UPBox and DataNotes: a collaborative data management environment for the long tail of research data

João Rocha da Silva
INESC TEC, FEUP,
Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto Portugal
joaorosilva@gmail.com

José Pedro Barbosa
FEUP, Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto Portugal
ei08036@fe.up.pt

Mariana Gouveia
FEUP, Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto Portugal
ei10124@fe.up.pt

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

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

ABSTRACT

Current research data management workflows are often an *a posteriori* process, with research datasets being targeted for preservation actions after the whole research process is completed. This approach works well for research publications but not for research datasets due to their dynamic nature. It is important to gather data production contexts, so the data management process should be present since the start of the research, effectively becoming a part of the workflow. Due to their rigid workflow-based deposit approach, widely used repository solutions are not intended to support the fast-paced evolution of datasets as they are produced. In this paper, we present a collaborative data management environment designed to help a small research group store and describe their datasets in preparation for later deposit in a data repository. It is built on two integrated, open-source components: UPBox—a private cloud and web-based file storage environment—and DataNotes—a solution tailored for researchers to collaboratively describe their data, based on Semantic MediaWiki. Preliminary usage tests have shown that the features of this solution respond to data management needs in research groups.

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, Semantic MediaWiki, digital curation

1. INTRODUCTION

Research data management is assuming an increasingly relevant role in the research workflow. The adoption of appropriate research data management practices presents advantages for research funding institutions (e.g. international recognition of their project's results) but it is ultimately the researchers who must realise the potential improvements to their work that may come from the adoption of such practices. These have already been extensively discussed and include increased citation rates for articles that provide access to base data, reproducibility of research results, formulation of new research questions [10, 3, 7] and also the wider goal: faster advancement of science [2]. These goals, while important, are often seen by researchers as unclear long-term benefits of a process that requires a substantial time investment on the researcher's part.

Current research data management workflows usually rely on a dataset description process performed by professional curators. While this process is effective for producing high-quality generic metadata, the inclusion of domain-specific metadata in the description of research datasets requires the close collaboration of the dataset creators, which are experts in the domain but often lack the data management skills required to perform comprehensive descriptions of their datasets [13]. Only through this cooperation can we produce rich domain-specific descriptions for research datasets [6]. However, this approach tends to require too much time from researchers, who often do not realise any immediate advantages in the data management process. At the same time, data curators become the bottleneck in the curation process—the end result is a process that can turn into a series of sporadic contacts and lost opportunities for describing those datasets as their authors move to pursue other research questions.

While community-supported research data repository direc-

tories are already a reality—an example is DataBib, a directory for research data repositories [14]—collaborative environments for curators and researchers to describe datasets are still in their early stages. In 2013, the DataUP project [12] has shown how a self-deposit tool built directly into Microsoft Excel can help researchers deposit spreadsheets directly from their working environment. An interesting aspect of the project is that it focuses on guiding researchers through the description of the spreadsheets, pointing out possible mistakes in their formatting and organisation, while making it easier to describe them using standardised metadata.

With this work, we present an approach at data management that has the primary goal of making it an ongoing process that supports the everyday activities of a researcher—a view that has already been expressed in a recent report [5]. This more dynamic environment relaxes some interoperability requirements and strict metadata production workflows in favour of capturing the data and its context as it is produced and processed. At the same time, it provides a set of features that immediately reward researchers for their efforts in describing and organising the datasets. By using UPBox—a "Dropbox" for research datasets—and DataNotes—a data description wiki—to manage the data, they gain access to a safe and simple file storage area for the datasets, easier data sharing within their research group and a collaborative wiki-based data description tool for the datasets.

Since researchers can be reluctant to deposit research datasets on infrastructures outside of their control, we have designed this environment to work completely under the research institution's control, that is, in its own servers. We see this environment as a "staging area" to prepare datasets for later repository ingestion. An important part of the data description work will already be done by the time the final data is produced and research results published, effectively making it an easier task and encouraging researchers to complete the data management process with the assistance of data curators. This work is oriented towards the "long-tail of data" as we are not trying to manage the very large datasets created in some research areas—which are often supported by appropriate infrastructures—but rather the myriad of small datasets produced by diverse research groups [9], which tend to be more at risk due to the scarce data management resources available in their projects.

2. COLLABORATION: THE KEY FOR USER ENGAGEMENT

Most current research data management workflows take an *a posteriori approach*. This means that researcher involvement in the process is reduced to a certain point in time when the datasets are curated and deposited in a repository platform. Some advantages of the *a posteriori* process are its simplicity in terms of planning (for both researcher and host institution), a relatively reduced effort and easy learning curve for researchers. More importantly, it yields comprehensive, standardised dataset metadata. However, our past experience has shown that this approach also has some drawbacks concerning the number of datasets that are actually preserved.

Another issue surrounding dataset description is *timing*. Interaction between researchers and curators usually takes place at a relatively late phase of the research activities, after researchers have gathered and processed their datasets, obtained results and published them. While currently this is the most common practice in publication management, it is clear that research data curation should start as early as possible in the research process [8, 5]; datasets should be described as soon as the researcher possesses adequate domain knowledge and has created them, since that is when the data production context is completely available [1].

In 2012, the UPData project [11] provided insight on the features that researchers find interesting in a data management workflow. Ensuring the reproducibility of research findings and relating publications to their base data is interesting, but researchers tend to focus on more immediate benefits of integrating the datasets in the research data management workflow. Among these are, for example, easy sharing among research colleagues—sending an URL to a resource where the dataset is available is a basic but clear example. One of the main reasons that make researchers reluctant to produce metadata for datasets is the work involved in filling in descriptors that they often see as irrelevant in their own domain. To make the process less tedious (albeit with a compromise in interoperability) metadata schemas can be replaced with application profiles: "schemas which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application" [4]. It is hard, however, for curators to manually design specific application profiles on a research group/project basis, so these profiles should emerge naturally through descriptor reuse on each domain—a collaborative description environment is a pre-requisite for this to happen.

In most cases, research activities are performed by groups of researchers in close collaboration, so it makes sense to reduce duplicate efforts and make research data management a collaborative effort as well. In fact, data management can even be useful to research teams by helping them share data within the research group, while encouraging the team to share description efforts as well. Metadata production in a collaborative context becomes *rewarding in the short term*, allowing the data management environment to become a central hub of the research activities. As a side effect, application profiles may surface as the descriptors from different metadata schemas are reused in different domains.

3. COMBINING A PRIVATE CLOUD WITH A SEMANTIC WIKI

Our proposed research data management environment is built on two main components, interconnected by a set of web-based communication endpoints or *web-services*.

The architecture of the system is shown in Figure 1. UPBox (1) uses the server's local storage, which can be mapped to a RAID-based storage (our selected solution), network volume, or distributed storage layer. A possible alternative would be a Hadoop File System (HDFS) mountable volume[1] to provide abstraction over a private cloud for hor-

[1] `http://wiki.apache.org/hadoop/MountableHDFS`

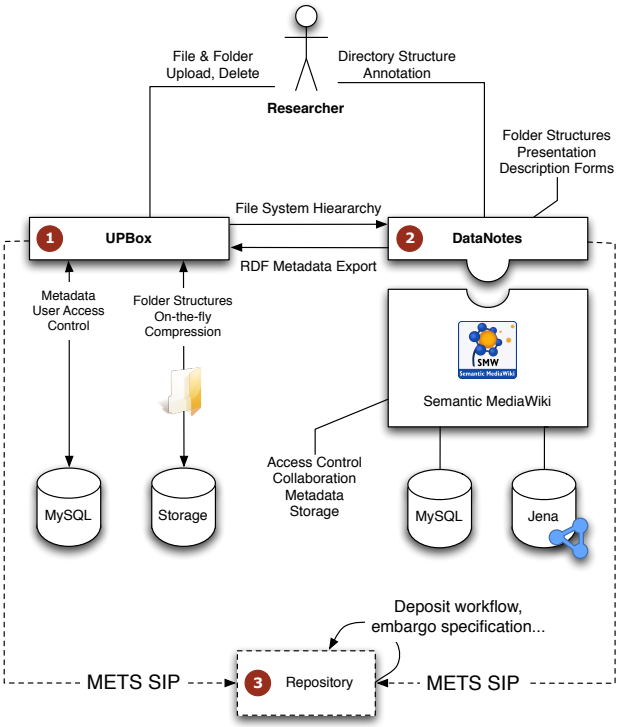

Figure 1: Architecture of UPBox and DataNotes

izontally scalable storage. A MySQL database is used to save the data required for user management, access control and metadata concerning the directory structures. All files are compressed and decompressed *on-the-fly* when users upload/download them to/from the server, to minimise the storage space required to support the system. UPBox is connected to U.Porto's central information system (SIGARRA) via an LDAP (Lightweight Directory Access Protocol) plugin, enabling U.Porto staff to log into the system using their SIGARRA credentials. External users can also register in the system, enabling inter-university collaborative work. The platform allows researchers to create "projects", areas where folders can be created and files can be deposited much like *Dropbox* folders. A project can be shared with team collaborators by adding members (the system provides suggestions from the list of registered users). Members of a project can upload files, as well as create folders or delete them. Several files can be uploaded simultaneously to facilitate the migration of existing datasets.

DataNotes (number 2 in Figure 1) is a wiki-based directory structure annotation platform, built on top of Semantic MediaWiki[2]. It allows users to quickly produce wiki pages containing the metadata for their datasets. The goals fulfilled by DataNotes are:

1. Providing a collaboration environment for describing directory structures, supporting version control, locking, concurrent edition management, namespaces and user access control.

2. Helping researchers in the group to find datasets via text-based search over the metadata.

3. Offering a friendly user interface, albeit with sophisticated capabilities to capture relationships between parts of the dataset and also semantic inter-dataset links for those cases where such detail is required.

4. Easy sharing of dataset descriptions (ideally the ability to send a direct link to a described folder or file).

5. Absence of dependencies on closed source solutions, modules or libraries that may endanger the access to the data stored in the solution as it becomes deprecated and there is no way to update or review its business logic.

6. Ease of installation, making it easy for any research institution to host their own DataNotes instance to support the work of their research groups.

7. Preparing datasets for long-term preservation by easing the export of dataset metadata records in a standard format (e.g RDF), ensuring the survival of the data even in the event of DataNotes being replaced with another platform.

8. Providing programmatic search capabilities that enable resource retrieval from the wiki, based on criteria specified by external systems.

Since DataNotes is based on a wiki platform, namespace management and access control features are already present, along with concurrent editing capabilities and continuous versioning of the wiki pages which contain file and folder metadata. Free text search is also present, allowing users to retrieve dataset pages via a global search function. The interface can be considered user-friendly as most of the standard MediaWiki components are maintained in Semantic MediaWiki and remain unchanged—keeping the easy learning curve that continues to make it possible for non computer-savvy users to write and review Wikipedia pages. The system also allows users to share dataset descriptions easily, since every description is a wiki page with its own unique URL—these URLs are shown in the web browser during navigation and can simply be copied and pasted in a message for sharing with other users that have permissions to access the resource.

The "Repository" module (number 3 in Figure 1) represents an existing repository (such as DSpace). After datasets are deposited in UPBox and their descriptions produced in DataNotes, researchers should be able to automatically package the existing state of a folder (for example) and send it in to the repository, where a new ingestion workflow will be started. The metadata for the new dataset will be subjected to all the usual validations by a curator (including embargo specifications) and then deposited in the repository. At that time, and due to the "static" nature of the resulting repository resource, it can be cited safely in publications via a persistent identifier (URI).

[2] `http://semantic-mediawiki.org`

4. CONCLUSIONS AND FUTURE WORK

Observation of current practices with research data suggests that data management should accompany researchers in their everyday activities instead of being performed *a posteriori*. The goal is to maximise the opportunities for gathering datasets, allowing their later ingestion into a repository for long-term preservation. Another goal is to make it possible for institutions to maintain complete control over the data produced by their researchers. To address these needs, we have designed and constructed a fully open-source collaborative environment for data sharing and description among research groups. The system allows researchers to deposit their datasets in Dropbox-like folders and then describe them using an integrated wiki interface.

Presently, there is no support for versioning in the UPBox platform—unlike DataNotes, which already offers versioning capabilities for the metadata pages of each file/folder since it is built on Semantic MediaWiki. Disk quotas for UPBox users are also in the list of possible improvements to provide control over the server's storage space. A more sophisticated access control system could also be implemented, allowing project owners to specify the actions to be performed by each team member for each folder (and subfolders) in the project. An UPBox desktop client to enable seamless synchronisation with the remote storage (much like *Dropbox*) is also planned, as well as a folder upload feature, to make it easier to migrate an entire existing directory structures to UPBox.

The possible improvements for DataNotes have to do with making dataset description easier by automating repetitive tasks. By allowing researchers to reuse sets of descriptors from a folder to annotate another, we can encourage the creation of application profiles for each domain through community reuse. Also, to complete the data management cycle, datasets described in this environment should be handed over to a repository in a transparent way, at a moment chosen by the project owner. To achieve this, a connection to a data repository must be available, and we plan to use DSpace to build upon previous work on DSpace extensions for managing research datasets. In the future it will be possible to start DSpace deposit workflows directly from UPBox or DataNotes; given DSpace's OAIS-compliant endpoints, these systems must be capable of building METS SIP packages and submitting them to DSpace. Our goal is to make this process fast enough for researchers to be able to cite their datasets at the time of the publication of results, making it easier for their audience to find the corresponding base data.

A small validation experiment with a group of researchers from the FEUP Mechanical Engineering department was performed; thus far, the feedback on the improvements introduced by this platform has been positive, and has provided some insight on further development. For example, the decision to allow external users to register in the system was taken due to the fact that this research group included members from UTAD (University of Trás-os-Montes and Alto Douro), and they wanted to use UPBox to share datasets in the group—a situation that we found very likely to occur in the future. As the tools start to be used by different research groups, we will also determine if these tools should act only as a "staging area" or if they should be extended to satisfy long-term preservation requirements as well.

5. ACKNOWLEDGEMENTS

This work is supported by research grant "SFRH/BD/77092/2011", provided by the FCT—Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) and by the ERDF—European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), supported by National Funds through the FCT—Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project "FCOMP-01-0124-FEDER-022701".

6. REFERENCES

[1] T. Alan and M. Peter. SPECTRa-T Final Report July 2008. 2008.

[2] C. Borgman. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 2012.

[3] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers. Examining the Challenges of Scientific Workflows. *Computer*, 40(12):24–32, Dec. 2007.

[4] R. Heery and M. Patel. Application profiles: mixing and matching metadata schemas. *Ariadne*, (25), 2000.

[5] L. Jahnke, A. Asher, and S. D. C. Keralis. *The Problem of Data*. Number pub154. Council on Library and Information Resources, 2012.

[6] S. Jones, S. Ross, and R. Ruusalepp. Data Audit Framework Methodology, 2009.

[7] P. Lord, A. Macdonald, L. Lyon, and D. Giaretta. From Data Deluge to Data Curation. In *eScience All Hands Meeting 2004*, pages 371–375, 2008.

[8] L. Martinez-Uribe and S. Macdonald. User engagement in research data curation. In *13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Proceedings*, volume 5714 of *Lecture Notes in Computer Science*, pages 309–314. Springer, 2009.

[9] C. Palmer and M. Cragin. Data curation for the long tail of science: The case of environmental sciences. *Proceedings of the 3rd International Digital Curation Conference*, 2007.

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

[11] J. Rocha da Silva, C. Ribeiro, and J. Correia Lopes. Managing multidisciplinary research data: Extending DSpace to enable long-term preservation of tabular datasets. In *iPres 2012 Conference*, pages 105–108, 2012.

[12] C. Strasser and P. Cruse. The DMPTool and DataUp: Helping Researchers Manage, Archive, and Share their Data. *Research Data Management Implementations Workshop*, 2013.

[13] A. Swan and S. Brown. The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. Report to the JISC. 2008.

[14] M. Witt and M. Giarlo. Databib: An Online Bibliography of Research Data Repositories. *ALA Annual Conference*, (Paper 2), 2012.