Data Curation at U.Porto:

Identifying current practices across disciplinary domains by

Cristina Ribeiro, Maria Eugénia Matos Fernandes [1]

U.PORTO

Abstract

The University of Porto is the largest Portuguese university with more than 60 research centers that generate a significant part of Portuguese scientific production. U.Porto is currently concerned with the curation of and the access to the scientific data generated by its researchers. Researchers are motivated to keep their data assets alive as integral part of their published results, and the scientific impact derived from open datasets is also becoming apparent. We have followed the recommendations from well-known actions in research dataset auditing to lead a short study on available data at U.Porto. The study has involved researchers from a diversity of disciplines, collecting their views on data curation and sample data. As a result we have identified some generic use cases to inform the development of a data repository prototype. Our contacts with the researchers have revealed a great diversity of situations, from groups where data curation was already integrated in the research practice to others who were struggling to incorporate it into their workflows. Our experiment was focused on data auditing and use case identification, but we also concluded that in many groups there is a strong concern with the premature exposure of the data. The sample datasets provided by the researchers are being transformed into preservation-friendly archives to be part of a data repository. We will extend the repository infrastructure with data search facilities and expect feedback from the researchers to help define the research data management services at U.Porto.

Keywords: : Data curation, management of research data, data repositories

Introduction

The University of Porto (U.Porto)[2] is currently concerned with the curation of and the access to scientific data generated by its researchers. A steady growth in research activity in all domains, international cooperation initiatives and access to data that is either generated by local projects or available via joint projects has generated many ad-hoc data archives. Research cycles of projects and scholarships are very short-term from the data assets point of view: data generated in one project may, if there is no continuation project, be abandoned and lost in less than five years. The researchers' perspective on the longevity of such data is, in general, quite optimistic and the lack of national mandates for data curation favors the continuation of this state of affairs.

In this work we have followed the recommendations of pioneering actions in scientific dataset auditing to lead a short study on available datasets at U.Porto. An analysis of current initiatives in this area has shown that close contact with researchers is essential for getting a clear view on their needs (Ribeiro, et al. 2010). Our study involved researchers from a diversity of disciplines, collecting their views on data curation and sample data (Rocha da Silva, Ribeiro and Correia Lopes 2011). As a result, we have identified some generic use cases to inform the development of a data repository prototype.

Our contact with researchers revealed a great diversity of situations. There are areas where some form of data curation is already embedded in current practice, mainly due to the requirements of publication venues or the need to share data in international initiatives. Some researchers are motivated and aware of both the value of their data and the existing threats on it, but are still struggling to incorporate data curation into their workflows. Others, faced with the possibility of having their data curated in a repository, were extremely cautious with respect to privacy issues.

Our study focused on data auditing and use case identification, but we also asked researchers for samples of their data. The datasets provided by the researchers are being transformed into preservation-friendly archives to be part of a data repository. We are extending an existing repository infrastructure with data search facilities and expect feedback from the researchers to help define the data services for the U.Porto data repository (Rocha da Silva, Ribeiro and Correia Lopes 2011).

In this paper we provide a short overview of research at U.Porto, an outline of the goals for the data curation project at U.Porto and describe its preliminary results. We conclude with some reflections on the project results and the perspectives for the management of research data at U.Porto.

U.Porto: a research university

U.Porto is the largest Portuguese university. It comprises 14 schools, a business school, 30 libraries, 12 museums and about 70 R&D units, 31 of which have been regularly classified at the top ranks by a panel of international experts as part of the Portuguese research units evaluation. Its population consists of about 30,000 students, more than 2,366 teachers and researchers (76% PhD) and 1,689 technical and administrative staff.

U. Porto offers a large range of courses covering all levels of higher education and all the major areas of knowledge. There are over 670 training programs, including undergraduate, masters, integrated master, doctoral, continuing education and specialization courses. The number of foreign students under mobility programs represents more than 8% of the total number of students. U.Porto aims at becoming a national and international reference by the high level of its students and the production and dissemination of knowledge. It can be said that the target of being among the top 100 higher-education European institutions for its 100th anniversary in 2011 has been reached.

The physical dispersion is a characteristic of U. Porto, as the buildings of the University—schools, RD&I institutes, student residences, sports and cultural facilities—are located in three separate areas of the city of Porto. Moreover, there are research institutes and centers spread throughout the city and some of them even beyond its geographical boundaries.

The shortcomings of this geographical dispersion have been practically overcome by the SIGARRA Information System (Information System for the Aggregated Management of Resources and Academic Records). SIGARRA originated in the Engineering School in 1996 and its success led to the transversal implementation in the University from 2003 on. Currently, all the U.Porto schools, as well as the Rectorate, the Social Services and some of the research centers use SIGARRA.

This integrated system was conceived to facilitate the production, flow, storage and access to the information managed by the institution—contents of pedagogical, scientific, technical and administrative nature—and to promote internal cooperation and the cooperation with external academic, scientific and business communities. The SIGARRA system interacts with other applications and systems within the University, such as the libraries, the e-learning services, the student administration and the financial management systems and also with U.Porto institutional repository, built on a DSPACE platform. Figure 1

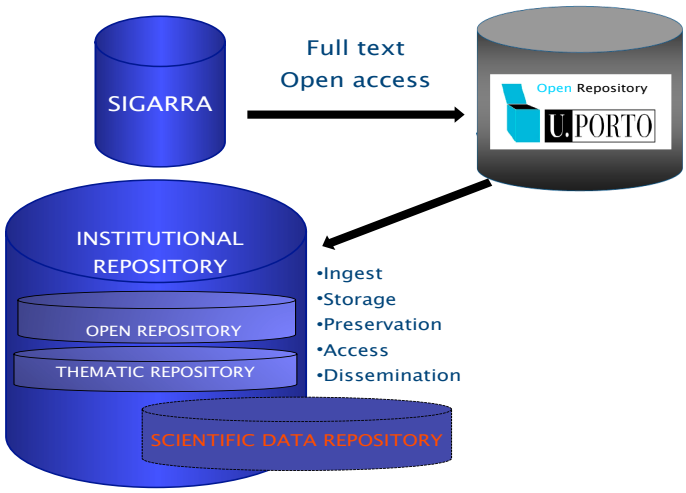

Figure 1 The SIGARRA information system and the institutional repository

illustrates the integration between the information system and the open repository.

The creation of the U.Porto Institutional Repository complemented the information management strategy by the end of 2007. The interface connecting the Information System and the Open Repository guarantees that the intellectual production of the academic and scientific community is transferred automatically from the Publications module of SIGARRA to the Open Repository. The authors just have to register and self-deposit the full text of their publications on their institutional pages, defining that they are public. The same interface also assures the connection between other applications used within the university to register and catalogue the library collections—such as Aleph—and the Open Repository, thus enforcing consistency of data across different applications and systems.

From the moment it was created, the number of publications of the Open Repository of U.Porto has grown steadily. At the beginning of 2008, the Repository had almost 1,800 full-text and open access publications. Three years later, the number of records has evolved to more than 18,000.

One of the missions of U.Porto is the creation of cultural, artistic and scientific knowledge within the academic community, composed by teachers, researchers and students. This concern has increased in the recent past due to a great variety of factors.

Beyond some aspects already mentioned above—such as the functionalities of the Publications module of the Information System and the benefits of the interconnection between SIGARRA and the Open Repository—, one cannot ignore the emphasis that has been placed on the recommendations made to the authors to make their intellectual outputs available, stressing the fact that these works are created in the context of their teaching and research activities. It is also important to highlight the suggestions made to the authors to have them consider, whenever possible, the "SPARC Author Addendum", when they sign contracts with publishers, so they maintain the right to self-archive their work in institutional open repositories, as well as the advice given to researchers to use the university-recommended format to register their affiliation.

Considering the last decade, 1/5 of the Portuguese scientific production was generated at U.Porto. Current figures show that U.Porto is responsible for more than 21% of the Portuguese scientific articles indexed in the ISI Web of Science.

Goals of the data curation project

U.Porto is currently concerned with the curation of and the access to scientific data generated by its researchers. There is a growing awareness of the fragility of personal digital archives and researchers feel that they need to keep their data assets alive as the research workflow becomes more sophisticated. The possibilities of scientific impact derived from open datasets are also becoming evident.

As a result of an identification task, we present a preliminary study on datasets that are being used in current research at U.Porto. The emphasis has been on diversity, picking examples from life sciences, engineering, social sciences and arts. The identification also provides insight on current models for data curation, both formal and informal, and on the sensitivity of researchers with respect to open access to their data (Scientific Data Curation at U.Porto, 2011).

The study has been complemented by the development of a data repository prototype. The purpose of the development is twofold: to provide services which address some of the requirements identified with researchers in a working tool and to establish the basis for a second round of interaction with the researchers, this time using the repository platform to illustrate the use cases in data curation and to test them with their end-users.

We were quite aware, from the start, of the many challenges of the project, but also of its strengths. The study conducted in the context of the national repository project (Ribeiro, et al. 2010) located similar initiatives (Rice 2009, Martinez-Uribe 2009) and existing recommendations that have helped to establish the main lines of the data audit experiment (University of Glasgow, DCC 2009). The commitment of the Rectorate and Digital University Services of U.Porto to the development of the Institutional Repository has provided a solid ground for supporting an experimental data repository. On the other hand, and in spite of the absence of mandates for data curation plans in national projects, we were able to find many researchers concerned with the management of their data and committed to sharing them within their research groups and in the context of international projects in which they are involved.

Data Auditing and Dataset Collection

IThe data audit at U.Porto has followed the recommendations issued by similar initiatives, namely the methodology proposed in the Data Asset Framework (University of Glasgow, DCC 2009). Considering that this is the first approach to data curation at the university level, we have decided to give preference to the diversity of domains. The choice of research groups to include in the study has followed a mixed strategy, selecting some groups due to personal contacts by the team members and others resulting from a call issued by the university Rectorate and Digital University Services to the directors of schools and research institutes. The first contact with the researchers led to an appointment of interviews at their laboratories, based on a script that allowed for many open questions. In cases where the researchers were willing to provide sample datasets, a follow-up interview was scheduled to discuss data formats, the definition of data and their terms of use. We adopted the recommendations of the Data Asset Framework (University of Glasgow, DCC 2009) to prepare an "Interview Guide" (U.Porto 2011) and a "Comprehensive Questionnaire" (U.Porto 2011) that were used to collect the researchers profiles, some general information on their datasets, preservation actions and expected use cases for the scenario of a university-level data repository.

There was no imposition on researchers to provide data, but most (8 out of 13) volunteered to provide sample datasets, knowing that the data would be used to design and prototype the system and that there was no agenda for a repository service, so they could not expect any immediate benefits from the collaboration.

Table 1 lists the nature of the collected datasets and the access conditions established by the researchers. Interviews were a rich source of information for their needs, where we can highlight data preservation and data exchange with research partners, either internally at U.Porto or externally in international projects and partnerships.
The collected datasets provide a first view on the research data at U.Porto, with data obtained from science, engineering and social sciences resulting from either automatic acquisition or direct collection by the researchers and access conditions ranging from open data to data useable in research but whose origin must be kept anonymous due to pending contracts. Most of the datasets under consideration were originally created as a result of research projects, but there were also data collected by external institutions with which U.Porto holds service contracts and data collected by national institutes, such as the census data created by the national statistics institute.

Table 1. Domains and access conditions for data

Domain	Dataset	Access
Astronomy	Gravimetry	Free
Chemical Engineering	Pollutant analysis	Contract pending
Mechanical Engineering	Material fracture	Embargoed
Civil Engineering	High-speed railways	Embargoed
Educational Science	Interviews	Embargoed
Psychology	Interaction records	Embargoed
Economy	Population	Embargoed
Ecology	Plant distribution	Embargoed

The interviews with the researchers confirmed our initial assumption that the design of a solution for a data repository should be determined by researchers needs, rather than by any abstract data management convenience (Borgman 2011). The interviews showed more concern with functionalities such as data browsing and querying than with strict data preservation or management.

For the 8 sample datasets provided by the researchers we created basic Dublin Core descriptions to ease their deposit into the upcoming data repository.

Future Directions for Data Management at U.Porto

The data audit at U.Porto has exceeded our expectations with respect to the commitment of researchers with data curation. In some areas

with established practices of deposit in international repositories, the data curation problem can be considered solved, but this is not the case in most domains. The resources required for this small-scale experiment are indicative of the effort required for setting up a data curation service at an institution with the size of U.Porto.

The use cases identified in this study are being used to define the requirements for the U.Porto data repository. The data samples are the basis for the design of data models where the tradeoff between generality and usefulness must be considered to make the curation process practicable. An experimental repository is being developed to test the requirements. As soon as we have a platform where some datasets are deposited and can be queried, researchers can explore it, detect the shortcomings of the proposed approach in their own domain and engage in future developments.

This work has raised even more issues than initially expected and many questions remain unanswered. We have observed that in several areas researchers are willing to participate in data curation, even in a scenario where they cannot expect any immediate benefits. This proves that we will be able to stimulate their cooperation in the following steps, but there must be some perceived gain for the researchers in order for this commitment to be sustained. A scenario where people are motivated to participate and get no practical results may ultimately compromise this and future initiatives.

The technological support for a research data repository is another open issue. The maturity of software for institutional repositories shows that we do not have to start from scratch and that basic functionality can be taken for granted. But, on the other hand, the use cases for research data are much less clear and less uniform than those for an institutional repository.

Another issue worth reflection and experimentation is the nature of data curation services. There are currently no data curation services in Portugal so there is no experience with respect to their integration in a research institution. Libraries are experienced with many of the issues in curation, but not equipped with the highly technological expertise it requires. Computing centers have complementary expertise, but their mission is centered in very different services.

Maybe the most critical aspect for the success of a data curation project is compliance with researchers needs. Institutional repositories have flourished due to the adoption of repository technology, originally created to satisfy very specific needs, by the more traditional library community. There are currently no well-established generic platforms for research data management but many custom-designed systems already exist. Experience and successful developments will show whether generic platforms can cater to the needs of researchers in different domains or if they have to be more specialized by discipline.

References

"U.Porto Comprehensive Questionnaire." UPData. 2011. http://science-data.up.pt/doc/ (accessed November 2011).

"Interview Guide (in Portuguese)." UPData. 2011. http://sciencedata.up.pt/doc/ (accessed November 2011).

Scientific Data Curation at U.Porto. Edited by João Rocha Silva. 2011. http://sciencedata.up.pt/updata/ (accessed November 2011).

University of Glasgow, DCC. The Data Asset Framework Implementation Guide. October 2009. http://www.data-audit.eu/ (accessed November 2011).

Borgman, Christine L. "The Conundrum of Sharing Research Data." Journal of the American Society for Information Science and Technology, 2011: 1-40.

Hey, Tony, Stewart Tansley, and Kristin Tolle, . The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft, 2009.

Martinez-Uribe, Luis. Using the Data Audit Framework: an Oxford case study. University of Oxford, http://ie-repository.jisc.ac.uk/300/, 2009.

OECD. OECD Principles and Guidelines for Access to Research Data from Public Funding. http://www.oecd.org/dataoecd/9/61/38500813.pdf, 2007.

Ribeiro, Cristina, Eloy Rodrigues, Maria Eugénia Matos Fernandes, and Ricardo Saraiva. Repositórios de Dados Científicos: Estado da Arte (in Portuguese). Project Report, Porto: RCAAP, 2010.

Rice, Robin. DISC-UK DataShare Project: Final Report. Project Report, Edinburgh: University of Edinburgh, 2009.

Rocha da Silva, João, Cristina Ribeiro, and João Correia Lopes. " UPData-A Data Curation Experiment at U.Porto using DSpace." Proceedings of 8th International Conference on Preservation of Digital Objects, iPRES 2011. iPRES, 2011.

NOTES

1. Cristina Ribeiro, DEI-Faculdade de Engenharia da Universidade do Porto/INESC TEC, Rua Dr. Roberto Frias, s/n, Porto, Portugal, mcr@fe.up.pt. Maria Eugénia Matos Fernandes, Reitoria da Universidade do Porto, Universidade Digital Praça Gomes Teixeira, Porto, Portugal, efernand@reit.up.pt.

2.U.Porto: Homepage. http://www.up.pt/