On research data publishing

Leonardo Candela¹ · Donatella Castelli¹ · Paolo Manghi¹ · Sarah Callaghan²

© Springer-Verlag Berlin Heidelberg 2017

1 Introduction

Data publishing is a practice where data are made public so that they can be reused by others, in accordance with associated policies, while guaranteeing that data provenance and attribution are tracked. There is not yet a general agreement on the importance of, and on the solutions for, data publishing [8–10]. While in some communities it has already solidified into codified policies, procedures and workflows, in others its motivations and realization shapes are still controversial [3].

In the reality, it is a multifaceted process that can occur at different stages of the scientific process, involves different levels of openness and controlled sharing and concerns data products at various levels of manipulation and maturity. There are two major schools of thought governing the approaches supporting data publishing, those aiming at exploiting scholarly communication practices (e.g., data journals [4]) and those proposing services and infrastructures for the dissemination of these artefacts (e.g., data repository

ries [1]). At the light of this wide range of possibilities, the data publishing process has to be analysed also from the new perspectives opened up by the availability of innovative technologies to understand if, and to what extent, it can/should be rethought.

The further development of data publishing requires a critical and in-depth analysis of the approaches so far developed by the different communities in order to identify the aspects actually facilitating data publishing, to distil best practices and promote solutions that enable effective cross-domain reuse of published data.

The goal of this special issue was to collect the concrete experiences and initiatives that have been developed to deal with the challenging, multifaceted and ever evolving issue of research data publishing.

2 Selected papers

The seven papers collected in this issue are addressing the data publishing research topic from different perspectives: (a) papers analysing the problem as a whole with the goal to identify a reference model for data publishing workflows and components; (b) papers proposing experiences and proof of concepts on how to produce innovative forms of publications either linking or embedding datasets and code; (c) papers focusing on specific data typologies (i.e., environmental data, social sciences, information retrieval experimental evaluations) and describing experiences as well as innovative solutions.

Austin et al. [2] analyse a set of data publishing workflows (e.g., in repositories, journals, projects) with the goal to propose a reference model identifying key components in data publication chains. They conclude the paper by giving a set of recommendations for organizations willing to establish new

✉ Leonardo Candela
leonardo.candela@isti.cnr.it

Donatella Castelli
donatella.castelli@isti.cnr.it

Paolo Manghi
paolo.manghi@isti.cnr.it

Sarah Callaghan
sarah.callaghan@stfc.ac.uk

¹ Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR), Via G. Moruzzi, 1, 56124 Pisa, Italy

² STFC - Rutherford Appleton Laboratory, Centre for Environmental Data Analysis, Chilton, Didcot, Oxfordshire OX11 0QX, UK

publication workflows including the importance to adhere to existing standards and the need to properly document the processes and the services underlying the workflow.

Kitchin et al. [7] propose an authoring workflow for integrating data and code into manuscripts and enabling the data and code to be published in reusable form. This workflow is actually a proof of concept that is demonstrated by a series of concrete examples.

Edmunds et al. [5] present the *GigaScience* experience in attempting to tackle the gaps in publishing large-scale, computationally intensive research. In particular, GigaScience focuses on biological data and promotes articles linking standard manuscript publication with an integrated database (GigaDB) that hosts all associated data, and provides additional data analysis tools and computing resources. The final goal is to move towards executable publications and to support reproducibility.

Van den Eynden and Corti [12] examine the data publishing practices for the social sciences. In particular, they report on the UK activities and present the lesson learnt and the developments resulting from the willingness to move the social science data publishing from an archivist responsibility to a researcher process.

Wright et al. [13] discuss the issues arising in the publication of environmental data in the field of ecology and hydrology. In particular, they present the workflow and the supporting infrastructure that is used by the Natural Environment Research Councils (NERC) Environmental Information Data Centre (EIDC). Harrison et al. [6] complement this work by discussing the successes, the challenges and the opportunities from a data centre perspective. In particular, they focus on the issues posed by dynamic datasets with respect to their citation and the recording and management of citation metrics. They conclude by highlighting that data centres are in a proper position to support data publishing (they might have infrastructure, tools, resources and processes) yet further work is needed to promote the uptake of data centres offering.

Finally, Silvello et al. [11] discuss the problems related to the publication of research data in the context of information retrieval experimental evaluation. They propose a description framework based on the linked open data paradigm to annotate and link dataset stemming from the specific application context thus to promote their dissemination and reuse. Moreover, they exploit this framework to automatically construct expert profiles that can be successfully used to perform topic-based experts search.

3 Concluding remarks

Publishing research datasets with the goal to favour their reuse both within the primary context the dataset is conceived

for, as well as in other contexts and scenarios, is certainly a challenging task. It requires technical solutions as well as organizational solutions where almost all the stakeholders involved in research data collection and exploitation play a key role. Funding agencies, researchers, repositories and infrastructure providers, citizen scientists are all contributing to having effective data publishing solutions. In fact, the act of publishing a dataset (e.g., in a repository or in a journal) should be just the beginning of large-scale collaborations. Researchers other than the data owners can “contribute” to the dataset by (a) producing new versions of it, e.g., by enriching the content or correcting errors, (b) reporting their concrete experiences, both successful and unsuccessful, in reusing the dataset for a certain investigation, e.g., by publishing a research paper to be added to the dataset documentation, (c) linking it with other datasets that are complementary or incompatible with respect to a given goal. Research dataset publication is an area that still deserves further investigation also following the parallel evolution of related technologies.

References

1. Assante, M., Candela, L., Castelli, D., Tani, A.: Are scientific data repositories coping with research data publishing? *Data Sci. J.* **15**(6), 1–24 (2016). doi:[10.5334/dsj-2016-006](https://doi.org/10.5334/dsj-2016-006)
2. Austin, C.C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V.K., Murphy, F., Nummerger, A., Raymond, L., Stockhause, M., Tedds, J., Vardigan, M., Whyte, A.: Key components of data publishing: using current best practices to develop a reference model for data publishing. *Int. J. Digit. Libr.* (2016). doi:[10.1007/s00799-016-0178-2](https://doi.org/10.1007/s00799-016-0178-2)
3. Borgman, C.L.: *Big Data, Little Data, No Data. Scholarship in the Networked World*. MIT Press, Cambridge (2015)
4. Candela, L., Castelli, D., Manghi, P., Tani, A.: Data journals: a survey. *J. Assoc. Inf. Sci. Technol.* **66**(9), 1747–1762 (2015). doi:[10.1002/asi.23358](https://doi.org/10.1002/asi.23358)
5. Edmunds, S.C., Christopher I Hunter, P.L., Xiao, S.Z., Davidson, R.L., Nogoy, N., Goodman, L.: Experiences in integrated data and research object publishing using gigadb. *Int. J. Digit. Libr.* (2016). doi:[10.1007/s00799-016-0174-6](https://doi.org/10.1007/s00799-016-0174-6)
6. Harrison, K.A., Wright, D.G., Trembath, P.: Implementation of a workflow for publishing citeable environmental data: successes, challenges and opportunities from a data centre perspective. *Int. J. Digit. Libr.* (2016). doi:[10.1007/s00799-016-0175-5](https://doi.org/10.1007/s00799-016-0175-5)
7. Kitchin, J.R., Van Gulick, A.E., Zilinski, L.D.: Automating data sharing through authoring tools. *Int. J. Digit. Libr.* (2016). doi:[10.1007/s00799-016-0173-7](https://doi.org/10.1007/s00799-016-0173-7)
8. Kratz, J., Strasser, C.: Data publication consensus and controversies. *F1000Research* (2014). doi:[10.12688/f1000research.3979.3](https://doi.org/10.12688/f1000research.3979.3). [v3; ref status: indexed, <http://f1000r.es/4ja>]
9. Lawrence, B., Jones, C., Matthews, B., Pepler, S., Callaghan, S.: Citation and peer review of data: moving towards formal data publication. *Int. J. Digit. Curation* **6**(2), 4–37 (2011). doi:[10.2218/ijdc.v6i2.205](https://doi.org/10.2218/ijdc.v6i2.205)
10. Parsons, M., Fox, P.: Is data publication the right metaphor? *Data Sci. J.* **12**, WDS31–WDS46 (2013)
11. Silvello, G., Bordea, G., Ferro, N., Buitelaar, P., Bogers, T.: Semantic representation and enrichment of information retrieval

- experimental data. *Int. J. Digit. Libr.* (2016). doi:[10.1007/s00799-016-0172-8](https://doi.org/10.1007/s00799-016-0172-8)
12. Van den Eynden, V., Corti, L.: Advancing research data publishing practices for the social sciences: from archive activity to empowering researchers. *Int. J. Digit. Libr.* (2016). doi:[10.1007/s00799-016-0177-3](https://doi.org/10.1007/s00799-016-0177-3)
13. Wright, D.G., Trembath, P., Harrison, K.A.: Meeting the challenge of environmental data publication: an operational infrastructure and workflow for publishing data. *Int. J. Digit. Libr.* (2016). doi:[10.1007/s00799-016-0176-4](https://doi.org/10.1007/s00799-016-0176-4)