Ontologies for research data description: a design process applied to Vehicle Simulation

João Aguiar Castro[1], Deborah Perrotta[2], Ricardo Carvalho Amorim[1], João Rocha da Silva[1], and Cristina Ribeiro[3]

[1] Faculdade de Engenharia da Universidade do Porto/INESC TEC
{joaoaguiarcastro, ricardo.amorim3, joaorosilva}@gmail.com
[2] Faculdade de Engenharia da Universidade do Porto/LIACC
deborahperrotta@gmail.com
[3] DEI—Faculdade de Engenharia da Universidade do Porto/INESC TEC
mcr@fe.up.pt

Abstract. Data description is an essential part of research data management, and it is easy to argue for the importance of describing data early in the research workflow. Specific metadata schemas are often proposed to support description. Given the diversity of research domains, such schemas are often missing, and when available they may be too generic, too complex or hard to incorporate in a description platform. In this paper we present a method used to design metadata models for research data description as ontologies. Ontologies are gaining acceptance as knowledge representation structures, and we use them here in the scope of the Dendro platform. The ontology design process is illustrated with a case study from Vehicle Simulation. According to the design process, the resulting model was validated by a domain specialist.

Keywords: Metadata Models; Research Data Management; Ontologies; Vehicle Simulation

1 Introduction

As research environments are capturing more and more diverse data, the management of such data becomes more challenging. In the long tail of science [4] where a large number of small research groups are producing a large quantity of heterogeneous data, management structures are more fragile and the problem is aggravated.

Recognizing that the value of research data goes way beyond the purpose of their creation, funding agencies are now issuing mandates that establish data deposit and publication as a requirement for project funding.

However, the support for data curation is not a common practice in most institutions, and researchers tend to store undocumented versions of their data in common storage devices. Research data is thus frequently at risk of being lost, sometimes permanently [10]. Sooner or later researchers have to deal with problems of data integrity, accuracy and accessibility, and therefore the creation of detailed metadata records is highly advisable, namely in the early stages of a project.

In this paper we deal with the problem of data description in small research groups. This is part of an ongoing effort to engage researchers in the management of their data, supporting it on Dendro, a prototype ontology-based data management platform, that offers researchers an environment to organize and document their data right from the beginning of a research project [9]. We elaborate here on the process of modelling domain-specific lightweight ontologies, to be loaded into Dendro as a source of descriptors. This process is instantiated with a case study from the Vehicle Simulation domain.

2 Data management workflow

Research data management is no more of a technological issue as it is a conceptual one. Sophisticated technological infrastructures are proposed every day [1] but the success of data management ultimately depends on the efforts of researchers.

Unlike publications, research data is not self-expressive about their content, thus requiring relevant contextual information in order to be accessed and fully interpreted, making metadata essential for its reuse [11]. However, data description needs are not the same for every scientific domain—datasets are heterogeneous and belong to different research cultures and interests [2]. Therefore, to obtain comprehensive and accurate metadata may prove hard. Our assumption is that research data reuse strongly depends on detailed descriptions that only their creators can provide.

In this context our research data platform, Dendro, aims to address both technical and conceptual data management issues. In Dendro we provide researchers with a collaborative environment to systematically capture timely metadata about the datasets they are creating [9]. Datasets together with metadata records can then be submitted to an external data repository for long-time preservation. We have a strong focus on the definition of domain-specific metadata models, formalized as ontologies, for research groups in the long tail of science, and our recent work includes the definition of those models for the fracture mechanics, analytical chemistry and the biodiversity domain [3,9]. These ontologies add new descriptors to be combined with those proposed by generic ontologies, or standards, that are already loaded in Dendro, such as `Dublin Core`[4], `Friend of a friend`[5] or `CERIF`[6]. Our process depends on the interaction between the data curator and the researchers who are the domain experts. Figure 1 depicts our vision for the process of combining generic and specific metadata models in a research data workflow.

The benefits of adopting metadata schemas for data description are obvious, but their use by researchers with no metadata skills can be a problem [8]. Vocabularies that capture common concepts in the researchers domain will most likely encourage them to document their data, but these are typically not available.

Ontologies are intrinsically incremental semantic representations, and this is their main advantage for supporting metadata records. Furthermore, in a landscape of vocabularies and metadata schemas, ontologies are being largely adopted for the sake of interoperability. They can be developed and refined with the collaboration of specific communities, and are therefore flexible enough to face the challenges of a fast-paced

[4] available at `http://bloody-byte.net/rdf/dc_owl/`
[5] `http://www.foaf-project.org/`
[6] `http://eurocris.org/cerif/main-features-cerif`

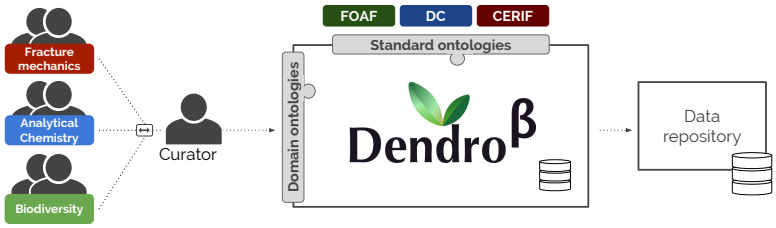

Fig. 1: Ontologies in the research data workflow

research data production system. They combine expressiveness, accuracy and non-ambiguous syntax, for both humans and machines, and these are essential in research data description.

3 The modelling process

In this paper we are proposing an agile ontology design process to face the challenge of data description in a multi-domain research environment. Being aware that every research domain, or experiment configuration, has different data description requirements, we are collaborating with a panel of researchers from several domains at the University of Porto. Our goal is to provide researchers with metadata models suited to their domains, with familiar terminology that mitigates their entry barriers to data description.

The interaction between data curators and researchers is crucial for metadata model design. Data curators acknowledge the value of metadata best practices, and can make good use of their data skills. However, their contribution in the long run can have less impact if researchers are not motivated to collaborate in the overall process. Data curators are not domain experts, or at least not in a wide variety of fields, and have limited know-how in experimental set-ups. On the other hand, researchers are domain experts and, as data creators, they are most apt to produce accurate metadata. Moreover, ideally metadata should be registered as soon as possible in the research workflow, where it is more likely that researchers hold full knowledge of the research context, otherwise the result can be lackluster descriptions [6].

A first moment in our process is a meeting between the data curators and the domain experts. This meeting consists in a preliminary interview supported by a script, adapted from the Data Curation Toolkit[7]. The interview provides insight of the current data management practices at the research team, as it informs on the way the research group organizes and shares their data, whether they are following standards to document the data, among other data management activities. This is also the opportunity to introduce them to research data management concepts, since in many cases researchers are not familiar with concepts like metadata or descriptor.

A second moment includes a content analysis based on researchers' publications, to manually extract the main operational domain concepts. The criteria for selecting these concepts is based on the many parameters at the core of a given research configuration, for instance collected materials, spatial and temporal variables.

[7] available at `http://datacurationprofiles.org/`

After a selection of possible descriptors we take another session to propose a set of domain-specific descriptors, and we also ask researchers to think about what contextual information must be provided to help retrieve and interpret their datasets. Finally the researchers are invited to validate the metadata model. Our interaction with the researchers takes a total of three sessions, with durations ranging between one and two hours.

The metadata model is then formalized as a lightweight ontology, that uses properties from standard vocabularies, if available, and otherwise purpose-built ones. To link together the domain-specific ontologies, we have defined a generic *Research* ontology, that models broad multi-domain concepts. This ontology comprises few classes that represent research types (such as `Experiment`, `Simulation`), and generic scientific properties like the `instrumentation`, `software` or `method` applied to data capture.

When creating a lightweight ontology for a concrete scenario, one can subclass `Experiment` with a specific type of experiment, using it as an extension point from which domain-specific properties can be devised. Our lightweight ontologies are then loaded into Dendro. Our process was already fully explored in two research domains– fracture materials and analytical chemistry experiments [3]. The next section presents a case study, in the Vehicle Simulation domain, to illustrate the use of the proposed process in a systematic way.

4 The Vehicle Simulation case study

At the time of the interview the Vehicle Simulation research group was dealing with data concerning the performance of electric buses in an urban context. In order to evaluate this performance the research group uses datasets containing the bus routes, files containing technical vehicle properties provided by the manufacturer and specific environmental information such as the air coefficient or the surface roughness. In the laboratorial context the data are loaded to run a simulation as close as possible to reality. As a consequence of this simulation new datasets are created, and those are liable to different interpretations and can be analyzed, or reused, according to any particular research goals. According to the researchers, data is mainly organized as Excel spreadsheets. When new external data arrives it is stored via Dropbox and personal e-mail to keep track of the new entries, while regular backups are maintained. The research group does not describe their data, although the simulation variables are part of a "ReadMe" file.

A mathematical model is in place to calculate specific electric bus performance parameters [7]. This model includes several subsystems; one computes the required energy for a vehicle to complete a driving cycle, another uses the kinetic energy of the vehicle to calculate the possible amount of energy that can be recovered from the regenerative braking. Other subsystems are related to the batteries and supercapacitors and evaluate if these are capable of absorbing the energy from the braking.

There are high-level entities that are essential to contextualize the electrical bus simulation set-up, like the vehicle itself, and the driving cycle from which all the vehicle calculations are based [5]. Both the tractive force, that compels the vehicle forward, and the kinetic energy have a great influence on the way the vehicle behaves.

Figure 2 shows the Vehicle Simulation lightweight ontology that domain experts can use to create their metadata records. This ontology uses properties related to the high-level entities. For instance we create the properties `vehicle` (corresponding to a vehicle category, like "electric bus", or other depending on the study) and `vehicleModel`,

so researchers can record the vehicle used in the simulation. Since there are available driving cycles, produced by different organizations, ready to be used in vehicle simulations, the `drivingCycle` property was also defined. These are properties with the potential to create access points to the dataset, as they can yield information that distinguishes a dataset from the rest. All the other properties deal with a set of variables that constrain the entire simulation and are tied to the calculation of the tractive force and of the kinetic energy. Values concerning the `aerodynamicDragCoefficient`, the `roadSurfaceCoefficient` or other variables are associated to many performance parameters, and therefore must be annotated to help one interpret, or reproduce, the output from a vehicle simulation.

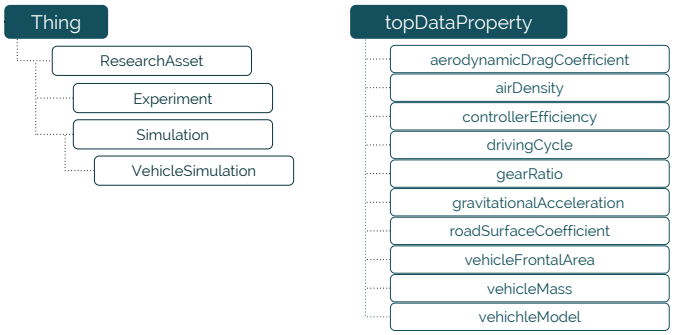

Fig. 2: Vehicle Simulation lightweight ontology

This ontology is not intended to fully represent the vehicle simulation domain, instead it captures the description needs of a group of researchers that run simulations to evaluate specific parameters on electrical bus performance. However, most of the properties defined are generic and can be widely applied to other vehicle simulation scenarios. If the researchers need to provide extra contextual information, namely the description of batteries and supercapacitors, the corresponding properties can easily be added to the ontology. By supporting our process in ontologies we are making sure that our approach is incremental and can easily account for any description needs.

5 Conclusions

In this paper we have presented a systematic process to design lightweight ontologies for the description of research data. We consider that metadata models must result from the collaboration of data curators and domain experts. The Vehicle Simulation ontology presented here served as an instantiation of our process, and was loaded to Dendro, together with those for other 10 domains. The experience of researchers describing their data with Dendro is currently running, as a part of an ongoing research on descriptor recommendation. We expect to show that the timely documentation of datasets will result in more data reaching the final stages of the research workflow and being reused by a broader community.

6 Acknowledgements

Project SIBILA-Towards Smart Interacting Blocks that Improve Learned Advice, reference NORTE-07-0124-FEDER000059, funded by the North Portugal Regional Operational Programme (ON.2–O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT). Deborah Perrotta is supported by FCT through PhD scholarship grant SFRH/BD/51256/2010, within the MIT-Portugal Program in Engineering Design and Advanced Manufacturing – Leaders for Technical Industries focus area. João Rocha da Silva is supported by research grant SFRH/BD/77092/2011,provided by the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

References

1. Amorim, R.C., Castro, J.A., Rocha da Silva, J., Ribeiro, C.: A Comparative Study of Platforms for Research Data Management: Interoperability, Metadata Capabilities and Integration Potential. In: New Contributions in Information Systems and Technologies, pp. 101–111. Springer International Publishing (2015)
2. Borgman, C.L.: The conundrum of sharing research data. Journal of the American Society for Information Science and Technology 63(6) (2012)
3. Castro, J.A., Ribeiro, C., Rocha da Silva, J.: 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. 0–3 (2014)
4. Heidorn, P.B.: Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2), 280–299 (2008)
5. Malcolm A. Weiss, John B. Heywood, E.M.D., Andreas Schafer, AuYeung, F.F.: On the road in 2020 - A life-cycle analysis of new automobile technologies. Energy Laboratory Report EL 00-003(October), 3–6 to 3–14 (2000)
6. Martinez-Uribe, L., Macdonald, S.: User engagement in research data curation. In: Research and Advanced Technology for Digital Libraries, 13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Proceedings. pp. 309–314 (2009)
7. Perrotta, D., Macedo, J.L., Rossetti, R.J., Sousa, J.F.D., Kokkinogenis, Z., Ribeiro, B., Afonso, J.: Route Planning for Electric Buses: A Case Study in Oporto. Procedia - Social and Behavioral Sciences 111, 1004–1014 (Feb 2014)
8. Qin, J., LI, K.: 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. pp. 25–34 (2013)
9. Rocha da Silva, J., Castro, J.A., Ribeiro, C., Lopes, J.C.: The Dendro research data management platform: Applying ontologies to long-term preservation in a collaborative environment. In: Proceedings of the iPres 2014 Conference (2014)
10. Smit, E., Van Der Hoeven, J., Giaretta, D.: Avoiding a Digital Dark Age for data: why publishers should care about digital preservation. Learned Publishing 24(1), 35–49 (Jan 2011)
11. Treloar, A., Wilkinson, R.: Rethinking Metadata Creation and Management in a Data-Driven Research World. 2008 IEEE Fourth International Conference on eScience pp. 782–789 (2008)