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

Juanle WANG*[*], *Jiulin SUN, Yaping YANG, Jia SONG, and Xiafang YUE

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

The 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 it is difficult to be archived and shared. To develop a 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 the 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 approach's data archiving process includes four phases: data plan development, data archiving preparation, data submission, and data sharing and management. In order to make data archiving operate more smoothly, a data archiving environment was established. This includes a uniform core metadata standard, data archiving specifications, a smart metadata register tool, and a web-based data management and sharing platform. During the last 3 years, 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 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 the International Geosphere Biological Program (IGBP), national or local research projects, and so on. It is difficult to collect and archive this type of data in comparison with the operational data because it is collected and hosted by different research teams or scientists. With the development of international, national, and regional science research activities, more and more research data is being generated. These data can serve as very important and sustainable data sources for other research in different and cross disciplinary fields. How to archive the data and make them accessible and reusable by others are challenging tasks for the scientific community, including the World Data System (WDS).

Based on developments of the former World Data Center (WDC) system in China, scientific data sharing has made sound progress in the past several years (Wang & Sun, 2007; Xu, 2003, 2007). With this background, the Ministry of Science and Technology (MOST) of China decided to keep investigations of data archiving for research projects funded by the government (Lin & Wang, 2008). The 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. Its initial projects are in the resources and environment field of the National Key Basic Research Program.. Through one years' design and preparation, data from these projects have been archived since 2008. This paper will introduce the new research data archiving approach and its progress in the past 3 years.

2 RESEARCH DATA ARCHIVING WORK FLOW

First of all, a research data archiving policy for research projects should be in place. After half a year preparation, the "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): the data plan development phase, the data archiving preparation phase, the data submission phase, and the 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 reviewed 5 years later.

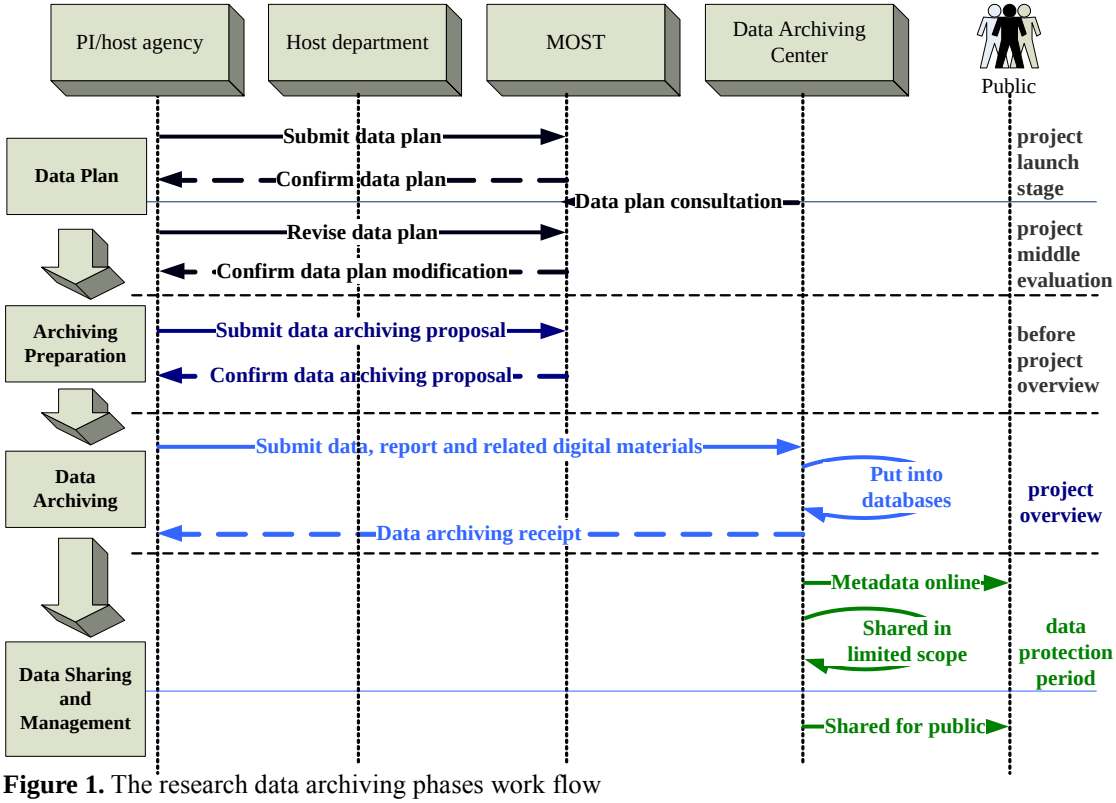

Figure 1. The research data archiving phases work flow

2.1 Data plan development phase

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

The data plan for science research data archiving should define the data output during the whole project period. The related output datasets information should be described in the data plan, including the datasets' 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

The data archiving preparation phase starts once a data plan is confirmed near the beginning of the project. In this phase, the data archiving center guides the project's preparation of the datasets collected during their research and provides related technology support for the datasets' management. All projects collect and manage their own data and metadata information during their research by using a software tool provided by the data archiving center. At the middle stage of this phase, the projects may need to revise their data plans according to the status and changes in the research projects. All revised data plans should be confirmed by MOST.

2.3 Data archiving phase

The data archiving phase takes place before the final project is reviewed. In this phase, all projects should submit their datasets according to their data plans. The phase includes 3 steps: (1) the data archiving center provides the related data archiving standards and specifications for each project, including the data archiving profile template, metadata standard, data document specification, data quality review specification, data submission format specification, etc.; (2) the projects submit their datasets under these standards and specifications to the data archiving center using CD-ROM media; (3) if the dataset and its quality are confirmed by the data archiving center, archiving receipts will be given to the related projects. Only those projects with archiving receipts are qualified for a final project review.

2.4 Data sharing and management phase

The data sharing and management phase is 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, respectively. Data providers can submit and edit their research data and review the data services report online; project managers can check their data archiving status online; public users can explore data and access those datasets without sharing restrictions online and may use those datasets with sharing restrictions (e.g., data protect period) offline.

3 DATA ARCHIVING ENVIRONMENT CONSTRUCTION

Research data has inherent interdisciplinary features. In order to integrate the data, a uniform data archiving environment is needed. This 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 attributes, 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 are designed and published by the data archiving center. These include project data plan specification, data archiving report specification, data archiving document format specification, data archiving CD ROM specification, and data quality review report specification.

3.3 Metadata collection and management tool

The metadata collection and management tool was designed and developed in a 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 a J2EE framework. All 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 October 2011, 49 projects in the resources and environment field of the National Key Basic Research Program had submitted their research data to the data archiving center. The size of the data accumulated is about 2.26TB, including more than 1000 datasets. The data storage and management types include attribute data, text data, vector data, remotely sensed data, raster data, picture data, and others. These data have their own individual disciplinary classifications. A more flexible and integrated data category by the data archiving center is under development and will be published in the data sharing platform in the near future.

The number of registered users in the data sharing platform has reached 103. The website has had 194704 hits. About 1.5GB data have been downloaded. The top 5 datasets downloaded are as follows: "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 a pilot initiative of the National Scientific Research Program in China. It will have far-reaching influence on the scientific research data archiving and sharing projects that are funded by the government. Encouraged by the implementation of data archiving in the resource and environment fields, MOST will promote research projects' data archiving and sharing in broader fields of nationally funded projects in China. It not only enhances the development of data holdings of the 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 the State Key Lab of Resources and Environment Information System, Science & Technology Basic Research Program of China (Grant No. 2011FY110400) and the Data Sharing Network of Earth System Science in China. Thanks for the contributions of Mr. Shen Jianlei and Dr. Chen Wenjun from MOST of China. Thanks to Mr. Wang Jiayi for Metadata tool development. Thanks to 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

Lin H. & Wang, J. (2008) Data archiving work was launched in national basic research program in resource and environment field. *Advances in Earth Science 23*(8), pp 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

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

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

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

(Article history: Available online 24 February 2013)