DOE/BC/14833--T3

NATIONAL GEOSCIENCE DATA REPOSITORY SYSTEM

PHASE II: PLANNING AND PILOT STUDY

PROGRESS REPORT DE-FG22-95BC14833

1st. Quarter
January - March, 1997

Submitted by the
AMERICAN GEOLOGICAL INSTITUTE

to the

Office of Fossil Energy, U.S. Department of Energy

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MASTER

April 1997

THE ATTACHED REPORTS HAVE BEEN
ENTERED INTO THE DTS AND DISTRIBUTED
ON 7-2-97
THIS IS THE COPY FOR THE AWARD FILE
DOCUMENT CONTROL CENTER

American Geological Institute, 4220 King Street, Alexandria, VA 22302-1502
(703) 379-2480 Fax: (703) 379-7563

DISCLAIMER

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

DISCLAIMER

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

NATIONAL GEOSCIENCE DATA REPOSITORY SYSTEM (NGDRS)

Phase II Activities Report — January - March, 1997

DE-FG22-95BC14833

Task 1. Planning and Specification

NGDRS DESIGN PRINCIPLES

The NGDRS goal is to preserve and make geoscience and engineering data readily accessible to those who have an interest in using such data. To achieve this goal requires the careful management of the process for transferring data from data providers to the public domain and the creation of the information systems organization and infrastructure to allow data consumers to find and access the data. The NGDRS was developed around the following key principles:

- Comprise a program involving industry, academia, and government entities to preserve large volumes of geoscience data in danger of being discarded or destroyed.
- Reduce cost of total repository services through sharing information, repository facilities and economies of scale.
- Consist of a centrally managed metadata repository and a network of many decentralized data repositories.
- Accept only the data contributions that conform to agreed-upon industry standard data models.
- Make publicly accessible data already on the public domain. Make publicly accessible all private donated data that is to be transferred to the public domain.
- Maintain the privacy of all private data that is to remain proprietary by imposing required levels of entitlement criteria.
- Facilitate the move of all participating repositories to the POSC industry standards environment to eliminate unwarranted data formatting operations.

THE NGDRS ORGANIZATION STRUCTURE

The NGDRS organization structure will consist of a centrally controlled facility for indexing, cataloging, browsing metadata and a large number of data stores, public or private, distributed across the USA. The NGDRS activities will be organized around three entities. The first entity is a steering committee which will provide guidance on policy issues. The second is AGI, which will establish a clearinghouse that manages the relationship between the data contributors and the receiving data repositories. Data contributors and receiving data repositories together constitute the third entity. AGI will also be responsible for the creation, utilization and routine update of the metadata repository, which will contain the catalog of information describing all source repositories in the NGDRS.

The NGDRS Clearinghouse will perform an overall coordinating and data transfer function. The Clearinghouse responsibilities will include:

- Identifying targets for data acquisition and transfer, and receiving repositories.
- Prioritizing those targets based on business, logistics and other criteria.

- Allocating the resources for data transfer, transcription, and preservation.
- Coordinating data transfer from the contributor to the receiving repository.
- Selecting and supervising service providers to handle the movement of physical data and conversion services, e.g., tape transcription and hardcopy scanning and digitizing.

Clearinghouse operations will be overseen by a NGDRS Steering Committee and a Clearinghouse manager. The Steering Committee will be made up of representatives from industry, academia, government and AGI Management. Processes will be in place to ensure that users and industry donors have an input into the Clearinghouse business decisions. AGI will be responsible for day-to-day activities relating to the coordination of data contributors, service providers, (data transcription, logistics, transportation) and receiving repositories.

THE NGDRS MARKET AND STAKEHOLDERS

The users of a data repository run the gamut from private petroleum companies, service companies, data brokers, and consultants to public government agencies and universities. All these stakeholders require (for different purposes) easy access to geoscience and engineering data. Some of the stakeholders are data providers, some are data consumers and some are data service providers. In many cases, each stakeholder may take on all three roles as each stakeholder determines its overall needs for data access, data distribution and data management. To fully implement the NGDRS goal, it is necessary to define a financial model that recognizes that each stakeholder has its own motivation for participation in the NGDRS.

The data providers must have some incentive to participate in the NGDRS, for example:

- **Petroleum and mining companies** want to donate a large volume of geoscience and engineering data to the public domain in those areas where, at present, they have no commercial interest. Their principal motivation is cost reduction. Since those areas, however, may again become commercially viable prospects in the future, companies want to retain access rights to the data donated.
- **State and federal agencies** would like to have the data being offered for donation by the petroleum and mining companies to stimulate the local petroleum and mining activities in areas where they may collect royalties in the future, perform resource assessment and for other commercial and non-commercial activities.
- **Data brokers** view NGDRS as a new market channel for data sales and possibly for data distribution. They are interested in displaying and advertising, selling and possibly distributing their data using the Internet powerful networking and electronic commerce capabilities.
- **Universities** view NGDRS as a vehicle to stimulate the usage of geoscience data in existing repositories already under their responsibility. They also hope the combination of many data types from a variety of repositories will enhance the capabilities for support of multi-disciplinary research projects within the academic environment.

The market environment created by the concentration of available information to be provided by the metadata facility will likely offer new and expanding business opportunities for private sector entrepreneurs. A plethora of ideas will arise on how to improve the value of the data through reprocessing with state-of-the-art signal processing and re-interpretation using modern workstations. A secondary market in reselling the enhanced information will evolve. The results

created by these activities will lead to new exploration opportunities in the USA in addition to new job opportunities in the expanding service business.

Task 2. National Directory of Geoscience Data Centers

A component of the Phase II project was development of a directory of existing geoscience data repositories to document the current holdings of data types in the public and private sector. A draft copy of the Directory is enclosed with the report.

Four tasks were required for the development of the National Directory of Geoscience Data Repositories:

- Selective Site Visits;**
- Survey Questionnaire;**
- Follow-up Contacts;**
- Directory Development.**

Selective Site Visits: As identified by the AGI Repository Steering Committee, site visits were made to approximately 15 public and private data centers across the United States. These site visits provided a cross-section of data-center operations and services that were used to help develop the scope and content of the survey questionnaire.

Survey Questionnaire: In addition to the site visits, letters were sent to all state geologists, to approximately 60 geological societies, and to members of the membership committee of the American Association of Petroleum Geologists. These letters requested names and addresses for all data centers known to these groups and individuals. A pilot questionnaire was then drafted and sent to approximately 25 of the identified data centers to gain insight into the questionnaire content. Feedback from that pilot test was incorporated into the final questionnaire (copy attached).

The questionnaire was then sent to all identified data centers as well as to private-sector data centers selected from several oil and gas directories. There were approximately 300 questionnaires mailed. The respondents were asked to return the questionnaires either via mail or e-mail via the Internet.

Follow-up Contacts: Approximately four weeks after the initial mailing, follow-up reminder postcards were sent to non-respondents. Two weeks following that mailing, phone calls were made to approximately 50 non-responding data centers. The final number of data centers included in the Directory is 124, although several additional respondents disqualified themselves as inappropriate for the Directory.

Directory Development: The Directory is organized alphabetically by and within state. The questionnaire was the model for the basic presentation of the data, which includes the following: name of center, address, contact person, with telephone and fax numbers and e-mail address, size of data center staff, type of repository, data access fee, catalog and index type, data delivery media, quantity and type of holdings, data center services provided, geographic areas of

holdings, by type and media, and comments. Appendix I of the enclosed Directory (Appendix II is the questionnaire) is a cross-referenced matrix of broad data types (geophysical data, drill-hole data, field/pool information, maps and location information, and other) by geographic areas of holdings and media. This matrix provides a fast look-up table that allows users to determine which data centers have the specific data types to meet their needs.

Task 1. Pilot Projects

THE FINANCIAL (BUSINESS) MODELS FOR DATA TRANSFER

During the process of creating the design and specifications for the NGDRS in Phase II, four pilot projects were undertaken to study the cost of implementation of the NGDRS and permit scale-up factors to be determined. Three of the pilots involved transferring data (wells, seismic, and core and cuttings) from the private sector to the public domain and were used to evaluate the time and cost involved in transferring the data.

Seismic Data Transfer

The following business objectives were identified as part of this financial model:

- Preserve digital seismic data on digital tape
- Reduce the cost of managing seismic tape
- Enhance data access capabilities

The following basic assumptions were adopted during the development of the model:

1. Majority of the current tape media is 1/2 inch 9-track magnetic tapes, no robotic units are available for data access.
2. Number seismic tapes involved in the transfer is 800,000, the total for Company X.
3. Technical strategy adopted is to transcribe and preserve digital well and seismic data on DD2 high density tape media.
4. The metadata will need to be collected before the data is transcribed to DD2. The metadata includes, but is not limited to, observers logs, acquisition parameters, processing parameters, and navigational position data.
5. Three (3) years will be required to transfer tapes.
6. Planning time frame for the model is twenty (20) years.
7. No tax write-offs were assumed.
8. Future value of money was not included — constant dollars are assumed.
9. DOE funds are available for the data transfer activities.

The financial model highlighted the following points:

- About 50% of existing costs are related to storage — physical space and other items. An additional 40% of the existing costs are related to direct and indirect labor to manage the tapes.
- Existing data management cost per tape range from \$1.05 to \$1.34.
- About 60% of the transfer costs are related to tape transcription, an activity requiring skilled labor and specialized technologies.
- Transfer costs per tape range from \$12.87 to \$16.67.
- Assuming Company X will pay only for the running of the existing repository during the transition (3 years) and that there are no unforeseen future costs, the payoff time for the project is three years.

Core Data Transfer

The following business objectives were identified as part of this financial model:

- Preserve core and cuttings for future use
- Reduce the cost of managing core and reduce future liability

The following basic assumptions were adopted during the development of the model:

1. Number of cores managed by Company Y is 220,000 boxes. Each box contains, on average, six (6) linear feet of core.
2. Technical strategy adopted is to transfer core to a repository already in existence and with sufficient floor space to house the collection of core being donated.
3. No core will be disposed and there is no cost provision in the model for the disposal of cores.
4. No core will be photographed before or after transfer. Model has no provision for transferring hardcopy data associated with the core, such as photos or laboratory analysis.
5. All cores are properly identified. The core metadata will need to be collected before the data is transferred.
6. One (1) year will be required to transfer cores.
7. Planning time frame for the model is twenty (20) years.
8. No tax write-offs were assumed.
9. Future value of money was not included — constant dollars are assumed.
10. DOE funds are available for the data transfer activities.

The financial model highlighted the following points:

- About 44 percent of existing costs are related to storage facilities. An additional 50% of the existing costs are related to direct and indirect labor to manage the cores.
- Existing data management cost per box range from \$2.2 to \$2.8.
- About 40 percent of the transfer costs are related to transportation of cores and the remaining 60 percent are related to direct and indirect labor to manage the core transfer process.
- Transfer costs per box range from \$2.5 to \$3.6.
- Assuming Company Y will pay only for the running of the existing repository during the transition (1 year) and that there are no unforeseen future costs, the payoff time for the project is 0.6 years.

Data Browser

In the process of transferring the data during the Phase II pilot projects, complete data stores were created on a computer system. In addition to well, seismic and core and cuttings data; cultural, political, and lease boundary data stores were created. These data stores provided the framework to create the fourth pilot — data cataloging and indexing.

The fourth pilot involved designing a data catalog and indexing system for implementing the AGI metadata repository that would support all the data repositories participating in the NGDRS. Mobil Oil Corporation provided a software system that enabled the design and implementation of a prototype software browser (named GeoTrek) . The browser provides a means to test the concept of accessing distributed data repositories with different data models and data access methods. The data models tested were the from the Texas Bureau of Economic Geology, United States Geological Survey, Central Intelligence Agency, Mobil Oil Corporation, British

Petroleum, Elf, Total, and Petroleum Geoscience Services. The data access methods included specific access to flat files, Standard Query Language, and Data Access and Exchange layer specified by POSC.

The data browser is a map-based user interface that provides a single point of entry to find geoscience and engineering data. A data broker software component registers and keeps track of the various data stores opened by a user during a particular working session. A data access layer provides a means to access different data models and different database management systems. The data stores are collections of geoscience and engineering data in different formats be they physical, electronic or hardcopy data.

In all, twenty-two data stores were connected to the data browser and made accessible through a single world map-based user interface. Three of the twenty-two data stores were Epicentre-compliant data stores. After the first POSC data store was connected, the other two were automatically connected, thus saving a considerable amount of time, cost and effort. One of the goals of the NGDRS is to facilitate the move of all participating repositories to the POSC standards environment to eliminate unwarranted data formatting operations. The POSC format will be used, as much as possible for data that is being donated, for data that is being distributed from the data repositories and for automation and the synchronization of all the repositories with the AGI metadata repository. Using POSC standards will eliminate cost and provide a single point of entry for the data consumer.

OVERALL ARCHITECTURE OF THE REPOSITORY SYSTEM

In the Phase II of the NGDRS project one of the main objectives was to deliver the specifications and components for the overall repository and in particular for the NGDRS metadata repository. Because of the complexity of the systems involved, it was necessary to define the general system architectural principles and models to assure the repository will perform as expected for many years. These same architectural considerations were used to facilitate the development of a comprehensive business model to utilize, operate, and manage the production-level repository scheduled to be created during the Phase III of the project.

The general vision statement for the NGDRS states that "potential data contributors and data users agree that the NGDRS should be composed of integrated network of regional data centers, rather than a single, centralized facility. The system should build upon existing state and regional data centers as a starting point for developing a distributed and interconnected national repository network". To explain this vision in some detail, it is convenient to conceptualize NGDRS as follows:

The System and the Source Nodes: The NGDRS will be made up of a set of computing and data storage regional centers (which are designated as NGDRS Source Nodes or simply as Nodes) connected via one or more networking means such as, telephone private lines or Internet public services providers. Every Source Node is managed and is the responsibility of the Source Node Entity (or just Entity) — such as for example the Bureau of Economic Geology (BEG), or The University of Texas, or the U.S. Geological Survey. We will call the collection of all NGDRS Source Nodes that will be networked as the System. The Entity with the overall responsibility for managing the System will be called NGDRS Management.

Data Stores: Entities may provide one or more data services to users. Each data service will be handled via a Data Store resident at each Node. The collection of data stores managed by a particular entity is a Repository. Data stores may contain technical data (such as Geosciences or Engineering), reference data (such as general reference maps or tables) or metadata. Metadata is defined as "data about data".

The NGDRS Metadata Repository Node: The System will include a NGDRS Metadata Node to provide the front-end entry level services required by all users of the repository. This Metadata Node will serve as the master on-line catalog of information for the System — known as a supercatalog. It will tell users what (geoscience, engineering, other) information is available across the System, its ownership, location, technology support environment and many other data items describing the characteristics of the available information — the so-called metadata.

Metadata Synchronization: Each Source Node will always contain the latest metadata on the information it provides — the so-called subcatalogs. Procedures will be defined between each Source Node and the NGDRS Management for periodic and frequent updates (synchronization) of the metadata on the NGDRS Metadata Node. The NGDRS Metadata Node will contain therefore the aggregate of the metadata of all Source Nodes in the System. However, not all metadata will be synchronized simultaneously; rather a polling process will be established to update periodically the metadata database for the Nodes. The periodicity of the update will be agreed between the NGDRS Management and the Source Node Entity.

User Validation Subsystem: The NGDRS Metadata Node will need to validate the user as part of a business entity (company, government organization, university, services, etc.) allowed to browse and to order information from the repository. Although the NGDRS is conceived fundamentally as a system open to all, because of marketing, general accounting, and billing functions it will be necessary for individual users to be validated using the validation subsystem. A single, unique password will be adopted (when viable) for all NGDRS-related business activities and transactions to minimize user efforts.

Usage Tracking Subsystem: The NGDRS Metadata Node must retain information relative to the usage of the resources of the System. This can only be partially achieved, since in most situations information in the usage of the resources may not be available from the Source Node. This subsystem will accumulate the usage of the System resources for each user on a Working Session basis. Software packages supported by each Source Node Entity will track the usage of the resources at each Source Node.

Transfer Control Subsystem: In certain situations the Metadata Node will also connect the user (meaning transfer control) to the Source Node where the target desired information resides. Depending upon arrangements agreed between the NGDRS Management and the participating parties, information may be physically retrieved (transferred) from the Source Node to user computing environment or alternatively to the Metadata Node.