Flexible metadata models and controlled vocabularies for research data description in multiple domains

Yulia Karimova
ylaleo@gmail.com

INESC TEC, Faculty of Engendering, University of Porto,
Dr. Roberto Frias, 4200-465 Porto, Portugal

Abstract. Research data management is required to enable data sharing and reuse. In this context, data description plays a central role, by providing researchers with sufficient information to interpret their data, yet it is a time consuming task. Therefore, it is important to provide them with appropriate tools that can facilitate the description process. Preliminary work with researchers shows that flexible metadata models help them to create detailed descriptions, making data easier to interpret. Given a metadata model, using controlled vocabularies can facilitate the introduction of descriptor values and improve metadata quality. However, few repositories offer support for flexible, domain-specific metadata and the corresponding controlled vocabularies. Thus, this proposal addresses the creation of metadata models, combining descriptors in multiple domains, and their articulation with controlled vocabularies. The expected results are methods for the development, implementation, and support of metadata models and the corresponding vocabularies, with application in research data repositories.

Keywords: research data management, data repository, controlled vocabularies, metadata models

1 Introduction

Nowadays, the value of the open research data is increasing day by day. According to the Committee on Data for Science and Technology (CODATA)[1], the International Council for Science (ISCU)[2] and the Open Data Institute[3], the potential of open science is clear and contributes to social and economic growth, promoting auditability of results, reuse and transparency [7]. Moreover, the deposit of research data is also required in the grant applications to most funding agencies [7]. In this context, and according to the FAIR principles[4], researchers

[1] http://www.codata.org/
[2] https://www.icsu.org/
[3] https://theodi.org/
[4] https://www.nature.com/articles/sdata201618

should manage their data and make them findable, accessible, interoperable and reusable.

However, research data management (RDM) activities, such as the creation of a Data Management Plan (DMP), the description and the deposit of research data, are time consuming tasks, requiring some knowledge of metadata standards, plus data publication experience and appropriate tools for data description and deposit [6, 16, 1]. These demanding activities may deter researchers from describing and sharing their data, therefore having a negative impact on metadata quality and data reuse. In this context, many institutions are developing tools to facilitate RDM activities, improve metadata quality and motivate researchers to publish their data [17, 22, 9].

In Library and Information Science, controlled vocabularies (CV) have proved useful. They are used for many purposes, for example to ensure coherence and reduce errors in the description of the content of digital objects, in manual and automatic translation, or as language tools in information retrieval [19]. They also help solve problems with synonyms and homonyms in natural language [19, 11] and control the diversity of conceptual and social variations of the terms in the description process [10]. Furthermore, CV can be used to improve information storage and web navigation systems. Thesauri, authority lists and subject heading lists are well-known examples of CV. They typically include preferred or authorized concepts to refer to ideas, people, places, events, and subjects [4].

The introduction of CV in RDM is expected to improve the quality of data description and, as a consequence, the findability and accessibility of data. Also, they can help to establish browsing strategies, serve as a basis for the personalization of resources and be used in the preparation of projects in knowledge and data management [19]. According to the ANSI/NISO z39.19 guidelines[5], CV can be used as the source for allowed terms for a specific metadata element and provide researchers with a list of selected values, based on concepts in use in their domains [5]. In general, well-designed and up-to-date CV can be regarded as part of a "conceptual map".

According Zhang et al. the data repository users are interested in using CV to select, annotate and generate subject terms, but few specific domains and few digital repositories have implemented them [23]. Therefore, more work is required to identify a suitable approach for the development, implementation, and support of metadata models and CV in data repositories, making the case for more efficient RDM.

2 Preliminary work

There are many initiatives aimed at improving the RDM processes [12]. The Metadata Standards Catalog Working Group of the Research Data Alliance[6], for example, developed the Metadata Standards Catalog[7], that includes metadata

[5] https://www.niso.org/publications/ansiniso-z3919-2005-r2010
[6] https://www.rd-alliance.org/
[7] http://rd-alliance.github.io/metadata-directory/standards/

standards for several specific domains, such as Social and Behavioral Sciences, and generic ones such as Dublin Core and CERIF. However, the complexity of standards and the lack of specific descriptors for many domains in the long tail of science complicates their use and makes researchers less willing to share their data [21, 14].

Recognizing these challenges for RDM, the TAIL project at the University of Porto, on which this doctoral proposal originates, is collaborating with researchers from different domains to analyze their needs and the challenges they face in data publication. It is essential to motivate researchers to publish their data and to develop tools to help them organize the data from the beginning of a project [15]. The Dendro platform [8] is one of these tools, that can be used to help researchers prepare and describe their data [18].

In order to help researchers create comprehensive metadata records and make their data easier to interpret, metadata models were created and added to Dendro platform for several scientific domains, such as Hydrogen Production and Biodiversity. Descriptors from widely adopted metadata standards such as Dublin Core were also added [2, 3]. This approach gives researchers flexibility in choosing descriptors and addresses the needs of different domains, making data more visible and reusable in the scientific community [13]. Moreover, during the description experiments they identified the convenience of controlled vocabularies, which help to improve the metadata quality and facilitate the description process in general [3, 8].

As a first experiment with controlled vocabularies, they were elaborated for the Hydrogen Production domain. Operationally, they were implemented on the Dendro platform, embedded as annotation properties in the corresponding ontology. This approach required no structural changes in Dendro, as CV appear in the interface as drop-down lists, allowing researchers to display the options and select one. To analyze the quality of metadata created with and without the use of CV, a series of data description experiments with researchers was carried out. The results showed that the CV can improve the quality of the descriptions and make the description task more appealing to researchers. Moreover, it has facilitated the description process and decreased the syntactic and semantic errors [8].

Despite the small scale of the experiments, there was a visible interest of researchers in using CV for data description. Therefore, more work is required to identify suitable approaches for the development, implementation, and support of CV in digital repositories.

3 Thesis proposal

My proposal is to integrate the design of CV in the development of metadata models. The following research questions provide the framework for the approach and the expected results:

[8] https://github.com/feup-infolab/dendro

1. Is it possible to design metadata models that can be used in multiple domains, and get a balance between descriptors from well-accepted standards and the needs of researchers from specific communities? If so, how can we design them?

2. Given a metadata model in some domain, and examples of its use, how can we identify the descriptors where a CV can be effective?

3. For a given descriptor, how can we define a CV, combining values obtained from existing vocabularies with values generated by researchers?

4. For a given data repository, how can a CV be incorporated, providing support for user interfaces and documentation for datasets?

While exploring these questions, the work is directed to promote the use of flexible metadata models and CV in multiple domains. Working with researchers, I expect to identify their RDM requirements and needs, encourage them to contribute to the development of metadata models with the corresponding CV, and support the publication of their data.

4 Research methodology

Designing the metadata models for a data repository and designing CV are integrated processes. To respond to the research questions and objectives we will use a participatory-design research method [20] that will involve researchers from different scientific domains in the process of creation of metadata models with CV. Figure 1 depicts the proposed research process. It include seven main steps: Collaboration with researchers (1), Requirements (2), Metadata model and CV (3), Method (4), Limitations (5), Evaluation (6), and Refinement (7).

Following an approach focused on the collaboration with researchers (step 1) [3], we will first get acquainted with their scientific domains. To achieve this, we will interview researchers and analyze their publications, any published data if available, the data repositories and metadata standards used, and any other relevant information related with their domain and research work. As a result of this preparation stage, we collect necessary information to identify the RDM requirements of specific groups (step 2), to design domain-specific metadata models (step 3), when needed, and to incorporate them on data repositories for data description experiments.

We can reuse or adapt existing models or develop new ones in case there are no models for a given domain or for some of the researcher requirements. This stage also provides the opportunity to help researchers in the creation of the data management plans that are required in the grant applications to most funding agencies. We expect this to motivate them to collaborate with us and encourage them to publish their data. The systematic process of accompanying the plans will help us detect new difficulties, propose new solutions, and improve the development of tools, according to their needs.

After the metadata models are implemented in the repository, we will carry out the series of data deposit and description experiments. These activities give us real metadata records with actual values assigned to the descriptors, which

Fig. 1. Research process

are the base material for the development of the CV (step 3). They allow us to identify the descriptors that are candidates to CV. Values for these can be assessed and compared to vocabularies in the corresponding scientific domains to see if the concepts can be reused. The creation of a CV also includes a design strategy, the implementation on a digital repository and evaluation sessions.

To assess the extent to which the controlled vocabularies fit researchers expectations, they will be invited to participate in another round of data description, after the implementation of the metadata models with the corresponding CV. This requires a separate set of participants from those in the first experiments, within the same scientific domain. At this point, the methods to design metadata models with CV in multiple domains can be systematized (step 4), identifying the limitations in their development (step 5).

Finally, the effectiveness of both the metadata models and the CV can be evaluated using two approaches (step 6): the observations of data reuse in repositories, using indicators such as data statistics, and consultation with domains experts to evaluate if the data are fit for reuse. During the stages of this work, interaction and evaluation templates are created to gather feedback from the researchers, contributing to define the common needs of their domains. This will also allow us to refine the approach used with researchers (step 7).

5 Expected contribution

In Library and Information Science there is significant work in methods for creating CV, using them and assessing their value in several databases and systems [19]. Likewise, they can be used in RDM to improve description accuracy and to ease metadata entry, making the data more easily interpretable and reusable.

Following this line, one of the objectives of this work is the definition of a method for constructing metadata models with CV for data repositories. This also includes the application of the models with CV in multiple domains and the definition of an approach for their reuse. Reusing metadata models and CV increases interoperability between repositories, reduces the development cost and facilitates RDM. The devised methods are expected to contribute to efficient RDM workflows, and to make them compliant with the FAIR principles, which in turn can motivate researchers to share their data and promote good RDM practices.

Acknowledgements

This work is financed by the ERDF European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project TAIL, POCI-01-0145-FEDER-016736. Yulia Karimova is supported by research grant SFRH/BD/136332/2018, provided by the FCT - Fundação para a Ciência e a Tecnologia.

References

[1] Carolyn Bishoff and Lisa Johnston. "Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University." In: *Journal of Librarianship and Scholarly Communication* 3.2 (2015), eP1231. DOI: 10.7710/2162-3309.1231.

[2] João A. Castro, João R. da Silva, and Cristina Ribeiro. "Creating lightweight ontologies for dataset description. Practical applications in a cross-domain research data management workflow". In: *IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. 2014, pp. 313–316. DOI: 10.1109/JCDL.2014.6970185.

[3] João A. Castro et al. "Involving Data Creators in an Ontology-Based Design Process for Metadata Models". In: *Developing Metadata Application Profiles*. IGI Global, 2017, pp. 181–214. DOI: 10.4018/978-1-5225-2221-8.ch008.

[4] Patricia Harpring. *Introduction to controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*. 2010, p. 259.

[5] Heather Hedden. "Taxonomies and Controlled Vocabularies Best Practices for Metadata". In: *Journal of Digital Asset Management* 6.5 (2010), pp. 279–284. DOI: 10.1057/dam.2010.29.

[6] P. Bryan Heidorn. "Shedding Light on the Dark Data in the Long Tail of Science". In: *Library Trends, Project MUSE* 57.2 (2008), pp. 280–299. DOI: 10.1353/lib.0.0036.

[7] Cynthia R. Hudson Vitale. "The Current State of Meta-Repositories for Data". In: *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository*. Ed. by Lisa R. Johnston. Association of College and Research Libraries, 2017, pp. 251–261. ISBN: 9780838988589.

[8] Yulia Karimova. *Vocabulários controlados na descrição de dados de investigação no Dendro*. Universidade do Porto, Faculdade de Engenharia. 2016. URL: http://hdl.handle.net/10216/85221.

[9] Dong J. Lee and Besiki Stvilia. "Practices of research data curation in institutional repositories: A qualitative view from repository staff". In: *PLoS ONE* 12.3 (2017), pp. 1–44. DOI: 10.1371/journal.pone.0173987.

[10] Kuang-Hwei Lee-Smeltzer. "Finding the needle: controlled vocabularies, resource discovery, and Dublin Core". In: *Library Collections, Acquisitions, and Technical Services* 24:2 (2000), pp. 205–215.

[11] Manikya Rao Muddamalle. "Natural language versus controlled vocabulary in information retrieval: A case study in soil mechanics". In: *Journal of the Association for Information Science and Technology* 49.10 (1998), pp. 881–887.

[12] Peter Murray-Rust. "Open Data in Science". In: *Serials Review* 34:1 (2008), pp. 52–64. DOI: 10.1080/00987913.2008.10765152.

[13] Jian Qin, Alex Ball, and Jane Greenberg. "Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data". In: *Proceedings of the International Conference on Dublin Core and Metadata Applications*. 2012, pp. 62–71.

[14] Jian Qin and Kai Li. "How Portable Are the Metadata Standards for Scientific Data? A Proposal for a Metadata Infrastructure". In: *Proceedings of the Internacional Conference on Dublin Core and Metadata Applications*. 2013, pp. 25–34.

[15] Cristina Ribeiro et al. " Research Data Management Tools and Workflows: Experimental Work at the University of Porto". In: *IASSIST Quarterly* 2.42 (2018).

[16] Djoko S. Sayogo and Theresa A. Pardo. "Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data". In: *Government Information Quarterly* 30 (2013), pp. 19–31. DOI: 10.1016/j.giq.2012.06.011.

[17] João R. da Silva, Cristina Ribeiro, and João C. Lopes. "The Dendro Research Data Management Platform: Applying Ontologies to Long-Term Preservation in a Collaborative Environment". In: *Proceedings of the 11th International Conference on Digital Preservation, iPRES*. 2014.

[18] João R. da Silva et al. "Dendro: Collaborative Research Data Management Built on Linked Open Data". In: *The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014. Lecture Notes in Computer Science*. Vol. 8798. Springer, Cham, 2014, pp. 483–487. DOI: 10.1007/978-3-319-11955-7_71.

[19] Elaine Svenonius. "Design of Controlled Vocabularies in the Context of Emerging Technologies". In: *Encyclopedia of Library and Information Science* (2003).

[20] Jakob Trischler et al. "The Value of Codesign: The Effect of Customer Involvement in Service Design Teams". In: *Journal of Service Research* 21.1 (2018), pp. 75–100. DOI: 10.1177/1094670517714060.

[21] Craig Willis, Jane Greenberg, and Hollie White. "Analysis and Synthesis of Metadata Goals for Scientific Data". In: *Journal of the Association for Information Science and Technology* 63.8 (2012), pp. 1505–1520. DOI: 10.1002/asi.22683.

[22] Joss Winn. "Open Data and the Academy: An Evaluation of CKAN for Research Data Management". In: *IASSIST 2013* May (2013), pp. 28–31. URL: http://eprints.lincoln.ac.uk/9778.

[23] Yue Zhang et al. "Controlled vocabularies for scientific data: Users and desired functionalities". In: *Proceedings of the Association for Information Science and Technology* 52.1 (2015), pp. 1–8. DOI: 10.1002/pra2.2015.145052010054.