Report on the First International Workshop on Reproducible Open Science

Paolo Manghi,
ISTI - Consiglio Nazionale delle Ricerche, Italy
paolo.manghi@isti.cnr.it

Jochen Schirrwagen
Bielefeld University Library, Germany
jochen.schirrwagen@uni-bielefeld.de

Oscar Corcho
Universidad Politecnica de Madrid, Spain
ocorcho@fi.upm.es

Amir Aryani
Australian National Data Service, Australia
amir.aryani@ands.org.au

1. INTRODUCTION

In the last decade, information and communication technology (ICT) advances have deeply affected the scientific process, which increasingly produces and relies on digital research products, such as publications, datasets, experiments, websites, software, blogs, etc. Accordingly, scientific communication has started mutating in order to adapt its mission (and business models) to such new scientific paradigms and benefit from the unprecedented Open Science opportunities that may arise from them: *reproducibility*, i.e., the ability of repeating a digital experiment and reusing its constituent products; and *transparent evaluation*, i.e., the ability of (*i*) effectively evaluating scientific experiments by means of reproducibility and (*ii*) assigning fine--grained scientific reward, based on effective authorship across the overall scientific process. Scientists, research institutions, and funders are pushing for innovative Open Science scholarly communication workflows (i.e., submission, peer-review, access, reuse, citation, and scientific reward), marrying a holistic approach where publishing includes in principle any digital product resulting from a research activity that is relevant to the evaluation and reproducibility of the activity or part of it. Defining, taking up, and supporting Open Science publishing workflows become urgent challenges, to be addressed by ICT solutions capable of fostering and driving radical changes in the way science is developed and disseminated.

The goal of the first International Workshop on Reproducible Open Science[1] was to provide a forum for constructively exploring foundational, organizational and systemic challenges towards the implementation of Open Science publishing principles. Its mission was to contribute to the actual picture of the state of the art approaches and solutions that researchers and practitioners active in these fields have investigated and realized: library and information scientists working on the identification of new publication paradigms, ICT scientists involved in the definition of new technical solutions to these issues, and scientists/researchers who actually demand tools and practices for transparent evaluation and reproducibility of science. The workshop has brought together skills and experiences focusing on the definition and establishment of the next generation scientific communication ecosystem, where scientists can publish research results (including the scientific article, the data, the methods, and any alternative product that may be relevant to the conducted research) in order to enable reproducibility (effective reuse and decrease of cost of science) and rely on novel scientific reward practices.

RepScience2016 has been organized in conjunction with the 20th edition of the International Conference on Theory and Practice of Digital Libraries[2]. Proceedings of the workshop are under publication as a special issue of the Open Access journal D-Lib Magazine[3].

2. WORKSHOP CONTRIBUTIONS

Each submitted contribution was peer-reviewed by three of the seventeen members of the Program Committee and ten were accepted, out of which three reported the results of RDA Working Groups. The workshop structure comprised two invited speak-

[1] *RepScience2016's web site* `http://repscience2016.research-infrastructures.eu`

[2] *TPDL2016's web site*, `http://www.tpdl2016.org`

[3] *D-Lib Magazine* `http://www.dlib.org`

ers and four sessions. In the following we shall first report on the invited talks, then group the presentations according to general topics they covered.

2.1 Invited talks

The workshop had two invited talks respectively covering the foundational and more theoretical aspects of reproducibility and the real case of reproducibility challenges currently studied at CERN's scientific information services.

Carole Goble in the talk entitled "What is Reproducibility? The R* Brouhaha" depicted the challenges of reproducibility in computational science by drawing an analogy between laboratory microscope experiments and e-infrastructure "datascope" experiments. The issues are similar, with "experiments" constituted by materials (e.g., datasets, parameters, algorithm seeds) and methods (e.g., techniques, algorithms, specifications of steps, models); and "set up" constituted by instruments (e.g., codes, services, scripts, libraries, workflows) and laboratory (e.g., e-infrastructure, system software, integrative platforms, engines). The definition of reproducibility is a non-trivial one, and weaker or stronger forms may be defined, depending on the intent of the researchers and the capabilities of the underlying e-infrastructure. Examples are "rerun", i.e., variations on experiment set up to enable robustness, "repeat", i.e., same experiment same laboratory to defend one's thesis, "replicate", i.e., same experiment, same set up, different lab to enable certification, "reproduce", i.e., variations of the same experiment, on different set ups and laboratories, and "re-use", i.e., different experiment using material, methods of the experiment. Overall, reproducibility has a cost, both social/cultural and technological, whose dimensions are portability/preservation, (packaging, containers), access (standards, licensing, PIDs), robustness/versioning (change, variation sensitivity, discrepancy handling), and description (standards, common metadata, ontologies). Finally, the Research Object framework[4] was presented as a possible solution to address these issues.

Sunje Dallmeier-Tiessen in the talk entitled "Enabling reproducible research: community practices, service needs and first lessons learnt" has presented CERN's challenges and vision to provide scientists with an e-infrastructure supporting Open Data principles and analysis preservation and reproducibility. CERN scientific information services serve today a variety of research communities each featuring different but also overlapping requirements on these matters. An important objective is to advocate and establish a culture of open sharing of data and algorithms, to get rid of the current "fear of losing control", by leveraging on the potential and concrete benefits and providing adequate technological support. To this aim, CERN Data Services are developing tools enabling linking data with data (subsets, versions, dynamic data), contributors (who, when, where), articles, institutions, and funders. On the side of analysis preservation and reproducibility, CERN is devising tools for supporting scientists at developing science, since its earliest phases, in such a way that the results will be reproducible, according to a model: save, retrieve, review/compare, and repeat/reproduce. So far, the challenges identified are those of (*i*) granularity, complexity, and dependencies of data and software, (*ii*) identification of solutions for data and software publishing, linking, and citation, and (*iii*) the demanding amount of manual work needed to make experimental material reproducible.

[4]Research Object, `researchobject.org`

2.2 Presentation of contributions

The following sections summarizes the workshop presentations[5] organized according to four themes: *Towards an enabling infrastructure*, *Models and languages*, *Systems*, and *Real-world experiences*.

Towards an enabling infrastructure. Enhancing the current scholarly communication infrastructure and workflows to support Open Science and reproducibility opens up to different visions and questions.

Stephan Pröll presented the paper "Enabling Reproducibility for Small and Large Scale Research Data Sets" where the authors have investigated the problem of guaranteeing transparent citation of subsets of data (e.g., results of queries) from dynamic data sources (e.g., databases). The most intuitive solutions (e.g., digital copies) raise a number of challenges (e.g., time, storage, DOIs/handles) which the framework identified by the authors helps at elegantly describe and solve at different extents, driven by a cost analysis.

Paolo Manghi presented "The Scholix Framework for Interoperability in Data-Literature Information Exchange". Scholix[6] is a framework for enabling exchange of information relative to links between scientific products across sources in the scholarly communication domain. The framework defines an information model and exchange formats for such links to transparently move across independent plat-

[5] *Workshop presentation slides* `http://repscience2016.research-infrastructures.eu/index.php?d=sessions`

[6] *Scholix Framework* `http://www.scholix.org`

forms, scientific domains, and stakeholders (e.g., repositories for data and publications, publishers, research infrastructures, libraries).

Models and languages. Enabling reproducibility requires ways to encode the elements composing the experiments, i.e., scientific products, and possibly the actions, i.e., the steps constituting the experiment.

On this respect, Markus Konkol reported on the paper "Opening the Publication Process with Executable Research Compendia". The authors propose the *executable research compendium* (ERC) as a means to publish and access computational research. ERC provides a new standardisable packaging mechanism which combines data, software, text, and a user interface description. As similar approaches to research objects or packages, ERC aims at satisfying needs of authors, readers, publishers, curators, and preservationists, in terms of scientific evaluation, reward, visibility, and reproducibility.

Markus Stocker presented his work "From Data to Machine Readable Information Aggregated in Research Objects". Data interpretation is an important process in scientific workflows, where scientists are called to interpret data (often) collected using large-scale environmental monitoring infrastructures to gain information about the monitored environment. Such information is typically represented to suit human consumption, while the authors propose an encoding into machine readable information objects that builds on the Research Object framework.

Paolo Manghi described the results of "FLARE: a flexible workflow language for research e-infrastructures", where the authors defined FLARE, a workflow language for the specification (and execution) of a scientific process in highly-heterogeneous environments, i.e., e-infrastructures whose workflow are partly manual and automated. FLARE lays in between business process modelling languages, i.e., high-level specifications of a reasoning, protocol, or procedure, and workflow execution languages, i.e., machine-readable specifications of computational steps executable by dedicated engines. FLARE tools allows the creation and sharing of hybrid workflows and their execution, via "web wizards" guiding the scientists through the manual and automated execution of the individual steps.

Systems. This session focused on new generation repositories required for depositing, sharing, and accessing the products of science, be them datasets or methods, in such a way they can be properly reused and experiments reproduced.

Vidya Ayer presented the article "Conquaire: Towards an architecture supporting continuous quality control to ensure reproducibility of research" reporting on the preliminary results of the project Conquaire, aiming at delivering an infrastructure based on subject-specific components offering functionalities for data deposition and versioning, enabling automated and discipline-specific quality checks over the data. The system architecture relies on a DCVS system for storing data and on continuous integration principles to ensure data quality.

Sheeba Samuel presented "Towards Reproducibility of Microscopy Experiments". The authors have realized an information system (based on the existing OMERO system[7]) that supports scientists in the domain of microscopy techniques at following a rigorous methodology for collecting documentation and research data to the level necessary to reproduce scientific experiments. Although the approach addresses the specific requirements of an interdisciplinary team of scientists from experimental biology to store, manage, and reproduce the workflow of their research experiments, it can also be extended to the requirements of other scientific communities.

Real-world experiences. Many scientists are today using tools to (*i*) publish their research products in order to achieve degrees of reproducibility or (*ii*) search out for the products needed to reproduce experiments. Such real-world experiences make a fertile ground where to identify common requirements for an open and reproducible science.

Jingbo Wang reported on two experiences in different contexts. The first was titled "Graph connections made by RD-Switchboard using NCIs metadata", where she demoed connectivity graphs linking datasets, papers, authors, and grants, built using the Research Data Switchboard[8] using NCIs metadata database[9]. By means of such graphs, the NCI database was enriched with critical but missing information in the network of researchers and article-dataset links, thereby enhancing the search capabilities of the system and enabling fit-for-purpose (e.g., research topic/context-driven) dataset discovery. The second experience was titled "Supporting Data Reproducibility at NCI Using the Provenance Capture System". The National Computational Infrastructure (NCI) of Australia has realised

[7] *Open Microscopy Environment Remote Objects* `http://www.openmicroscopy.org/site/products/omero`
[8] *The Research Data Switchboard* `http://www.RD-Switchboard.org`
[9] *The Australian National Computational Infrastructure* `https://nci.org.au/`

a system supporting researchers at modelling their workflows, including those that have been used to create data extracts (e.g., queries to databases), using a standards-based provenance representation. This information, combined with access to the original dataset and other related information systems, allows data extracts to be easily regenerated to support experiment reproducibility, limiting preservation of data extracts to very specific cases.

Finally, Jan H. Höffler presented the experience of "ReplicationWiki: Improving Transparency in Social Sciences Research", an attempt to compensate in the field of empirical social sciences the lack of scientific reward regarding authoring of "replicable studies" and authors of the required "replicable products". ReplicationWiki[10] today documents 2500 empirical studies and the relative replication products found in the literature, so far mainly in economics. The wiki is populated by professors and students in economics across several participating institutions, with the aim of establishing the culture of open reproducible science, as well as facilitating academic teaching, and setting incentives for replicability and replication.

3. WORKSHOP DISCUSSION

The concluding brainstorming session brought up two main relevant considerations and future issues with respect to open and reproducible science.

Experimental context (or set up). Computational reproducibility spins around the concept of experimental context (or set up), namely the components required to execute an experiment by applying computation over data, or to evaluate the quality of digital products, be them data or computation. The experimental context must be shared by scientists, to ensure a common ground of evaluation and execution, thereby enabling transparent evaluation, comparisons, and reproducibility. The ability of sharing an experimental context to the largest and agnostic audience entails a trade off between "ability to adopt" and "portability of experiments". On the one hand scientists can assume to share experiments and relative products based on common and agreed on experimental context and methodologies. This will allow them to compile minimal descriptions on how products are to be combined to reproduce and experiment, hence making it easy for scientists to adopt reproducibility practices, but in turn making the experiments effectively reproducible solely to scientists aware of the underlying "commons". On the other hand, scientists can instead share the components and the relative descriptions, so as to ensure the experiments can be reproduced beyond the borders of their community. The extent of details for such descriptions may ensure a broader coverage but in general hinders the adoption by scientists, e.g., tedious metadata provision or evolution of external software components. Identifying the optimal balance is not trivial and also depends on the maturity of a common experimental context, typically research e-infrastructures, and its components.

Roadmap to reproducibility. Open Science has become more and more relevant and appealing for all stakeholders of scientific communication, i.e., research and academic organizations, researchers, publishers, libraries, and funders. The first results are visible with the strong shift of funders and organizations towards mandates for Open Access to publications, which started less than a decade ago, and more recently to ensuring research data sharing (i.e., deposition, description, and preservation), e.g., Data Pilot of the European Commission. Countries, libraries, and research communities (research infrastructures) are moving towards economy of scale solutions for the storage of data, and several initiatives are suggesting methodologies and cost/sustainability analyses that may facilitate this highly expensive process. As reproducibility is gaining relevance and appeal among the very same stakeholders, it is reasonable to expect that similar initiatives will face the problem of how a research community or a library can initiate supporting reproducibility of science for a community or multiple communities starting from a given e-infrastructural setting.

4. ACKNOWLEDGMENTS

Special thanks are also due to the members of the program committee[11] whose research experience largely contributed in making this workshop an attractive venue and constructive experience. Moreover, our sincere gratitude goes to all participants, invited speakers and authors, whose enthusiasm and vision constitute the soul of this workshop and future research in the field. This event was funded by the RDA Europe project[12] under the H2020 program of the European Commission (grant agreement: 653194).

[10] *ReplicationWiki* `http://replication.uni-goettingen.de/`

[11] For *RepScience Program Committee* follow `http://repscience2016.research-infrastructures.eu`

[12] *RDA Europe*, `http://europe.rd-alliance.org`