Engaging researchers in data management with LabTablet, an electronic laboratory notebook

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

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

Abstract. Dealing with research data management can be a complex task, and recent guidelines prompt researchers to actively participate in this activity. Emergent research data platforms are proposing workflows to motivate researchers to take an active role in the management of their data. Other tools, such as electronic laboratory notebooks, can be embedded in the laboratory environment to ease the collection of valuable data and metadata as soon as it is available. This paper reports an extension of the previously developed LabTablet application to gather data and metadata for different research domains. Along with this extension, we present a case study from the social sciences, concerning the identification of the data description requirements for one of its domains. We argue that the LabTablet can be crucial to engage researchers in data organization and description. After starting the process, researchers can then manage their data in Dendro, a staging platform with stronger, collaborative management capabilities, which allows them to export their annotated datasets to selected research data repositories.

1 Introduction

With increasing amounts of research data being produced every year [3], institutions tend to implement guidelines and workflows to preserve them, in a similar way to what it is already the current practice with publications [6]. Nevertheless, this approach can pose some barriers to the dissemination and reuse of such datasets, as a consequence of the lack of metadata that is essential for other researchers to understand the production context of a specific dataset [12]. Likewise, gathering domain-level metadata at the deposit stage can be a very demanding—and time consuming—task for curators, often responsible for more than one research domain. At the same time, researchers play a key role in their data description [7], as they have the best knowledge of their production environment, and can add metadata to their data from the early stages. Existing platforms for research data management, such as Figshare[3] or Zenodo[4], already support simple descriptive metadata, but the barrier between them and the researchers'

[3] http://figshare.com/
[4] http://zenodo.org/

working environment is still high. It is therefore recognized that important data and metadata are still temporarily stored in frail locations such as personal computers and laboratory notebooks [10]. Ultimately, even with guidelines for data management in place, some of these resources never reach the deposit stage as they are susceptible to neglect.

In this paper, we present LabTablet as an application to help researchers gather data and metadata during experimental runs or field trips, and directly export them to a staging repository—in our case Dendro [5]—responsible for creating a collaborative, description-oriented approach to research data management. With this approach, we can provide a better handling of research data and provide conditions for capturing metadata as soon as it becomes available. At the end of a research project, Dendro is capable of creating and exporting the dataset package to existing platforms for data preservation, that can also take advantage of the included metadata to improve the visibility of the dataset.

2 Research data management

Amid the research activities, researchers produce both raw and processed data that support their conclusions towards the project goals. These resources are sometimes neglected after the publication of the results, weakening the link between project results and the data that supported them.

Managing research data has evolved to include tasks besides storage and preservation, ensuring a proper handling of research outputs to facilitate their retrieval and long-term preservation. Furthermore, similarly to what happens with research publications, the deposit of research assets in repositories has to be accompanied by a comprehensive description—also known as metadata records—to facilitate their retrieval and interpretation. Ideally, when a dataset is provided with sufficient metadata, others will be able to reuse it [12]. An equally measurable result is the credit that researchers get from publications citing their data, with side effects related, among others, with the possible reduction of costs inherent to the research activity.

2.1 Data description

Datasets and publications have different requirements concerning their description [11]. Considering the diverse scenarios in which datasets are produced, we can identify sets of possible metadata descriptors that can be directly related to each specific research domain, and at the same time extend the basic, high level ones, used to describe publications. For each research domain, the description possibilities vary, and thus, the data repositories are evolving to comply with this required flexibility [2].

Well-known metadata schemas, such as Dublin Core, have been considered fit to a broad scope of applications and allowed the emergence of protocols for exchanging metadata and enhancing publications visibility [9]. The OAI-PMH[5] is the best known, and is widely used to index different repositories, allowing their resources to be presented in publications search engines. Basic descriptors, such as title, description and author, can be added by a designated curator and provide the link between data and publications, but when considering the broad possibilities for description in each of the

[5] https://www.openarchives.org/pmh/

domains, this task has to include researchers. Actively involving researchers in the description of their data faces some limitations, as the platforms created for this purpose must also take into account usability requirements and offer features that meet their goals as researchers, such as receiving credit for their data and sharing them with their peers.

2.2 Researchers' engagement in data management

In the course of research activities, researchers often resort to personal computers to store collected data and to their laboratory notebooks to record any observations or context. With the increasing amounts of research data, these approaches pose some risks in terms of data preservation, which can later constrain data's availability.

In the past few years, several platforms emerged to integrate the research environment, with some of them being actively used by several communities [2]. These platforms aim to implement established protocols for data preservation and dissemination, while featuring easy to use interfaces along with collaborative environments. The assessment of several existing platforms showed that issues such as data ownership, dataset description and dissemination are already a concern, although these platforms are still considered as a final location for dataset deposit [2]. Staging platforms such as Dendro, on the other hand, aim at creating management tools closer to the researchers' daily routines and offer a place where they can collaboratively store and describe data. It is important to stress out that, for these platforms, all the managed data are private and unaccessible from the external community, as it can involve sensitive data that have to be adjusted prior to its disclosure. Only then they can be cited and reused. At the end of the research activity researchers can export the resulting resources to the final repositories, often aimed at long-term preservation.

3 Electronic laboratory notebooks

We have previously highlighted the importance of data management repositories, both as staging environments and as research data preservation solutions. As several researchers resort to field trips or experimental runs to gather data—often a typical approach to data production—there is still a gap between data production and their deposit in the mentioned platforms. Electronic laboratory notebooks can fill this gap, allowing researchers to record and directly deposit data, while mitigating the risk of loosing such records during the process [8]. Nevertheless, the existing solutions tend to focus on a particular domain or offer limited functionality, not taking advantage of some of the available sources of metadata, and excluding prospective users from other domains.

3.1 LabTablet

Taking advantage of the growing popularity of handheld devices, LabTablet was developed as an electronic laboratory notebook to help researchers describe their data as soon as the project starts. Besides having an easy to use interface, the underlying representation for each metadata record follows established standards, ensuring a streamlined curation process before the final deposit in a repository. The first version of this project was focused on gathering metadata in the field, relying on previously built

application profiles, and therefore using a set of descriptors for that specific domain. In any of the versions, LabTablet is capable of uploading each dataset to Dendro, or any other staging platform, from which it can later be included in preservation solutions. This approach allows curators to have standards-compliant metadata records upon deposit but, more importantly, domain-level metadata that would otherwise be lost is properly maintained.

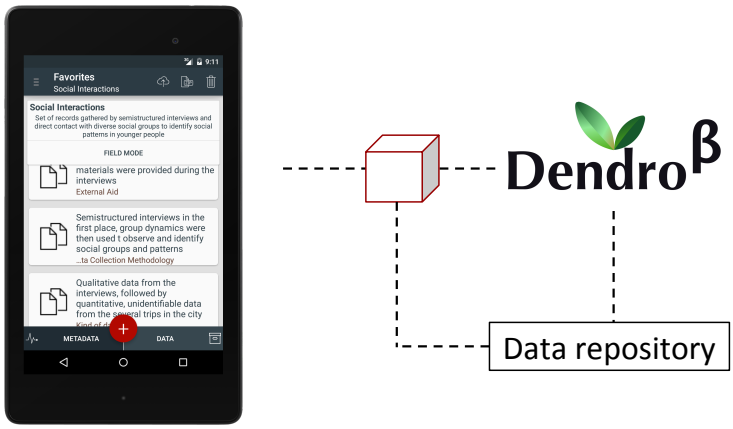

Fig. 1: View of a project with the gathered metadata.

Figure 1 the application's interface regarding an opened project, with the correspondent gathered descriptions that can later be exported to a designated platform. After preliminary evaluation with researchers from the biodiversity domain [1], a new approach was developed, extending the metadata capabilities of this application and including mechanisms to also gather opportunity data—observations collected by chance while performing some other activity. Opportunity data can be directly linked to the researchers' field trips and be enriched by the use of the tablet's built-in sensors to gather metadata from the available sources such as camera, GPS or accelerometer. In addition to those, LabTablet also allows voice recordings, sketches, and tracking a field trip, and is able to export the results to a compliant format[6]. Furthermore, researchers can also import other types of data (namely spreadsheets) from their computers, merging them into the workflow. To take advantage of the device's capabilities, and considering a wider set of research domains, additional input modes were also implemented, namely forms, used in surveys. Forms can be custom-designed and filled directly in the application.The workflow for such process relies on the researcher to create a model, and to instantiate it whenever a subject is interviewed. The same applies for other activities that require some kind of form or survey such as routine evaluations and observations. The gathered data is then exported to files that are compatible with common statistical analysis tools such as Excel or SPSS[7]. At this stage, the development of new LabTablet features is mainly dependent on the integration of existing workflows, as well as the

[6] A KML-based representation (`https://developers.google.com/kml/`), containing a set of connected coordinates, for instance.

[7] `http://www-01.ibm.com/software/analytics/spss/`

inclusion of standards that are already in use among the research workgroup or possible direct connection between the application and the researcher's tools (LabView[8] or SPSS, for instance).

At the end of each field trip (or when the researcher finds it convenient to do so), the application can sync the collected resources with a repository where researchers are able to share them with their team or the community. This ensures that data are stored in the appropriate location, under their institutions supervision. Additionally, as with metadata records, the created package can follow any guidelines, namely the structure of a Submission Information Package (SIP), from the Open Archival Information System model[9], provided the correct integration is done.

4 Social Sciences: a case study

As a part of an ongoing partnership, a researcher from the social sciences domain was interviewed to assess the different data management needs for this specific domain[10]. Initially, a set of questions was proposed to address metadata needs or possible constraints on data sharing. As the interview went on, several important aspects related to the workgroup's current practices allowed us to tailor the existing workflow to their needs. Previous work with researchers in engineering domains [4] showed that usually researchers deal with systematic data production which has features that are common to several domains: experimental data, for instance, tends to deal directly with the experimental setup and the physical properties of samples or compounds. In the social science domains, on the other hand, workflows are centered on temporal or spatial coverages, having their main focus on social traits that can differ greatly. As a result we have high heterogeneity of dataset structures and description needs across different research groups, that are highly dependent on the researcher's view of the event.

4.1 The social sciences domain

Our interview revealed the researcher's awareness of the recent evolution of data management guidelines on this area. However, these had never been put into practice. Studies in this group are mainly focused on evaluating phenomena in different social groups, directly interacting with them either through field observations, structured or unstructured interviews, or content analysis. During these activities, the produced data is mainly of qualitative nature, with a small portion of quantitative data as well. Qualitative data is, for this group, mostly related with observations or notes which contents are fully dependent on the producer, whereas quantitative data results from surveys and questionnaires.

Concerning the publication of research data, the researcher highlighted some limitations, as some projects are not expected to disclose data and some datasets are of a sensitive nature and need to follow ethical recommendations, or need to be anonymized before their disclosure, if applicable. Still, for some projects, pursuing data disclosure would benefit both parts, as they would be able to cite datasets in publications and their peers could access and reuse such data in subsequent analysis.

[8] `http://www.ni.com/labview/pt/`

[9] `http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284`

[10] The survey for this evaluation was based on the Data Curation Toolkit, available at `http://datacurationprofiles.org/`

LabTablet proved to be capable of handling all these needs in terms of data production, as well as helping researchers identify some equally important descriptors that could be added to provide extra context. During the course of this interview, a set of basic Dublin Core elements revealed to be satisfactory for the description needs in this domain[11]. Nevertheless, for a deeper data description, other schemas should also be included to achieve an extensive metadata record.

4.2 Preparing for data description

After identifying the basic description needs and suggesting an initial profile for this purpose, we proceeded to identify other domain-level descriptors. In this field, the Data Documentation Initiative (DDI)[12] proved to have a suitable set of descriptors for social sciences domains, namely[13]:

- `Data Collection Methodology`—to specify which methodology was used to collect the samples or questionnaires. This revealed to be a recurrent scenario as researchers often worked with a small set of methodologies;
- `Data Source`—to identify the source of the collected data, including the associated project. As some of the projects could include partnerships with other data providers, this descriptor was chosen to support such specification;
- `Sample Size`—to state the dimension of the sample or the number of interviewees during a field session;
- `External Aid`—a reference to any support given during the experiment, such as text cards or multimedia support;
- `Kind of data`—a specification of the dataset's content type. This allows researchers to specify whether the packaged data is of a qualitative, quantitative or mixed type;
- `Universe`—a description of the referenced population, if applicable. This can include informations related to age, gender or income classifications.

The selected descriptors allow a better understanding of the dataset in question. A clear description of the population will, for instance, enable other researchers to search for datasets that were obtained from specific social communities, and the same happens for the other descriptors such as the `Sample Size`. According to the researchers, identifying the methodology was considered to be a key item in the description process. This identification was often extensively done and it was a common item to be mentioned in each project. According to the schema specification, this item is expected to mainly consist of a brief description of the involved methodology, but in this case—and considering related work in this area—this field can sometimes be very extensive, which led the researcher to suggest that other descriptors should also be present to promote a structured representation of this information.

After this selection of descriptors, we proceeded to create the ontology for this domain. Along with the descriptors from the Data Documentation Initiative, we included high level descriptors from the Dublin Core profile as well. This ontology can be loaded at any time into the LabTablet application and be used to describe data in this area. The same is true for Dendro, our staging repository.

[11] These consist of the base Dublin Core elements profile, namely abstract, contributor, creator, subject, title, description, publisher, date, type, and others, as specified in `http://dublincore.org/documents/usageguide/elements.shtml`

[12] `http://www.ddialliance.org/`

[13] Not all the descriptors are depicted here

5 Conclusions

By analyzing different research domains, we can identify many differences concerning data management practices. While some groups have data management procedures already in place, most are still far from addressing the issue, mostly due to the nature of their data rather their motivation.

The researcher from our case study recognized the added value in automatically exporting the daily produced data to a centralized location, where it could be properly handled and edited. Additionally, some specialists in the field advise against using any kind of note taking tools during the interviews, not to influence the interviewee; however, the researcher considered very important to be able to record or transcribe the interviews in the background.

We are testing the collection of metadata throughout the entire research workflow with several research teams. It is clear by now that devices and tools to make the process easier on the researchers can make the difference between a process regarded as an extra burden on researchers and one where they perceive the benefits and get involved.

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). João Rocha da Silva is also supported by research grant SFRH/BD/77092/2011,provided by the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

References

1. Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. LabTablet: semantic metadata collection on a multi-domain laboratory notebook. In *Metadata and Semantics Research*, pages 193–205. Springer, 2014.
2. Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. A Comparative Study of Platforms for Research Data Management: Interoperability, Metadata Capabilities and Integration Potential. In *New Contributions in Information Systems and Technologies*, pages 101–111. Springer International Publishing, 2015.
3. Christine L. Borgman. Advances in Information Science: The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, 63(6):1059–1078, 2011.
4. João Aguiar Castro, João Rocha 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*, pages 0–3, London, 2014.

5. João Rocha da Silva, João Aguiar Castro, Cristina Ribeiro, and João Correia Lopes. The Dendro research data management platform: Applying ontologies to long-term preservation in a collaborative environment. In *Proceedings of the iPres 2014 Conference*, 2014.
6. Clifford A. Lynch. Institutional repositories: essential infrastructure for scholarship in the digital age. *Association for Research Lybraries. Bimonthly Report no.226*, 2003.
7. Liz Lyon. Dealing with Data: Roles, Rights, Responsibilities and Relationships,(2007). *Consultancy Report. UKOLN*, (June):1–65, 2011.
8. Jason T Nickla and Matthew B Boehm. Proper laboratory notebook practices: protecting your intellectual property. *Journal of neuroimmune pharmacology*, 6(1):4–9, March 2011.
9. Robin Rice. Applying DC to Institutional Data Repositories. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, page 2008, 2009.
10. Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, 6(6):e21101, 2011.
11. Andrew Treloar and Ross Wilkinson. Rethinking Metadata Creation and Management in a Data-Driven Research World. *2008 IEEE Fourth International Conference on eScience*, pages 782–789, 2008.
12. Craig Willis, Jane Greenberg, and Hollie White. Analysis and Synthesis of Metadata Goals. *Journal of the American Society for Information Science and Technology*, 63(8):1505–1520, 2012.