DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013

Designing an Application Profile Using Qualified Dublin Core: A Case Study with Fracture Mechanics Datasets

João Aguiar Castro
FEUP & INESC TEC
Universidade do Porto
Portugal
mci11008@fe.up.pt

Cristina Ribeiro
DEI-FEUP & INESC TEC
Universidade do Porto
Portugal
mcr@fe.up.pt

João Rocha da Silva
FEUP & INESC TEC
Universidade do Porto
Portugal
joaorosilva@gmail.com

Abstract

Metadata production for research datasets is not a trivial problem. Standardized descriptors are convenient for interoperability, but each area requires specific descriptors in order to guarantee metadata comprehensiveness and accuracy. In this paper, we report on an ongoing research data management experience at the University of Porto (U. Porto), which led to the proposal of a domain-specific application profile. We presented two curation tools to a group of researchers from mechanical engineering, to help them manage and describe their datasets. After monitoring their interactions with the solutions and analyzing the needs of the group, we were able to select a subset of qualified Dublin Core (DC), as well as a set of complementary descriptors, to capture the main aspects of their experiments. The resulting application profile combines generic, standardized DC descriptors with descriptors from a different experimental standard, and introduces extra domain-specific ones. The profile has been validated by the researchers and is now being used in the description of their datasets.

Keywords: Research data management, application profile, Dublin Core, experimental data, fracture mechanics.

1. Introduction

The importance of research data management is widely recognized, with metadata production playing a central role. While some advocate strict conformance to metadata standards for the sake of interoperability, others argue that, due to the diversity of domains, a "mix-and-match" is the best alternative, as demonstrated by application profiles (Heery and Patel, 2000).

Current initiatives related to data repository services are aligning metadata description best practices with the research data life cycle. ANDS[1], for instance, has adopted the ISO RIF-CS as a data interchange format and DANS[2] recommends Dublin Core and DDI schemas, among others, depending on the research domain. The DataONE[3] and EDINA[4] repositories implemented data management plans that emphasize the need for researchers to document their data from early stages, and also recommend standards. Other well-documented examples include the Dryad Application Profile[5], developed in conformance with Dublin Core (DC) guidelines, and the work led by Robin Rice (2008) applying DC metadata to institutional data repositories at the University of Edinburgh. The need to describe datasets from many different domains has led to the creation of application profiles. These are better at capturing the lifecycle context of a resource but may create artificial barriers for metadata semantics, preventing interdisciplinary research (Willis et al., 2012). For publications, content indexing is effective; but the contents of a dataset may

[1] Australian National Data Service – http://ands.org.au/index.html
[2] Data Archiving and Networked Services - http://www.dans.knaw.nl/en
[3] DataOne - http://www.dataone.org/
[4] EDINA - http://edina.ac.uk/
[5] Dryad Application Profile - http://wiki.datadryad.org/Metadata_Profile

https://doi.org/10.23106/dcmi.952136244

provide no clue as to its nature. Metadata is therefore especially important for datasets to be indexed and retrieved.

It is recognized that the data creation process varies from domain to domain; in the specific case of experimental research datasets, this process should be reproducible, i.e. others should be able to replicate the data given a reasonable description of the experimental conditions. Keeping record of the methods and instruments used to capture data gives the experiment context to those willing to verify results—otherwise, the data re-user is left to obtain it from the data creators, which is time-consuming for both parties.

The amount of effort required to describe data is a limitation to researchers' willingness to share them (Van House, 2003, quoted by Akmon, 2011). Thus, we believe that, instead of making researchers adhere to strict metadata schemas, they should have access to an application profile tailored to their own domain. As defined by Heery and Patel (2000), an application profile is a set of data elements optimized for an application and drawn from one or more schemas combined together by implementers. By taking elements from existing metadata schemas, we keep a certain degree of interoperability in the dataset records, allowing research communities to share digital materials efficiently (Wira-Alam et al.,2012).

In this paper, we describe the first steps of the deployment of a set of data curation tools with a group of researchers working on fracture mechanics. The goal was to assist them in the management and description of their datasets, and to use the conclusions gathered from the study to improve our curation tools. After considering the needs of the group, we have also designed an application profile especially for this domain. The profile combines generic, qualified DC descriptors with those from another metadata schema (EML), as well as a set of new ones, designed to capture the main aspects of their experiments in particular. The new *FM* profile has been validated by the researchers and will be applied to their datasets.

2. Research Data Management Workflow

To assess researchers´ needs with respect to the management of their data, a data audit experiment conducted in 2012 at the U. Porto gathered research data management requirements, as well as a sample of datasets from diverse domains (Ribeiro and Fernandes, 2011). Some datasets have been used to test a prototype data repository built as an extension to DSpace, a widely used repository software platform. The datasets were described using elements from several existing metadata schemas at the dataset level and using Dublin Core at the DSpace *Item* level (Rocha da Silva et al., 2012).

The repository is currently in a prototype phase, as we are trying to encourage researchers to deposit their data and make it visible in the community. U. Porto already provides repository related services, such as the *Open Repository* for exposing or sharing publications, but data resulting from the various steps in the research process are not yet present. The *Data Repository*, currently a prototype, aims to provide a multidisciplinary platform to store, preserve and give access to research-related datasets.

The university repositories are designed to preserve artifacts after they have been completed. For data management, however, it is widely recognized that efforts should start early, ideally as the datasets are created (Tonge and Morgan, 2008). To achieve this dynamic data management environment, we have designed and implemented two closely integrated data management tools, to support the daily data management activities of a research group. *UPBox*, a data management platform with an interface similar to Dropbox, aims to simplify data storage while keeping the data fully under the control of the research group (Barbosa, 2013). *DataNotes* is a semantic wiki designed to support the annotation process (Gouveia, 2013). These two platforms are expected to provide greater researcher autonomy in the management and description without dispensing with the support of a data curator.

A typical use case scenario for these platforms is a researcher creating a *Project* in DataNotes and uploading a file for a dataset. The researcher then proceeds to describe it in a collaborative

https://doi.org/10.23106/dcmi.952136244

interaction via the DataNotes web application. Projects have a folder structure and files are annotated by the original submitter or other researchers associated to the project, making the new data easy to interpret by other team members. To facilitate the description process, it is interesting to have the service provide an application profile or a standard metadata scheme (Gouveia, 2013).

3. Data Management In The Research Group— Current Practices And Requirements

The two data management applications were introduced to a mechanical engineering research group working in fracture experiments at Faculdade de Engenharia da Universidade do Porto (FEUP) to assist them in the management of the datasets they produce. After the interaction that followed, we suggested a set of elements from established metadata standards and provided some insight in data curation and management, making the researchers familiar with the concepts of "metadata" and "descriptors".

Our first approach was to gather information about the research group's current data management practices. A script was designed to support our analysis, adapted from the Data Curation Profile Toolkit[6]. The researchers explained the details of their fracture mechanics experiments as well as their data collection procedures. The experiments are cantilever fracture tests, wherein a force is applied to a sample of the material being studied; the force is increased up to the point where the material fractures. The evolution of the force and the corresponding cantilever displacement is measured by specialized equipment. According to the head researcher, data collection may be divided in two phases—first, data is captured by proprietary software that produces an Excel spreadsheet and then the data is processed through domain-specific analytical methods, where the force and displacement measurements are converted into energy values. The group includes researchers from U.Porto and U. Trás-os-Montes e Alto Douro (UTAD). For each stage, data is shared among the members of the research group via file sharing platforms, but without an established procedure. There is therefore the need for inter-university data sharing in a controlled manner, meaning that datasets must initially be accessible only to group members.

One of the main points that we assessed was the small amount of detailed information associated to each dataset. The head researcher explained that any new researchers joining the group are usually already prepared to interpret the datasets, so there was apparently no need to produce detailed descriptions. Also, it was initially stated that the kind of data produced is not hard to understand for people trained in this domain. Processed data, on the other hand, requires more expertise regarding the particular method of collection in order to be fully interpreted, and those methods are only described in the published papers. Despite the small amount of information that is required (according to the researchers) to interpret these datasets, we felt a certain lack of awareness regarding the need for their long-term preservation. The interviewee noted, however, that metadata might be useful for future dataset retrieval within the research group itself only—since presently they believe that the data may not be interesting for third parties. Also, the research group did not use any standardized metadata schema nor is there a mandate by funding agencies or publishers for a formal data management plan.

At first, when confronted with simple and qualified DC elements, the head researcher pointed out that only a few elements were in fact needed to document their data, and that he would be satisfied with a small subset of elements (title; creator; subject; date). Nevertheless he stated that finding a particular document was a major time-consuming activity. This opinion began to change as he became aware of the opportunities brought by more detailed data description, particularly when consider data sharing and retrieval within the research group.

[6] See http://datacurationprofiles.org/

https://doi.org/10.23106/dcmi.952136244

4. Application Profile for Fracture Experiments

To make the data retrievable, two steps are needed when adding metadata: ensuring interoperability (by incorporating simple or qualified DC) and adding domain-specific descriptors to meet specific research data management needs. To maintain dataset record interoperability, the latter might be ignored by automated harvesting solutions, or transposed to the *dc:description* element during automated harvesting, for example.

The proposed Fracture Mechanics application profile (shown in Table 2 with the "fm" prefix) was built to satisfy some fundamental requirements for the documentation of research data described by Willis et al. (2012). It was designed to be *comprehensive* in the sense that at this time, it provides *all the necessary descriptors* for a research group from this domain to describe their datasets—the profile is subject to continuous improvement, so this may obviously change. It is also *simple*, enabling users without sophisticated data management skills to describe their datasets. Finally, it promotes *data interchange* among the working group, enhancing *data* documentation as well the *discovery and retrieval* of these datasets for later reuse. This application profile combines simple and qualified DC terms (see Table 1), along with a few elements from the Ecological Metadata Language (EML) schema[7]. It provides research context that, in the particular case of experimental science, should include the methods or instruments used to produce the data (Michener, 2006, Willis et. al., 2012).

TABLE 1 – Dublin Core terms used in the Fracture Mechanics application profile

DC		Qualified DC
dc:title (required)	dc:date (required)	dcterms:references (if available)
dc:subject (required)	dc:identifier (if available)	dcterms:isReferencedBy(if available)
dc:description (required)		dcterms:format:medium (recommended)

Domain-specific elements for fracture mechanics experiments were designed in order to fully describe the experiments, while still being applicable in similar domains. The term «specimen», for example, was selected in detriment of a «sample» term because this research group was more comfortable with the concept of specimen. We then designed elements to cover all the properties related to the specimen—«height», «width», «length» and «initial crack length». The «specimenProperties» element allows researchers to complement dataset descriptions with additional and complex information that does not lie within the scope of any other element (thus the need for its type to be free-text). The application profile also includes elements that describe the conditions that can influence the experiment results, namely the ambient «temperature» and «moisture» as well the «testVelocity». We recommended that all these specific elements be filled in when describing a dataset from this domain.

TABLE 2- Fracture Mechanics application profile

Descriptor	Description	Example
eml:methods	Procedures that are used in the creation or the subsequent processing of a dataset	Free text
eml: instrumentation	Description of the instruments used in the data collection or quality control and quality assurance	INSTROM-1125
fm:specimen	Type of specimen used in the experiment	Pinuspinaster (Wood)
fm:specimentLenght	Speciment geometric length	L= 400 mm

[7] http://knb.ecoinformatics.org/software/eml/eml-2.1.1/eml-methods.html

https://doi.org/10.23106/dcmi.952136244

fm:specimenWidth	Specimen geometric width	B=20mm
fm:specimenHeight	Specimen geometric height	2h = 20 mm
fm:specimenInitialCrackLenght	The crack in the double cantilever beam specimen prior to the fracture test	ao= 150mm
fm:specimenProperties	Specimen specific properties	Free text
fm:temperature	The ambient temperature of the experiment location	18ºC
fm:moisture	The moisture percentage at the experiment location.	55
fm:testVelocity	Velocity at which the sampling machine pressed in the sample during the experiment	3mm/m

5. Conclusions and Future Work

To promote the deposit of research data, we have developed dynamic, lightweight tools that favour the initial management of datasets by research groups at U. Porto. The goal is to have researchers take an active part both in the organization of project data and on their description. To perform a preliminary evaluation of the tools, we have established a cooperation effort with a fracture mechanics group from our University and worked together to settle on an application profile for capturing the relevant metadata. The application profile has been validated by the domain experts and is ready for application to recently collected datasets. Preliminary usage experiments within the research group demonstrate improvements in the data management workflow through the use of the profile, paving the way for its use by groups working in the same domain.

We expect that the continuation of this work, possibly with groups in different domains, will provide further insight on the data management practices and help to improve the management tools. This experiment is showing a clear improvement in the awareness of the value of well-described research data. As more and more researchers start to describe their datasets, their transition to the public data repository will become easier and more widespread, hopefully providing a rich view of the existing data. At the same time, the growth of the data repository may motivate other researchers to value data curation and the increased visibility gained with data citation.

Acknowledgements

We would like to express our thanks to the Mechanical Engineering research group (namely Prof. Marcelo Moura and Eng. NunoDourado) for their availability, patience and interest in cooperating with this work. João Rocha da Silva is supported by Ph.D. grant SFRH/BD/77092/2011, provided by the FCT (Fundaçãopara a Ciência e Tecnologia). This work is financed by ERDF – European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and 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»

References

Akmon, Dharma, Ann Zimmerman, Morgan Daniels and Margaret Hedstrom (2011). The Application of archival concepts to a data-intensive environment: Working with scientists to understand data management and preservation needs. Archival Science, 11, 3: 329-348.

Barbosa, José (2013). UPBox: ArmazenamentonaNuvempara Dados de Investigação da U.Porto (Master´s Thesis). Faculdade de Engenharia da Universidadedo Porto, Portugal.

Gouveia, Mariana (2013). DataNotes – um sistemacolaborativoparaanotação de estruturas de directórios (Master´s Thesis). Faculdade de Engenhariado Porto, Portugal.

https://doi.org/10.23106/dcmi.952136244

Heery, Rachel and Manjula Patel (2000). Application Profiles: Mixing and Matching Metadata Schemas. Ariadne 25, September 2000. Retrieved March, 2013 from http://www.ariadne.ac.uk/issue25/app-profiles.

Michener, W.K, (2006). Meta-information concepts for ecological data management. Ecological Informatics, 1(1), 3-7

Retrieved March, 2013 fromhttp://www.sciencedirect.com/science/article/pii/S157495410500004X.

Ribeiro, Cristina and Maria Eugénia Matos Fernandes (2011). Data Curation at U. Porto: Identifying current practices across disciplianarydomais. IASSIST Quarterly, Winter 2011: 14-17.

Rice, Robin (2008). Applying DC to Institutional Data Repositories. Proceedings of the International Conference on Dublin Core and Metadata Applications, North America, 2008. Retrieved March 2013 from http://dcpapers.dublincore.org/pubs/article/view/945.

Rocha da Silva, João, Cristina Ribeiro, JoãoCorreia Lopes (2012). Managing multidisciplinary research data: Extending Dspace to enable long-term preservation of tabular datasets. iPRES 2012 Conference,Toronto, Canada

Tonge, Alan, and Peter Morgen (2008). SPECTRa-T Final Report July 2008. Retrieved March 2013 from http://ie-repository.jisc.ac.uk/387/1/SPECTRa-T_Final_Oct08.doc.

Willis, Craig, Jane Greenberg, Holly White (2012). Analysis and Synthesis of Metadata Goals.Journal of the American Society for Information Science and Technology 63, 8: 1505-152.

Wira-Alam. Andias, DimitarDimitrov, Wolfgang Zenk-Möltgen (2012) Extending Basic Dublin Core for an Open Research Data Archive.Proceedings of the International Conference on Dublin Core and Metadata Applications, North America, 2012, 56-61. Retrieved March 2013 from http://dcpapers.dublincore.org/pubs/article/view/3664.

https://doi.org/10.23106/dcmi.952136244