Research Data Management Tools and Workflows: Experimental Work at the University of Porto

Cristina Ribeiro, João Rocha da Silva, João Aguiar Castro, Ricardo Carvalho Amorim, João Correia Lopes, Gabriel David[1]

Abstract

Research datasets include all kinds of objects, from web pages to sensor data, and originate in every domain. Concerns with data generated in large projects and well-funded research areas are centered on their exploration and analysis. For data in the long tail, the main issues are still how to get data visible, satisfactorily described, preserved, and searchable.

Our work aims to promote data publication in research institutions, considering that researchers are the core stakeholders and need straightforward workflows, and that multi-disciplinary tools can be designed and adapted to specific areas with a reasonable effort. For small groups with interesting datasets but not much time or funding for data curation, we have to focus on engaging researchers in the process of preparing data for publication, while providing them with measurable outputs. In larger groups, solutions have to be customized to satisfy the requirements of more specific research contexts.

We describe our experience at the University of Porto in two lines of enquiry. For the work with long-tail groups we propose general-purpose tools for data description and the interface to multi-disciplinary data repositories. For areas with larger projects and more specific requirements, namely wind infrastructure, sensor data from concrete structures and marine data, we define specialized workflows. In both cases, we present a preliminary evaluation of results and an estimate of the kind of effort required to keep the proposed infrastructures running.

The tools available to researchers can be decisive for their commitment. We focus on data preparation, namely on dataset organization and metadata creation. For groups in the long tail, we propose Dendro, an open-source research data management platform, and explore automatic metadata creation with LabTablet, an electronic laboratory notebook. For groups demanding a domain-specific approach, our analysis has resulted in the development of models and applications to organize the data and support some of their use cases. Overall, we have adopted ontologies for metadata modeling, keeping in sight metadata dissemination as Linked Open Data.

Keywords

Research workflows, research data management, data publication, metadata, e-science

Introduction

Research data are created and used in diverse contexts. They may be generated specifically for a research project, such as sensor data captured in some experiment or interview data from a survey. They include data that are systematically captured for some purpose and then also used in research, such as meteorological data or the security logs of a computational facility. They may be ordinary data, such as a set of web pages, collected ad hoc to assess performance of a search engine. This

 Ribeiro, Cristina; da Silva, João Rocha; Castro, João Aguiar; Amorim, Ricardo Carvalho; Lopes, João Correia and David, Gabriel (2018) Research Data Management Tools and Workflows: Experimental Work at the University of Porto, *IASSIST Quarterly* 42 (2), pp. 1-17. DOI: https://doi.org/10.29173/iq925

diversity makes research data management (RDM) a rather elusive task, for which researchers have neither well-established processes nor a clear intuition for its usefulness.

Data generated in large projects within well-funded research areas are already being curated in disciplinary infrastructures. The NCBI (NCBI Resource Coordinators, 2013) in the life sciences and ICPSR (Doty et al., 2015) in social sciences are good examples of stable infrastructures supporting large communities and used by researchers to get base data and to contribute new research outcomes.

For data on the so-called long tail of science, there are still no general-purpose solutions and even when researchers recognize the value of data management, they typically have no support for making data visible, satisfactorily described, deposited, and searchable. Efforts in this area are currently at the project stage, as illustrated by two large EU-funded initiatives, OpenAIRE (Manghi et al., 2010) and EUDAT (Lecarpentier et al., 2013).

One important aspect to take into account is the stakeholders in RDM and the conditions that will foster their collaboration (Ribeiro et al., 2015). Figure 1 shows stakeholders and tools together with the steps in the research workflow. We concentrate on the roles of researchers, research managers and curators, in their collaboration with developers to build an effective support for dataset description and publication.

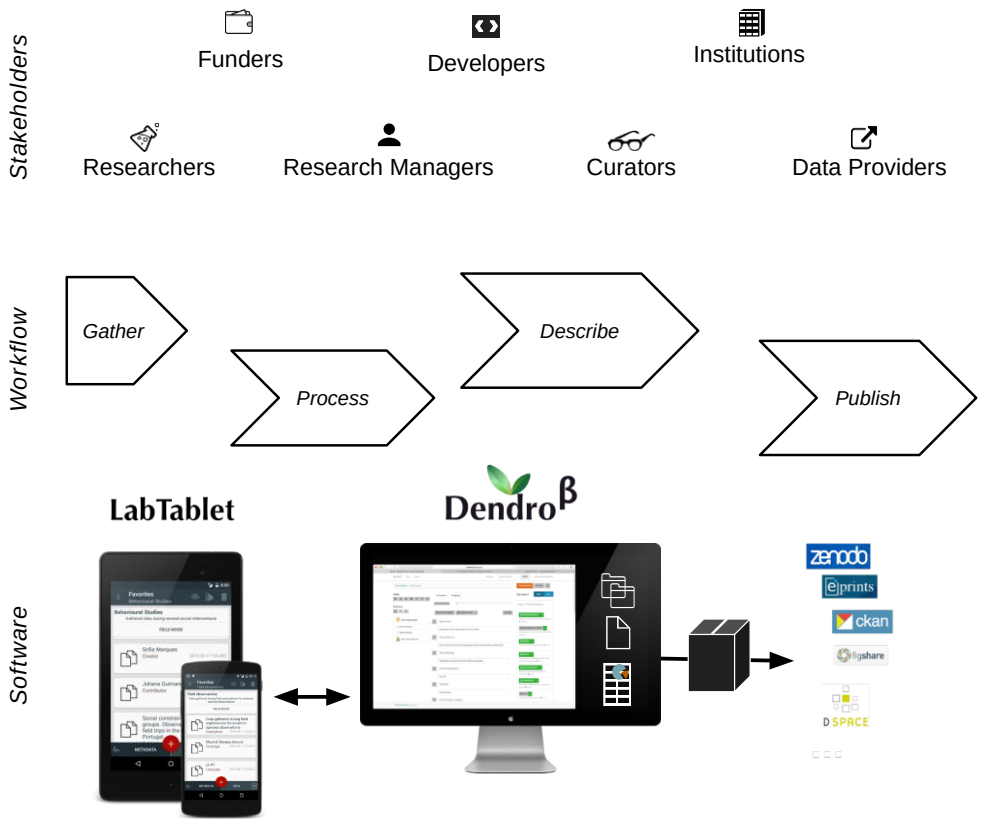

Figure 1 RDM Stakeholders, workflow stages and supporting software

 Ribeiro, Cristina; da Silva, João Rocha; Castro, João Aguiar; Amorim, Ricardo Carvalho; Lopes, João Correia and David, Gabriel (2018) Research Data Management Tools and Workflows: Experimental Work at the University of Porto, *IASSIST Quarterly* 42 (2), pp. 1-17. DOI: https://doi.org/10.29173/iq925

The concern with RDM in the long tail is relatively recent, and several entities provide support for multi-domain datasets (DCC, 2017; ANDS, 2017; DANS, 2017; DataONE, 2017; Dash, 2017). However, in some areas where large datasets are at the core of research there is a longer record of initiatives around research data, such as large databases identifying individual contributions, with a strong connection with publications; NCBI and ICPSR were mentioned before as examples in the life sciences and social sciences, respectively. In other areas, as international groups grow, data sharing becomes more important, and research assessment requires the link to source material, new international infrastructures are being set up. This is illustrated by the projects promoted by ESFRI, which is supporting a network of long-term Research Infrastructures of pan-European interest (European Strategy Forum on Research Infrastructures, 2016).

The work in RDM at the University of Porto started with a scoping study, with the collaboration of 8 research groups, using existing recommendations and covering aspects such as the awareness with respect to data curation, the pressing needs regarding existing datasets, current solutions for data storage, the perceived value of legacy data, and the required support for RDM actions (Ribeiro and Fernandes, 2011). Based on this study, we started to design a workflow for research groups, and the tools to improve the effectiveness of the publication and dissemination of research.

The tools available to researchers are decisive for their commitment. For groups in the long tail, we focus on data preparation, namely on dataset organization and metadata creation. The workflow proposed for the long tail includes Dendro, an open-source research data management platform, and LabTablet, an electronic laboratory notebook for automating metadata creation. For groups demanding a domain-specific approach, our analysis has resulted in the development of models and applications to organize the data and support their primary use cases. Overall, we have adopted ontologies for metadata modeling, keeping in sight metadata dissemination as Linked Open Data.

Research groups and RDM requirements

There is an overall recognition of the need for data access and reuse. Low data reuse can be related to the difficulties faced by researchers in creating rich contextual metadata (Faniel and Yakel, 2011). This may be related to the fact that researchers tend to produce data documentation that is more focused on their own personal needs (Mayernik, 2011).

With respect to data documentation, it is possible to follow a traditional approach, describing data according to the well-established models used in publications, and depositing them in existing repositories, as yet another kind of publication. This is the line followed in initiatives such as the European project OpenAIRE+, promoting the Zenodo repository for gathering and interlinking papers, datasets, software, and the projects where they originate. This approach makes data description a lighter task, but the lack of domain-specific data description can also make datasets harder to find, to understand and to reuse. Researchers interested in a dataset, even if they locate it, will probably need to contact the authors to make sense of the data.

As an alternative, we propose to empower researchers with expressive metadata models to describe their data. This requires tools, workflows and specialized support. Our approach is to define a distributed multi-domain metadata model as part of the curator workflow, assessing RDM requirements at the domain level. For each domain, this includes the identification of domain

concepts, their representation as descriptors, and their incorporation in the Dendro interface, used by researchers to describe their data with comprehensive contextual information (Rocha da Silva et al., 2016).

We have worked with a growing set of research groups from different domains, whom we asked for a moderate commitment to RDM activities. These contacts started with a scoping study (Ribeiro and Fernandes, 2011) sent through the deans of the 15 schools in the University of Porto. People selected by the deans have answered and we conducted 13 interviews with researchers in a broad diversity of areas, following the Data Asset Framework recommendations (DAF, 2017). We asked people to identify the nature and goals of their datasets, introduced them to RDM concepts, selected some relevant datasets for further analysis and evaluated RDM tools in the context of the regular activities of the groups. 8 out of these 13 groups provided sample datasets. 3 groups from this subset provided feedback on the prototype data repository we have built based on DSpace (Rocha da Silva et al., 2012). The contacts with these 3 groups continued to the phase where metadata requirements were assessed. In the meantime, through personal contact and project collaborations, other groups got notice of our work and a panel of 11 groups was available by the time the first version of Dendro was tested. Currently we articulate the more in-depth work with groups, where new metadata models are considered, with a regular collaboration with researchers who need support in the use of standard metadata models, such as Dublin Core, to make datasets part of their research outputs.

Considering the areas where we have committed to the design of metadata models, they range from groups that produce experimental data, such as double cantilever beam experiments, to more computational-intensive scenarios, like vehicle performance simulations (Castro et al., 2014; Castro et al., 2015), or the combinatorial optimization problems in operations research (Toledo et al., 2013) and predictive ecology studies in the biodiversity domain (Rocha da Silva et al., 2014). Table 1 shows the groups that collaborated in data management contacts and their domains.

Table 1. Groups involved in domain-specific metadata identification

Domain	Experiment	Affiliation
Fracture Mechanics	Double Cantilever Beam	Engineering School, Department of Mechanical Engineering
Analytic Chemistry	Pollutant Analysis	Engineering School, Department of Chemical Engineering
Vehicle Simulation	Bus performance in urban route	Engineering School, Department of Informatics Engineering
Hydrogen Production	Hydrogen production via chemical hydrides	Engineering School, Department of Chemical Engineering
Biological Oceanography	Interaction between marine	CIIMAR (Marine Environmental Research) and Federal University Rio Grande (FURG), Brazil

	and estuarine organisms	
Biodiversity	Species dynamics and distribution, biological communities and ecosystems	CIBIO (Research Center in Biodiversity and Genetic Resources)
Social Sciences	Social groups behaviors and beliefs	Psychology School
Computational Fluid Dynamics	Wind studies	Engineering School, Department of Mechanical Engineering
Cutting and Packing		Engineering School, Department of Industrial Management

To engage with these groups we applied different techniques, namely interviews, content analysis and data description experiments. Non-scripted meetings also took place to get as much feedback as possible from the researchers.

Semi-structured interviews are usually the technique applied in the first interaction with the researchers, providing knowledge about the domain, their RDM practices, and the overall research data lifecycle used by the researcher or group. The interview script was adapted from the Data Curation Profile Toolkit and translated into Portuguese. In the first case studies the interview was conducted with little forethought from the subject perspective (fracture mechanics, analytical chemistry). In the remaining cases the interview form was sent by e-mail beforehand, so the researchers had time to read and better prepare their answers, ensuring richer information.

Content analysis on researchers' publications serves as a complement to the interview, since publications are good sources of information regarding data collection and analysis. A methodology section traditionally articulates pieces of information that may contextualize data, like experimental configurations, environmental features, and considering the many input variables; this was particularly interesting in the vehicle simulation domain (Castro et al., 2015). There were exceptions, as with the hydrogen production group, where content analysis was performed before the first meeting to gain some domain knowledge. In the computational fluid dynamics case it was the main technique used for concept identification, since we did not have the researchers available for interview at the time. Here, domain experts were consulted to validate the selected concepts in the first meeting.

All the research groups have participated in a data description experiment using Dendro, where they were invited to create a folder and upload a file, or more, and pick descriptors from the many metadata models previously loaded. At least two researchers from each group were involved, one

using a version of Dendro with all the available descriptors to choose from. The other used a version with a recommendation system where the interaction with the first set of researchers was used to select a better ranking for descriptors (Rocha da Silva, 2016).

As to the non-scripted meetings, some were upstream, while others were downstream, depending on researcher's schedule or workplace proximity. We schedule preliminary meetings to present the objective of the following interactions, and to gather preliminary insight on RDM requirements and researchers' perspectives. With research groups where a more in-depth focus was not possible we run opportunity meetings after the interview took place, or simultaneously. These downstream meetings were useful to discuss and detail the main points from the interview or to specify the information needed to describe the datasets researchers were working on at the time.

Overall, the level of the engagement with the researchers was not uniform, and depended on factors such as their metadata expertise, the relevance of using a different RDM tool depending on the research group, and the time that researchers have to dedicate to RDM activities.

Concerning metadata, while most groups were not familiar with the concept, the biodiversity research group was already working on metadata guidelines from the INSPIRE directive in the context of a running project (Pôças et al., 2014). For this case we were able to map the INSPIRE concepts and develop a suitable domain metadata model (BIOME) (Rocha da Silva et. al, 2014). In other groups, not so familiar with metadata concepts, we took descriptors from metadata standards, such as the Data Documentation Initiative (Vardigan et al., 2008) for the social science domains, and the Ecological Metadata Language (Fegraus and Andelman, 2005) and Darwin Core vocabularies (Wieczorek et al., 2012) for the biological oceanography domain. For others, no ready-to-use vocabularies were identified, and a more in depth collaboration with the domain experts was necessary.

Concerning the use of tools (Dendro and LabTablet, detailed in the next section), while all groups carried out descriptions tasks using Dendro, in some cases it was pertinent to also experiment with LabTablet, depending on the context of the data. A researcher collecting field data or running an experiment in a controlled environment is more likely to use LabTablet than one performing computational studies. Therefore, researchers from the social sciences and biodiversity domains used LabTablet.

The time dedicated to work with each group is a dimension where we have no complete control. The collaboration timeline is difficult to predict and our experience has shown that the same amount of interaction or level of engagement can be accomplished over very different periods of time in different groups. For instance, in one of our cases we had three meetings over a period of three months, while the same number of meetings was performed over two weeks in another case, to achieve similar results. Other researchers were contacted but our interactions were sporadic, or of a more informal nature. Nevertheless, these groups are good complements to the previous ones and help us assemble a more representative set of research domains.

Data organization and metadata creation with Dendro + LabTablet

The uptake of RDM depends on the existence of clear processes for researchers with respect to data collection, organization, description, deposit and publication. Interviews with researchers on their

RDM practices showed a large gap between the processes used in paper preparation and the routines required to organize data and make them publication-ready. It was clear that appropriate tools might ease the task for researchers in small groups; this led us to design, implement and test tools to support data preparation in the long tail. These tools can have a very broad field of application—data collection, data cleaning, processing, organization, description, deposit, search, are some examples. As RDM is evolving as an integral part of research, we expect that repository platforms, either disciplinary or generic, will become mainstream. However, one should not assume that datasets can be managed as publications already are, particularly when it concerns metadata. Research data are often purely numeric—unlike a paper, for example—and emerge from very specific and advanced studies, so the tools for data collection and data processing are likely to be domain dependent. Any tools that attempt to capture the production context of a research dataset with these characteristics must be able to capture domain-specific metadata as well, adopting established metadata standards in the corresponding community whenever possible. Also, repository managers may be experts at metadata and data management, but they are not the domain experts and thus do not know what domain-specific information is needed to retrieve and reuse the data. To respond to these two issues, we have approached data organization and data description with the goal of simplifying the interface between researchers and repository managers. Any tools for preparing and describing datasets should therefore establish a consistent RDM discourse with researchers, while giving them the openness to interface with several repository platforms, so that they can share their data everywhere they want without having to fill in metadata multiple times.

Dendro is an open-source, ontology-based RDM platform currently in development at INESC TEC and the University of Porto, whose dependencies are also entirely open-source. It targets researchers as its main users, and helps them deposit and share data both within their research group and with external elements. Dendro uses the 'Dropbox' metaphor for data upload and adds sophisticated data description capabilities. The concepts in Dendro include 'Project'—a Project is a shared folder where every member can deposit files of any kind, 'Folder'—project members create folders and subfolders to organize resources, and 'Descriptor'—metadata are associated to resources using both domain-specific and generic descriptors. Dendro is integrated in the research workflow to cover the steps between dataset creation and publication, and can export data and metadata, or just the metadata, to major repository platforms when the research group is ready to publish them (Rocha da Silva et. Al, 2018).

Mobile devices have evolved to include advanced capabilities that render them suitable for an array of research-related activities. Numerous cases of researchers using their own devices to improve the research workflow prove that this is an important trend. Besides an increasing storage capacity, mobile devices are often permanently connected to the Internet and have several built-in sensors that can provide contextual data on the researcher's environment effortlessly. This has been the motivation in the development of LabTablet, a notebook application with an emphasis on RDM.

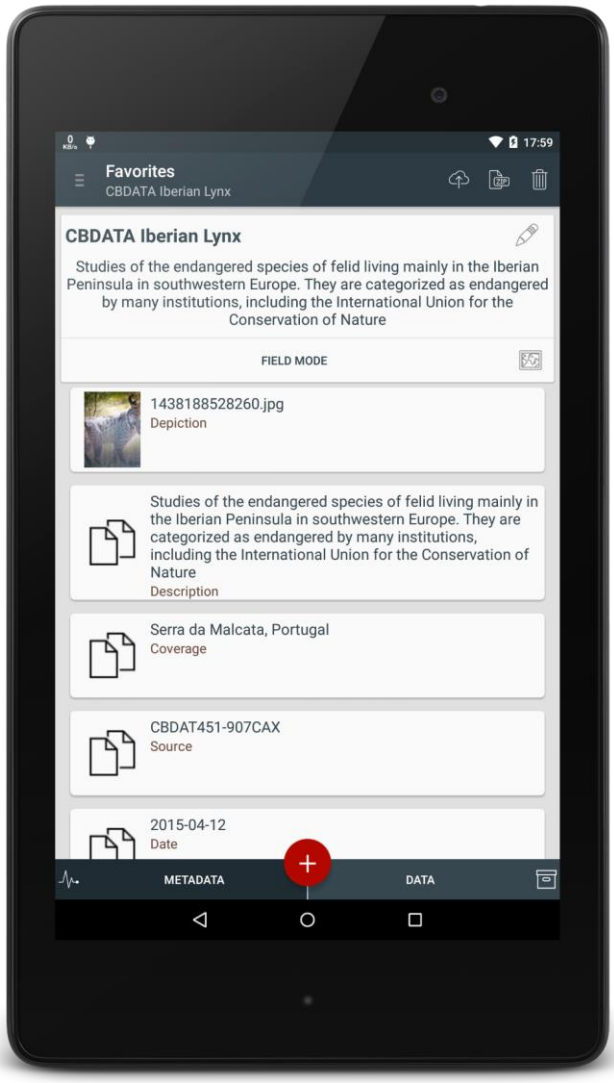

Figure 2: Production of metadata records using LabTablet

LabTablet is an electronic laboratory notebook, i.e. an application that takes advantage of sensors onboard the mobile device to help researchers describe their data. In some cases the description with the mobile device replaces a process that used paper-based notebooks, making sure metadata are not lost but instead recorded, associated to data and deposited. Figure 2 shows the LabTablet interface, in a situation where it is used to collect data in a field experiment. A simple tool like this, integrated into frequently used devices, can contribute to get more metadata associated to datasets. Moreover, with the mobile device, metadata creation becomes a seamless part of data collection, distributing the effort over the duration of the process and avoiding time-consuming task at the end of a project.

The first step in data organization and description workflow using Dendro and LabTablet is illustrated in Figure 3, with a project in the biodiversity domain. After the metadata are synchronized with Dendro, we can see the structure of the project folders and files, the descriptors used for this domain (generic Dublin Core plus a domain-specific subset of the INSPIRE descriptors) and the communication between Dendro and LabTablet. The generic workflow is as follows: metadata models are passed from

Dendro to LabTablet, metadata collection takes place in LabTablet, and descriptor values are added to Dendro when the two platforms synchronize.

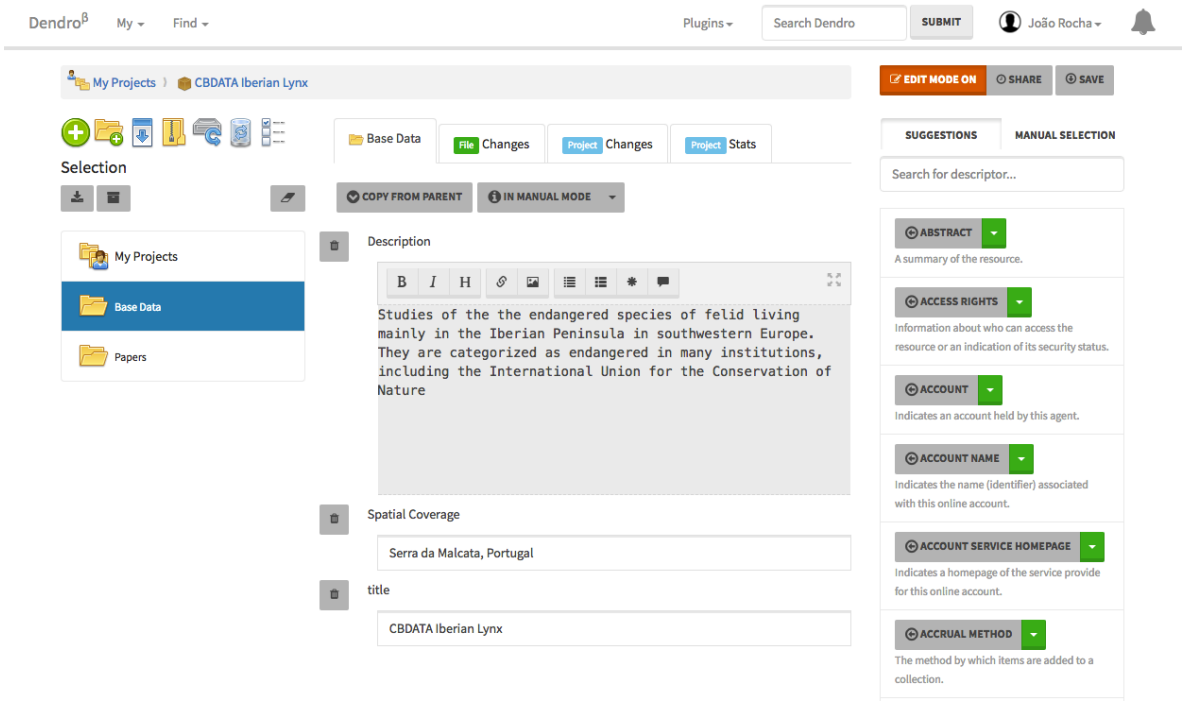

Figure 3: Reviewing and improving metadata records in Dendro

The availability of domain-dependent descriptors is one of the principles in Dendro. As the number of domains grows, so does the number of available ontologies (representing metadata models), making it harder for researchers to locate the more appropriate ones. To cope with this information overload, Dendro also offers descriptor recommendation as an advanced feature, assisting researchers (who are not expected to be research data management experts) in the discovery and selection of descriptors that match their description needs. Manual descriptor selection is still possible, as the platform allows the user to restrict the ontologies from which the user will add descriptors from, controlling cross-domain descriptor usage.

The effectiveness of the Dendro platform has been tested at different times. Three experiments were run in different conditions, and the results are already available for two of them. A fourth experiment is in the final stage of data collection. The first experiment, *DendroDC1*, used a preliminary version of Dendro, after it had been tested with some researchers from our panel. The subjects, a set of students from an Information Science course, were asked to fill in Dublin Core descriptors for datasets published by several organizations. This experiment had two goals: to test Dendro in realistic load conditions and to observe the use of a generic metadata model in the annotation of datasets with descriptor recommendation. The system proved robust in lab conditions and showed some improvement in description when recommendation was turned on (Rocha da Silva et al., 2018). The second experiment, *DendroPanel*, was a more realistic one, and involved the 11 groups in Table 1 and a Dendro instance configured with the ontologies for the domains in case there were specific ones.

 Ribeiro, Cristina; da Silva, João Rocha; Castro, João Aguiar; Amorim, Ricardo Carvalho; Lopes, João Correia and David, Gabriel (2018) Research Data Management Tools and Workflows: Experimental Work at the University of Porto, *IASSIST Quarterly* 42 (2), pp. 1-17. DOI: https://doi.org/10.29173/iq925

For each group (each domain) we had two participants, each describing a dataset from the corresponding domain: one used Dendro with and the other without recommendation. Results were in favor of the use of descriptor ranking, providing a recommendation based on usage data (Rocha da Silva, 2016). The third experiment, *DendroDC2*, replicated *DendroDC1* with a new set of subjects and the results are not analyzed yet. A fourth experiment, currently collecting data, uses a combination of Dublin Core and the Biodiversity ontology in the description of datasets in this domain, to gain insight on the importance of domain-dependent descriptors.

Several tests were also conducted using LabTablet and researchers in Engineering and Biodiversity domains. Some comprehensive experiments with LabTablet are yet to be planned. In the meantime, the application has also been used as the basis for two data-related applications: one for displaying sea conditions for nautic sports (Amorim et al., 2016) and the other to support the collection of specimen data in the SeaBioData project, mentioned in the sequel (SeaBioData, 2017).

Disciplinary solutions: dedicated data and metadata models

As we explore the requirements of research groups in a large research institution, focusing on small groups where data curation needs to be established as an agile process, we were faced with cases that do not fit into the long tail. The solutions for these areas required specific projects supporting the analysis, design and implementation of solutions that satisfy more specific requirements. We now look at three of these projects that ran as separate initiatives, involving teams from the corresponding disciplines, but still have RDM at the core. They provide a view on the different functions covered by a domain-specific system, and extend the focus from organization and description to functions such as data collection, processing and archiving.

Sensor data from concrete structures

The research method in the structural health monitoring, a subfield of civil engineering, is well established. To monitor a bridge, a tower, or other large structure, a research project is set up, the data collection system is designed and deployed on the structure, and data streams with periods of the order of milliseconds start to be produced during months or years. Each sample in a data stream contains values from the various sensors in a specific data acquisition system (temperature, acceleration, etc.). Each stream is segmented in files for transfer and storage purposes. These raw data files are cleaned and the result is subject to one or more processing routines. Both the cleaned data files and the result files are visualized, and conclusions on the health of the structure are incorporated into reports and research papers. After the publications are issued, the data are sometimes discarded.

The growing awareness of the value of data as evidence supporting the published conclusions and as source for further studies motivated the research team to launch a digital archive project tailored to their needs. Our team has elicited the requirements with the strong involvement of the disciplinary experts and has designed an architecture for the digital archive with three components: a systematic directory structure and naming rules organize all the data files in the file system; a database collects all the metadata; and an application adds management and visualization functionalities.

The metadata are organized in five packages:

1) Contextual metadata on the structure characteristics and project details, dates, and funding body;
2) System information on the design, implementation and parameterization of the data acquisition system and its sensors and components;
3) Data file structure, measured variables and units, sampling frequency, timestamp, and file location;
4) Stakeholders like the structure owner and designer, the research team, and external researchers;
5) Documents of any kind, from the project contract and the data acquisition system design to any pictures, as well as published reports and papers.

The project resulted in a Web application, its functionalities including the management of the five packages of metadata, the automatic ingestion of new data files, a graphic visualization tool to browse the data on intervals, and an asynchronous facility to export data files using OAI-PMH.

The system developed (da Costa et al., 2014) is generic enough for monitoring systems in other fields besides structural health monitoring.

Wind data from LiDARs

WindS@UP is a prototype of the e-infrastructure developed in the scope of the WindScanner.eu project and includes a repository for experimental data sets, consisting of georeferenced time series data generated by LiDAR sensors used in open-air field tests (field campaigns). The platform manages the data and corresponding metadata, provides processing capabilities for in-situ data processing and is deployed in the WindScanner Hub (Gomes et al., 2014).

The WindScanner device is a wind (short and long-range) LiDAR with scan-head and control software that, in coordinated operation of three units, can be used to measure 3D wind vector field with high accuracy. This is the core technology of the WindScanner infrastructure to be used by the corresponding ESFRI project. WindScanners are deployed at existing or planned test facilities, covering different climate conditions and terrains. Standard procedures may be applied to collected data and the resulting time-series are stored for further use.

Data from the time series have GPS timestamps used to align values for signals obtained from different devices. Besides raw and processed data, the Winds@UP platform provides storage for metadata and for other resources—research objects—created by researchers in the course of their work. Research objects include datasets, archives, photos, charts and scientific papers.

Metadata for the research objects are organized in three categories:

1) Descriptive—title, author, abstract and keywords, which help discovery through searching and browsing;
2) Administrative—preservation, rights management, and technical aspects such as format or experimental setup; and
3) Structural—how different components of a set of associated data objects relate to one another (e.g. datasets, procedures or results).

The experiments' raw data are transferred from the devices or data-loggers to the campaign site where a quality assurance process validates the received data. Data are also transformed to standard formats, such as the Network Common Data Form (NetCDF)[2]. The quality assurance process may require, depending on the campaign, the processing of raw data to produce "clean data". These data are similar to the raw data—time series—but some portions may be purged due to detected errors, and some pre-processing operations may be performed (e.g. 2 or 10 minutes averages). In addition, further documentation should be added, e.g. detailing the process used to clean data.

The data and associated metadata, packed in self-describing, machine-independent data formats in NetCDF files, are uploaded to the platform repository. At this point, these research objects are described using the same metadata model (captured as an ontology) used for other research objects in the domain. Raw data, clean data and processed data are used together with related research objects by a Web portal with search facilities suitable for researchers, enabling research collaboration and the reproducibility of results.

WindS@UP is designed as an e-infrastructure, providing dedicated in-house storage for a community. This is one of the differences from the long-tail cases, where external repositories are required to store data. The concepts of campaign, experiment, observed phenomena, and time-series data are common to many domains, and so are the generic metadata elements used to capture them.

WindS@UP is currently a solid prototype, which underwent three development cycles and contributed to the preparation phase of an ESFRI research infrastructure. Further development is expected in the context of the next phase of the ESFRI national and European wind research infrastructure (windscanner, 2017). The most recent version of the prototype (windsP) has been used recently to plan the experiments of the European project NEWA (NEWA, 2017) and has followed the execution of the field campaign in Perdigão (Witze, 2017), a double-hill field experiment in Portugal.

Seamounts physical and biological data

We have been approached by a large team on marine research in order to build a digital archive for their data. Marine research is typically organized in campaigns aboard a research ship. A campaign is composed by a set of stations in predefined locations. In a station samples are taken from the soil and the water column using appropriate devices and procedures. The samples may be subject to an on-site study that may be later complemented with physical, chemical and biological analysis, using predefined procedures. The collected data are very diverse in nature, ranging from variables measured in specified units to the identification of the presence or density of biological species, to pictures, video and sound, to actual captured specimens. The georeferencing and the detailed log of the actual sample collection process and involved researchers are very important. At the same time, data streams are received from instruments installed in buoys, from the positioning of ships, and from satellite data out of the campaign scenario.

This marine research field, as a subfield of environment research, is heavily controlled by European and Portuguese regulations like INSPIRE (European Commission Joint Research Centre, 2013) and those issued by SNIMAR (SNIMAR, 2016). The INSPIRE European directive imposes a rather strict API on environmental repositories in order to improve interoperability. This API embodies an abstract data model centered on observations with values for properties, associated procedures and geographic

references. There are implementations for this abstract data model and we have chosen two: the 52North Sensor Observation Service (SOS) implementation and the Geoserver Web Feature Service (WFS), Web Map Service (WMS), and Web Coverage Service (WCS) implementation.

The paradigm for this case is different from the two previous ones, where the data resided in data files with specific formats stored in the file system. In this case, each observation is stored in a database, along with the metadata, thus making it easier to search the data using ad hoc queries.

In this project the goal was to develop a client (called SeaBioData) for those two servers that maps well with the concepts, data sheets and methods of the marine research team and is able to ingest the data already available in digital form and to present them in visually effective ways.

Requirements in disciplinary infrastructures

In disciplinary platforms, such as the WindScanner.eu and SeaBioData, besides the immediate requirements we are dealing with, it is likely that, as more research groups participate and use the platforms, more specific functionalities will be requested. This can be regarded as a natural evolution of the infrastructures and even as a sign of their successful adoption.

One effect that may result from the definition of disciplinary infrastructures stems from their specialization: datasets are managed in specific structures that, even if they are open, may become hard to discover and search. Two current lines counter this effect. One is the existence of research data aggregators, such as re3data, supported by DataCite (Brase et al., 2015). The other is the trend of Linked Open Data (LOD) as applied to metadata. If metadata can be harvested by LOD services, the possibilities for data to be available to different kinds of applications increase.

Conclusions

As we look at research data management, we cannot help recognize it is a daunting task. In a way, this is common to all archival endeavor: how can we estimate the valuable assets, when not all can be curated and preserved? With respect to traditional archives, web archiving initiatives have already dealt with the problem of making selection and description a more lightweight, semi-automated task. A similar problem confronts research data archives, where probably some collections will be way better curated than others. At the current point in RDM efforts, we have to struggle to achieve a balance between the effort required for data description, which will render datasets part of the research trail, as publishable and re-usable artifacts, and the immediate rewards that researchers get from their investment in data curation. We argue strongly in favor of tools that can alleviate the researchers' tasks, embed data curation in the overall research activities, and provide trust on the resulting research outputs. Infrastructures and tools appear at a faster pace than researchers can handle, and there is a large gap between their functionalities and the requirements as perceived by researchers. The second main line of action is therefore the analysis of requirements for RDM, the matching of requirements with existing tools, and the identification of the missing links.

In the work reported here, there is a focus on solutions for the long tail, but also some cases concerning RDM in areas that require custom-designed solutions. It is important to assume from the start that not all research groups need the same kind of solutions, while looking for common issues. We take the examples of a civil engineering group collecting stream data from sensors, a wind research

community designing the infrastructure to collect and process large datasets, and a marine and atmosphere institute dealing with the diversity of data used in forecast products, marine research and specimen collection. For each of these groups, we have designed and implemented custom systems dealing with some parts of the research workflows.

The field work with diverse researchers and the controlled experiments performed to evaluate the tools have both confirmed the perceived need for a more solid support to the full RDM workflow. In interviews, experiments and informal conversation, researchers have been curious about the concepts they are not familiar with—metadata, ontologies, repositories, preservation— and provided many clues on the ways to address their requirements. Moreover, RDM is becoming a strong concern as national and international policies and funding bodies move towards open science. Both kinds of motivation are essential to take RDM forward. Without genuine interest from researchers, RDM actions may become just another administrative burden with no real commitment to data organization and description. But the institutional requirements are also crucial to provide short-time rewards and bring RDM to the attention of researchers in all areas.

With the data organization and description tools reaching maturity, our concern is now the complete research workflow, from data collection to data preservation. In the continuation of the work with the research groups we are setting up partnerships to collaborate on the definition of RDM strategies starting with Data Management Plans and continuing to their execution and evaluation (Active DMP, 2017). Another important aspect is the identification of researchers and research outputs, and the connection to systems such as ORCID for researcher identification and DOI for outputs. This brings interoperability with institutional systems, and therefore visibility at institutional level and in international repositories, and enables the management of links between different research results. Metadata models, their evolution in communities and the promotion of metadata harvesting and aggregation in data repositories is also a long-time effort where early engagement will contribute to motivate and reward researchers.

References

ACTIVE DMP. 2017. *Research Data Alliance- Active Data Management Plans Interest Group* [Online]. Available: https://www.rd-alliance.org/groups/active-data-management-plans.html [Accessed July 2017].

AMORIM, R. C., ROCHA, A., OLIVEIRA, M. & RIBEIRO, C. 2016. Efficient Delivery of Forecasts to a Nautical Sports Mobile Application with Semantic Data Services. *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering.* Porto, Portugal: ACM.

ANDS. 2017. *ANDS- Australian National Data Service* [Online]. Available: http://ands.org.au/ [Accessed July 2017].

BRASE, J., SENS, I. & LAUTENSCHLAGER, M. 2015. The Tenth Anniversary of Assigning DOI Names to Scientific Data and a Five Year History of DataCite. *D-Lib Magazine,* 21.

CASTRO, J. A., PERROTTA, D., AMORIM, R. C., ROCHA DA SILVA, J. & RIBEIRO, C. 2015. Ontologies for Research Data Description: A Design Process Applied to Vehicle Simulation. *Metadata and Semantics Research - 9th Research Conference, MTSR 2015.*

CASTRO, J. A., RIBEIRO, C. & ROCHA DA SILVA, J. 2014. Creating lightweight ontologies for dataset description. Practical applications in a cross-domain research data management workflow. *IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2014.*

DA COSTA, F. P., CUNHA, A. & DAVID, G. 2014. ViBest SHM: an Information System and Data Repository for Structural Health Monitoring. *9th International Conference on Structural Dynamics (EURODYN).*

DAF. 2017. *Data Asset Framework* [Online]. Available: http://www.data-audit.eu/ [Accessed July 2017].

DANS. 2017. *Data Archiving and Networked Services* [Online]. Available: http://www.dans.knaw.nl/en [Accessed July 2017].

DASH. 2017. *Dash- Data Sharing made easy* [Online]. Available: https://dash.cdlib.org/ [Accessed July 2017].

DATAONE. 2017. *DataONE* [Online]. Available: https://www.dataone.org/ [Accessed July 2017].

DCC. 2017. *DCC- Digital Curation Centre* [Online]. Available: http://www.dcc.ac.uk/ [Accessed July 2017].

DOTY, J., HERNDON, J., LYLE, J. & STEPHENSON, L. 2015. Learning to curate. *Bulletin of the American Society for Information Science and Technology,* 40.

EUROPEAN COMMISSION JOINT RESEARCH CENTRE 2013. INSPIRE Data Specification on Bio-geographical Regions – Technical Guidelines 10.12.2013.

EUROPEAN STRATEGY FORUM ON RESEARCH INFRASTRUCTURES 2016. Strategy Report on Research Infrastructures.

FANIEL, I. M. & YAKEL, E. 2011. Significant properties as contextual metadata. *Journal of Library Metadata,* 11.

FEGRAUS, E. H. & ANDELMAN, S. 2005. Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America,* 86**,** 158-168.

GOMES, F., LOPES, J. C., PALMA, J. L. & RIBEIRO, L. F. 2014. WindS@UP: The e-Science Platform for WindScanner.eu. *Journal of Physics: Conference Series,* 524.

LECARPENTIER, D., MICHELINI, A. & WITTENBURG, P. The building of the EUDAT Cross-Disciplinary Data Infrastructure. EGU General Assembly Conference Abstracts, 2013. EGU2013-7202.

MANGHI, P., MANOLA, N., HORSTMANN, W. & PETERS, D. 2010. An infrastructure for managing EC funded research output - The OpenAIRE Project. *The Grey Journal (TGJ): An International Journal on Grey Literature,* 6.

MAYERNIK, M. S. 2011. Metadata realities for cyberinfrastructure: Data authors as metadata creators. *ProQuest Dissertations and Theses***,** 338.

NCBI RESOURCE COORDINATORS 2013. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research,* 41**,** D8-D20.

NEWA. 2017. *NEWA, New European Wind Atlas* [Online]. Available: http://www.neweuropeanwindatlas.eu/ [Accessed July 2017].

PÔÇAS, I., GONÇALVES, J., MARCOS, B., ALONSO, J., CASTRO, P. & HONRADO, J. P. 2014. Evaluating the fitness for use of spatial data sets to promote quality in ecological assessment and monitoring. *International Journal of Geographical Information Science,* 28**,** 2356-2371.

RIBEIRO, C. & FERNANDES, M. E. M. 2011. Data Curation at U.Porto: Identifying current practices across disciplinary domains. *IASSIST Quarterly,* 35**,** 14-17.

RIBEIRO, C., ROCHA DA SILVA, J., CASTRO, J. A., AMORIM, R. C. & FORTUNA, P. Motivators and Deterrents for Data Description and Publication: Preliminary Results (Short Paper). On the Move to Meaningful Internet Systems: OTM 2015 Workshops, 2015. 512-516.

ROCHA DA SILVA, J., RIBEIRO, C. & CORREIA LOPES, J. 2018. Ranking Dublin Core descriptor lists from user interactions: a case study with Dublin Core Terms using the Dendro platform.

International Journal on Digital Libraries, April 2018, pp 1-20, https://doi.org/10.1007/s00799-018-0238-x

ROCHA DA SILVA, J. 2016. *Usage-driven Application Profile Generation Using Ontologies.* Ph.D., Universidade do Porto.

ROCHA DA SILVA, J., CASTRO, J. A., RIBEIRO, C., HONRADO, J., LOMBA, A. & GONÇALVES, J. 2014. Beyond INSPIRE: An Ontology for Biodiversity Metadata Records. *On the Move to Meaningful Internet Systems: OTM 2014 Workshops.*

ROCHA DA SILVA, J., RIBEIRO, C. & CORREIA LOPES, J. 2012. Managing multidisciplinary research data: Extending DSpace to enable long-term preservation of tabular datasets. *iPres International Conference on Digital Preservation.*

ROCHA DA SILVA, J., RIBEIRO, C. & LOPES, J. C. 2016. Usage-Driven Dublin Core Descriptor Selection. *Research and Advanced Technology for Digital Libraries: 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016.* Springer International Publishing.

SEABIODATA. 2017. *SeaBioData- Portuguese Seamounts Biodiversity Data Management* [Online]. Available: http://eeagrants.org/project-portal/project/PT02-0017 [Accessed July 2017].

SNIMAR. 2016. *The SNIMAR Metadata Profile* [Online]. Available: http://www.snimar.pt/ar/ficheiros/perfilSNIMar.pdf [Accessed July 2017].

TOLEDO, F. M. B., CARRAVILLA, M. A., RIBEIRO, C., OLIVEIRA, J. F. & GOMES, A. M. 2013. The Dotted-Board Model: A new MIP model for nesting irregular shapes. *International Journal of Production Economics,* 145**,** 478-487.

VARDIGAN, M., HEUS, P. & THOMAS, W. 2008. Data Documentation Initiative: Toward a Standard tor the Social Sciences. *The International Journal of Digital Curation,* 3**,** 107-113.

WIECZOREK, J., BLOOM, S., GURALNICK, R., BLUM, S., DORING, M., GIOVANNI, R., ROBERTSON, T. & VIEGLAIS, D. 2012. Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE,* 7.

WINDSCANNER. 2017. *WindScanner.eu- A new European Distributed Research Infrastructure* [Online]. Available: http://www.windscanner.eu/ [Accessed July 2017].

WITZE, A. 2017. World's largest wind-mapping project spins up in Portugal. *Nature,* 542**,** 282-283.

---

1 Cristina Ribeiro (corresponding author, mcr@fe.up.pt) and Gabriel David are Associate Professors with the Informatics Engineering Department, Faculty of Engineering, University of Porto. João Correia Lopes is an Assistant professor with the same Department. João Rocha da Silva, João Aguiar Castro and Ricardo Carvalho Amorim are senior researchers at INESC TEC. All authors are affiliated with INESC TEC.

2 http://www.unidata.ucar.edu/software/netcdf/