UPData

A Data Curation Experiment at U.Porto using DSpace

João Rocha da Silva
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto PORTUGAL
joaorosilva@gmail.com

Cristina Ribeiro
DEI — Faculdade de Engenharia da Universidade do Porto / INESC Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto PORTUGAL
mcr@fe.up.pt

João Correia Lopes
DEI — Faculdade de Engenharia da Universidade do Porto / INESC Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto PORTUGAL
jlopes@fe.up.pt

ABSTRACT

UPData is a scientific data curation experiment currently under development at University of Porto which aims to determine the main digital preservation needs of several research groups at the university. In the course of the experiment, eight datasets have been collected from diverse scientific domains. After conducting several interviews with researchers working at U.Porto, we have concluded that from their point of view, flexible data access is the most valued capability when analysing a preservation solution and that offering such access it is the best way to involve them in the preservation workflow. We propose an extension to the DSpace repository platform to complement it with data curation capabilities. In the proposed solution, the system ingests Excel spreadsheets containing scientific data and translates them into XML documents which can then be queried via automatically generated XQuery statements. Researchers use a search webpage designed for displaying deposited data and applying various filters to it, retrieving the parts they need without having to scan each file. The collected datasets will be used as test cases for data deposit, and also to evaluate the effort required by the curation procedure.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: On-line Information ServicesScientific Data Preservation

General Terms

Digital Preservation, Scientific Data Curation, Repositories

Keywords

Scientific Data, Preservation, Repository, DSpace Extensions, Digital Curation

1. INTRODUCTION

Nowadays, large institutions all over the world are realising the usefulness and potential of digital preservation practices when applied to scientific data. Projects such as the Data Asset Framework [4], the Edinburgh DataShare [10] or the DANS Data Archive [8] are good examples of such efforts towards better preservation of digital data assets.

It is in this context that a scientific data curation experiment named *UPData* [7] is currently being developed at the University of Porto (U.Porto). This experiment involves the university central services and a research group from the Engineering School, and has the following goals:

1. Gathering a series of heterogeneous datasets from several research domains;
2. Determining the needs of several researchers working at U.Porto and writing a use case report to document those needs;
3. Developing an extension to the *DSpace* platform [2], complementing it with scientific data management tools;
4. Depositing the gathered datasets in this extended platform and seeking feedback from previously interviewed researchers.

Building on the experiences from the Data Asset Framework, the first step of the work was to analyse the current data management reality at U.Porto. This analysis helped determine the current data preservation concerns within 13 different research groups, belonging to 7 schools within the University. The research domains are heterogenous, including Engineering, Psychology, Economics and Education Sciences.

The dataset gathering procedures included a series of interviews with the research data creators. This step was essential to ensure the correct interpretation of the supplied datasets and provided valuable insight on the potential role of a scientific data repository in ensuring the proper preservation of this kind of data. Possible improvements in backup, annotation and sharing were enumerated and prioritised. Those which were most frequently pointed out by researchers were selected as the use cases for the proposed repository extension.

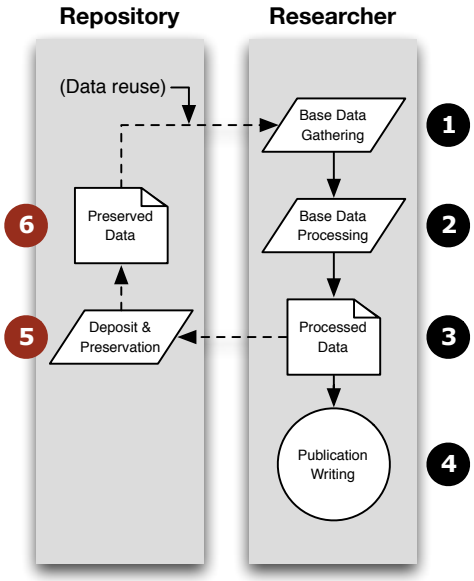

Figure 1: The research workflow and the additional data deposit steps

This solution is not intended to support the whole research workflow—it is intended to complement the publication of scientific discoveries with a data deposit and preservation step. Researchers publish their results after they complete the data gathering and analysis; however, the base data which supported the publication's findings could be better preserved, so we propose the inclusion of an additional data deposit step as shown in Figure 1. Step 1 includes the creation or gathering of base data to be processed according to the goals of the research group (Step 2). After the base data is processed (Artefact 3), conclusions are drawn and published (Step 4). We propose the inclusion of an additional data deposit step (Step 5), in which research data is gathered, annotated and translated into a preservable format (Artefact 6). Other researchers may then reuse the preserved data as base data for secondary research.

This is in line with recent policies adopted by some main scientific publications which are starting to require the inclusion of base data along with submitted articles in order to enhance the replicability of published results and provide that data to researchers in the same domain.

2. A SCIENTIFIC DATA REPOSITORY

A scientific data repository must preserve the data on a bit level and also ensure that the data is accessible and interpretable for future use. In fact, we have concluded that better accessibility is often regarded as the most interesting trait of a data repository. Researchers regard simple data backup as something they can perform on their own at a very low cost, and state that interesting data access features are important to encourage the self-deposit of data.

2.1 Use Cases

The focus on accessibility and reuse was present in many of the interviews conducted during the *UPData* experiment. Many researchers pointed out that useful repository tools should incorporate the following capabilities:

1. Easily sharing annotated datasets with their peers, reducing the need for individual follow-up contacts with researchers interested in the data;
2. Finding datasets by their dimensions, regardless of their production domain. Such dimensions are mensurable quantities or characteristics such as age, length, substances, latitude or height;
3. Exploring and querying deposited datasets through domain dimensions;
4. Retrieving query results, which involves the partial retrieval of datasets.

The data representation formats used by researchers are, in most cases, not suitable for direct data retrieval and querying. As a consequence, traditional approaches such as saving whole files and associated metadata are ill-suited for this purpose. Finding a way to reduce the granularity of data beyond the file level is a pre-requisite for building automated data manipulation capabilities.

2.2 A curation step in the data deposit workflow

There are several open-source repository solutions currently available. For this experiment, DSpace was the selected platform because U.Porto already has two operational public repositories built using this solution—the Open Repository and the Thematic Repository [11]. Contributing to DSpace can also help raise awareness on the topic of scientific data curation since the platform already has a large user base, with more than 1000 running instances [3] mainly at educational institutions. It is also open-source software, meaning that contributions can be submitted to the developer community and integrate future releases.

As part of the UPData experiment, we are designing and implementing an extension for the DSpace repository platform. This extension aims to provide users with the most requested data preservation features—easy data sharing, better searching and sub-dataset querying.

After analysing datasets gathered during the course of the experiment, we have concluded that researchers use many different formats for storing their data, which makes it difficult to develop tools to automate its processing. We have also determined that the main cause of data loss is the common lack of annotation and the use of proprietary file formats. The analysed data has, in general, quite simple models and multidimensional or hierarchical data are not very common. Most scientific data can therefore be organised into tables because the most prevalent types of files are spreadsheets, text files or other formats which can be converted into such formats by the original programs used to create the data. For these reasons, creating a better way

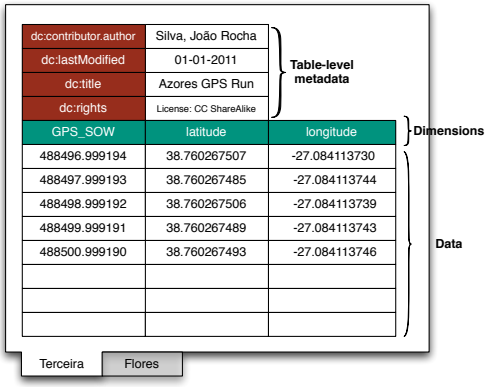

Figure 2: Example Excel spreadsheet layout

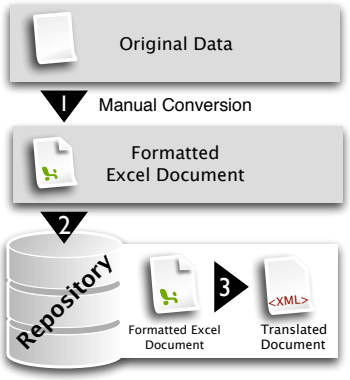

Figure 3: Data ingestion process

to manage tables in a repository platform was considered a good starting point towards better preservation of scientific data.

To make it possible for the repository to extract the relevant information from a dataset, we are designing a system which ingests specially formatted Excel spreadsheets to facilitate the interaction between the repository and the end users (either data curators or researchers). We decided to adopt this format because Excel is a common format among researchers at U.Porto and also because the implementation of a dedicated web-based deposit interface would mean heavier implementation efforts. A sample layout for a data deposit spreadsheet is shown in Figure 2. The spreadsheet is to be filled in manually by the repository curators, starting from the data which must be previously self-deposited by the original creator in its original format and layout. Since the data annotation process requires specific domain knowledge, it must be conducted in strict cooperation with the data creator. If an original data file contains several conceptual tables, each must be placed in a separate sheet of the Excel document—in this example, these are labeled "Terceira" and "Flores".

The dataset deposit and annotation workflow is depicted in Figure 3 and includes the following steps:

1. Gathering of the research data in a table-oriented format and filling in the dataset submission spreadsheet by specifying the appropriate header columns; Filling in table-level metadata, and pasting the data values—this process must be carried out by a data curator in strict cooperation with the dataset creator;

2. Submitting the spreadsheet to the repository system via the dataset ingestion page;

3. The system analyses the uploaded spreadsheet, processing each sheet for table-level metadata and column headers. Then, it matches the metadata fields with those parametrized in the repository and converts the data into an XML Document which is saved in the core DSpace database.

Since it not viable to develop automated conversion tools for all types of dataset representations, Step 1 is necessarily a manual process.

2.3 DSpace Data Explorer extension

DSpace includes a workflow engine designed to support *item* self-deposit by researchers. This workflows supports the necessary steps for the upload of dataset files and also the inclusion of relevant metadata. Such annotation can be carried out through qualified Dublin Core elements as well as other elements from additional metadata namespaces which can be parametrized in DSpace.

In DSpace terms, `Item`s are the smallest annotatable elements. These include a series of `Bitstream`s—the files contained in the `Item`. Newly submitted `Item`s must be assigned to a `Collection`. Finally, for each `Collection` there must be a group of system users, or `ePerson`s which are responsible for validating submitted `Item`s before they are published in the repository [1]—a dataset `Curator` must be a member of this group.

The deposit and indexing of datasets pose several challenges to the DSpace platform. Since dataset tables can have many different structures depending on their domain subject, a conventional relational model for such a heterogeneous reality might probably resemble the Associative Model of Data [6], with clear performance issues. XML Documents, on the other hand, have the required flexibility to represent all these different table formats and can also be queried through the XQuery language.

The high-level architecture of the DSpace extension is depicted in Figure 4, and is made up of the following modules:

1. The ingestion page can be accessed through the item viewing page in DSpace. Collection curators can upload a single formatted spreadsheet representing the data content of each of the files that make up the deposited item.

2. The XML Manager module takes care of all the operations on the XML-represented data. These include the

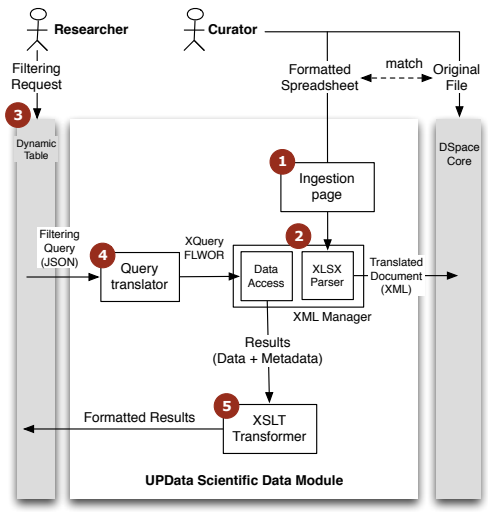

Figure 4: The UPData module architecture

translation of specially formatted spreadsheets—using the *Apache POI library* [9] for manipulating the Microsoft OOXML[1] format—and running XQuery FLWOR[2] statements over deposited data to select parts of the dataset.

3. When the user interacts with the dynamic table, a filtering statement (in the form of a JSON[3] request) is sent to the server. These filtering statements contain the column to filter by, the operator to be applied and the argument value, and can be combined using `OR`/`AND` operators to make up more complex queries. The server must then implement the required business logic to filter the data. In this case, the server side querying logic generates XQuery statements to be executed over the XML data stored in the repository.

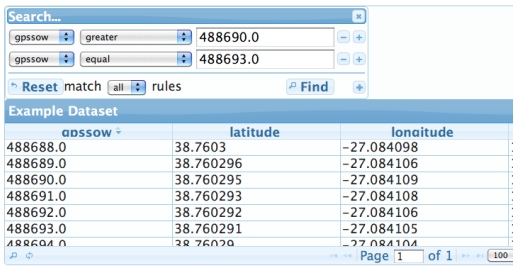

Figure 5: Web interface for a dataset table

4. The Dynamic Table component presents an XML document to the user in the form of an interactive table which is generated by the *jqGrid* library [5]. It supports basic data manipulation functionality, such as ordering data rows by specific columns in the dataset and also more complex column filtering features (numeric and string-based operators). An example table generated by the developed DSpace extension using this library is shown in Figure 5.

5. The XSLT transformer module was created to provide flexible means for presenting the data stored in the repository as well as the results of data selection. At the present time, it is used to transform the preserved data (which is stored in a rich format, complete with the relevant metadata) into a generic XML format which the *jqGrid* Javascript library can understand, so that it can create the dynamic tables shown to the repository users (see Figure 5). In the future, other transformation scenarios can be added, such as data exporting in XML-based formats or metadata-only exporting features.

[1]Office Open XML
[2]For, Let, Where, Order By, Return
[3]JavaScript Object Notation

3. ONGOING WORK

One of the planned objectives for this experiment is to obtain a reasonable estimate of the effort involved in the curation of an individual dataset. These estimates may prove to be valuable insight when considering the implementation of such practices in academic and research institutions and are to be determined in the course of this work.

4. REFERENCES

[1] P. Dietz. *DSpace 1.7.1 - System Documentation*. Duraspace, 2011.
[2] Duraspace. About DSpace. http://www.dspace.org/about.
[3] Duraspace. DSpace Registry. http://www.dspace.org/whos-using-dspace.
[4] HATII, University of Glasgow. Data Asset Framework. http://www.data-audit.eu/index.html.
[5] jQuery Grid Plugin. jQuery Grid Plugin - Grid plugin. http://www.trirand.com/blog/.
[6] T. A. Model. The Associative Model of Data White Paper technology Lazysoft. *Bookseller*, 2003.
[7] J. Rocha, C. Ribeiro, and J. C. Lopes. UPData - Scientific Data Curation at U.Porto. http://joaorosilva.no-ip.org/updata/wiki/doku.php.
[8] Royal Netherlands Academy of Arts and Sciences and Netherlands Organisation for Scientific Research. Data Archiving and Networking Services - About DANS. http://www.dans.knaw.nl/en/content/about-dans.
[9] The Apache Software Foundation. Apache POI - the Java API for Microsoft Documents. http://poi.apache.org/.
[10] University of Edimburgh. What is Edinburgh Datashare? http://datashare.is.ed.ac.uk/.
[11] University of Porto. Open Repository and Thematic Repository - repositorio.up.pt. http://repositorio.up.pt/repos.html.