Chapter 10
Managing Research Data at the University of Porto: Requirements, Technologies, and Services

João Rocha da Silva
University of Porto, Portugal

Cristina Ribeiro
University of Porto, Portugal

João Correia Lopes
University of Porto, Portugal

ABSTRACT
This chapter consists of a solution for the management of research data at a higher education and research institution. The chapter is based on a small-scale data audit study, which included contacts with researchers and yielded some preliminary requirements and use cases. These requirements led to the design of a data curation workflow involving the researcher, the curator, and a data repository. The authors describe the features of the data repository prototype, which is an extension to the widely used DSpace repository platform and introduced a set of features mentioned by the majority of the interviewed researchers as relevant for a data repository. The data repository platform contributes to the curation workflow at the university, with XML technology at its core—data is stored using XML documents, which can be systematically processed and queried unlike its original-format counterpart. This system is capable of indexing, querying, and retrieving, in whole or in part, datasets represented in tabular form. There is also the possibility of using elements from domain-specific XML schemas for the cataloguing process, improving the interoperability and quality of the deposited data.

DOI: 10.4018/978-1-4666-2669-0.ch010
INTRODUCTION

It is currently recognized in most research areas that data is not only an essential component of the scientific process but that research is more and more driven by the data itself. The so-called “data deluge” (Hey & Trefethen 2003) has prompted research institutions, funding agencies, data specialists, editors and researchers in many areas to find solutions for managing research data and to comply with the requirements of their research processes.

Current technologies and devices generate an immense flow of data, most of it even prior to the application of any scientific process; but even data that are not created in the scope of research can become the subject of research if it is recorded, making it available and interpretable. This “data deluge” is raising concerns about the possibilities that may be missed if all this data cannot be stored, explored, and made available.

This increasingly widespread use of powerful computing infrastructures has led to the coinage of the term “e-Science.” E-Science is a general term used to designate research activities that are supported by large quantities of data and substantial computational resources. The growth in computational capability in the recent decades has led to the emergence of the so-called “Fourth Paradigm” of research, through which scientific research is built on massive quantities of data captured by instruments or generated by simulations. These data assets are leveraged using federated resources designed to support researchers in their collaborative efforts, aiding them in the analysis, visualization and dissemination of their results (Hey, et al., 2009).

The publication of research, which relies on extensive exploration of data, tends to be limited without the publication of the data itself and major publishers have made very concrete proposals in this direction. Nature, for instance, provides for the availability of “supporting online material” on the published papers, and even considers software in this category (Hanson, et al., 2011).

Another line of inquiry related to the wider availability of data is data sharing and reuse. The Open Data movement advocates free access to data as the path to improving research, policy making and transparency in many domains (Foundation, 2011). The concern with data reuse and the possibilities of promoting new “digital data products” has also been raised (Faniel & Zimmerman, 2011).

A recent overview (Borgman, 2011) has shown that preserving research data is relevant for many reasons—but there are four main drivers for the implementation of these practices. These are ensuring the reproducibility and verifiability of the research findings, making the results of publicly funded research available to the public, enabling others to ask new questions based on existing data, and finally, advancing the state of research and innovation.

Research data is especially hard to preserve because of its complex underlying semantics and its heterogeneity. Researchers use many file formats, often relying on proprietary software or even software that they wrote themselves. Datasets can also vary greatly in volume depending on the domain and on the type of data that is to be stored. However, it is important to point out that big data storage issues are not within the scope of this work—we are instead tackling the issue of how to provide better ways for researchers to retrieve the preserved data.

The underlying semantics of research datasets must be captured in a set of relevant metadata that can help researchers who wish to reuse those datasets (“re-users”) to understand their original context of production. If a dataset is not sufficiently well annotated, it may be difficult for the potential re-user to evaluate that data in terms of its relevance to their work or verify its authenticity and correctness—which would most certainly mean that such data would not be reused at all.

Managing research data at a research institution requires knowledge on many different research fields, from ethics to low-level data representation formats. In this chapter, we will start with an overview of the requirements that have been identified.
Related Content

Continuous and Progressive XML Query Processing and its Applications
[www.igi-global.com/chapter/continuous-progressive-xml-query-processing/27782?camid=4v1a](www.igi-global.com/chapter/continuous-progressive-xml-query-processing/27782?camid=4v1a)

A Reuse Definition, Assessment, and Analysis Framework for UML
[www.igi-global.com/chapter/reuse-definition-assessment-analysis-framework/4940?camid=4v1a](www.igi-global.com/chapter/reuse-definition-assessment-analysis-framework/4940?camid=4v1a)

A Framework for Managing Consistency of Evolving UML Models
Tom Mens, Ragnhild Van Der Straeten and Jocelyn Simmonds (2005). *Software Evolution with UML and XML* (pp. 1-30).
[www.igi-global.com/chapter/framework-managing-consistency-evolving-uml/29608?camid=4v1a](www.igi-global.com/chapter/framework-managing-consistency-evolving-uml/29608?camid=4v1a)

UML Modeling Support for Early Reuse Decisions in Component-Based Development
[www.igi-global.com/chapter/uml-modeling-support-early-reuse/30572?camid=4v1a](www.igi-global.com/chapter/uml-modeling-support-early-reuse/30572?camid=4v1a)