World Journal of Social Science Research
ISSN 2375-9747 (Print) ISSN 2332-5534 (Online)
Vol. 5, No. 2, 2018
www.scholink.org/ojs/index.php/wjssr

Original Paper

Study on California Digital Library's DataUp Project

Guoqiang Liu[1*]

[1] Library of Tianjin Polytechnic University, Tianjin, China

[*] Guoqiang Liu, Library of Tianjin Polytechnic University, Tianjin, China

Received: May 8, 2018 Accepted: May 17, 2018 Online Published: May 21, 2018

doi:10.22158/wjssr.v5n2p136 URL: http://dx.doi.org/10.22158/wjssr.v5n2p136

Abstract

Introduce the background and survey process of CDL's DataUp project, analyze scientists' research data management needs and the requirements for DataUp tool, display the project's outcome, reveal the enlightenments for domestic research data management.

Keywords

DataUp, Research data management, Metadata standards, Data identifier

The DataUp project of California Digital Library (CDL)through the research and development of data management tools, improve the efficiency of data management of scientific research personnel, and then promote the process of scientific research and technology development in universities.

1. DataUp Research Background

The DataUp project for research data management originated from CDL's understanding of the data management models for geoscientists, environmentalists and ecologists, CDL discovers that the three types of researchers will use the Spreadsheet function in Excel during scientific data management. In order to make this dataset creation and management practice more efficient and convenient, CDL has initiated the DataUp project for the development of scientific data management tools in conjunction with the Gordon and Betty Moore Foundation and Microsoft Research.The project is divided into two phases: the first phase is funded by the Gordon and Betty Moore Foundation and Microsoft Research, and the second phase is funded by Microsoft Research and NSF in the form of the Data ONE Data Network Supplementary project. In other words, Dataup is an integral part of the Data ONE project toolbox.

2 DataUp Pre-Development Requirements Research

2.1 DataUp Project Research Design

Before designing the DataUp data tools, in order to get a better understanding of the research data management model of the earth, the environment, and ecologists, and capture their management requirements, the project team conducted an investigation visit to 133 scientific researchers from August to December 2011. In addition, the project team has extensively collected a large number of professional recommendations from data management agencies such as academic pavilions and data centers, etc. In summary, the DataUp project research design has the following characteristics: (1) The multi-channel collection of information. The research team used project-specific sites, Data Pub blogs, Twitter, interviews, conferences, webinar and professional seminars to collect information. (2) The professional relevance of the questionnaire design. In order to understand the Excel data management mode of the researchers in different fields, and to develop highly practical tools, the survey team designed a variety of questionnaires applied to different subject fields. (3) Multiple selection of survey objects. 133 respondents from 84 different institutions represented multi-level researchers including students, junior, intermediate and senior researchers. The project team eventually collected 112 valid questionnaires and 30 interview records.

2.2 DataUp Project Research Results

2.2.1 Three Kinds of Researcher Data Management Level

The survey results show that the data management capabilities of the three types of researchers in the Earth, the environment, and the ecology are generally flawed. They are: (1) Lack of practical experience in scientific data management. (2) Little is known about data center and metadata standards. (3) Not fully aware of the value that data management and data sharing can bring.

2.2.2 Three Kinds of Researcher Data Management Mode

(1) In operating system selection, the Windows operating system occupies an overwhelming majority, with the 74% used by the interviewed users, and other 23% and 2% of the users choose to use Mac and Linux operating system. (2) In the use frequency survey of Excel, 80% of users said that they use it every day, and another 8% and 12% of users said that they often and infrequently use it. (3) In the survey on the use of Excel features, 97% of the interviewed users stated that they often use the title line to create functions, 83% of users often use inline formula functions, and 74% often use cell shading as temporary metadata form. In addition, 50%, 41%, and 32% of users often use Excel's cells, pivot tables, and note features. (4) A survey of the usage of the Excel assistant software shows that, in addition to the use of the Excel data function, the interviewed researchers will also use software such as Microsoft Access, MATLAB, Sigma Plot, GIS, and SAS as a supplement to the Excel data management function in different proportions .

3. DataUp Function Positioning Based on User Demand Analysis

3.1 Scientific Researcher Excel Data Management Requirements Analysis

Many researchers spontaneously use Excel data management and storage, some results have been achieved, and some drawbacks are increasingly highlighted, Excel is not a specialized tool for scientific data management, and it does not provide an overview tool in the critical computing, management, and storage areas of research business. What's more, the use of simple spreadsheets can also lead to errors in processing results based on scientific data. Based on the analysis of the EuSpRIG (European spreadsheet Risk Interest Group), Excel data management specifically has the following deficiencies: (1) The irrationality of the data table structure. (2) Metadata missing or metadata standard is not unified. (3) The presence of embedded numbers, charts, and annotations makes spreadsheets incompatible with other non-Excel systems. (4) Lack of data procedures for calculating, counting, and using formulas .

3.2 DataUp Data Service Function Positioning

Combining the flaws in Excel data management and the in-depth analysis of the needs of researchers during the previous investigation, the DataUp project team developed the detailed functional development requirements for the management tools shown in Table 1. (1) Check the data file format, users can generate and download custom reports for specific files, detailing the factors in the data set that may affect archiving or generating CSV form files. (2) Use the DataUp tool to generate a recognizable metadata file, embedded into the original data file, displayed in the new options menu in the spreadsheet, and can be saved independently. (3) The data citation file under the standard format is generated and becomes an integral part of the metadata. The citation file supports the download under the standard format. (4) Use the DataUp tool to perform the selected database-wide authentication. The authentication must be performed under the user's permission to use the database. The authenticated data file can also be stored in the database through DataUp. (5) Users can use DataUp to give them a certain identification symbol to store files, so as to facilitate long-term preservation and retrieval of documents. (6) Before the data file is officially saved to a specific database, DataUp will check whether the file to be saved has passed the three-step process of uniform format, metadata, and reference file. After that, it will generate the technical metadata required for the development of the database. (7) File submission and save to specified database. (8) Ensure that the stored data file is compatible with the Excel system where the DataUp plug-in user is not installed. In other words, whether or not there is the plug-in, the user can open and access the file through Excel.

Table 1. DataUp Data Management Tools Function Detail

Number	Functional Features	Function Description
1	Uniform data file format	CSV compatibility check and create CSV data files
2	Metadata	Generate, embed, and save metadata files
3	Referencing data files	Building data citations in standard formats to become metadata components
4	Database authentication	Authentication and storage of data files
5	Data file identifiers	Identifiers assigned to data files for saving and retrieving
6	Check before the formal preservation	Check whether the save file passes the above processing steps to generate the technical metadata required for the specified library
7	Data file submission and save	File commits and save to the specified database
8	Ensure compatibility of storage files	Users can open and access stored data files via Excel with or without plug-ins

4. DataUp Research and Development Results

4.1 DataUp Tool Publishing Form

The project team faces two choices in deciding the format of the DataUp software: the download and installation of the Excel add-in program, and the use of the Web application program. Although the former is more convenient and faster, it can only be downloaded and run in the Windows environment. There are problems such as software compatibility and downloading updates in the future. The latter method has drawbacks in implementing Excel functions. The project team consulted more than 200 researchers on the recommendations of social networks, questionnaires, and other channels. 95% of them are willing to download add-ins, but 83% of them are assuming that the programs can also be run in the Mac environment. 72% of people mentioned various obstacles to the download and installation of add-ins. Based on the survey results, the project team believes that both of the two distribution methods have their own demand user groups, and determine the two-pronged DataUp tool release model.

4.2 DataUp Tool Operation Process

The operation process of the DataUp scientific data management tool strictly follows the established objectives of the project, i.e., best practice detection-establish standardized metadata-determine the dataset unique identifier-upload metadata to the repository. The specific implementation of each step will be described one by one.

4.2.1 Best Practices for Data Format Detection

The purpose of the best practice test is to ensure that the data format is well-formed and consistent with the best management practices. The key to detection is to identify hidden issues that are not conducive to data storage and management. Some normal display formats such as annotations, embedded charts, graph cells, etc., cannot be in non-Excel programs. Based on extensive research, the project team summarized 11 types of hidden dangers and corresponding suggestions for modification, as shown in Table 2.

Table 2. DataUp Data Format Detection and Modification

Project	Format Problems	Modification Suggestions
1	Embedded table image cannot be displayed in non-Excel programs	Move to a separate form of a file or create a separate file to store
2	Embedded notes cannot be displayed in non-Excel programs	Enter the content of the note into the cell
3	Content after comma cannot be displayed in non-Excel programs	Move comma-separated content to other cells
4	Special characters that may not display correctly in other programs	Try to use only letter and number
5	Shadow or color cells, formatting files are lost in export	Use descriptive text, letter, and number encoding schemes in new cells
6	Columns containing mixed contents of numbers, text, etc. are not easily recognized by other programs	Ensure that the column contains only a single form of content
7	Blank columns or blank rows, typically used to separate different graphs	move the chart to a different Excelsheet
8	Merging cells may result in loss of information in file conversions	Ungroup cells, adding annotations
9	Blank cells may cause read problems in other programs	Explain the blank cells using a specific coding scheme
10	The header line does not exist or is too many, there is a problem with the use of the document	Create a unique header row that reflects the primary content of the document
11	There are multiple Excel sheets that cannot be converted to a single CSV file	Convert an Excel single table to a CSV file

4.2.2 Standardize Metadata Creation

The DataUp project regulates the metadata standards from the following two aspects: (1) Data standards at the file level, including specifications for file names, e-mail addresses, organizations, and data set titles; (2) Metadata attributes Specification, including variable information, units of measure, and column data descriptions in the data set. The DataUp project selectively used the EML (Ecological Metadata Language) standard that is more common in the academic world when defining the metadata standard. This choice is based on two reasons. One of the standards is widely used in the DataUp target customer group. The EML is a metadata standard that combines the characteristics of flexibility and extensibility. The project team can modify the optimal metadata normalization mode according to the project needs.

4.2.3 Data Set Identifier Generation

In order to expand data sharing and quotation, the DataUp project uses CDL's ARK (Archival Resource Key) uniform resource locator to identify data sets. The ARKs have many advantages such as simplicity, versatility, transparency, and identifiability. The identifier will be saved as part of the metadata.

4.2.4 Metadata Upload and Storage

Once the metadata is created, users can directly connect to the selected database through the DataUp tool to upload and store metadata files. Currently DataUp has built a project counterpart database ONE Share, which is a dedicated public data that anyone can use to store table data.

4.3 DataUp's Physical Connection System

As a network data management tool, DataUp inevitably generates links with various coding systems and storage systems, and forms a network data management system, as shown in Figure 1. DataUp's data encoding library is based on the NET application framework written using Visual Basic, while the online application version is provided through the Windows Azure cloud platform. Both the add-in terminal and the network service requester establish a link with one or more databases through a unified network transit service station. The add-in program runs directly through the Windows environment of the user's computer. The network service application and the transfer server use a unified OData service communication protocol. At the same time, all ports use the EML metadata standard, and the management system on Azure conducts network transit service management.

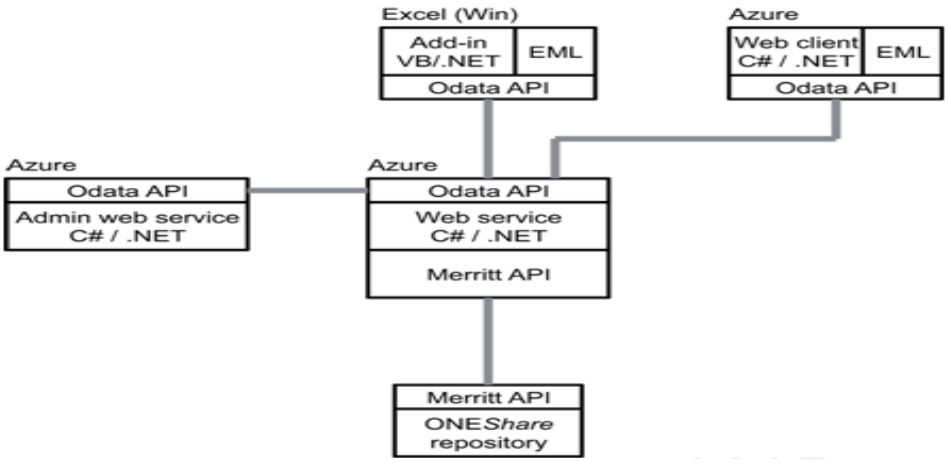

Figure 1. Dataup Data Management Tools Feature Detail

5. Revelation of DataUp project of Scientific Data Management Tool

5.1 In-Depth Rsearch and Development of Scientific Data Management Requirements and Practices

The DataUp project team of CDL starting from the actual demand model and management model of scientific data users, insists on extensive research and in-depth analysis methods for each details, in an effort to make the tool design maximally satisfy the user's demand. The specific performance is as follows: (1) In order to understand the data management level and management model of the target scientific researchers, the project team conducted investigations through Microblog, Twitter, questionnaires, etc., and deeply understood the data management status of the scientific researchers. (2) When determining how to provide DataUp tools, the project team adopted a two-pronged approach to service provision on the basis of extensive recommendations. (3) When summarizing the 11 types of hidden problems and corresponding modification suggestions, the project team passed interviews, questionnaires, etc. The method has collected a large number of practical experiences of scientific research scholars, database administrators, and academic curators who have rich experience in data management, and has listed the potential data management hidden problems and modification opinions of the form files to the greatest extent.

5.2 Based on Existing Tools to Achieve the Further Development of Scientific Data Management

CDL's DataUp tool is actually an extension of Excel data management tools. The development of practical tools based on Excel has the following advantages: (1) Facilitate the convergence and transition of old and new data management tools. (2) With the extensive mass base used by Excel in the past, it is easy to promote and popularize DataUp. (3) In the process of DataUp, the practical experience of scientific data users can be extensively learned, and first-hand data of scientific data management can be provided, thereby improving the hit rate and efficiency of improvement.

5.3 Solve Problems in Data Management from the Microscopic Point of View

From the 11 types of hidden dangers and suggestions for the corresponding amendments, it can be seen that the project team holds a rigorous work attitude and treats every microscopic detail in the design of data management tools. In the data management process, carefully consider the impact of the design of each minor link on the final result, make bold assumptions, and careful verification, through repeated practice and testing, not missed any minor potential factors that does not meet the requirements of the best practice model, be improved, and achieve refined management of scientific data.

5.4 Unified Management, Expanding the Sharing of Research Data and Improving the Efficiency

CDL's DataUp scientific data management tool project has established scientific data in the academic world through the establishment of uniform formats, metadata standards, and data identifiers for research data. The existence of a uniform format enables scientific data files to be identified and accessed by different programs. The formulation of the unified metadata standard can universalize the description of digital resources, display the essential nature, detailed information and characteristics of resources, and promote the sharing and utilization of resources. The generation of unique identifiers gives unique identities to data resources and facilitates the long-term preservation, retrieval, and use of data. Domestic scientific data management also needs to establish unified data standards, and languages, smooth data flow, expand data sharing, and provide data usage efficiency.

References

California Digital Library Resource. (2014). Retrieved from http://dataup.cdlib.org/resources.html

CDL ARK. (2014). Retrieved from https://confluence.ucop.edu/display/Curation/ARK

Dataup: A tool to help researchers describe and share tabular data. (2014). Retrieved from http://f1000research.com/articles/3-6/v1

Original Dataup Requirments. (2014). Retrieved from http://dataup.cdlib.org/resources.html

Spreadsheet Mistakes-News Stories. (2014). Retrieved from http://eusprig.org/horror-stories.html