the conversion between different contexts
  - Convert: Converts data between contexts
  - Get Description: Gets description of the conversion module

9.B.5 Data Source Layer



Data Source (DS)

Possible Source Types:

* Database Exports
  * MS Access
  * Oracle
* Tables
  * MS Excel
  * Text (CSV)
* Shape Exports
  * GIS Shapes (ESRI, MapInfo)
  * Oracle Spatial
* Grids and Rasters
  * GIS Rasters
  * Images (GIF, JPEG, TIFF)

Figure 9.B.5: Data Source Layer Diagram

There are no defined modules in DS, as it represents a variety of possible data sources.

Appendix C: System Implementation Supplement

Hardware-Software configuration of DIMS System

- Integration node (E40-482-1.mit.edu)
    - Pentium, GHz
    - A Layer Software
        - ESRI ArcGIS 8.1: Display and basic analysis
        - VB scripts in ArcGIS: Analysis
        - Programs in Oracle: Analysis
    - K Layer Software
        - DIMS programs in Oracle: Database level integration
    - DI Layer Software
        - ESRI ArcSDE for Oracle: Data interface program for ESRI products
        - DIMS programs in Oracle: Database level interfaces
    - DS Layer Software
        - Oracle 9i Enterprise: Local database
        - MS Office - Access, Excel
- User Interface Node
    - Pentium, GHz
    - UI Layer
        - Oracle 9i Application Server with Apache: Information hosting
        - ESRI ArcIMS: Map Display
    - A Layer Software
        - ESRI ArcGIS 8.1: Display and basic analysis
        - VB Scripts in ArcGIS: Analysis
        - Programs in Oracle: Analysis
    - K Layer Software
        - DIMS programs in Oracle: Database level integration
- Miscellany
    - Networked on 10 Mbps Ethernet

Appendix D: Data Source Supplement

Data Sources (DS):

- Gas Information System (GASIS)
    - Description: The Gas Information System combines information from six previous gas atlases with information from Dwight's Energy Data and other sources to produce a database with powerful capabilities for exploration, development, planning, economic analysis, and market assessment
    - Source: NETL, DOE
    - Timeliness: 1999, no plan for further updates
    - Internet resource:
      http://www.netl.doe.gov/scng/projects/model/r-d/rdp28139.html
- Geographic Names Information System (GNIS)
    - Description: The Geographic Names Information System (GNIS), developed by the USGS in cooperation with the U.S. Board on Geographic Names (BGN). The Federally recognized name of each feature described in the data base is identified, and references are made to a feature's location by State, county, and geographic coordinates.
    - Format: Relational Table
    - URL:
      http://geonames.usgs.gov/
- Emissions and Generation Resource Integrated Database (eGRID)
    - Description: The Environmental Protection Agency (EPA) has gathered and distributed a database on aspects of all power plants in the US in order to track emission levels of compounds of interest from the plants.
    - Source: Environmental Protection Agency (EPA)
    - Format: Relational Tables (Excel)
    - Internet Resource:
      http://www.epa.gov/cleanenergy/egrid/index.html
- U.S. Streams and Water Bodies
    - Description: Map layer portraying the streams and waterbodies of the United States with associated official geographic names.
    - Source: US Geological Survey
    - Format: Shapefiles
    - Internet Resource:
      http://nationalatlas.gov/hydrom.html
- Mid-continent Interactive Digital Carbon Atlas and Relation dataBase (MIDCARB)
    - Description: Aggregation of five state geological survey databases used to evaluate the potential capacity for geologic sequestration of CO_2 in the member states.
    - Source: Mid-continent Interactive Digital Carbon Atlas and Relation dataBase (MIDCARB)
    - Format: Relational Table

    - Internet Resource:
      http://www.midcarb.org/
- GESTCO Carbon Source Database
    - Description: Database on carbon emission sources around the world. Estimates of CO_2 emissions are generated from many journals and databases.
    - Source: IEA, Ecofys
    - Format: Relational Table (Excel)
    - Internet Resource:
      N/A
- Electronic Topography, 5 minute gridded elevation data (ETOPO5)
    - Description: ETOPO5 was generated from a digital data base of land and sea- floor elevations on a 5-minute latitude/longitude grid
    - Source: NOAA, National Geophysical Data Center (NGDC/NOAA)
    - Format: Spatial Raster
    - Internet Resource:
      http://www.ngdc.noaa.gov/mgg/global/etopo5.HTML
- States and Counties
    - Description: Map layers portraying the 2000 state and county boundaries of the United States. Compiled by the U.S. Geological Survey from a variety of sources.
    - Source: US Geological Survey (USGS)
    - Format: Shapefiles (ArcGIS)
    - Internet Resource:
      http://nationalatlas.gov/statesm.html
      http://nationalatlas.gov/county00m.html
- U.S. Census Database, 2000
    - Description: This data table contains 2000 population information for total population counts, population density values, gender and age statistics, and various statistics on race and ethnicity distributions in the United States and Puerto Rico. The information was provided by the U.S. Census Bureau.
    - Source: Census
    - Format: Relational Table (DBF)
    - Internet Resource:
      http://nationalatlas.gov/census2000m.html
- Total Oil Recovery Information System (TORIS)
    - Description: Database developed by the National Petroleum Council (NPC) for its 1984 assessment of the nation's enhanced oil recovery (EOR) potential. The technical data description is at the reservoir level.
    - Source: National Petroleum Technology Office (NPTO/DOE)
    - Format: Relational Table
    - Internet Resource:
      http://www.npto.doe.gov/Software/dbindx.html
- Coal Quality Database (COALQUAL)

  - Description: A subset of the 13,035 samples contained in the NCRDS (National Coal Resources Data System) USCHEM (US geoCHEMical) database, and contains coal quality data in which a complete record represents a coal sample with a possible total of 136 fields.
  - Source: USGS
  - Format: Shapefiles
  - Internet Resource: http://energy.er.usgs.gov/products/databases/CoalQual/intro.htm
- Brine Database (UTBEG Brine)
  - Description: Developed data on brine databases determined to be high potential for CO_2 sequestration
  - Source: Univerity of Texas, Bureau of Economic Geologists
  - Format: Shapefiles, Raster files
  - Internet Resource: http://www.beg.utexas.edu/environqlty/co2seq/finalreport.pdf