Managing multidisciplinary research data

Extending DSpace to enable long-term preservation of tabular datasets

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

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

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

ABSTRACT

In a recent scoping study we have inquired into the data management needs of several research groups at the University of Porto and concluded that data quality and ease of on-line data manipulation are among the most valued features of a data repository. This paper describes the ensuing approach to data curation, designed to streamline the data depositing process and built on two components: a curation workflow and a data repository. The workflow involves a data curator who will assist researchers in providing meaningful descriptions for their data, while a DSpace repository was customised to satisfy common data deposit and exploration requirements. Storing the datasets as XML documents, the repository allows curators to deposit new datasets using Excel spreadsheets as an intermediate format, allowing the data to be queried on-line and the results retrieved in the same format. This dedicated repository provides the grounds for collecting researcher feedback on the curation process.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Digital Libraries; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Human Factors, Management, Standardisation

Keywords

Research data management, data repositories, DSpace extension, digital curation

*Supported by research grant SFRH/BD/77092/2011, provided by the FCT (Fundação para a Ciência e Tecnologia).

1. INTRODUCTION

The need to adopt effective data management procedures as part of the research workflow is currently assuming great importance, with data management requirements being imposed by research funding institutions. An example is the NSF, which now requires the inclusion of a data management plan in every research grant proposal [1]. In the UK, JISC has recently launched the Managing Research Data programme covering aspects such as infrastructures, data management plans and supporting technologies [3]; also, the Digital Curation Centre[1] provides resources and consultancy for researchers. Besides official policies, researchers are also becoming aware of the scientific impact of their data assets [5]. Universities are realising the benefits of exposing their research in institutional repositories and are seeking to extend them to research data. Several scoping studies and projects such as the DAF (Data Asset Framework) [9], the Edinburgh DataShare [11] or the DANS (Data Archiving and Networked Services) [2] have yielded data management workflows and recommendations. The management of research data requires a deep involvement of the researchers, since they are both creators and consumers of datasets. Thus, they must be involved in the dataset preparation and description process necessary to make the data available in the repository. Datasets pose hard problems regarding preservation formats, an issue that has been addressed by the MIXED project with a rich XML schema that can be used for the preservation of Excel spreadsheets through intermediate XML formats [12]. More general solutions for format identification and validation include the JHOVE[2] and the DROID[3]. After learning from the conclusions of large projects in this field, we designed our repository as a tool to support the researcher throughout the research workflow, continuing a previous work [8] in which we have detailed the architecture of a data management repository for U.Porto (University of Porto). Since then, a set of open-source modules have been combined with the DSpace platform, yielding a prototype that can be used by researchers from different domains for recording, sharing and preserving tabular data, allowing us to gather additional researcher feedback.

[1] `http://www.dcc.ac.uk/`
[2] JSTOR/Harvard Object Validation Environment : `http://hul.harvard.edu/jhove/`
[3] Digital Record Object Identification : `http://sourceforge.net/apps/mediawiki/droid`

2. SCOPING STUDY AND DEVELOPMENT CONTEXT

In order for research data repositories to become an integral part of the research workflow they must provide an added value to researchers throughout their research in return for their assistance in the curation of the datasets that they produce. With this in mind, we have designed a data curation workflow to handle the data management needs of active research groups at U.Porto. It is recognised that data curation should not be performed only at the end of the research workflow but rather follow the research process and start as early as the raw data is gathered from sensors or other equipment [4]. Later on, as researchers fully understand the concepts necessary to describe their domain of study [10], detailed annotation should be arranged between researchers and curators, quickly enough to make it possible for the researcher to cite the datasets in a publication. After the work is published, it should be made available at a publication repository with a direct link to the data—the latter being stored in a specialised repository offering data exploration and annotation features. This workflow was designed starting from the experience of previous work in this field [2, 9, 11] combined with our results from a scoping study involving several research groups at U.Porto[6]. Researchers interviewed in our study often stated that the added value of a repository depends on the ability to query and explore parts of the deposited data using their web browser, and also on the existence of metadata comprehensive and accurate enough to make it possible for them to reuse the data. At the organisational level, this means that research activities should be supported by a curation service that helps researchers maintain their data—a process that must begin early in the research process. From an engineering point of view, this poses several challenges that we have started to study at U.Porto by taking a standard DSpace repository and extending the underlying data model to allow for finer-granularity data access and metadata annotation, while maintaining the user-friendliness of its user interface.

3. DATA ACCESS AND PRESERVATION

Implementing a data repository using DSpace proved to be a challenge because its underlying data model is not designed to handle querying and exploring tabular datasets at table row granularity. In the standard DSpace data model, the smallest-granularity entity to which metadata can be added is an `Item`, which groups a set of data files representing the authors' work. This does not allow the system to retrieve parts of the data inside a file, requiring the user to download it as a whole and then explore the data using the program that the researcher originally used to create and manipulate it. This dependency on the original software used to prepare the data is often a reason for data loss as that software may become obsolete. The curation process starts with the standard DSpace workflow for self-deposit, after which an `Item` containing all the data files pertaining to the new dataset is created. Curators may then access a curation page (that we implemented) to upload custom-defined Excel workbooks [8] containing the tabular data originally present in the `Item`'s files, associating those tables with the files. When an Excel workbook is uploaded, the repository translates the data into an XML document and stores it in the DSpace database. This provides DSpace with the flexibility to have as many columns in a data table as necessary, regardless of their datatypes (`integer`, `string`...), something not originally possible given DSpace's relational model. XML documents are also easy to query using XQuery and are much more suitable for long-term preservation than their original counterparts. After a file is curated in this way, it becomes accessible to all registered users through a data exploration tool that allows users to restrict the visible rows by applying various filters on the columns directly from the Web browser. It is also possible to download the filtered data as an Excel workboook—the Excel format is only used as an intermediate format, not as the core storage of the repository.

3.1 An XML format for tabular data

The structure of the XML representation of the data tables is presented in a systematic manner in Figure 1.

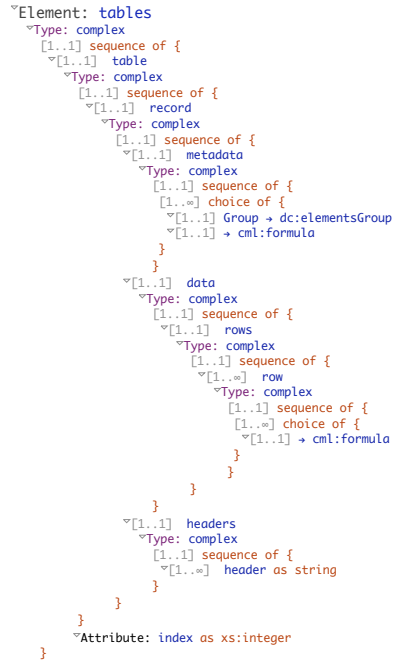

Figure 1: An example of the XML documents stored in DSpace, containing tabular data and their metadata

The structure of the documents includes a root element called `tables`, and contains a series of `table` elements. Each `table` contains a `metadata` section delimiting a sequence of qualified elements and their respective values. We need to incorporate elements from different XML schemas depending on the domain of the research dataset, so the `metadata` section can include elements from the Dublin Core schema as well as others from different metadata profiles, as is illustrated by the inclusion of the `cml:formula` element from the CML[4] schema. The `headers` section contains a list of all the table headers for this table (qualified metadata elements from arbitrary schemas), and the `data` section contains a series of `rows` from the current `table`. Each `row` contains a series of cells that match qualified elements and corresponding

[4]Chemical Markup Language `http://www.xml-cml.org/`

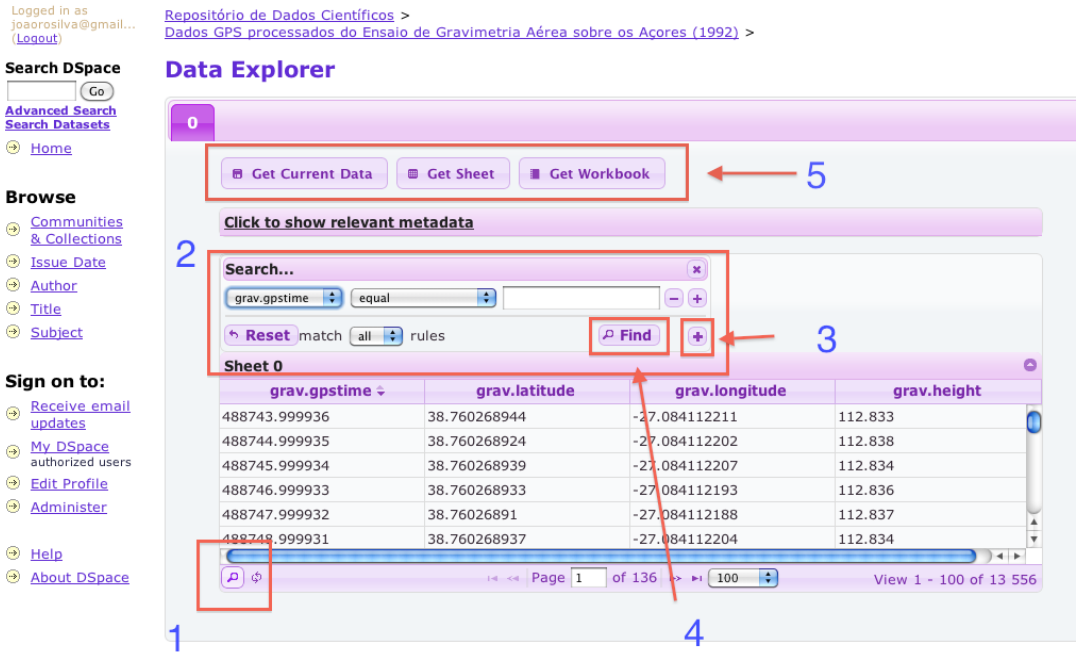

Figure 2: Dynamic grid interface used to explore datasets

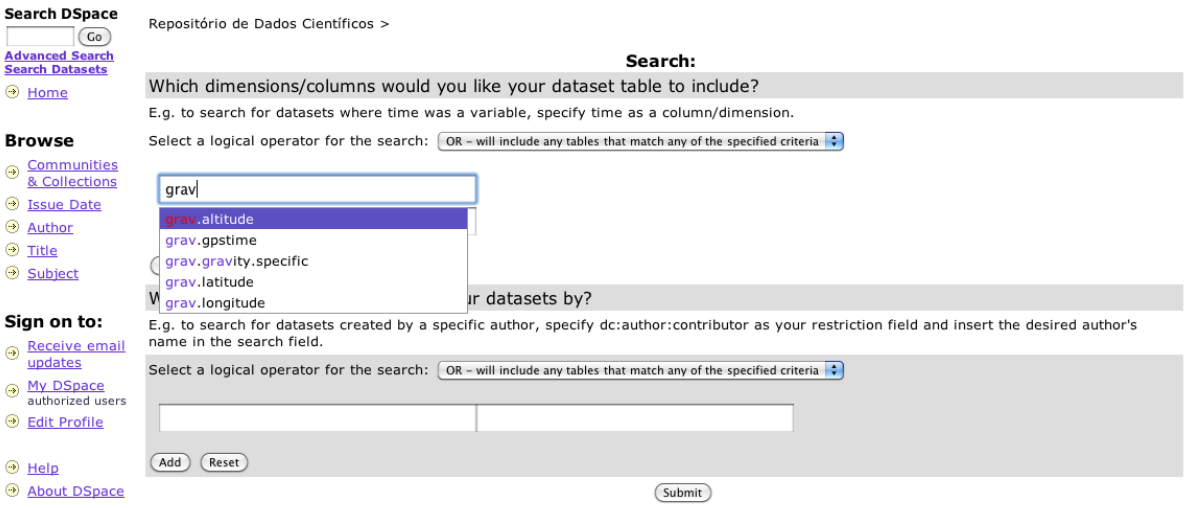

Figure 3: Search interface used for retrieving selected data tables

values. The format does not currently differentiate string, integer or any other datatypes for the contents of each cell, like existing profiles such as MIXED by DANS[12]. The MIXED project has proposed a richer XML schema for tabular data; we have chosen a lighter XML model for our data, for the sake of building a complete workflow where we can provide a fast prototype and use it to evaluate with our users the satisfaction of their requirements. Since the visualisation component relies on XML Stylesheets to convert the documents stored in the database to the format accepted by jqGrid (a jQuery component used for data presentation) it is possible to make changes to the internal schema without major changes in the code.

4. A WALKTHROUGH OF THE SOLUTION

The data curation workflow begins with a meeting between the researcher and the curator. In this meeting, the dataset that the researcher wishes to deposit goes through the standard DSpace depositing workflow, and a new `Item` is created. This `Item` will have its own metadata and all the files created by the researcher, exactly as they were originally produced. The next stage in our proposed workflow is the intermediate curation step for tabular data, in which the curator accesses the newly created DSpace `Item`'s page to retrieve each file and build an Excel workbook containing the data inside it, as well as any relevant metadata—each of the sheets of this Excel workbook contains a single table, as well as matching

table-level metadata. A workbook is built for each file and sent to the repository through a specific link (see area 3 of Figure 4), after which it is parsed by the system and translated into an XML document that is stored in the DSpace database. The Excel spreadsheet is discarded since it is only the intermediate format. After a file is curated, its tables can be explored through the data navigation interface (shown in Figure 2) and retrieved by specifying some of their columns on the querying interface (shown in Figure 3).

This interface is accessible from the `Item` page through the "Explore Data" link that we have added (Figure 4, area 2), next to the option that is normally used to download the file (area 1). When the "Explore Data" option is selected, the tables contained in the file are shown in a dynamic grid—shown in Figure 2— allowing the user to filter the data directly from the Web browser. In Figure 2, button 1 allows users to specify combinations of restrictions on each of the table's columns (which will appear in area 2). The user may add more restrictions by selecting button 3 or execute the filtering by selecting button 4. At any time, the user may download the selected data, the currently selected table or the whole workbook with all tables in the file. Both the data and metadata are provided in Excel format when the user selects the desired option from area 5. These use cases can be seen at the project's documentation wiki [6], including several videos [7] designed to demonstrate how curators and researchers can interact with the developed system.

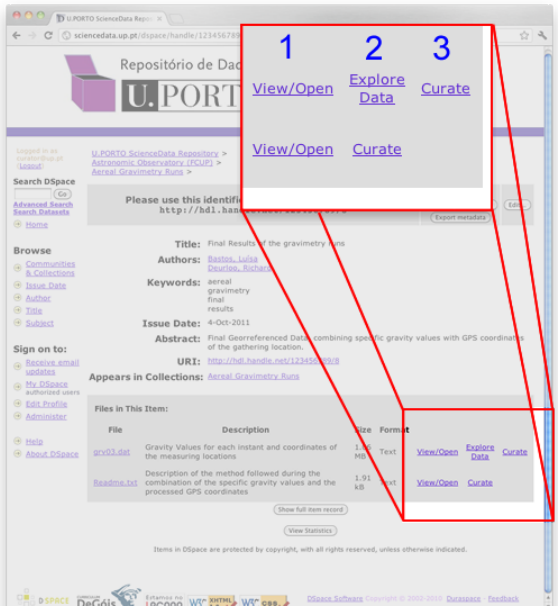

Figure 4: Extra "Curate" and "Explore" options were added to the DSpace Item exploration page

5. CONCLUSIONS AND FUTURE WORK

We have successfully implemented a curation workflow based on the needs of several U.Porto research groups. It uses an extended DSpace instance as the supporting platform, taking advantage of its effective built-in workflow engine for the self-deposit of datasets. The implemented extensions provide additional curation features designed to access the data at a fine granularity. The end result is a data exploration interface that allows users to query the data directly from their Web browsers and, if they wish, download the results in Excel format. The core data storage uses an XML format for enhanced long-term preservation, while Excel workbooks are used as vehicles for data transfer, therefore improving the user-friendliness of the system. The "look-and-feel" that users are accustomed to finding in a DSpace repository was maintained in the interface design, in an effort to provide a consistent user experience throughout the whole extended platform.

We are now in the process of validating the prototype through a second round of interviews with the researchers that participated in the scoping study. As for the continuation of the developments, we are now focusing on improving the methods through which researchers can retrieve datasets (by extending the DSpace free search capability to index the contents of our research datasets) and also in finding similarities between them using their metadata.

6. REFERENCES

[1] N. S. Foundation. NSF Data Management Plan Requirements. `http://www.nsf.gov/pubs/policydocs/grantsgovguide0111.pdf`, January 2011.

[2] Ingrid Dillo and Peter Doorn. *The Dutch data landscape in 32 interviews and a survey*. 2001.

[3] JISC. Research Data Management Infrascructure Projects, 2011.

[4] L. Lyon. Dealing with Data: Roles, Rights, Responsibilities and Relationships, 2007.

[5] H. A. Piwowar, R. S. Day, and D. B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, 2(3):e308, 03 2007.

[6] J. Rocha da Silva, C. Ribeiro, and J. Correia Lopes. UPData - Scientific Data Curation at U.Porto. `http://joaorosilva.no-ip.org/updata/wiki/doku.php`.

[7] J. Rocha da Silva, C. Ribeiro, and J. Correia Lopes. UPData - Scientific Data Curation at U.Porto - Demonstration Videos. `http://sciencedata.up.pt/doc/doku.php?id=demo_videos`.

[8] J. Rocha da Silva, C. Ribeiro, and J. Correia Lopes. UPData A Data Curation Experiment at U.Porto using DSpace. In *iPRES 2011 Proceedings*, pages 224–227, 2011.

[9] Sarah Jones. Data Audit Framework lessons learned report: GUARD Audit.

[10] A. Tonge and P. Morgan. SPECTRa - Submission, Preservation and Exposure of Chemistry Teaching and Research Data. Technical report, Cambridge University, Imperial College (JISC Digital Repositories Programme), 2007. http://lib.cam.ac.uk/spectra/FinalReport.html.

[11] University of Edimburgh. What is Edinburgh Datashare? `http://datashare.is.ed.ac.uk/`.

[12] R. van Horik and D. Roorda. MIXED: Repository of Durable File Format Conversions. In *iPRES 2009 Proceedings*, 2009.