THE UNSTRUCTURED DATA SHARING SYSTEM FOR NATURAL RESOURCES AND ENVIRONMENT SCIENCE DATA OF THE CHINESE ACADEMY OF SCIENCE *

Dafang Zhuang[1], Wen Yuan[2]*, Jiyuan Liu[3], Dongsheng Qiu[4], Tao Ming[5]

Institute of Geographical Sciences and Natural Resources Research, CAS, Beijing, PRC, 100101
*Emails: [1]zhuangdf@lreis.ac.cn,[*2] yuanw@lreis.ac.cn, [3]liujy@igsnrr.ac.cn,[4] qiudsh@lreis.ac.cn, [5]mingt@lreis.ac.cn*

ABSTRACT

The data sharing system for resource and environment science databases of the Chinese Academy of Science (CAS) is of an open three-tiered architecture, which integrates the geographical databases of about 9 institutes of CAS by the mechanism of distributive unstructured data management, metadata integration, catalogue services, and security control. The data tiers consist of several distributive data servers that are located in each CAS institute and support such unstructured data formats as vector files, remote sensing images or other raster files, documents, multi-media files, tables, and other format files. For the spatial data files, format transformation service is provided. The middle tier involves a centralized metadata server, which stores metadata records of data on all data servers. The primary function of this tier is catalog service, supporting the creation, search, browsing, updating, and deletion of catalogs. The client tier involves an integrated client that provides the end-users interfaces to search, browse, and download data or create a catalog and upload data.

Keywords: Data sharing, Metadata, Catalog service, Unstructured data, Environment data, Natural resource data

1 INTRODUCTION

The development of spatial information technologies has promoted research methods in the fields of natural resources and environment. In developed countries, especially the United States, an infrastructure has provided researchers with the ability to find and acquire the desired data quickly by data-sharing among large natural resource and environment databases. Such an infrastructure is also in needed by Chinese researchers. In the Chinese Academic of Science, there are 14 institutes related to the fields of natural resources and environment, and each institute has built up its own large spatial databases, which are limited to one or more sub-fields and cover only a part of China. However, there are no bridges among those isolated databases, and thus researchers are hindered from acquiring and integrating the data distributed in those isolated databases. Recently a new organization, the Center for Resource and Environment Science Data, consisting of a main center and 9 sub-centers (see Figure 1), was set up in August, 2003. The role of the organization is to direct natural resource and environment

* This research is supported by the Important Knowledge Innovation Project of CAS: Construction of Exchange and Sharing Platform of Resource and Environment Science Data (NO. KZCX3-SW-357) and National Natural Science Foundation of China Project: A study on new data structure for raster based on global discrete grids(NO.40501057)

science data sharing among the 14 CAS institutions. This paper will address the infrastructure and technologies used to provide the technical support for this innovation.

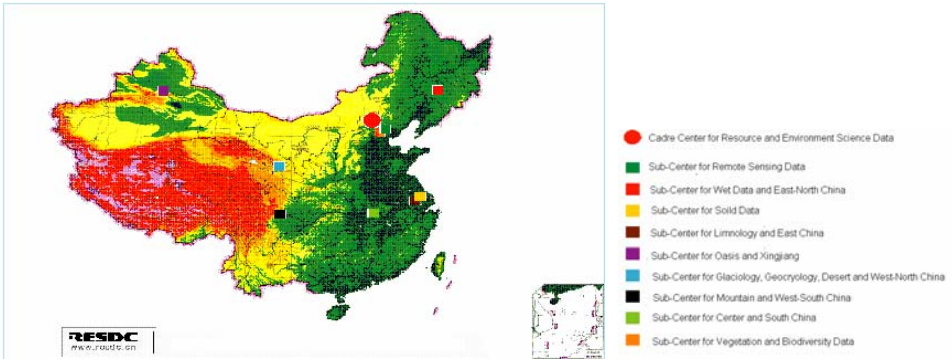

Figure 1. Distribution of Data Centers of Resource and Environment Science Data Sharing System

The main targets of the data-sharing system include: 1) a networking architecture bringing all databases of the institutes of CAS into a whole, 2) a distributed data-sharing architecture based on centralized metadata, and 3) powerful user interfaces for users to query and download or upload data and thus to archive and make all valuable datasets accessible to every staff member at CAS, to reduce the duplication of data creation and maintenance, and to realize the full value of the data by increasing the usage of generated data.

The system has a three-tiered, data-sharing architecture using a metadata-drive approach for various unstructured file formats, which consists of a centralized metadata server and several distributive data servers. In the middle tier, every dataset has a metadata registry stored in the metadata server, and a catalog service is provided to manage dataset searches, uploading, and downloading. The datasets are archived into a file in the distributive data server in the data tier, and the catalog service keeps the coherence of the dataset address with the metadata registry. In the client tier, a powerful client is provided for users to query, acquire, and upload their desired datasets.

2 THE DATA-SHARING SYSTEM FRAMEWORK

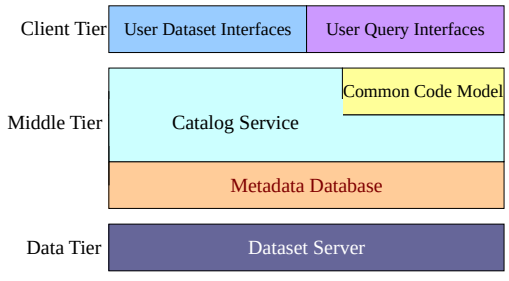

Figure 2. System Architecture

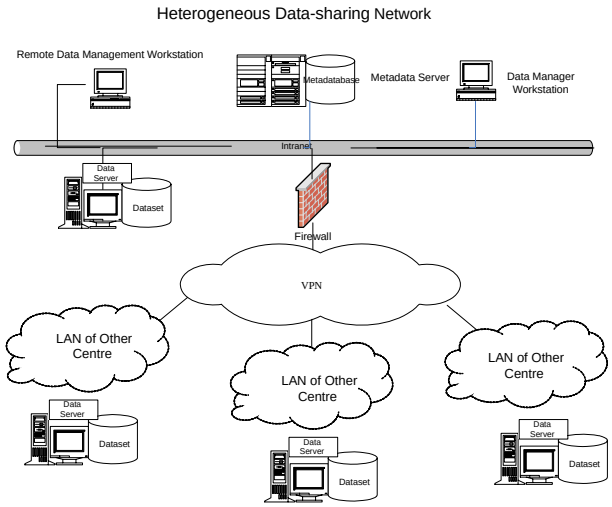

Figure 3. Networking Architecture

For the data-sharing system, there are several common issues, including data location, how to understand data, how to access data, data safety, and data readability. The popular solution is to use metadata (FGDC, 1997; SDI Cookbook, 2001; GIS, 1999), which is also the basis of the natural resources and environment data sharing system.

The data-sharing system has a three-tiered architecture, including the data tier, middle tier, and client tier (see Figure 2). In the data tier, various datasets are stored at distributive data servers for sharing. Datasets can be in any format, either structured or non-structured. New types of data can be added at any time. Data servers can be expanded whenever needed. The middle tier manages metadata for the datasets in the data tier. Datasets can be searched in one step through a unified server and accessed remotely if authorized. In the middle tier, every registered dataset gets a corresponding metadata registry stored in the metadata database, and this tier's main function is catalog service. Within the middle tier, the client tier provides a powerful interface for users to query and use all the data resources. Users may search and download the dataset to their local computers for processing.

Figure 3 shows the networking architecture of the data-sharing system. All LANs of sub-centers are connected directly by VPN and protected by a firewall. The data-sharing system is a network consisting of a centralized metadata server and 10 distributive data servers. The centralized metadata server is located in the Main Center, and data servers are in individual institutes.

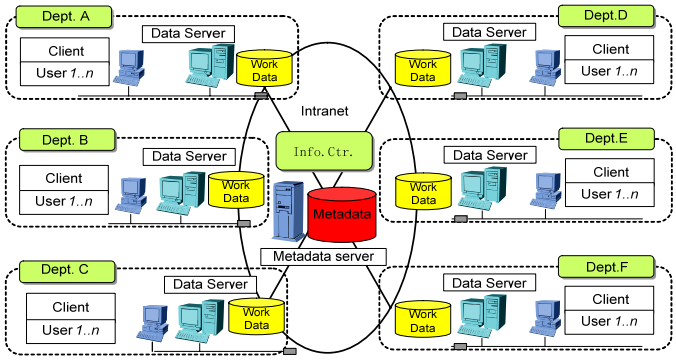

Figure 4. Data-sharing Among 14 CAS Institutes

The system enables distributive and unstructured data resources to be shared in an integrated way as follows (see Figure 4): 1) all datasets are associated with metadata registries, which are stored in the centralized metadata server, recording important descriptive information and enabling the understanding and interpreting of the data when it is exchanged among different users. 2) Valuable data is stored and secured at designated distributive data servers to avoid unintended deleting and unauthorized access. The data servers are geographically remote from each other. 3) User access of metadata servers is verified at a single access point by private keys. Dataset and information traveling between metadata servers and clients are encrypted. 4) Wherever the datasets are stored, they are searchable through a unique point and accessible from anywhere by each member of the organization or the public. Valuable data can be shared and utilized efficiently. Duplicated labor costs for data production and maintenance are avoided. Work efficiency is improved by saving data preparation time and making datasets available to researchers. 5) Data can be published and updated in real-time by the owner to increase the value of the data.

Figure 5 shows the data-sharing procedure process. There are three roles within the system. The first is the system administrator, who is responsible for managing, maintaining, and monitoring the system.

This person is the only one to manage user accounts and metadata or who can view definitions. The second is the data owner who can publish his/her own datasets by creating a new metadata registry and updating it to the data server. The last is the data user. The system provides data users with the interface to query catalogs, browse metadata registries, and access and download their desired datasets. The procedure of dataset search and acquirement is transparent for the data user and data owner.

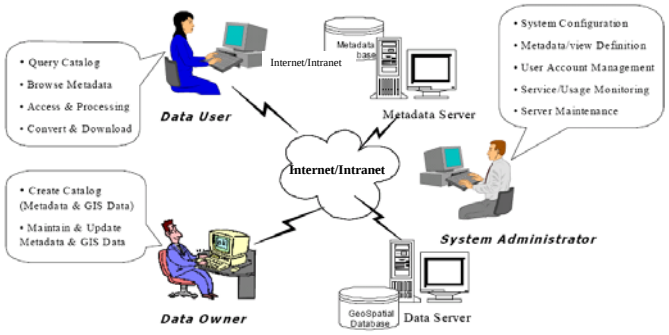

Figure 5. Data-Sharing Procedure Perspective

3 METADATA STANDARDS FOR RESOURCE AND ENVIRONMENT SCIENCE DATA

Metadata are data that describe the content, quality, condition, and other characteristics of datasets (FGDC, 1997; SDI, 2001). Metadata, the core of the system, are necessary for users to locate, understand, and process the shared data. The system bridges the spatial databases of the CAS institutes, each of which puts its heterogeneous data in different file formats with object-level metadata (OLM). There are multiple metadata standards defined to support documents, vector data files, grid or remote sensing images, multi-media files, tables, and other kind of files, based on related Chinese GB standards and FGDC metadata standards (FGDC, 1997).

For geospatial datasets, the relative metadata standard has the following descriptive attributes: identification, data quality, spatial data organization, spatial reference, entities and attributes, distribution, metadata reference, citation, time period, and contact. For other datasets, the relative metadata standards have several basic descriptive attributes, such as file name, file format, file generation time, and owner information.

In the system, the metadata are developed for the following purposes: 1) for a catalog system to search precisely for datasets of interest; 2) for users to browse the contents and relevant information of the dataset, in order to judge if the data satisfy their application purposes; and 3) for providing technical information for correct processing and use of the datasets.

Because there is not a fixed standard for all datasets, the system allows users to define multiple metadata models for different purposes. The metadata model is either a flat list or hierarchical: 1) For each meta-database, multiple views can be defined for searching or quick catalog listing. Each view can contain full or partial metadata items within the metadata model. This feature allows users to concentrate on certain metadata items for convenience in different situations. 2) Spatial location metadata can be defined by points or rectangles. Index maps can be registered to such spatial items showing the location on a map or by collecting coordinates from a map. Word (or location name) lists can be defined as strings or spatial items. Users can select a word from a word list for strings or get the

location of spatial items from a location name list.

4 CENTRALIZED METADATA SERVER AND DISTRIBUTIVE DATA SERVERS

In the system, datasets from multiple data servers are registered to the centralized metadata server, which is located in the Main Center. The metadata server stores only the metadata catalogs, and the corresponding datasets can be located in an arbitrary data server. The queries for datasets are performed first by the metadata server and then are redirected to the desired data server according to the dataset address in the metadata registry. In the system, the metadata server is built based on an Oracle database and hosts multiple meta-databases. There are 54 meta-databases hosted in the centralized metadata server (see Figure 6) and 9 data servers (see Figure 7). For datasets stored in a data server, there are 6 meta-databases.

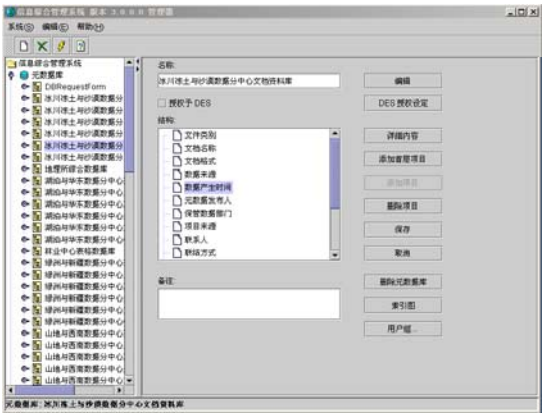

Figure 6. Meta-database Configuration

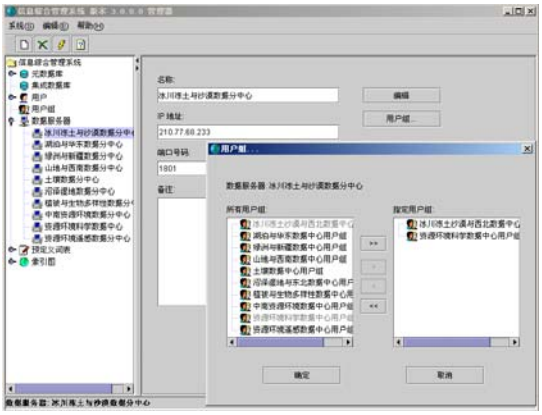

Figure 7. Data Server Configuration

5 DATA ACHIEVEMENT IN DISTRIBUTIVE DATA SERVERS

In the data-sharing system, the datasets are stored in isolated data servers. In each data server, there are not only geospatial datasets but also various other files, such as text documents, electronic spreadsheets, presentation slides, pictures, and videos, which are also important for resource and environment science research. These files are in binary formats without a standard structure and must be viewed with specific programs. In the system these unstructured files are directly controlled by the file system of the operating system.

A dataset with a metadata registry can consist of more than one file, and when creating a catalog, the files are compressed into a file with a unique name. The unique name is used as the dataset address in the metadata registry. When users browse or download a dataset, the system will uncompress it into its original state automatically before delivering to the client.

The dataset address and access code are updated and available in the corresponding metadata registry when creating or editing the metadata entry, so the user is directed to the correct dataset after searching and browsing. Once registered in the catalog, the dataset is archived for immediate access. The data server has auto-service. When a new metadata registration is created, the dataset is automatically uploaded to the desired data server. Archiving data to a data server makes the data always available and accessible as long as the system is in service.

6 DATA SECURITY

The system adopts traditional two-level security control, including user count and user group, which assigns or denies access to catalogs and datasets. When registering a dataset, the owner is able to assign authorized persons to modify or use (access or download) the data, instead of just issuing and certifying user accounts at login. The users can be assigned to one of three groups: Data Owners, Data Users, and System Administrators. Data owners create datasets and register the created dataset to any available data server. When registering the dataset, the dataset owner can make metadata and a preview to input into the catalog and assign security constraints. The owner is responsible for data maintenance and can come back at any time to update or modify the dataset, metadata, and catalog settings. During data creation or updating, the data owner can also be a data user. For example, the data owner can search and add maps registered by another data owner as a portion of their own dataset. In such cases, the original dataset uploaded by the owner stays unchanged and unmoved, but it will be linked to the newly created dataset. When the new dataset is downloaded or accessed, the linked datasets will be functionally equivalent to the non-linked datasets. This feature ensures there is only one valid copy of the dataset that is under direct maintenance by the original owner. Other users will use the updated data.

A data user is a consumer of the catalog and registered datasets, who queries the catalog and browses the metadata and previews. Once the desired dataset is found, the data user can download the data to a local computer in the required format. Alternatively, the data user can directly access the dataset on the server, without necessarily knowing where the data are located or what the original format is. A system administrator can use administration tools to configure system settings, define metadata models and optional views, create meta-databases, register data servers, index maps, prepare predefined word lists, validate and maintain user accounts, observe system and data usage, and maintain the efficiency of the entire system.

7 CATALOG SERVICES

The open GIS Consortium (SDI, 2001: GIS, 1999) defines a catalog as the set of service interfaces which support the organization, discovery, and access of geospatial datasets. Catalog services help users or application software to find information that exists anywhere in a distributed computing environment. A catalog can be thought of as a specialized database of information about geospatial datasets available to a group or community of users. Because the system links with vast numbers of datasets, it is only through catalog searches that users can efficiently find the data suitable to their purposes. Without a catalog, it is almost impossible to locate the correct data, even when the data exist.

The catalog service is the main function of the system, provided through its metadata server that allows remote searches on the metadata database and access to distributive datasets linked with the catalog. Registration, editing, and maintenance can be done remotely by dataset owners. Catalog services can be accessed through the clients. Catalog services also look after the coherence of metadata registries with the linked datasets. In the system, the catalog service is implemented by middle software as a module of the metadata server, and it provides the only interface to access and interpret the registry records in the Oracle metadata database.

The main functions of catalog services are:
1) Searching the catalog: Allows clients to submit queries to search the catalog of metadata servers and

get the metadata, preview, and location of the dataset of interest.

2) Creating the catalog: Allows clients to create a new catalog with metadata, data files, and preview on metadata servers. A catalog may have only metadata without an on-line dataset. This is useful for publishing just the information about data for searches, and owners will supply data by off-line processing. If the data files are chosen, the files will be uploaded to the data server automatically.

3) Updating the catalog: Allows clients to modify (delete or update) an existing catalog (metadata, dataset, preview, and security settings).

4) Managing a dataset: Allows clients to create, delete, or update a dataset (group of data files).

5) Accessing a dataset: Allows clients to download, upload, read, or write a dataset.

7.1 Metadata View By User-definition

In the system, for each meta-database, multiple views can be defined for searching or quickly finding a catalog listing. Each view can contain full or partial metadata items in the metadata model. This feature allows users to concentrate on certain metadata items for convenience in different situations.

7.2 Common Metadata View

Figure 8. Common Metadata Model Configuration

To support powerful search functions, the system introduces a common metadata model to overcome the different nicknames of a common term resulting from different principles, which turn maps containing common information and names into fields having special terms. The system supports multiple metadata standards based on different principles, and different names in different metadata standards may refer to a common term. The common metadata model bridges the different names for a common term and is the basis for simple and advanced searches. In the client, users can configure the map of a metadata standard field with a common term (see Figure 8).

8 USER QUERY INTERFACES

The system provides powerful user query interfaces, including simple search, advanced search, and expert search. Simple search is in the Google style, and users can define arbitrary combinations of keywords. Advanced search is a multi-matching of key metadata fields, while expert search needs the user to choose a concrete meta-database and define the conditions field by field. Simple search and advanced search are both accomplished through multiple meta-databases by the common metadata model. The system also provides the interfaces to create, edit, delete, and browse metadata catalogs. For dataset downloading and uploading, the function of literal translation is provided. The query procedure can be export into XML files for re-usage. Figure 9 shows the main user query interfaces.

8.1 Simple Search

It is a Google-style search tool. User can define several keywords or a compound of keywords. The function will automatically search all meta-databases according to a common metadata view

8.2 Advanced Search

Users can limit the search by document type, owner, time, and geographical location. The function will automatically search all meta-databases according to the common metadata view.

8.3 Expert Search

This is the original search interface. To use this query, the user must have advanced knowledge about the databases to be able to define which meta-database and metadata view to use.

Figure 9. Main User Interfaces

9 CONCLUSION

In this paper, we introduce the data-sharing system for resource and environment science data. The data sharing system has an open three-tiered architecture, which integrates the geographical databases of 9 institutes of CAS by the mechanism of distributive unstructured data management, metadata integration, catalogue services, and security control. The data tier consists of several distributive data servers, which are located in each CAS institute and supports such unstructured data formats as vector files, remote sensing image or other raster files, documents, multi-media files, tables, and other format files. For spatial data files, a format transformation service is provided. The middle tier (or data sharing tier) involves a centralized metadata server, which stores metadata records of data on all data servers. The primary function of this tier is a catalog service, which creates, searches, browses, updates, and deletes catalogs. The client tier involves an integrated client that provides the end users with interfaces to search, browse, and download data or create catalogs and upload data. The system provides a group of tools at the server and client tiers respectively. Server tools, including server administrator and log viewer, are provided for users to configure data servers, define metadata, manage user accounts, and monitor system and data usage. A client tool is provided for registration, maintenance, and application of catalogs and datasets. The main functions of the server administrator include metadata and view configuration, meta-database creation and updating, user accounts and group management, data server registration, index map registration, and word list creation. The functions of the log viewer include metadata server access logs, data server access logs, statistics on frequently accessed data (statistics on owner or time period), and statistics on data creation.

10 REFERENCES

Federal Geographic Data Committee (FGDC) (1997) Content Standards for Digital Geospatial Metadata. Retrieved from the WWW, September 26, 2007: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/index_html.

The OpenGIS™ Abstract Specification Topic 13 Catalog Services (1999) Retrieved from the WWW, September 26, 2007: http://www.opengis.org/techno/abstract/99-113.pdf.

The SDI Cookbook (2001) Retrieved from the WWW, September 26, 2007: http://www.gsdi.org.