A NEW APPROACH TO RESEARCH DATA ARCHIVING FOR WDS SUSTAINABLE DATA INTEGRATION IN CHINA

WANG Juanle[*], SUN Jiulin, Yang Yaping, Song Jia, and Yue Xiafang*

State Key Lab of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resource Research, Chinese Academy of Sciences, Datun Road, 100101 Beijing, China
Email: wangjl@igsnrr.ac.cn

ABSTRACT

World Data System (WDS) requires that WDS data centers have significant data holdings and sustainable data sources integration and sharing mechanism. Research data is one of the important science data resources, but difficult to be archived and shared. To develop long term data integration and sharing mechanism, a new approach to data archiving of research data derived from science research projects has been developed in China. In 2008, the host agency of World Data Center for Renewable Resources and Environment, authorized by the Ministry of Science and Technology of China, began to implement the first pilot experiment for research data archiving. The data archiving process of the approach includes four phases, i.e., data plan development, data archiving preparation, data submission, and data sharing and management. In order to make data archiving more smoothly, a data archiving environment was established. It includes a uniform core metadata standard, data archiving specifications, a smart metadata register tool, and a web-based data management and sharing platform. Through the last 3 years practice, research data from 49 projects has been collected by the sharing center. The datasets are about 2.26 TB in total size and have attracted over 100 users.

***Keywords**:* World Data System, Data Sharing, Research Data, Data Archiving, China

1 INTRODUCTION

Data is one of the most important bases for science research. In general, science data can be divided into two types. One type is operational data derived from operational observation systems, such as meteorology data, seismology data, oceanography data, and so on. These data can be easily collected and shared under the national or departmental data sharing policies. Another type of science data is research data, which is collected and/or produced from scientific research programs or projects, such as International Geosphere Biological Program (IGBP), some national or local research projects, and so on. It is difficult to collect and archive this type of data comparing with the operational data, because the data is collected and hosted by different research teams or scientists separately. With the developments of international, national or regional science research activities, more and more research data will be generated. These data can serve as very important and sustainable data sources for other researches in different and crossing disciplinary fields. How to archive the data and make them to be accessible and reusable by others are a challenge tasks for the science communities including the World Data System (WDS).

Based on developments of the former World Data Center (WDC) system in China, scientific data sharing has been making sound progresses in the past several years (Wang & Sun, 2007; Xu, 2003 and 2007). Under this background, Ministry of Science and Technology (MOST) of China decided to keep investigations on the data archiving for research projects funded by the government (Lin & Wang, 2008). Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, which is the host agency of WDC for Environment and Renewable Resources in Beijing, is authorized by MOST to design and implement the science research project data archiving experiment. Projects in the resources and environment field of National Key Basic Research Program are specified as the initial participant projects. Through one years' design and preparation, outputting data from these projects has been archived since 2008. This paper will introduce the new research data archiving approach and its progresses in the past 3 years.

2 RESEARCH DATA ARCHIVING WORK FLOW

First of all, research data archiving policy for research projects should be in place. After half a year preparation, "National Key Basic Research Program Data Archiving Management Specification on Resource and Environment Field" was published by MOST on 20 March, 2008. This specification not only defines the responsibilities and duties of data owners, managers and users, but also specifies the data archiving work flow.

There are 4 phases in the data archiving process (shown in Figure 1), i.e., data plan development phase, data archiving preparation phase, data submission phase, and data sharing and management phase. The 4 phases cover the whole project research cycle from the beginning when the project was launched to the end when the project would be overviewed 5 years later.

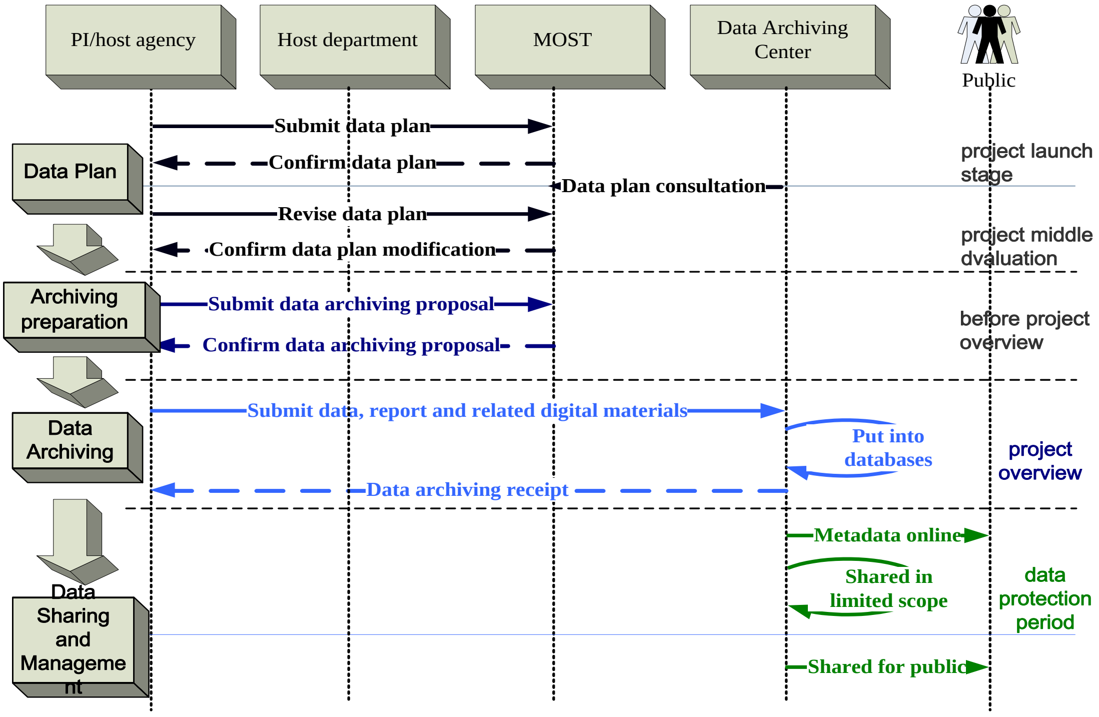

Figure 1. The research data archiving phases work flow

2.1 Data plan development phase

Data plan is the guideline of the whole data archiving procedure. Many agencies require their science research projects to manage science research data based on data plans, such as National Institute of Health (NIH, 2003) and National Science Foundation (NSF, 2011).

Data plan for science research data archiving should define the data outputs during the whole project period. Those related output datasets information should be described in the data plan, including data sets' name, main data content description, data types, data formats, data security classification, data protection time period, sharing styles, related software tools, funding sources, etc.

2.2 Data archiving preparation phase

Data archiving preparation phase starts once a data plan confirmed and towards the near end of project. In this phase, data archiving center will guide projects to prepare their datasets collected during the research, and provide related technology support for their datasets management. All the projects will collect and manage their data and metadata information in the process of research by using a software tool provided by the data archiving center. At the middle stage of this phase, projects may need to revise their data plans according to status and changes of research projects. All revised data plans should be confirmed by MOST.

2.3 Data archiving phase

Data archiving phase will be taken place before the final project overviewed. At this phase, all projects should submit their datasets according to their data plans. It includes 3 steps. (1) Data archiving center provides the related data archiving standards and specifications to each project, including data archiving profile template, metadata standard, data document specification, data quality review specification, data submission format specification, etc. (2) Projects submit their datasets under these standards and specifications to data archiving center using CD-ROM media. (3) If the dataset and its quality are confirmed by data archiving center, archiving receipts will be given to the related projects. Only those projects with archiving receipts have qualification for final project overview.

2.4 Data sharing and management phase

Data sharing and management phase will be conducted under a data management and sharing platform at the data archiving center. The platform provides customized functions for data providers, project managers and the public users respectively. For the data providers, they can submit and edit their research data, review the data services report online; for project managers, they can check the data archiving status online; for public users, they can explore data and access those datasets without sharing restriction online, and may apply those datasets with sharing restriction (e.g., data protect period) offline.

3 DATA ARCHIVING ENVIRONMENT CONSTRUCTION

Research data has inherent interdisciplinary features. In order to make these data integration together, a uniform data archiving environment is needed. It includes data archiving standards, data management specifications, related data archiving tools and sharing platform, etc.

3.1 Core Metadata Standard

A core metadata standard has been designed for research data archiving. Its metadata elements are listed in Table 1.

Table 1. The core Metadata elements for research data archiving

Metadata element	Metadata content definition
Dataset name	Dataset's specified name, which contains information about data thematic attribute, time period and region of data content
Project number	Specified project number allocated by MOST
Abstract	General and brief introduction of data content
Keyword	Significant or descriptive words for datasets
Dataset time	Time period of data content
Dataset format	Description of data storage format
Dataset quality	General evaluation information of dataset quality
Contact information	Contact information of producer(s) or the person(s) who is in charge for data publication or management
Usage restriction	Data copyright or privacy protection
Dataset web link	Website for data accessing

3.2 Data archiving specification

According to the requirements of research data archiving, a series of data management standards and specifications were designed and published by data archiving center. These include project data plan specification, data archiving report specification, data archiving document format specification, data archiving CD ROM specification, data quality review report specification.

3.3 Metadata collection and management tool

The metadata collection and management tool was designed and developed in Microsoft .Net environment. Its core functions include metadata records collection, review, appending, modification, delete and search. This tool is disseminated to all the projects and used for data preparation and archiving.

3.4 Data management and sharing Platform

The science research data management and sharing platform was developed in J2EE framework. All the data management and shared functions will be integrated in the platform, including the functions for data providers, data managers and data users mentioned above.

4 APPLICATION AND CONCLUSION

4.1 Application

By the end of Oct, 2011, 49 projects in resources and environment field of National Key Basic Research Program have submitted their research data to data archiving center. The size of the data accumulated is about 2.26TB, including more than 1000 datasets. According to the data storage and management types, these data can be divided into attribute data, text data, vector data, remotely sensed data, raster data, picture data and others. These data has their own individual disciplinary classification. A more flexible and integrated data category is under development by the data archiving center and will be published in the data sharing platform in the near future.

The number of registered users in the data sharing platform reached to 103. The website hits are 194704. About 1.5GB data has been downloaded. The top 5 datasets downloaded are listed as follow, “Tibetan plateau GDP change serials datasets (1970-2006)”, “Tibetan plateau ground temperature serials datasets (1951-2006)”, “Tibetan plateau livestock number change serials datasets (1970-2006)”, “Tibetan plateau population change serials datasets (1970-2006)”, and “China palmer drought index datasets”.

4.2 Conclusion

This science research data archiving experiment is the pilot initiative for the National Scientific Research Programs in China. It will have far-reaching influence to the scientific research data archiving and sharing projects which are funded by the government. Encouraged by the implementation of data archiving in resource and environment field, MOST will promote the research projects’ data archiving and sharing in broader fields of national funding projects in China. It not only enhances the developments for the data holdings of WDS data centers in China, but also contributes a robust and approved approach to the world community of science data integration and sharing.

5 ACKNOWLEDGEMENTS

This work is partially supported by State Key Lab of Resources and Environment Information System, Science & Technology Basic Research Program of China (Grant No. 2011FY110400) and Data Sharing Network of Earth System Science in China. Thanks the contribution of Mr. Shen Jianlei, Dr. Chen Wenjun from MOST of China. Thanks Mr. Wang Jiayi for Metadata tool development. Thanks Dr. Erjiang Fu for English language revision. We also thank the reviewers for their constructive suggestions and the editors for their careful check and revisions, which further improved this article.

6 REFERENCES

Wang Juanle, & Sun Jiulin (2007) Development of China WDC Systems for Data Sharing. *China Basic Science Research,* pp 36-40.

Guan-Hua Xu (2007) Open Access to Scientific Data: Promoting Science and Innovation. *Data Science Journal*, Volume 6, Open Data Issue, pp OD21-OD25.

Xu guanhua (2003) Advance for enhance China’s science and technology innovation capacity by data sharing. *China Basic Science Research*, pp 5-9.

Lin Hai, & Wang Juanle (2008) Data archiving work was launched in national basic research program in resource and environment field. *Advances in Earth Science* 23(8), 895-896.

National Science Foundation (2011) Chapter II - Proposal Preparation Instructions, Retrieved January 1, 2011 from the World Wide Web: Http://www.nsf.gov

National Institutes of Health (2003) NIH Data Sharing Policy and Implementation Guidance. Retrieved March 5, 2003 from the World Wide Web: Http://grants.nih.gov