Giovanni Esposito, Davide Reverberi, Giovanni Romagnoli and Riccardo Ghinzelli

Research Data Management for Laboratory Services: the DigiLab4U Use Case of Dataverse

Abstract

The ongoing digitalization of academia and research institutes has led to an increasing need for suitable processes of data management and dissemination. As a result, research is increasingly asking for standardized data management processes. Research data management (RDM) has emerged as an important concern for the whole scientific community, and several platforms to support data deposits have been designed and released. Actually, (i) rules of management and (ii) the curation of advanced data catalogues seem to be generally lacking. Moreover, one of the major challenges is encouraging consultation and 'buy-in' from researchers and senior managers. This paper presents the implementation of Dataverse by the DigiLab4U consortium from this standpoint. The benchmarking process among research data management (RDM) platforms available on the open-source market, and the hierarchical structure for storing and managing data are introduced and discussed.

Keywords

Research Data Management (RDM), Dataverse, Servitization

1 Introduction

The ongoing digitalization of academia and research institutes has led to an increasing need for suitable processes of data management and dissemination. As a result, important scientific journals, academia, but also third-party funding institutions are increasingly asking for standardized data management processes (Wilms et al., 2018). Research data management (RDM) has emerged as an important concern for the whole scientific community, and several platforms to support data deposits have been designed and released (Amorim et al., 2015). Technological progress, especially under the advent of the Internet of Things (IoT) era, created a state of the art involving

high-level performances, with respect to the possibility of storing huge amounts of data, and accessing them everywhere by means of cloud services, for instance. This allowed providers to supply advanced data management services. Although these important results were achieved, the management and curation of advanced data catalogues seem to be lacking (Cox et al., 2017). As a consequence, institutions that eventually share the data must establish their own rules for the suitable management and promotion of research data (Wilms et al., 2018), otherwise, data catalogues could result in a mess. In addition, major challenges include (i) resourcing, (ii) adaptive capacities and communicability with other services, and especially (iii) encouraging consultation and 'buy in' from researchers and senior managers (Cox et al., 2017). The 'buy in' formula for the provisioning of data especially is a new and important topic in research, and it is gaining attention, especially from business entities that are interested in buying research data and scientific knowledge from academia and research institutes (Esposito et al., 2021). Therefore, this paper presents the implementation of Dataverse by the DigiLab4U consortium. The DigiLab4U consortium, under the Open Digilab4U project funded by the German Federal Ministry of Education and Research (BMBF), aims at creating a network of digitalized labs via the Internet of Things, towards hybrid education, cross-institutional research, and cross industrial cooperation. Dataverse is a type of open-source software for the management of files in the academic field. It has been selected within a benchmarking assessment, with respect to some requirements required by the DigiLab4U, and identified according to Amorim et al., (2015). This paper provides a standardized hierarchical structure for organizing the research material in a RDM system and hence answers the research question about a possible solution to managing the curation of content in research data catalogues. The remainder of the paper is as follows: section 2 provides an overview of the literature on digital online labs and RDM in this field. Section 3 briefly introduces the DigiLab4U case and provides the benchmarking assessments of the RDM platforms and software. In section 4, the hierarchical structure is presented and tested, and then its validation is discussed. Finally, section 5 addresses conclusions and outlooks for future works.

2 Literature review

The literature on digital online labs has been increasing even more over the last decade (Heradio et al., 2016). Basically, two research lines have arisen: one about the didactical perspective, and one about technical implementati-

on (Zappatore et al., 2015). What is missing is a deep analysis of financial and organisational aspects for making labs and networks in which they are inserted robust from a life cycle perspective (Esposito, Kammerlohr, et al., 2021). In this regard, Esposito, Mezzogori, et al. (2021) have showed that most digital online experiences last only for the time in which they are funded by institutions and organizations under the programs of national or international projects. From this point of view, Esposito, Kammerlohr, et al. (2021) analyzed the possibility of using research data for business partnerships upon payment, with generally positive results from several Italian companies. As a result, platforms for data curation and sharing are needed of course. Although advisory and consultancy services have been recently stressed, technology and data deposit assistance seem still to be lacking and are only forecast in the near future (Corrall et al., 2013). As a counterproof, a single result is obtained by querying Scopus with the following search string: (TITLE-ABS-KEY ("Research Data Management" OR RDM)) AND ((("data curation")) AND ("data curation")) AND ("data catalogue"), and the only work by Cox et al. (2017) attests to the lack of works and research on data catalogues and the active curation of data.

3 Dataverse within the DigiLab4U environment

In this chapter, firstly the DigiLab4U case is described. The technical system is just referenced here using the work by Galli et al. (2020) and Kammerlohr et al. (2021) since it is not of interest in this paper and is not discussed further. A deeper overview of the services provided is discussed, instead. Secondly, the requirements of the DigiLab4U for the selection of the RDM system are introduced. There are four key aspects to identify the DigiLab4U requirements, according to Amorim et al. (2015): (1) architecture, (2) metadata handling capabilities, (3) interoperability, content dissemination, and search features, and finally (4) community acceptance. Lastly, commercial solutions are analyzed in a benchmarking assessment, resulting in the selection of Dataverse.

The DigiLab4U case and its services

DigiLab4U is the cross-Institutional network of Industry 4.0 lab infrastructure. The consortium is led by the Hochschule für Technik Stuttgart (HFT), and joined by the other four founding members: the Bremen Institute for Production and Logistics (BIBA), the Institut für Wissensmedien

(IWM) of Koblenz-Landau, the Rheinisch-Westfälische Technische Hochschule (RWTH) of Aachen, and the University of Parma. Nowadays, it has another nine partners all around the world (see https://digilab4u.com/consortium/). The network was funded by the German Federal Ministry of Education and Research (BMBF) for developing the project 'Open Digital Lab for You', which created an integrated, hybrid learning and research environment consisting of a large variety of lab technologies offering digital services and reaching all kinds of possible users. The digitized lab environment intends to enable a real IoT learning marketplace, consisting of both digital labs from several suppliers and users who access to the labs. The cooperation between universities, research institutions and industry allows the suppliers to be pooled so that users have access to a larger variety of digital courses based on different IoT labs. As a consequence, one of the main features of such a network is the availability of tracked and traceable data repositories for research data, and suitable systems for retrieving and accessing them. Hence, an RDM was needed.

Requirements of the DigiLab4U for the RDM system

The proposed system must comply with two main functionalities of storing data: their ownership is assigned and they can be retrieved with a suitable reference system. First, from the user's perspective, people have access to old batches of data or new data generated by remote, performing analysis without collecting data physically in a lab. Second, from a content uploader's perspective, while they are uploading a data set, a new DOI must be automatically created, making data referencing easier and faster when using them and the information generated by them. Lastly, an efficient and organized framework to store all the data is required to promote data consultation and referencing. These translate into the following seven requirements, and each one is related to the four key aspects of Amorim et al. (2015): (i) the tool must support REST API, (ii) it possibly needs to be open source, and (iii) data must be hosted within Germany—these three requirements meet the key aspect (1); (iv) the ability to visualize the stored data must be accomplished, (v) it must support a wide range of data formats, based on the recommendations by the German association "Verbund Forschungsdaten Bildung", and (vi) the facility to generate edit and back up data is required—these three requirements meet the key aspects (3) and (4); finally, (vii) extraction of metadata from data must be simple and structured—this requirement meets the key aspects (2) and (3).

Benchmarking commercial solutions

Several RDM systems are commercially available, and in this phase, some have been analyzed with respect to the requirements of the DigiLab4U. Software identified from a market investigation performed on the Web are listed here: CKAN, RADAR, bwScienceToShare, Freidok plus, Dataverse, EdShare, dSPACE, and DKAN. Functionalities and characteristics have been mapped through ten main features, according to the requirements. These are (i) support for file formats, REST API, service support, and community, (ii) suitable data visualization; (iii) consistently building technology; (iv) type of license and price; finally, (v) host and service provider, and main users. If a single requirement is missing from the above features, the software is neglected. As a result, two types of software have been selected for comparison here: Dataverse from the Dataverse project, and dSPACE by dSPACE GmbH. Dataverse is rich in features and has an active service and user community to support users and service providers. Hence, the benchmarking has been concluded selecting Dataverse, which meets all the DigiLab4U requirements.

4 The structure and its transposition

Dataverse is an open-source Web application from the Dataverse Project, developed to share, preserve, cite, explore, and analyze research data. Mainly used by academia, it allows researchers, journals, data authors, publishers, data distributors, and affiliated institutions to access and replicate data from research, ensuring academic credit and Web visibility. A Dataverse repository (wordplay for data-and-universe) is the software installation, which then hosts multiple virtual archives called Dataverse collections, administrated by its creator, who has access to managing all the settings. A Dataverse collection is a container for data sets, each one containing data files (e.g., research data, code, documentation) and related descriptive metadata (e.g., tag and keywords, including documentation and a code that accompanies the data). Once a file is uploaded into a data set, it is no longer possible to eliminate it from the Dataverse. Also, the Dataverses can hold one or more Dataverse collections, which can be set up for individual researchers, departments, journals, and organizations. Dataverses and data sets within the main Dataverse can be created and placed arbitrarily, and they can also be categorized by means of Dataverse categories that identify the type of data hub (e.g., institution, laboratory) and address possible query strings. Hence, the need for a standard framework for uploading and managing con-

tent, ensuring ease of retrieving and accessing data. The proposed structure consists of a five-level structure in a father-son manner, from the Dataverse of the single institution within the main DigiLab4U Dataverse (at the top of the hierarchy) to the file attribute (at the bottom of the hierarchy). The structure is provided in Figure 1, with reference to an example in which two partners upload data from experiments performed in their respective remote laboratories. Each level is described in the following. At level 0 it is possible to find every key partner in the Digilab4U network. Every institution will have its own Dataverse collection in the main DigiLab4U Dataverse. Level 1 contains all the labs of every institution. Level 2 refers to all the specific experiments or analyses that can be performed by a laboratory. Level 3 contains all the data sets of a specific type of experiment or analysis. Lower level 4 contains all the files of a specific data set. Three actions are envisaged when uploading content. First, before uploading data, they all need to be renamed using a formatting standard consistent with the uploading session. Second, specific tags need to be applied to every single file. Third, if the files need to be collected and organized in folders according to their characteristics, compressed files can be used, which enables tree visualization.

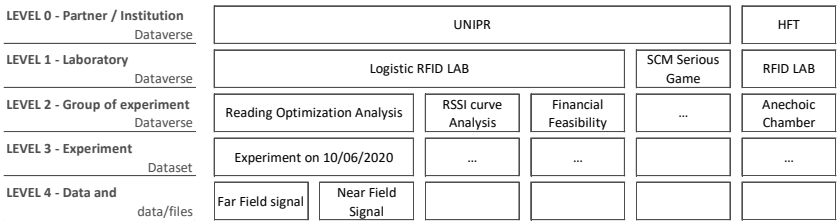

Figure 1: Hierarchical structure for storing Dataverses and data sets

The hierarchical structure so formalized has been discussed by a panel of 8 experts in the field of education and data management from HFT, BIBA, and Parma. The installation and the efficiency of the structure have been discussed, and no concerns arose. Therefore, the demo of the DigiLab4U Dataverse has been officially presented and validated by the experts. The verification of its functionalities and its approval has been achieved, and the system has been recognized as suitable with respect to the original requirements of the DigiLab4U and is now approaching the Go-Live phase.

5 Discussion and Conclusions

In this paper, a hierarchical structure for RDM in Dataverse is presented, referring to the DigiLab4U case. The novelty presented in this paper refers to a more efficient and organized way to store data on the RDM platform Dataverse by means of a five-level structure. This has been chosen as the best compromise between the redundancy of Dataverses and levels of detail in the data catalog. This structure has been verified with respect to (i) the simplicity of the query for retrieving data, and (ii) the suitability of the data catalog structure fostering data consultation. Although this is not evidenced by the paper, it has been discussed, with the several experts involved in the validation and verification process, that companies could be interested in acquiring research results that mostly fit their needs, creating room for the supposed financial sustainability of labs and the network. Future works could analyze this topic, and authors are working on this.

References

Amorim, R. C., Castro, J. A., da Silva, J. R., & Ribeiro, C. (2015). A comparative study of platforms for research data management: Interoperability, metadata capabilities and integration potential. *Advances in Intelligent Systems and Computing*, *353*, 101–111. https://doi.org/10.1007/978-3-319-16486-1_10

Corrall, S., Kennan, M. A., & Afzal, W. (2013). Bibliometrics and research data management services: Emerging trends in library support for research. *Library Trends*, *61*(3), 636–674. https://doi.org/10.1353/lib.2013.0005

Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology*, *68*(9), 2182–2200. https://doi.org/10.1002/asi.23781

Esposito, G., Kammerlohr, V., Reverberi, D., Rizzi, A., Romagnoli, G., & Bisaschi, F. (2021). Business Model validation for a marketplace of lab network initiatives. *Proceedings of the 26th Summer School "Francesco Turco"*.

Esposito, G., Mezzogori, D., Reverberi, D., Romagnoli, G., Ustenko, M., & Zammori, F. (2021). Non-Traditional Labs and Lab Network Initiatives: A Review. *International Journal of Online & Biomedical Engineering*, *17*(5).

Galli, M., Mezzogori, D., Reverberi, D., Uckelmann, D., Ustenko, M., & Volpi, A. (2020). Digilab4u: General architecture for a network of labs. *Proceedings of the 25th Summer School "Francesco Turco"*.

Heradio, R., de La Torre, L., Galan, D., Cabrerizo, F. J., Herrera-Viedma, E., & Dormido, S. (2016). Virtual and remote labs in education: A bibliometric analysis. *Computers & Education*, *98*, 14–38.

Kammerlohr, V., Pfeiffer, A., & Uckelmann, D. (2021). Digital Laboratories for Educating the IoT-Generation Heatmap for Digital Lab Competences. In M. E. Auer & D. May (eds.), *Cross Reality and Data Science in Engineering. REV 2020. Advances in Intelligent Systems and Computing, vol. 1231* (pp. 3–20). Springer, Cham. https://doi.org/10.1007/978-3-030-52575-0_1

Wilms, L., K., Stieglitz, S., Buchholz, A., Vogl, R., & Rudolph, D. (2018). Do researchers dream of research data management (2018). *Proceedings of the 51st Hawaii International Conference on System Sciences*, 4411–4420.

Zappatore, M., Longo, A., & Bochicchio, M. A. (2015). The bibliographic reference collection GRC2014 for the online laboratory research community. *Proceedings of 2015 12th International Conference on Remote Engineering and Virtual Instrumentation (REV)*, 24–31.

Authors

Giovanni Esposito, PhD
University of Parma
Parco Area delle Scienze, 181
43124 Parma
https://orcid.org/0000-0001-5150-0855
giovanni.esposito@unipr.it

Giovanni Romagnoli, PhD
University of Parma
Parco Area delle Scienze, 181
43124 Parma
https://personale.unipr.it/en/ugovdocenti/person/96588
giovanni.romagnoli@unipr.it

Davide Reverberi
University of Parma
Parco Area delle Scienze, 181
43124 Parma
https://orcid.org/0000-0001-6768-3932
davide.reverberi@unipr.it

Riccardo Ghinzelli
University of Parma
Parco Area delle Scienze, 181
43124 Parma
riccardo.ghinzelli@studenti.unipr.it