Report from Dagstuhl Seminar 16041

Reproducibility of Data-Oriented Experiments in e-Science

Edited by

Juliana Freire[1], Norbert Fuhr[2], and Andreas Rauber[3]

1 **New York University, US,** juliana.freire@nyu.edu
2 **Universität Duisburg-Essen, DE,** norbert.fuhr@uni-due.de
3 **TU Wien, AT,** rauber@ifs.tuwien.ac.at

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 16041 "Reproducibility of Data-Oriented Experiments in e-Science". In many subfields of computer science, experiments play an important role. Besides theoretic properties of algorithms or methods, their effectiveness and performance often can only be validated via experimentation. In most of these cases, the experimental results depend on the input data, settings for input parameters, and potentially on characteristics of the computational environment where the experiments were designed and run. Unfortunately, most computational experiments are specified only informally in papers, where experimental results are briefly described in figure captions; the code that produced the results is seldom available.

This has serious implications. Scientific discoveries do not happen in isolation. Important advances are often the result of sequences of smaller, less significant steps. In the absence of results that are fully documented, reproducible, and generalizable, it becomes hard to re-use and extend these results. Besides hindering the ability of others to leverage our work, and consequently limiting the impact of our field, the absence of reproducibility experiments also puts our reputation at stake, since reliability and validity of empiric results are basic scientific principles.

This seminar brought together experts from various sub-fields of computer science to create a joint understanding of the problems of reproducibility of experiments, discussing existing solutions and impediments, and proposing ways to overcome current limitations.

Seminar January 24–29, 2016 – http://www.dagstuhl.de/16041
1998 ACM Subject Classification A.1 Introductory and Survey
Keywords and phrases Documentation, Reliability, Repeatibility, Replicability, reproducibility, Software
Digital Object Identifier 10.4230/DagRep.6.1.108
Edited in cooperation with Daniel Garijo

1 Executive Summary

Norbert Fuhr
Juliana Freire
Andreas Rauber

In many subfields of computer science, experiments play an important role. Besides theoretical properties of algorithms or methods, their effectiveness and performance often can only be validated via experimentation. In most of these cases, the experimental results depend

on the input data, settings for input parameters, and potentially on characteristics of the computational environment where the experiments were designed and run. Unfortunately, most computational experiments are specified only informally in papers, where experimental results are briefly described in figure captions; the code that produced the results is seldom available.

This has serious implications. Scientific discoveries do not happen in isolation. Important advances are often the result of sequences of smaller, less significant steps. In the absence of results that are fully documented, reproducible, and generalizable, it becomes hard to re-use and extend these results. Besides hindering the ability of others to leverage our work, and consequently limiting the impact of our field, the absence of reproducibility experiments also puts our reputation at stake, since reliability and validity of empiric results are basic scientific principles.

Reproducible results are not just beneficial to others – in fact, they bring many direct benefits to the researchers themselves. Making an experiment reproducible forces the researcher to document execution pathways. This in turn enables the pathways to be analyzed (and audited). It also helps newcomers (e.g., new students and post-docs) to get acquainted with the problem and tools used. Furthermore, reproducibility facilitates portability, which simplifies the dissemination of the results. Last, but not least, preliminary evidence exists that reproducibility increases impact, visibility and research quality.

However, attaining reproducibility for computational experiments is challenging. It is hard both for authors to derive a compendium that encapsulates all the components (e.g., data, code, parameter settings, environment) needed to reproduce a result, and for reviewers to verify the results. There are also other barriers, from practical issues – including the use of proprietary data, software and specialized hardware, to social – for example, the lack of incentives for authors to spend the extra time making their experiments reproducible.

This seminar brought together experts from various sub-fields of Computer Science as well as experts from several scientific domains to create a joint understanding of the problems of reproducibility of experiments, discuss existing solutions and impediments, and propose ways to overcome current limitations.

Beyond a series of short presentations of tools, state of the art of reproducibility in various domains and "war stories" of things not working, participants specifically explored ways forward to overcome barriers to the adoption of reproducibility. A series of break-out sessions gradually built on top of each other, (1) identifying different types of repeatability and their merits; (2) the actors involved and the incentives and barriers they face; (3) guidelines for actors (specifically editors, authors and reviewers) on how to determine the level of reproducibility of papers and the merits of reproduction papers; and (4) the specific challenges faced by user-oriented experimentation in Information Retrieval.

This led to the definition of according typologies and guidelines as well as identification of specific open research problems. We defined a set of actions to reach out to stakeholders, notably publishers and funding agencies as well as identifying follow-up liaison with various reproducibility task forces in different communities including the ACM, FORCE11, STM, Science Europe.

The key message resulting from this workshop, copied from and elaborated in more detail in Section 6.5 is:

> *Transparency, openness, and reproducibility are vital features of science. Scientists embrace these features as disciplinary norms and values, and it follows that they should be integrated into daily research activities. These practices give confidence in the work; help research as a whole to be conducted at a higher standard and be undertaken more*

efficiently; provide verifiability and falsifiability; and encourage a community of mutual cooperation. They also lead to a valuable form of paper, namely, reports on evaluation and reproduction of prior work. Outcomes that others can build upon and use for their own research, whether a theoretical construct or a reproducible experimental result, form a foundation on which science can progress. Papers that are structured and presented in a manner that facilitates and encourages such post-publication evaluations benefit from increased impact, recognition, and citation rates.

Experience in computing research has demonstrated that a range of straightforward mechanisms can be employed to encourage authors to produce reproducible work. These include: requiring an explicit commitment to an intended level of provision of reproducible materials as a routine part of each paper's structure; requiring a detailed methods section; separating the refereeing of the paper's scientific contribution and its technical process; and explicitly encouraging the creation and reuse of open resources (data, or code, or both).

2 Table of Contents

3 Overview of Tools

3.1 noWorkflow

Vanessa Braganholo (Fluminense Federal University, BR)

Main reference L. G. P. Murta, V. Braganholo, F. S. Chirigati, D. Koop, J. Freire, "noWorkflow: Capturing and Analyzing Provenance of Scripts" in Proc. of the 5th Int'l Provenance and Annotation Workshop (IPAW'14), LNCS, Vol. 8628, pp. 71–83, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-16462-5_6
URL https://github.com/gems-uff/noworkflow

Capturing provenance in scientific experiments has been a major concern both for result comprehension and reproducibility. Although the scientific community often writes experiments using script languages, most of the existing provenance capture approaches require scientists to change the way they work, by wrapping their experiments in scientific workflow systems, installing version control systems, or modifying and instrumenting their scripts, which may be laborious and error prone. As a solution to these problems, noWorkflow is a non-intrusive tool that transparently captures provenance of scripts, keeping data about how they evolve over time, as well as about their execution. It systematically monitors script execution without requiring any modification to the source code. Provenance data can then be analyzed using graphical interfaces, SQL or Prolog queries. We also provide ways of comparing two different executions, highlighting their differences, and support Jupyter notebooks[1].

3.2 ReproMatch

Fernando Chirigati (NYU Tandon School of Engineering, US) and Juliana Freire (New York University, US)

Joint work of Fernando Chirigati, Tommy Ellqvist, Juliana Freire
URL http://repromatch.poly.edu/

ReproMatch stands for Reproducibility Match and it was designed as a search engine to help you find the tool (or tools) that best matches your reproducibility needs. The tools in the ReproMatch catalog are classified according to different reproducibility tasks, which we organized in a taxonomy[2]. Researchers can submit information about new tools, or corrections to existing information.

[1] http://jupyter.org/
[2] http://repromatch.poly.edu/task-descriptions/

3.3 ReproZip: Computational Reproducibility With Ease

Fernando Chirigati (NYU Tandon School of Engineering, US) and Juliana Freire (New York University, US)

Joint work of Fernando Chirigati, Rémi Rampin, Juliana Freire, Dennis Shasha
Main reference F. Chirigati, R. Rampin, D. Shasha, J. Freire, "ReproZip: Computational Reproducibility With Ease", in Proc. of the 2016 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD'16), Demo Session, pp. 2085–2088, ACM, 2016
URL http://dx.doi.org/10.1145/2882903.2899401
URL https://vida-nyu.github.io/reprozip/

ReproZip provides a lightweight solution that makes experiments reproducible without forethought. Researchers can create an experiment without thinking about reproducibility and use ReproZip to make it reproducible and portable to other machines. In a nutshell, ReproZip automatically and transparently captures the provenance of an existing experiment by tracing system calls, and uses this information to create a lightweight reproducible package that includes only the required files needed for its reproduction. It also adds important features and contributions, including: (1) portability – ReproZip provides unpackers that allow researchers to automatically create a VM or a Docker container encompassing the experiment, thus allowing it to be reproduced in different operating systems; it also generates a workflow specification for the experiment, which can be used to easily change parameters or modify the original dataflow; (2) extensibility – by implementing new unpackers, researchers can easily extend ReproZip to port experiments to other environments and systems while keeping compatibility with existing packaged experiments; (3) modifiability – ReproZip automatically identifies input files, parameters, and output files, allowing researchers to easily modify these for reuse purposes; and (4) usability – researchers have control over the collected trace and can customize the reproducible package; the tool also provides command-line interfaces that make it easier to setup, reproduce, and modify the original experiment.

ReproZip has been recommended for the SIGMOD Reproducibility Review[3], and listed on the Artifact Evaluation Process guidelines[4].

3.4 Janiform: Intra-Document Analytics for Reproducible Research

Jens Dittrich (Universität des Saarlandes, DE)

URL https://github.com/uds-datalab/PDBF

Peer-reviewed publication of research papers is a cornerstone of science. However, one of the many issues of our publication culture is that our publications only publish a summary of the final result of a long project. This means that we put well-polished graphs describing (some) of our experimental results into our publications. However, the algorithms, input datasets, benchmarks, raw result datasets, as well as scripts that were used to produce the graphs in the first place are rarely published and typically not available to other researchers. Often they are only available when personally asking the authors. In many cases, however,

[3] http://db-reproducibility.seas.harvard.edu/
[4] http://www.artifact-eval.org/guidelines.html

they are not available at all. This means from a long workflow that led to producing a graph for a research paper, we only publish the final result rather than the entire workflow. This is unfortunate and has been criticized in various scientific communities. In this demo we argue that one part of the problem is our dated view on what a "document" and hence "a publication" is, should, and can be. As a remedy, we introduce portable database files (PDbF). These files are jani-form, i.e. they are at the same time a standard static pdf as well as a highly dynamic (offline) HTML-document. PDbFs allow you to access the raw data behind a graph, perform OLAP-style analysis, and reproduce your own graphs from the raw data – all of this within a portable document. We demo a tool allowing you to create PDbFs smoothly from within LATEX. This tool allows you to preserve the workflow of raw measurement data to its final graphical output through all processing steps. Notice that this pdf already showcases our technology: rename this file to ".html" and see what happens (currently we support the desktop versions of Firefox, Chrome, and Safari). But please: do not try to rename this file to ".ova" and mount it in VirtualBox.

3.5 DIRECT and LOD-DIRECT

Nicola Ferro (University of Padova, IT)

Main reference M. Agosti, N. Ferro, "Towards an Evaluation Infrastructure for DL Performance Evaluation", in G. Tsakonas, C. Papatheodorou (eds.), "Evaluation of Digital Libraries: An Insight to Useful Applications and Methods," Chandos Publishing, Oxford, 2009.

Distributed Information Retrieval Evaluation Campaign Tool (DIRECT[5]) is a system which models IR experimental data and manages all the steps of an IR evaluation campaign, like creation of the topics, submission of system runs, creation of relevance judgements, computation of performance measures and so on. DIRECT not only supports IR evaluation campaigns but takes also care of archiving the IR experimental data in order to make the accessibile and referenceable for future re-use. At the time of writing, DIRECT counts about 35 millions documents, 14 thousands topics, around 4 million relevance judgements, 5 thousands experiments and 20 millions measures. This data has been inserted and used by about 1,500 researchers from more than 70 countries world-wide. Overall, DIRECT counts around 650 visitors who accessed and downloaded the data. LOD-DIRECT[6] is an evolution of DIRECT to model and make available a subset of its IR experimental data as Linked Open Data.

[5] http://direct.dei.unipd.it/
[6] http://lod-direct.dei.unipd.it/

3.6 Research Objects, FAIRDOM and SEEK4Science

Carole Goble (University of Manchester, GB)

Main reference S. Bechhofer, J. Ainsworth, J. Bhagat, I. Buchan, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, C. Goble, D. Michaelides, P. Missier, S. Owen, D. Newman, D. De Roure, S. Sufi, "Why Linked Data is Not Enough for Scientists", in Future Generation Computer Systems, Vol. 29(2):599–611, 2013.
URL http://dx.doi.org/10.1016/j.future.2011.08.004
URL http://www.researchobject.org/

Making scientific experiments FAIR – findable, accessible, interoperable, reusable – is hard. To be reproducible means bundling, along with the narrative, the experimental methods, computational codes, data, algorithms, workflows, scripts – some of which might be hosted remotely, in many different repositories and with the potential to change. In this talk I presented a framework for Research Objects[7] – a metadata framework for bundling, porting and linking resources and representing the context of experiments. Research Objects have a manifest and a container. The manifest uses off the shelf standards and ontologies to construct the manifest and describe the content held in a container. The description is tailored to the type of Research Object, for example a Systems Biology Experiment or a computational workflow. The description broadly covers provenance, dependencies, versioning and checklists (aka reporting guidelines). Containers are off the shelf packaging platforms like Zip, Docker, Bagit or bespoke platforms that are "RO native".

In the talk I presented FAIRDOMHub[8], a Systems Biology Commons for supporting the reporting and sharing of models, data and Standard Operating Procedures arising from projects. It is built on the RO-compliant SEEK4Science[9] commons and cataloguing platform. The system gathers the metadata needed for reproducible modelling. Moreover it supports the packaging up of content to be exported and deposited into other repositories like Zenodo.

Finally I presented other implementations of the RO framework: the COMBINE Archive for Systems Biology models which uses zip, Workflow RO bundles using Bagit, which is part of the Common Workflow Language, the STELAR Asthma eLab which uses Docker and ATLAS LHC Experiments, which uses Docker and CDE.

3.7 Moore/Sloan Data Science Environments Projects

Randall J. LeVeque (University of Washington – Seattle, US)

URL https://reproduciblescience.org

The Moore and Sloan Foundations are funding a joint project between the University of Washington, NYU, and Berkeley on creating better environments for data science research within academics. There is a joint working group on Reproducibility and Open Science[10] that is developing several tools of possible interest. This repository[11] contains short descriptions

[7] http://www.researchobject.org/
[8] http://www.fairdomhub.org/
[9] http://www.seek4science.org/
[10] https://reproduciblescience.org/
[11] https://github.com/BIDS/repro-case-studies/tree/submissions/case-studies

and diagrams of workflows as examples of how researchers from many different disciplines have approached collaboration, data management, and sharing of code and data. I also gave a status report on a project to develop a system of badges to acknowledge the steps people take to make their work open and reproducible[12], and as means to collect links of examples others can follow. The main goals are to provide incentives and education about what is possible.

3.8 YesWorkflow

Bertram Ludäscher (University of Illinois at Urbana-Champaign, US)

Main reference T. McPhillips, T. Song, T. Kolisnik, S. Aulenbach, K. Belhajjame, R. K. Bocinsky, Y. Cao, J. Cheney, F. Chirigati, S. Dey, J. Freire, C. Jones, J. Hanken, K. W. Kintigh, T. A. Kohler, D. Koop, J. A. Macklin, P. Missier, M. Schildhauer, C. Schwalm, Y. Wei, M. Bieda, B. Ludäscher, "YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts", in Int'l Journal of Digital Curation, 10(1):298–313, 2015; pre-print available as arXiv:1502.02403v1 [cs.SE].
URL http://dx.doi.org/10.2218/ijdc.v10i1.370
URL http://arxiv.org/abs/1502.02403v1
URL http://yesworkflow.org/

Traditional workflow automation approaches are based on scripting languages and remain very popular, e.g., due to the availability of countless libraries, a gentle learning curve (e.g. for Python), and – last not least – the high productivity that users of scripting languages experience. New interactive environments such as iPython/Jupyter add further to the popularity of script-based approaches. The YesWorkflow toolkit aims to bring some of the advantages of scientific workflow systems to researchers who use scripting languages such as Python, R, or Matlab. YesWorkflow enables script writers to reveal the computational steps and flow of data within the scripts they write (i.e., prospective provenance) by annotating their code with special comments. YesWorkflow extracts and analyzes these comments, represents the scripts in terms of entities based on a typical scientific workflow model, and provides graphical renderings of this view of the scripts. YesWorkflow additionally enables researchers to reconstruct retrospective provenance of data products used by scripts, and to query both prospective and retrospective provenance, allowing users powerful insights into their script-based models and simulation runs.

[12] http://uwescience.github.io/reproducible/badges.html

3.9 Process Migration Framework

Rudolf Mayer (SBA Research – Wien, AT)

Main reference A. Rauber, T. Miksa, R. Mayer, S. Proell, "Repeatability and Re-Usability in Scientific Processes: Process Context, Data Identification and Verification", in Proc. of the 17th Int'l Conf. on Data Analytics and Management in Data Intensive Domains (DAMDID'15), CEUR Workshop Proceedings, Vol. 1536, pp. 246–256, 2015.
URL http://ceur-ws.org/Vol-1536/paper33.pdf

The process migration framework (PMF) aims to make a process such an e-Science experiment repeatable, by extracting the process from its original environment, and enabling to redeploy it in a dedicated virtual machine. To this end, PMF logs the resources a process is utilising during execution time. Based on these logs, required files (executables and data) are identified and copied to the new environment, where the process can be run again. In addition, the PMF creates a higher-level semantic description of the process. Based on the execution trace of the process, the PMF also creates a basic process model visualising the sequence of steps identified.

By further analysing and aggregating the identified resources to for example Linux software packages, the PMF creates a smaller and more human-readable description of the process dependencies.

Finally, specific emphasize is put on identifying and discovering resources external to the original system, such as calls to web services e.g. for data processing, or the connection to a database server.

This analysis informs the user on resources that are outside of his direct control, and on which manual emphasize on ensuring the long-term availability needs to be put on.

3.10 ROHub

Raul Antonio Palma de Leon (Poznan Supercomputing and Networking Center, PL)

Main reference R. Palma, O. Corcho, J. M. Gómez-Pérez, C. Mazurek, "ROHub – A Digital Library of Research Objects Supporting Scientists Towards Reproducible Science", in Semantic Web Evaluation Challenge at ESWC'14, Communications in Computer and Information Science, Vol. 475, pp. 77–82, 2014.
URL http://dx.doi.org/10.1007/978-3-319-12024-9_9
URL http://www.rohub.org

ROHub is a digital library system supporting the storage, lifecycle management, sharing and preservation of research findings via Research Objects. It includes different features to help scientists throughout the research lifecycle: (i) to create and maintain ROs that are compliant with predefined quality requirements so that they can be interpreted and reproduced in the future; (ii) to collaborate along this process; (iii) to publish and search these objects and their associated metadata; (iv) to manage their evolution; and (v) to monitor and preserve them through time ensuring that they will remain accessible and reusable. ROHub has a modular structure, comprising a backend (rodl) that exposes a set of RESTful APIs and a SPARQL endpoint; and a frontend Web Portal providing a graphical interface for end users (scientists/researchers/etc.)

3.11 TIRA

Martin Potthast (Bauhaus-Universität Weimar, DE) and Benno Stein (Bauhaus-Universität Weimar, DE)

Main reference A. Hanbury, H. Müller, K. Balog, T. Brodt, G. V. Cormack, I. Eggel, T. Gollub, F. Hopfgartner, J. Kalpathy-Cramer, N. Kando, A. Krithara, J. Lin, S. Mercer, M. Potthast, "Evaluation-as-a-Service: Overview and Outlook", arXiv:1512.07454v1 [cs.CY].
URL http://arxiv.org/abs/1512.07454v1
URL http://www.tira.io

The TIRA experimentation platform is a web service that supports organizers of shared tasks in computer science to accept the submission of executable software [1]. TIRA automates software submission to a point at which it imposes no significant overhead on organizers and participants alike. From the start, TIRA has been in active use: since 2012, TIRA is employed for the PAN shared task series on digital text forensics [2], and as of 2015, TIRA hosts the annual shared task of the CoNLL conference. TIRA's technology stack relies primarily on a combination of low-level (LXC, Docker) and high-level (hypervisor) virtualization technology, server-side control software, and a web front end that allow for the remote management of shared tasks. TIRA distributes virtual machines across a number of TIRA hosts, which are remote-controlled by a master server. Every virtual machine is accessible from the outside by participants via SSH and remote desktop, and both Linux and Windows are supported as guest operating systems. This allows for a variety of development environments, so that participants in a shared task can directly work as they usually would. TIRA further hosts the datasets used in a shared task, split into training datasets and test datasets. The former are publicly visible to participants, including ground truth data, whereas the latter are accessible only to participant software in a secure execution environment that protects the test datasets from leaking to participants. Before executing the software on a test dataset, TIRA clones its virtual machine into the secure execution environment, where Internet access is disabled. After the software successfully executed on the test dataset, its output is copied, whereas the cloned virtual machine is deleted to prevent any potentially private files on its virtual hard disk from exiting the execution environment. In this way, participants in a shared task can run their software on the shared task's test datasets, whereas its organizers need not worry about the data leaking. TIRA also enables the use of proprietary and sensitive data as evaluation data. Finally, TIRA hosts a special purpose virtual machine for each shared task, where the organizer deploys software for performance measurement. The output of participant software that was executed on a training dataset or a test dataset is fed directly into the performance measurement software at the click of a button. The results are displayed on a dedicated web page for the shared task on TIRA's web front end.

References

1 Tim Gollub, Benno Stein, and Steven Burrows. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12), pages 1125–1126, August 2012. ACM.

2 Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In Evangelos Kanoulas et al, editors, Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14), pages 268–299, Berlin Heidelberg New York, September 2014. Springer.

3.12 CodaLab

Evelyne Viegas (Microsoft Research – Redmond, US)

Main reference https://github.com/codalab

CodaLab is an open source platform which goal is to accelerate the rate of research by enabling collaboration among researchers and scientists across disciplines and make science truly reproducible. CodaLab Worksheets[13] focuses on accelerating data-driven research and making it more sound while enabling scientists to publish their research as executables papers with full provenance on data and code. CodaLab competitions[14] is a powerful framework for running data-driven competitions that involve result and/or code submission. Users can either participate in an existing competition or host a new competition as an organiser. CodaLab enables coopetitions, a new collaborative framework where users with different expertise can work together in a new environment favouring cross-pollination of ideas.

4 State of the Art in Different Areas of CS

4.1 State of the art trade-offs in IR Research

Shane Culpepper (RMIT University – Melbourne, AU)

URL https://github.com/lintool/IR-Reproducibility

This talk briefly presented a state-of-the-art comparison of ad-hoc search engines for a common TREC task. By aggregating results from the IR Reproducibility Challenge in the 2015 ACM SIGIR Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR), we contrast fully reproducible baseline runs and "best known" submissions from the TREC Adhoc Search Task between 2004–2006.

4.2 Managing and Curating IR Experimental Data

Nicola Ferro (University of Padova, IT)

Information Retrieval (IR) is a discipline deeply rooted in experimentation since its inception and, over the time, it has developed robust and shared methodologies for conducting experiments, relying on the so-called Cranfield Paradigm. In particular, the adoption of large-scale and shared experimental collections, typically used in international evaluation

[13] https://github.com/codalab/codalab-worksheets/wiki
[14] https://competitions.codalab.org/

campaigns like TREC[15], CLEF[16], and NTCIR[17] and then available for further re-use by the community, provide the means for running comparable experiments. This experimental paradigm gives raise to three targets for reproducibility:

- **experimental collections**: they consists of documents, topics, which surrogate real user information needs, and relevance judgements, which determine which documents are relevant to which topics. Experimental collections are an integral part of the experimental design and they are often used for many different purposes after their creation. It is thus important to understand their limitations and their generalizability as well as the reproduce the process that led to their creation. This is not always trivial since, for example, topics may be sampled from real system logs or relevance judgements are made by humans and, more and more often, using crowdsourcing.
- **system runs**: they are the most common target for reproducibility since they are what is discussed in papers proposing new methods and algorithms.
- **meta-evaluation experiments**: IR has a strong tradition in assessing its own evaluation methodologies, such as robustness of the experimental collections, reliability of the adopted evaluation measures or appropriateness of the adopted statistical analysis methods. All these investigations strongly rely on existing experimental collections and gathered systems runs and their reproducibility should be a key concern, since they probe our own experimental methods.

All the above mentioned three targets for reproducibility heavily depend on experimental data. Unfortunately, even if IR has a long tradition in ensuring the due scientific rigor is guaranteed in producing such data, it has not a similar tradition in managing and taking care of such valuable data. There currently are several barriers to proper data curation for reproducibility. There is a lack of common formats for modelling and describing the experimental data as well as almost no metadata (descriptive, administrative, copyright, etc.) for annotating and enriching them. The semantics of the data themselves is often not explicit and it is demanded to the scripts typically used for processing them, which are often not well documented, rely on rigid assumptions on the data format or even on side effects in processing the data. Finally, IR lacks a commonly agreed mechanism for citing and linking data to the papers describing them.

All these issues may be addressed by adapting solutions developed in other fields with similar problems but the biggest issue is the community itself, which would need to evolve its experimental methodologies to take into account reproducibility and the actions needed to guarantee it. This calls for an orchestrated effort and a cultural change which are the most compelling challenges towards a proper management and curation of experimental data.

[15] http://trec.nist.gov/
[16] http://www.clef-initiative.eu/
[17] http://research.nii.ac.jp/ntcir/index-en.html

4.3 Reproducibility in Databases

Juliana Freire (New York University, US), Fernando Chirigati (NYU Tandon School of Engineering, US), Jens Dittrich (Universität des Saarlandes, DE), and Tanu Malik (University of Chicago, US)

While some authors in the community make their code available, reproducibility has not been widely adopted. In 2008, SIGMOD instituted a reproducibility review: authors of accepted papers are invited to submit their experiments for an independent review. Over the years, the rate of participation has varied, usually fewer than 35% of the accepted papers were evaluated for reproducibility. Authors have argued that making papers reproducible requires too much work. Reviewers have faced many challenges to install and run the experiments, due to incomplete instructions, missing dependencies, incompatible operating system [1]. Most of the reproducibility submissions include source code and data, together with instructions on how to build them. However, it is often the case that some of them fail the set-up phase because dependencies are not made explicit, which puts a high burden on reviewers; few submissions include a VM and a workflow. Authors complained that the process requires too much work for the benefit derived – and the process does require a substantial amount of work if there is no planning for reproducibility since the beginning of the project. The SIGMOD reproducibility review was revamped in 2015[18] by Stratos Idreos, and ACM now allows papers to be marked as reproducible in ACM Digital Library. Elsevier's Information Systems Journal now also has a Reproducibility Section, led by Dennis Shasha and Fernando Chirigati, where some accepted papers are invited to submit a reproducibility paper, explaining in detail how to run it and what the effort was to make it reproducible. The paper is submitted together with source code (GitHub), data (GitHub or Mendeley Data), and a Docker container/ReproZip package/VM (Mendeley Data) to ease the review process. Reviewers also become co-authors of the report, as they describe in the paper the efforts in reproducing the published experiment. This approach provides incentives for both authors and reviewers: by making their experiment reproducible, authors have a new paper, and by having to review and reproduce the experiment, reviewers are also included in the same publication [2, 3].

Some technical challenges must still be solved to help databases experiments to be reproduced. First, it is still unclear how to reproduce experiments that access networked resources, including Web services, remote databases, and HDFS. Tools such as LDV (Lightweight Database Virtualization) are a step towards this, but must be made more general to broaden the adoption, since currently users must use Postgres. A second challenge is how to enable reproducibility for distributed applications (e.g.: MPI, Hadoop, Spark). Many more variables and configurations are involved, and performance results are often important. Using a different compiler, compilation flags, or architecture may change these results. There are reproducibility challenges in data integration and data analysis, where in data from a large variety of sources are either aggregated into a database and/or modified/analyzed respectively along the process. Here reproducibility tools for database must interact with file-system tools to ensure cross-system reproducibility.

A noteworthy effort in the database community is the inauguration of the experiments&analyses track at VLDB 2008. This is a special conference track allowing researchers

[18] http://db-reproducibility.seas.harvard.edu/

to submit experimental studies, rebuttals as well as negative results. Accepted papers on this track are not distinguished from standard "research" papers in the final conference program and the proceedings. Papers on E&A may run code used in other papers as blackboxes. However, in general, also often whiteboxing is done (in the sense of algorithm and implementation analysis as well as re-implementations) to make the experimental comparison and the comparability of algorithms and systems stronger. The E&A track has become quite popular in recent years. So far at least two best paper awards have been given to E&A track papers.

References

1 P. Bonnet, S. Manegold, M. Bjørling, W. Cao, J. Gonzalez, J. Granados, N. Hall, S. Idreos, M. Ivanova, R. Johnson, D. Koop, T. Kraska, R. Müller, D. Olteanu, P. Papotti, C. Reilly, D. Tsirogiannis, C. Yu, J. Freire, and D. Shasha. Repeatability and workability evaluation of sigmod 2011.SIGMOD Record, 40(2):45–48, 2011.

2 A. Wolke, M. Bichler, F. Chirigati, V. Steeves. Reproducible experiments on dynamic resource allocation in cloud data centers Information Systems, Available online 7 January 2016.

3 A. Wolke, B. Tsend-Ayush, C. Pfeiffer, M. Bichler. More than bin packing: Dynamic resource allocation strategies in cloud data centers Information Systems, Volume 52, August–September 2015, Pages 83–95

4.4 Reproducibility using semantics: An overview

Daniel Garijo (Technical University of Madrid, ES)

URL http://www.slideshare.net/dgarijo/reproducibility-using-semantics-an-overview

The Semantic Web has helped to create knowledge bases that link and facilitate accessibility to research data. However, how can it help scientists to make their experiments reproducible? This talks introduces an overview of the different initiatives led by the Ontology Engineering Group (UPM) to address reproducibility by using semantics. The initiatives are distributed among several disciplines, including the formalization of laboratory protocols to detect ambiguity and missing descriptions [1], documentation and publication of scientific workflows and their resources [2], capturing the infrastructure needed to reproduce a scientific experiment [3], achieving long term preservation of research objects and conditional access to resources based on their intellectual property rights[19].

References

1 O. Giraldo , A. García, J. Figueredo and O. Corcho. Using Semantics and NLP in Experimental Protocols. 8th Semantic Web Applications and Tools for Life Sciences International Conference (SWAT4LS 2015). Cambridge, UK. 2015

2 D.Garijo and Y. Gil. A new approach for publishing workflows: abstractions, standards, and linked data. Proceedings of the 6th workshop on Workflows in support of large-scale science (WORKS11), pp 47–5, Seattle, 2011.

3 I. Santana-Perez, R. Ferreira da Silva, M. Rynge, E. Deelman, M. S. Pérez-Hernández and O. Corcho. Reproducibility of execution environments in computational science using Semantics and Clouds. Future Generation Computer Systems, 2016.

[19] http://licensius.com/

4.5 Reproducibility in Earth Science: aspects and ongoing work

Raul Antonio Palma de Leon (Poznan Supercomputing and Networking Center, PL)

URL http://www.slideshare.net/rapw3k/aspects-of-reproducibility-in-earth-science

The "Earth Science Research and Information Lifecycle" can be regarded as a continuous, iterative and ongoing process used by scientists for conducting, validating and disseminating scientific knowledge. It can undergo an unlimited number of iterations that lead to the development of new and innovative ideas, concepts, techniques and technologies, which ultimately benefit both science and society. The life cycle can be briefly summarized into four main phases that involve multiple categories of stakeholders: (i) scientists access information and (usually) share results; (ii) shared results and information are analysed and interpretative models are generated and discussed with other colleagues; (iii) discussions lead to novel ideas and concepts which might need validation through further experimentation or data acquisition; (iv) new results are validated and shared so that other scientists can access them and start the process again.

This presentation introduces the ongoing work of the EU project EVEREST that aims at establishing a Virtual Research Environment (VRE) e-infrastructure for Earth Science. The VRE is being validated in four communities: sea monitoring, natural hazards, land monitoring and supersites, and is applying the Research Objects concepts and technologies as the mean for sharing information and establish more effective collaboration in the VRE. Regarding the reproducibility in their domain, they have a slightly different vision as other disciplines like experimental science that often aims at testing a hypothesis. For instance Supersite community can be described as an historical science that is mostly based on past observations. For such community, the main goals involve measuring geophysical parameters in the natural environment, derive information on the effects of the phenomena, model this information to generate space/time representations and provide these representations to risk management and other relevant stakeholders, and only complementary scientists may use this information to develop theories or confirm hypothesis. Hence, in such communities, reproducibility is mainly concerned about the execution of common or community-agreed workflows for data analysis and modelling, and for testing algorithms and data products. Nevertheless, there are still several limitations for achieving reproducibility in these communities: they are not yet using formalised (computational) workflows, the data necessary is not always available or known, workflows usually require considerable human intervention, etc. These are some of the challenges currently being addressed in EVEREST.

4.6 Reproducible Data Sets in Dynamic Settings: Recommendations of the RDA Working Group on Dynamic Data Citation

Andreas Rauber (TU Wien, AT)

URL https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf

One key requirement to enable repeatability in data-driven processes is the ability to specify precisely the data that was used as input, and to be able to re-create this identical input data for any re-execution of an analysis. This proves challenging due to two reasons:

1. granularity of the data: Databases contain massive amounts of data, of which specific subsets are selected and used for analysis. To enable reproducibility or comparability of results we need to be able to specify precisely, which subset was extracted from a larger dataset. Providing a verbal description of the extraction process (i.e. specifying the rows and columns selected and filter criteria used, or describing an arbitrary geographic region in natural language is error-prone and requires significant effort to recreate exactly the same data set again. Another solution encountered frequently, i.e. storing a backup dump of each subset used, does not scale and leads to massive data management overheads.
2. Dynamic data: In many settings,the data available for analysis changes, by new data being added continuously, erroneous data being deleted or corrected. Again, we need to ensure that we can obtain earlier "versions" of data as used in a study to enable repeatable and comparable results. Solutions such as artificially defining e.g. annual batch releases of data delay the availability of current data and again lead to massive overhead in storing duplicate batches of unchanged data.

The Working Group on Dynamic Data Citation of the Research Data Alliance (RDA WGDC) has elaborated a set of recommendations [1] to solve these challenges. In a nutshell, the solutions is based on (1) time-stamped and versioned data to ensure that earlier versions of data remain available, and (2) storing the queries used to select arbitrary subsets of data with a timestamp. A persistent identifier (PID, e.g. a DOI) is added to such a query together with additional metadata such as hash keys for fixity information, to ensure the time-stamped query can be re-executed against the time-stamped database to retrieve an identical subset. This approach allows retrieving the data both as it existed at a given point in time as well as the current view on it, by re-executing the same query with the stored or current timestamp, thus benefiting from all corrections made since the query was originally issued. This allows tracing changes of data sets over the time and comparing the effects on the result set. The query stored as a basis for identifying the data set provides valuable provenance information on the way the specific data set was constructed, thus being semantically more explicit than a mere data export. The query store also offers a valuable, central basis for analyzing data usage. Metadata such as checksums support the verification of correctness and authenticity of data sets retrieved.

The recommendations are applicable across different types of data representation and data characteristics (big or small data; static or highly dynamic; identifying single values or the entire data set). Pilot implementations have been used to evaluate this approach in different settings including data stored in relational databases (RDBMS, e.g. MySQL), XML databases (e.g. X-Base), and comma separated value files (CSV). More details are available from the homepage of the working Group[20].

References

1 Andreas Rauber, Ari Asmi, Dieter van Uytvanck and Stefan Pröll. Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). Research Data Alliance, October 20 2015.

[20] https://rd-alliance.org/node/141

4.7 Reproducibility in Visualization

Paul Rosenthal (TU Chemnitz, DE)

During the last decades, visualization has made its way towards a serious science. However, finalizing this procedure would also require that publications and results based upon a verifiable basis, i. e. that they are reproducible. Introducing a culture of reproducibility within this community would also increase the acceptance of visualization methods in other communities and applications, make contributions more well-grounded, and speed up the development of the community by making comparisons and advancements much easier.

In this talk, the recent efforts around the EuroRV^3, the EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization[21], which goes into its fourth year in 2016, were presented. The important role and the benefits of a strong culture of reproducibility have been discussed over the last years in the EuroRV^3 workshops in detail. And there were only a few cases identified where very special requirements prevent the achieving of basic reproducibility at all. However, it was also observed that, in current publications, providing reproducibility is often limited to the fraction that is needed to be accepted to a venue. This seems to be due to the fact that reproducibility is not rewarded at all.

Consequently, efforts arise to encourage good reproducibility in future publications by introducing a badge of honor or a similar sign of appreciation for all papers fulfilling a set of criteria. The main goal is to point out and honor such work and strengthen the community in its effort to establish a common culture of reproducibility. The efforts in this direction are still in the phase of planning and negotiations.

4.8 Research Data Alliance: State of the Art

Rainer Stotzka (KIT – Karlsruher Institut für Technologie, DE)

URL https://rd-alliance.org

The Research Data Alliance (RDA) is an international organization focused on the development of infrastructure and community activities that reduce barriers to data sharing and exchange, and the acceleration of data driven innovation worldwide. With more than 3,200 members globally representing more than 100 countries, RDA includes data science professionals from multiple disciplines, including but not limited to academia, library sciences, earth science, astronomy and meteorology. RDA is building the social and technical bridges that enable open sharing of data to achieve research reproducibility and transparency.

[21] http://www.eurorvvv.org/

5 War Stories

5.1 Reimplementation study "Who wrote the Web?"

Martin Potthast (Bauhaus-Universität Weimar, DE)

Main reference M. Potthast, S. Braun, T. Buz, F. Duffhauss, F. Friedrich, J. M. Gülzow, J. Köhler, W. Lötzsch, F. Müller, M. E. Müller, R. Paßmann, B. Reinke, L. Rettenmeier, T. Rometsch, T. Sommer, M. Träger, S. Wilhelm, B. Stein, E. Stamatatos, M. Hagen, "Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval", in Proc. of the 38th European Conf. on IR Research – Advances in Information Retrieval (ECIR'16), LNCS, Vol. 9626, pp. 393–407, Springer, 2016; pre-print available from author's webpage.
URL http://dx.doi.org/10.1007/978-3-319-30671-1_29
URL http://www.uni-weimar.de/medien/webis/publications/papers/stein_2016d.pdf

We revisited author identification research by conducting a new kind of large-scale reproducibility study: we selected 15 of the most influential papers for author identification and recruited a group of students to reimplement them from scratch. Since no open source implementations have been released for the selected papers to date, our public release will have a significant impact on researchers entering the field. This way, we lay the groundwork for integrating author identification with information retrieval to eventually scale the former to the web. Furthermore, we assess the reproducibility of all reimplemented papers in detail, and conduct the first comparative evaluation of all approaches on three well-known corpora.

5.2 Repeatability in Computer Systems Research

Christian Collberg (University of Arizona – Tucson, US)

Main reference C. Collberg and T. Proebsting, "Repeatability in Computer Systems Research", Communications of the ACM, Vol. 59(3):62–69, 2016.
URL http://dx.doi.org/10.1145/2812803
URL http://repeatability.cs.arizona.edu/

We give anecdotal as well as empirical evidence that Computer Systems researchers are generally unwilling or unable to share research artifacts (code and data) and that, when they do, their code often does not build. As a result, the minimum requirements for the reproducibility of applied Computer Systems research (that code used in experiments is available and that it builds) are generally not met.

We give an account of our own failed attempt at reproducing the results in a published research paper, a description of an empirical study into the reproducibility of the research published in 601 papers that appeared in the last two years in ACM publications, and a recommendation for how researchers, public funding agencies, and academic publishers can improve the reproducibility of Computer Science research.

6 Working groups

6.1 PRIMAD – Information gained by different types of reproducibility

Andreas Rauber (TU Wien, AT), Vanessa Braganholo (Fluminense Federal University, BR), Jens Dittrich (Universität des Saarlandes, DE), Nicola Ferro (University of Padova, IT), Juliana Freire (New York University, US), Norbert Fuhr (Universität Duisburg-Essen, DE), Daniel Garijo (Technical University of Madrid, ES), Carole Goble (University of Manchester, GB), Kalervo Järvelin (University of Tampere, FI), Bertram Ludäscher (University of Illinois at Urbana-Champaign, US), Benno Stein (Bauhaus-Universität Weimar, DE), and Rainer Stotzka (KIT – Karlsruher Institut für Technologie, DE)

6.1.1 What is Reproducibility

What is "reproducibility" anyways? And how is it different from "repeatability", "replicability", or any of the other r-words? There are already a number of attempts at defining and sorting out these different notions. De Roure [1] lists 21 different r-words grouped into 6 categories, stating that reproducibility means reusing a research object with a change to some circumstances, inputs, resources or components in order to see if the same results are achieved independent of those changes. Often these notions are context-sensitive (e.g., validation vs verification have rather precise and very different meanings in different communities.

As an alternative approach to sort out terminological confusions, we attempted to look at a different perspective. When trying to reproduce a study, what are the things that are kept the same (e.g., the overall method or algorithm) and what is changed (e.g., the input data or implementation language, etc.)? More importantly, while changing these things, what information is gained by successfully reproducing (or failing to reproduce) a study?

6.1.2 The PRIMAD Model

As a starting point, we defined a preliminary list of "variables" that could potentially be changed:

- (R) or (O) Research Objectives / Goals
- (M) Methods / Algorithms
- (I) Implementation / Code / Source-Code
- (P) Platform / Execution Environment / Context
- (A) Actors / Persons
- (D) Data (input data and parameter values)

This spells: OMIPAD. Rearranging the letters that we use to represent the several aspects that can be changed, it can be remembered as PRIMAD: (P)latform, (R)esearch Goal, (I)mplementation, (M)ethod, (A)ctor, (D)ata (both input and parameter data), which allows us to ask: What variables have you "primed" in your reproducibility study?

As a concrete example of the meaning of these variables, let's assume our (R)esearch objective is to sort a data set. We could use Quick Sort as the sorting (M)ethod (algorithm), which could be (I)mplemented as a script in Python and run over a Python 2.7 compiler on an iMac running MacOS 10 (and this would be the execution (P)latform). We could run this

over a specific (D)ataset (data.csv) using 0 as the pivot parameter. The (A)ctor, in this case, is the researcher that is executing the sorting. Summarizing:

- Research goal: sorting the input
- Method: quick sort
- Implementation: script in Python
- Platform: Python 2.7, MacOS, iMac, etc
- Input Data: the data that is to be sorted
- Parameter: the position of the pivot
- Actor: user that is executing the experiment

As a more concrete example, we can take Tandy Warnow's statistically binning paper and the controversy around it[22]. In this case, the controversy was that her initial approach (we will call it method M, proposed by team T1) was claimed (by team T2) to be non-reproducible. More specifically, team T2 implemented method M and could not reproduce the original results obtained by team T1. So, in this example, we have the following scenario:

- Research Objective: Improve state-of-the-art in phylogenetic tree construction
- Method: Statistical binning (supposedly M = M', but one side is arguing that M != M')
- Implementation: two available, by team T1 and by the "opposing" team T2
- Platform: various (we suppose)
- (input) Data: different datasets – some arguments were made about the suitability here as well, since apparently team T2 did not respect some premises of how the input data should be organized.

To describe this reproducibility study in terms of these variables, only the research objective R and the method M are fixed; everything else is varied (team T1 actually argues that the implementation I2 isn't of the method M, but of another method M'). To represent what changed, we use primed variables.

In this case, T1 argues: P'RI'M'A'D', while T2 argues P'RI'MA'D' (variables with apostrophe were changed, and non-apostrophe variables were kept the same). Thus, both teams actually disagree on whether M = M' or not!

6.1.3 Gains from different types of reproducibility

Reproducibility in its various forms, however, is never a goal in itself. We do it in order to gain something. By changing some (or several) of these variables, we gain different kind of knowledge. For example, if one keeps R, M, and I fixed, but varies the platform $P \longrightarrow P'$, then the reproducibility study tests the portability, stability, or platform-independence of the experiment.

Figure 1 shows an attempt to categorize and label the various types of reproducibility and to summarize the gain they bring to a computational experiment. The precise terminology to use is still subject to further debate and no final agreement could be reached, specifically with respect to the labels and the mapping to the terminology found in the literature to describe different types of reproducibility. This may be partially due to the fact that many of the terms used describe repeatability settings refer to combinations of the above, e.g. to differentiate between obtaining a certain level of repeatability within the same lab or by an external lab. But even independent of this combinatorial issues the exact terminology proves to be difficult to agree upon already within a computing setting, not to mention beyond this domain.

[22] https://youtu.be/-0jd0x7Kg90?list=PLO8UWE9gZTlAgHZPaxQbpUNY0T26zeL_f

Label	Data		Platform / Stack	Implementation	Method	Research Objective	Actor	Gain
	Parameters	Raw Data						
Repeat	-	-	-	-	-	-		Determinism
Param. Sweep	x	-	-	-	-	-		Robustness / Sensitivity
Generalize	(x)	x	-	-	-	-		Applicability across different settings
Port	-	-	x	-	-	-		Portability across platforms, flexibility
Re-code	-	-	(x)	x	-	-		Correctness of implementation, flexibility, adoption, efficiency
Validate	(x)	(x)	(x)	(x)	x	-		Correctness of hypothesis, validation via different approach
Re-use	-	-	-	-	-	x		Apply code in different settings, Re-purpose
Independent *x* *(orthogonal)*							x	Sufficiency of information, independent verification

Figure 1 PRIMAD Model: Categorizing the various types of reproducibility by varying the (P)latform, (R)esearch Objective, (I)mplementation, (M)ethod, (A)ctor and (D)ata, analyzing the gain they bring to computational experiments. x denotes the variable primed i.e. changed, (x) a variable that may need to be changed as a consequence, whereas – denotes no change.

We now elaborate on the various aspects that can be changed, and how we could "label" reproducibility studies that use such combinations of changes.

1. ϵ equals to not changing anything, simply repeating an earlier experiment within the same computational environment, using the same code and data, allows to verify that the computed results are deterministically consistent. **Suggested label:** [**repeat**]
2. **Data → Parameters:** changing the parameter settings (e.g. parameter sweep, 10-fold cross-validation, etc.) allows to determine the: robustness/sensitivity of an experiment wrt. the specific parameters. Suggested label: [**rerun: robustness check, parameter sweep**]
3. **Data → Raw (Input) data:** changing the raw data processed by an experiment allows to verify how far the statements made hold across a larger part of the input space. Depending on the degree of similarity/difference in the input data, statements on the generality can be made. It also allows to evaluate whether the data originally used is representative/comparable for a given domain. **Suggested label:** [**rerun: check generality**]
4. **Platform:** changing the execution platform (i.e. the context, execution environment, including the software and hardware stack, i.e. a Java virtual Machine, running on a specific version of some operating system, within some hypervisor, running on specific HW) allows to test the platform independence/portability of an experiment. It may gain wider adoption or higher stability by being runnable on a wider range of platforms. **Suggested label:** [**port**]
5. **Implementation:** changing the implementation allows to verify the correctness of the previous implementation. It may also gain you higher efficiency, provide broader set of execution platforms, leading to higher adoption in different communities. Note that changing the implementation may incur a change of the execution platform. **Suggested label:** [**re-code**]

6. **Method:** changing the method allows to validate the correctness of a hypothesis using a different methodological approach. This provides a method-independent verification, or may provide a more efficient method to support the claims made. Note that a change in the method by definition will incur a change in the implementation, and possibly also of the execution platform. **Suggested label:** [**validate**]
7. **Research Objective:** changing the research objective (hypothesis) basically constitutes a re-purposing / re-use of an earlier experiment, allowing science to progress faster, opening new avenues for research. It requires trustworthy results/components to offer a solid basis. **Suggested label:** [**repurpose**]
8. **Actor:** changing the actor is orthogonal to all changes discussed above. It allows both independent verification of the characteristics, and also determines whether the information provided is sufficient to achieve such independent verification. **Suggested label:** [**experimenter-independent <activity>**]

Consistency: success or failure of a reproducibility study has to be evaluated wrt. the consistency of the outcomes. The criterion to apply thus is not whether the outcomes of priming any of the above variables leads to identical results, but whether results are consistent with the previous ones. Depending on the setting, this may require identity of results, but may also be lessened to consistency within certain error bounds or allow differences that are not statistically significant.

Transparency: Another dimension to be considered is transparency. It denotes the ability to look into all necessary components to be able to understand the path from the hypothesis to the results. While many of the changes above can be performed on a black-box level (repeating a run using binary code, performing the repeatability evaluation on a virtual machine provided by the original authors) it does not allow to make qualified inspections on the internal functioning on the respective levels. Thus, the degree of transparency should be used as a measure for the degree of inspection possible.

6.1.4 Variations on PRIMAD

After analyzing the various aspects that can be changed, we realized that using just one letter to represent both input data and parameters may not be enough. We are also aware of the fact that the differences between these attributes may not always be very clear-cut, as e.g. the fuzzy distinction between parameter and data to be supplied to an algorithm, or the boundary between an implementation and the execution platform becoming less clear-cut via the use of static or dynamically linked libraries. Yet, we find that the current set of variables helps in distinguishing core concepts and challenges to repeatable experiments relying on computation. Thus, we tried to identify possible letters we could use to represent each of the aspects we discussed:

- (O,R,G) Research Objectives / Goals
- (M,A) Methods / Algorithms
- (I,C,S) Implementation / Code / Source-Code
- (E,C) Platform / Execution Environment / Context
- (D,I,R) Input Data (“raw” data)
- (P) Parameter values
- (A) Actors / Persons

In the future, we may define a new acronym using these letters to better represent all the possible variations. Some possibilities are APDEIMO, PDEIMOA, AOMIEDP, OMIEDPA,

OMIEPAD. We may also need a deeper analysis of the various attributes and their changes, seeing in how far these can be mapped to, first of all, the different definitions of types of reproducibility being used in different communities. Furthermore, with most scientific work today spanning several disciplines and crossing methodological boundaries we need to investigate, in how far the concept of fixing and changing various attributes can be applied in more general settings. However, while the precise labels being used may change, we have the feeling that having a precise definition and understanding of the attributes that are fixed or changed is essential to define the various types of reproducibility studies and, specifically, to understand the benefit we gain from them. Reproducibility is not a means to its own end. While showing deterministic results by simply repeating a computation without changing anything may already be an exciting fact in some settings we very likely will want to go beyond such basic settings of reproducibility studies, gaining deeper insights into scientific work and establishing trust in results, methods and tools for the benefit of science.

References

1 De Roure, D., (2014). The future of scholarly communications. Insights. 27(3), pp. 233–238.

6.2 Reproducibility Tools and Services

Tanu Malik (University of Chicago, US), Vanessa Braganholo (Fluminense Federal University, BR), Fernando Chirigati (NYU Tandon School of Engineering, US), Rudolf Mayer (SBA Research – Wien, AT), and Raul A. Palma de Leon (Poznan Supercomputing and Networking Center, PL)

Sharing code and data increase reproducibility, but such sharing may not reflect the overall method, which is typically published in research papers. The current format of research papers (text-based) does not link code and data at finer granularity, the page-limit restricts detailed description of analyses and/or reporting of negative results, and authors have little motivation to describe in detail on a companion website. The consequence is built-up of scientific bias, which can be hard to break, given long cycles of publishing and funding.

Consequently, there is a critical need for reproducibility tools that, along with the changing culture of reproducibility, can also help researchers achieve the desired state of reproducibility in an efficient manner. However, before developing and/or applying a tool-suite to solve a reproducibility problem, several issues at hand must be understood. These range from:

1. **Precise identification of gaps in the research lifecycle.** A precise identification of gap in the research life-cycles is needed to understand which tool is applicable for solving the problem. Three gaps are often identified in the research lifecycle. The first one is related to the lack of motivation from researchers to apply reproducibility on their research. Better methods to incentivize reproducibility are needed, e.g.: having regulations and funding agencies to "force" the practice of reproducible research.
   A second possible gap is due to the poor linking between computational assets and text-based research outputs: there is rarely a connection between computational artifacts (research material, data, samples, software, models, methods, etc.) and the published results (paper and review process). This gap is very much discipline specific: some disciplines have developed standards on how to handle these artefacts and document the

procedures (e.g.: systems biology), while others have done less so (e.g.: computational sciences).
Last, a third possible gap is the lack of sufficient tools to help researchers do reproducible research according to their requirements, which, again, are domain-specific.
While the development of tools and services helps fill the last two gaps, the lack of motivation remains a barrier that must be addressed to broaden the adoption of reproducibility. It does not matter if we can build the most useful and easy-to-use reproducibility tools: unless there are proper incentives, these tools will be pointless.

2. **Domain-related issues.** In several domains, the term "reproducibility" is vaguely defined, and when defined, can substantially differ among communities. The main reason for such disparity is due to the fact that different domains have different requirements regarding reproducible research. For instance, while numerical analysis does not often handle large amounts of data, a tool or service constructed for the databases area must take into account how to share terabytes of data for reproducibility purposes. Therefore, a distinction must be made between common and discipline-specific reproducibility requirements. Also, different domains use different technologies (e.g.: programming languages, protocols, and types of data), which often influences the development of domain-specific tools.
3. **Are there tools that can solve the problem at hand?** It is unclear whether we need to build new tools to solve the existing problems, and the reason is twofold: first, we do not know what the problems are, as the requirements from different domains must be better understood; second, it is hard to know all the tools available and all the features they provide. Assuming there are sufficient tools, it may be a matter of just improving existing tools, either by creating new features or tackling new requirements. In addition, since different tools address different issues, a question that comes up is whether they are integrated enough with the existing environment/infrastructure to achieve the desired reproducibility.

6.2.1 Tool Landscape

The primary challenge for the user is to how to navigate the tool landscape with minimal effort but improved reproducibility. Thus, given L reproducibility levels, and N dimensions of assessing the reproducibility, the objective in tool landscape is to guide the user to move from tool A to tool B such that there is minimal effort but a gain in the reproducibility level along one or more dimension.

Reproducibility is a continuous process that is achieved over time, but in several cases it can be discretized to various levels that provide a different state of reproducibility of the experiment. For instance, if we do not change any of the PRIMAD attributes, then at the bare minimum we demonstrate determinism and consistent behavior. If the user moves from simple scripting – an environment in which the user is already satisfied – to a "workflow" environment (e.g.: YesWorkflow, noWorkflow, VisTrails, Taverna, Wings), then an effort is required to transition depending on the domains and experiment, while gaining both a demonstration of tool idependence / correctness of the implementations, as well as higher portability or easier adoption / re-use in different settings (e.g.: transitioning from Python scripting to noWorkflow is straightforward). Packaging tools such as CDE, PTU, Docker, and ReproZip may significant improve reproducibility with small effort.

6.2.2 Addressing the gap

First, we need to understand **what is needed**, rather than **what is possible**: not everything that can be done and developed is needed (and wanted) when it comes down to reuse and reproducibility. Therefore, it is of crucial importance to collect the different common and domain-specific requirements, and then recognize what we are missing. Related to the requirements, another important issue is knowing the target group that is interested in reproducibility, which helps determining the needs and requirements of appropriate tools and services.

Regarding the gap associated to the poor linking between computational resources and dissemination reports, papers, etc., the concept and approach of research objects may be one of the means to support scientists and the research community in filling this gap. Research objects (ROs) are aggregating objects that bundle together resources that are essential to a computational scientific study or investigation (data used/produced, methods applied, results, publications, people, etc.), along with semantic annotations on the bundle or the resources needed for the understanding and interpretation of the scientific outcomes, including provenance and evolution information, descriptions of the computational methods, dependency information and settings about the experiment execution. There is also a plethora of tools that can create executable papers, such as Janiform, Galaxy, and VisTrails. In addition, there is a set of literate programming tools that help linking documents to code and data (e.g.: Jupyter notebooks).

In terms of gaps in the existing suite of tools and services, there might be enough of them available. However, these may still not serve the intended purpose and may need to be improved according to the collected requirements. Such an approach may be preferred rather than developing and implementing yet new ones. In addition, integrating such tools may be useful for not having to reinvent the wheel, or at least enabling their interoperability through common or established models and formats.

One of the main issues is understanding **which tools are available, what each of them support, and which types of problem they solve**. There is a plethora of available tools for reproducible research, and these can be categorized in different ways: (1) features provided (provenance capture, representation, portability, archiving, longevity, access to remote services, etc.); (2) target domain areas; (3) reproducibility modes (planning for reproducibility, such as scientific workflow systems and configuration management tools, vs. reproducibility as an afterthought, such as packaging tools); and many others. Depending on these different tags, researchers may need different tools. A search engine for reproducibility tools, for instance, would be useful. ReproMatch[23] is a step towards this.

There are some well established and widely **accepted infrastructures**, such as DropBox, GitHub, and Zenodo. It is questionable, however, how effective these are for the intended reproducibility and whether the granularity of the stored artifacts is sufficient. For instance, these examples are entitled to provide pure storage and versioning, and not curation; also, the lineage between code and data, and the reasoning behind the versioning are not captured. If this is a requirement, certainly these tools do not address the needs.

It may be helpful to reflect on a **basic research environment**, which allows to automatically track and record individual steps and milestones during the research and developing process. This may include an electronic notebook providing automatic documentation support. Such infrastructure would provide stable and standardized code, including a version

[23] http://repromatch.poly.edu/

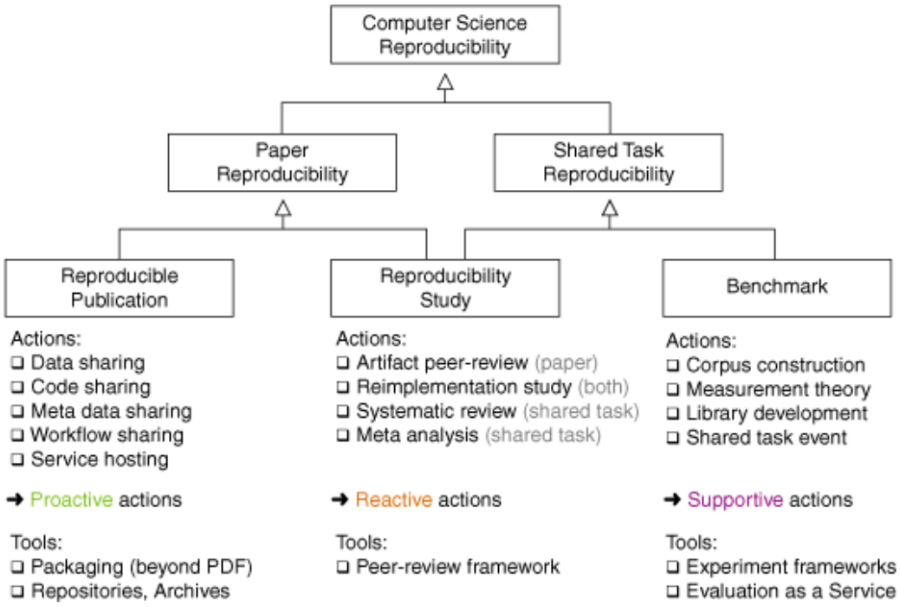

Figure 2 Taxonomy of actions towards improving reproducibility in computer science.

control and a notification service in case anything changes in any related software package, library, or operating system. It is unclear, however, if a single environment can be developed for different domains.

6.3 Taxonomy of Actions Toward Reproducibility

Martin Potthast (Bauhaus-Universität Weimar, DE), Fernando Chirigati (NYU Tandon School of Engineering, US), David De Roure (University of Oxford, GB), Rudolf Mayer (SBA Research – Wien, AT), and Benno Stein (Bauhaus-Universität Weimar, DE)

There are a number of well-known actions that researchers may take today to improve the reproducibility of computer science, whereas many of them are at best partially supported by tools, or not at all. Figure 2 organizes these actions within a taxonomy.

The taxonomy comprises two levels, where the first divides reproducibility into two disjunct categories: "Paper reproducibility" comprises only actions that serve to improve the reproducibility of individual papers, whereas "shared task reproducibility" comprises actions aimed at improving the reproducibility of groups of independent papers that address a common problem of interest (i.e., a shared task). The distinction is important, since the actions that can be taken in each case significantly differ.

On the second level of the taxonomy, the artifacts that may result from taking action toward improving reproducibility are given. There are basically three categories, namely a "reproducible publication," a "reproducibility study," and a "benchmark." Reproducibility studies can be done for both, individual papers as well as shared tasks, whereas some actions apply only to one, the other, or both.

Below these categories, specific actions are listed that ultimately yield artifacts belonging to their respective category. All of the actions can be distinguished by who takes them, and their relation to reproducibility: actions that lead toward reproducible publications must be taken proactively by authors before publication. Reproducibility studies, however, are always in reaction to publication of new results. They can be done at publication time, e.g., by peer-reviewers, but are more often conducted long after publication by researchers working on the same problem. For shared tasks, supportive actions are often taken in order to create standardized benchmarks, and to ensure comparability across papers.

These actions and the required information can be nicely mapped to the PRIMAD model of distinguishing different types of reproducibility by varying (priming) specific aspects of a study while keeping others unchanged. This requires the unchanged aspects (data, code, execution environment) to be shared and deliver different gains in knowledge on the original study.

Proactive actions: to share the data and to share the code underlying a paper are perhaps the most important actions that authors can take to create a reproducible publication. Moreover, hosting a prototype service and providing metadata and workflow information may prove to be key assets to understanding the runtime behavior of a given contribution.

All of the aforementioned artifacts can be shared today with some extra effort, since there are places on the web where data, code, and other artifacts belonging to a given paper can be hosted. However, spreading artifacts across different platforms is hardly straightforward to follow up, much less maintainable. Rather, scientists are used to obtaining research at one place, namely the PDF hosted at publisher site. Since no common standards for sharing scientific artifacts beyond the PDF have emerged to date, the landscape is disorganized, with a number of individual solutions as well as a number of community-specific tools. Two particular kinds tools can be identified in this respect, namely packaging technology that envelopes all artifacts resulting from a given piece of research, and repositories and archives that allow for retrieval, long-term storage, and maintenance of published artifacts.

Reactive actions: ideally, submitted publications would be immediately checked for reproducibility, e.g., within an artifact peer-review, but this not, yet, commonplace. Otherwise, reproducibility studies are the most effective way to ensure independent reproducibility. Specifically, reimplementing a given paper including all of its experiments, or reimplementing individual approaches proposed in a group of papers on a shared task without reproducing all experiments of all papers. In addition, for shared tasks, systematic reviews and meta analyses may shed light onto the state of reproducibility in a given shared task. In this connection, we emphasize the distinction between systematic reviews and literature reviews (i.e., surveys): systematic reviews abstract over a subject matter (e.g., by unifying terminology, by organizing existing contributions with regard to previously unconsidered criteria, or by recasting the problem of interest in a new way), whereas literature reviews merely collect the existing contributions with little to no abstraction. Hardly any tool support beyond the existing academic search engines has been invented so far, since systematic reviews and meta analyses rely on abstract thinking over the original papers that are studied. Also the peer-review of artifacts is currently hardly supported within conference management systems.

Supportive actions: benchmarks for software-driven and data-driven computer science are perhaps one of the few cases where computer science already excels: once a given shared task is studied more frequently, researches often build specific evaluation corpora, they study the theory of measuring performance of proposed solutions, and they develop software libraries to collect state-of-the-art algorithms for the shared task. If the community surrounding a shared

task agrees on a benchmark, papers published henceforth are comparable without further need for coordination among researchers. In some cases, shared task events are organized, where researchers compete to build the best solution for the shared task's underlying problem, and where submitted solutions are evaluated within a standardized evaluation setup that often becomes a new benchmark.

Regarding tools for supportive actions, experiment frameworks allow for the structured execution of experiment series for particular tasks. Such frameworks are often tailored to particular research domains and shared tasks. Moreover, since not all datasets can be shared publicly for reasons of privacy and copyright, among others, this prevents some important benchmarks from becoming widely available. To mitigate these limitations, it has been proposed to move evaluation to the cloud under the recently proposed evaluation as a service paradigm [1]: under this paradigm, software that solves a given shared task is deployed within a cloud infrastructure, and the software's processing rights for the sensitive datasets are managed and controlled, so that data cannot leak.

Automation: all of the aforementioned tools together point into an interesting new direction for experimental, data-driven and software-driven science, namely automation. Much of the process of optimizing scientific software to a problem includes parameter optimization, which often boils down to an (informed) search in hyperparameter space. The expanding capabilities of cloud computing can be exploited to tune scientific software deployed under the evaluation as a service paradigm, maximizing expected performance compared to manual or semi-automatic optimization. Moreover, considering individual papers, a more standardized way of annotating the scientific process and its outcome in the form of papers and other artifacts, will allow for their inclusion in the linked data cloud and, eventually, inference on top of that.

Social interaction: the tool support to improve reproducibility will not be based solely on standardized interfaces. Rather, the web services that will eventually emerge will likely include social networks. Unlike papers, artifacts may not always be perfectly documented, which question to be answered and solutions to be discussed, in order to complete a documentation or fix issues with previously published artifacts. This is especially true when the original authors of the artifacts are not available, anymore, long after publication.

References

1 Allan Hanbury, Henning Müller, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Ivan Eggel, Tim Gollub, Frank Hopfgartner, Jayashree Kalpathy-Cramer, Noriko Kando, Anastasia Krithara, Jimmy Lin, Simon Mercer, and Martin Potthast. Evaluation-as-a-Service: Overview and Outlook. ArXiv e-prints, December2015.

6.4 Actors in Reproducibility

Justin Zobel (The University of Melbourne, AU), Shane Culpepper (RMIT University – Melbourne, AU), David De Roure (University of Oxford, GB), Arjen P. de Vries (Radboud University Nijmegen, NL), Carole Goble (University of Manchester, GB), Randall J. LeVeque (University of Washington – Seattle, US), Mihai Lupu (TU Wien, AT), Alistair Moffat (The University of Melbourne, AU), Kevin Page (University of Oxford, GB), and Paul Rosenthal (TU Chemnitz, DE)

Reproducibility is a component of a greater activity (e.g. reviewing, reusing) undertaken by actors (e.g. reviewer, author) who have their own behaviours (inherent or induced by external drivers). Interventions to motivate reproducibility behaviours, through positive incentives or the removal of obstacles, requires us to first classify actors and then layout behavioural standard

6.4.1 Actors

- **Creators**: authors, academic leaders/lab directors, research software engineers, thesis supervisors
- **Consumers**: readers, authors, students, policy makers, educators, adopters, technical communities, IT services, industry, user, research software engineers, PhD students
- **Moderators**: editors
- **Examiners**: reviewers, thesis examiners, research evaluation committees,
- **Enablers**: funders, publishers, institutions, academic leaders/lab directors, data providers, thesis supervisors, digital archives, professional societies, industry, research software engineers
- **Auditors**: funders, policy makers, institutions, professional societies

6.4.2 Questions

- What are the properties of reproducibility for each actor?
- What are the interventions they can invoke?
- What are the current behaviours, and how might they shift?
- What aspects of behaviour are important to whom?
- What timeframes apply?
- What are the obstacles to good behaviour?
- What are the incentives to encourage change in behaviour?
- What are the interventions to action change in behaviour?

6.4.3 Authors

This section summarizes the main obstacles and expectations for an author.

6.4.3.1 Obstacles (real or perceived) to good behaviour for authors

Obstacles may be external drivers over which the authors have limited control, or internal where the authors can be responsible for their own behaviour. Table 1 describes the obstacles in detail.

Table 1 Obstacles for authors.

Recognition	Lack of explicit recognition of the need for reproducibility within a lab Lack of credit for achieving reproducibility
Cultural pressure	Lab culture Publication (volume) pressure Time pressure
Ambition/Personal Pressure	Paranoia – fear of losing competitive advantage Embarrassment, limitations as a developer Fear of having mistakes exposed (security through obscurity)
Awareness	Ignorance of the benefits of reproducibility, lack of mentoring and guidance Misjudgement of the difficulty of achieving reproducibility Lack of planning for reproducibility – it cannot be an afterthought Perception of achievability
Intention	Code/data was meant to be disposable (ephemeral)
Resources	Lack of access to appropriate resources Inertia, apathy, lack of incentives
Institutional restrictions	Legal and licensing issues, Corporate privacy requirements
Innate restrictions	Code or data cannot be encapsulated

Three tiers of standard – sufficient, better, exemplary – set out a rubric of expected behaviour. Interventions and incentives have the capacity to move up the reproducibility ramp.

6.4.3.2 Standards: Sufficient

These elements, if present in a paper and appropriate to that paper, represent a minimum expectation of authors – with regard to both ethical requirements and the demands of reproducibility.

- Methods section – to a level that allows imitation of the work
- Appropriate comparison to appropriate benchmark
- Data accurately described
- Can re-run the experiment
- Verify on demand (provide evidence that the work was done as described)
- Ethical considerations noted, clearances listed
- Conflicts noted, contributions and responsibilities noted
- Use of other authors' reproducibility materials should respect the original work and reflect an attempt to get best-possible results from those materials

6.4.3.3 Standards: Better

Addition of elements such as these represent a substantial increment beyond sufficient, while not yet being best practice.

- Black/white box
- Code is made available, in the form used for the experiments
- Accessible or providable data

Table 2 Obstacles to good behaviour for reviewers.

Recognition	Lack of explicit recognition of the need for reproducibility within the discipline Lack of credit for examining reproducibility
Cultural pressure	Time pressure Volume pressure
Ambition/Personal Pressure	Embarrassment, technical limitations Lack of understanding of why reproduction failed – is it really the fault of the reviewer or authors?
Awareness	Ignorance of the benefits of reproducibility, lack of mentoring and guidance Misjudgment of the difficulty of examining reproducibility Perception of achievability
Intention	None
Resources	Lack of access to appropriate resources – technical, personnel Inertia, apathy, lack of incentives
Institutional restrictions	None
Innate restrictions	None

6.4.3.4 Standards: Exemplary

Addition of these elements, in or accompanying a paper, represent best practice for authors.

- Open-source software
- Engineered for re-use
- Accessible data
- Published in trustworthy, enduring repository
- Data recipes, to allow construction of similar data
- Data properly annotated and curated
- Executable version of the paper; one-click installation and execution

6.4.4 Reviewers

Noting the potential for reviewers to be explicitly assigned to provide either technical review or scientific review:

6.4.4.1 Obstacles (real or perceived) to good behaviour for reviewers

Table 2 describes the obstacles in detail.

6.4.4.2 Standards: Sufficient

- Assesses reproducibility
- Fair assessment, respect of strengths and weaknesses
- Clarity on what was assessed and what the limits of the review are
- Conflicts noted

6.4.4.3 Standards: Better

- Checks that reproducibility is in fact possible

Table 3 Obstacles to good behaviour for editors.

Recognition	Lack of explicit recognition of the need for reproducibility within the discipline Lack of credit for examining reproducibility
Cultural pressure	Time pressure Volume pressure
Ambition/Personal Pressure	None
Awareness	Ignorance of the benefits of reproducibility, lack of mentoring and guidance Misjudgment of the difficulty of examining reproducibility Perception of achievability
Intention	None
Resources	Inability to find technically accomplished reviewers
Institutional restrictions	None
Innate restrictions	None

6.4.4.4 Standards: Exemplary

- Reproducible, within limits of materials and resources
- Timely reviews

6.4.5 Editors

6.4.5.1 Obstacles (real or perceived) to good behaviour for editors

Table 3 describes the obstacles in detail.

6.4.5.2 Standards: Sufficient

- Find reviewers who can assess the science
- Have reviewing policies that require examination of reproducibility/methodology
- Have instructions for authors on expectations with regard to reproducibility/methodology
- 'Reproducibility compacts' (or contracts) for authors, in which they must state availability of code and so on [1]

6.4.5.3 Standards: Better

- Find reviewers who can assess the technical contribution
- Separation of assessment of papers on science grounds from reproducibility/methodology grounds
- Have processes for working with authors to improve reproducibility

6.4.5.4 Standards: Exemplary

- Advocacy to the publisher of requirements for reproducibility
- Advocacy of standards
- Leadership regarding all aspects of reproducibility
- Participation in relevant advocacy bodies

Table 4 Obstacles to good behaviour for institutions.

Recognition	Lack of explicit recognition of the need for reproducibility Lack of credit for achieving reproducibility
Cultural pressure	Publication (volume) pressure Fear of having mistakes exposed (security through obscurity)
Ambition/Personal Pressure	Lack of enduring commitment – long-term budgeting Lack of communication plansResistance to openness Paranoia – fear of losing competitive advantage Fear of having mistakes exposed (security through obscurity)
Awareness	Ignorance of the benefits of reproducibility, lack of mentoring and guidance Misjudgment of the difficulty of examining reproducibility Perception of achievability Legal and licensing issues
Intention	None
Resources	Resources, services, infrastructure, repositories Lack of standards and tools Lack of access to appropriate resources Lack of understanding of the resources requiredInertia, apathy, lack of incentives
Institutional restrictions	Confused lines of responsibility, mixed ownership of the problem Human resources structures: mentoring, training, staffing Mismatch between academic and organizational goals Conflicting or missing or ill-informed policies Legal and licensing issues Corporate privacy requirements
Innate restrictions	None

6.4.6 Institutions (also as transmitted via academic leaders)

6.4.6.1 Obstacles (real or perceived) to good behaviour for institutions

Table 4 describes the obstacles in detail.

6.4.6.2 Standards: Sufficient

- Clear policies on reproducibility, ethic

6.4.6.3 Standards: Better

- Compliance framework
- Resourcing of reproduction – technical, financial
- Constructive environment with recognition of demands of reproduction

6.4.6.4 Standards: Exemplary

- Trusted, enduring repository
- Reproduction as a primary research goal

References

1 C. Collberg, T. Proebsting and A. M. Warren. Repeatability and Benefaction in Computer Systems Research. University of Arizona TR 14-04.

6.5 Guidelines for Authors, Editors, Reviewers, and Program Committee Chairs

Alistair Moffat (The University of Melbourne, AU), Shane Culpepper (RMIT University – Melbourne, AU), Arjen P. de Vries (Radboud University Nijmegen, NL), Carole Goble (University of Manchester, GB), Randall J. LeVeque (University of Washington – Seattle, US), Mihai Lupu (TU Wien, AT), Andreas Rauber (TU Wien, AT), and Justin Zobel (The University of Melbourne, AU)

A framework for explaining the need for reproducibility is to describe it in three separate elements, the *what*, the *why*, and the *how*. The *what* consists of articulating the authors' goals in the context of the instructions provided to reviewers and editors (or PC Chairs). The *why* consists of communicating an understanding of desirability of reproducibility, and of helping to convey the distinctions in the key terminology (reproducibility, repeatability). The *how* entails guidelines as to the means by which papers are assessed as a consequence of introduction of expectations regarding reproducibility.

The SIGMOD reproducibility guidelines[24] are a description of *what*: it is stated that it is desirable that papers' materials have shareability, coverage, and flexibility; noting that in some cases mechanical descriptions (recipes) rather than code may be necessary to couple shared resources to the external world. Another extension is the need for authors to re-use open materials in an appropriately scientific way – reusability materials are a scientific resource, not conventional open-source software.

An explanation of *why* appears in Section (see Section 6.7) of this report, and can be summarized as providing:

- Confidence in the work
- Acceleration of the science and of the state-of-the-art
- Verifiability
- Falsifiability
- Participation in the community, contribution to the community

A key requirement to achieve these outcomes is in the instructions that are provided to referees and authors. The European Conference on Information Retrieval 2016 has provided guidance in this regard[25], to which we add a final sentence (cf. also the PRIMAD model and the gains of different types of reproducibility in Section 6.1):

> *Reproducibility is key for establishing research to be reliable, referenceable and extensible for the future. Experimental papers are therefore most useful when their results can be tested and generalized by peers. This track specifically invites submission of papers reproducing a single or a group of papers, from a third-party where you have* ***not*** *been directly involved (e.g.* ***not*** *been an author or a collaborator). Emphasize your motivation for selecting the paper/papers, the process of how results have been attempted to be reproduced (successful or not), the communication that was necessary to gather all information, the potential difficulties encountered and the result of the process. A successful reproduction of the work is not a requirement, but it is important to provide*

[24] http://db-reproducibility.seas.harvard.edu/
[25] http://ecir2016.dei.unipd.it/call_for_papers.html

> *a clear and rigid evaluation of the process to allow lessons to be learned for the future."* ***It is not sufficient for a reproduction to be a simple re-execution of the existing code on the original data.***

A key *how* is the encouragement and definition of papers that have as their primary objective the reproduction and extension of previous work, and the provision of a refereeing process that recognizes the merits of such papers, and evaluates them accordingly. Encouraging behaviors that facilitate independent reproduction is another key goal. SIGMOD again provides an illustration of how this is done, with an independent evaluation process, different from the regular scientific acceptance review process. Authors should be encouraged to plan for reproducibility right from the commencement of each investigation, with a clear plan in place for how they will develop methods and artifacts that can be communicated to and reused by others.

Collberg et al.'s [1] evaluation describes the concept of a compact (or contract) that authors are expected to add to their paper at submission time and retain in the final version, in which they make claims about the likely reproducibility of their work. We regard the routine adoption of such a statement in published work as a low-cost but effective instrument for reproducibility.

Other key interventions that might benefit one or more of the various actors involved in the processes of scientific research, funding, and publishing include:

- A template for journals of instructions for authors and reviewers, in a form that allows adaptation; include explanation of how to respond when attempts to reproduce struggle or fail; include explanations in a digestible form that promotes reproducibility
- A template for thesis examination
- Managerial appraisal questions; compliance requirements at all levels
- A consolidated and maintained resource of information about reproducibility in computing.
- Recruitment of advocates and champions
- Information packs targeted at particular categories of actor
- Recognition systems – effective, visible – promotion, tenure, within institution, within journal/conference

Drawing on similar statements (including Science [2]; the SIGMOD[26] and ICTIR[27] Calls for Papers; and other sources), **we suggest the following message be sent to a wide pool of journal editors and conference chairs**, to be then used by them in communications with other senior members of the various communities:

> Transparency, openness, and reproducibility are vital features of science. Scientists embrace these features as disciplinary norms and values, and it follows that they should be integrated into daily research activities. These practices give confidence in the work; help research as a whole to be conducted at a higher standard and be undertaken more efficiently; provide verifiability and falsifiability; and encourage a community of mutual cooperation. They also lead to a valuable form of paper, namely, reports on evaluation and reproduction of prior work. Outcomes that others can build upon and use for their own research, whether a theoretical construct or a reproducible experimental result, form a foundation on which science can progress. Papers that are structured and presented in a manner that facilitates and encourages

[26] http://db-reproducibility.seas.harvard.edu/
[27] http://ictir2015.org/cfp

such post-publication evaluations benefit from increased impact, recognition, and citation rates.

Experience in computing research has demonstrated that a range of straightforward mechanisms can be employed to encourage authors to produce reproducible work. These include: requiring an explicit commitment to an intended level of provision of reproducible materials as a routine part of each paper's structure; requiring a detailed methods section; separating the refereeing of the paper's scientific contribution and its technical process; and explicitly encouraging the creation and reuse of open resources (data, or code, or both).

- This document provides links and resources to the following:
- template instructions for authors
- examples of authorial statements of commitment
- template guidelines for reviewers
- lists of resources (such as trustworthy repositories and tools)
- lists of examples of publication venues that have implemented such measures
- list of exemplary papers

The list of objects in the bulleted points would need to be assembled and made available.

6.5.1 Template Instructions for Authors

Following are two simple examples that capture the "pay it forward" benefit to the community of having papers that are explicitly designed with reproducibility in mind.

Version 1

The [insert name of journal/conference] encourages authors to provide their work in a way that enables reproduction of their outcomes. Just as you have benefited as an author from the work you cite in your paper, and the tools and resources that others have provided, your efforts will also assist the community, including your future collaborators, if you provide access to and understanding of the tools and resources that you have used and created while carrying out your project. We therefore encourage authors to include in their papers detailed explanations of how their work might be reproduced by others in the field, and to accompany their papers with links to data and source code.

Version 2

The [insert name of journal/conference] encourages authors to provide their work in a way that enables reproduction of their outcomes. Just as you have benefited as an author from the work you cite in your paper, and the tools and resources that others have provided, your efforts will also assist the community, including your future collaborators, if you provide access to and understanding of the tools and resources that you have used and created while carrying out out your project. We therefore request that authors include in their papers detailed explanations of how their work might be reproduced by others in the field, and to accompany their papers with links to data and source code if it is possible to do so. Authors can request separate reviewing of the reproducibility of their work, a category of publication that we explicitly acknowledge.

Example of supplementary statement

In order to support these expectations authors are encouraged to include a detailed methods section in their paper that describes the techniques, tools, data resources, and code resources that enables readers to easily reproduce the work. Such a methods section is of greatest benefit to the reader when it is linked to materials stored in a trusted open repository, and these materials include illustrative or complete data, and code that can easily be re-used and understood.

References

1 Christian Collberg, Todd A. Proebsting: Repeatability in Computer Systems Research. Communications of the ACM, Vol. 59 No. 3, Pages 62-69.

2 B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, T. Yarkoni: Promoting an open research culture. Science 26 Jun 2015, Vol. 348, Issue 6242, pp. 1422-1425

6.6 What Makes A Reproducibility Paper Publishable

Alistair Moffat (The University of Melbourne, AU), Shane Culpepper (RMIT University – Melbourne, AU), Arjen P. de Vries (Radboud University Nijmegen, NL), Carole Goble (University of Manchester, GB), Randall J. LeVeque (University of Washington – Seattle, US), Mihai Lupu (TU Wien, AT), Andreas Rauber (TU Wien, AT), and Justin Zobel (The University of Melbourne, AU)

The nature of science is to improve our understanding, and to build on the work that precedes our own. This means that there will be occasions when of necessity we repeat, review, re-implement, or re-execute experimental work that has been published by others. The goals of such a review may include: extending, establishing limitations/scope/applicability of, confirming, validating, and in some cases invalidating, the previous work. Particular cases might include ones in which analysis is provided that yields a mathematical underpinning to empirical results.

Figure 3 provides guidance as to what might be regarded as being interesting findings in such work, and hence publishable. In some ways, what is being described in the diagram is the essence of the scientific method. It should not be regarded as being a failure, or even unusual, for previous work to be shown to be of limited applicability, or inaccurate when viewed through a more precise lens. Authors who improve on the work of others should always acknowledge their debt to the authors of the prior work and be graceful in their criticism, even when pointing out flaws and errors. The key attitude must be one of "standing on the shoulders" rather than "kicking in the shins".

To warrant publication, a paper must do more than merely take existing code and existing data and repeat a previous experiment to obtain the same result. Work becomes interesting when it adds value or insights or defines applicability by documenting extensions, or in a

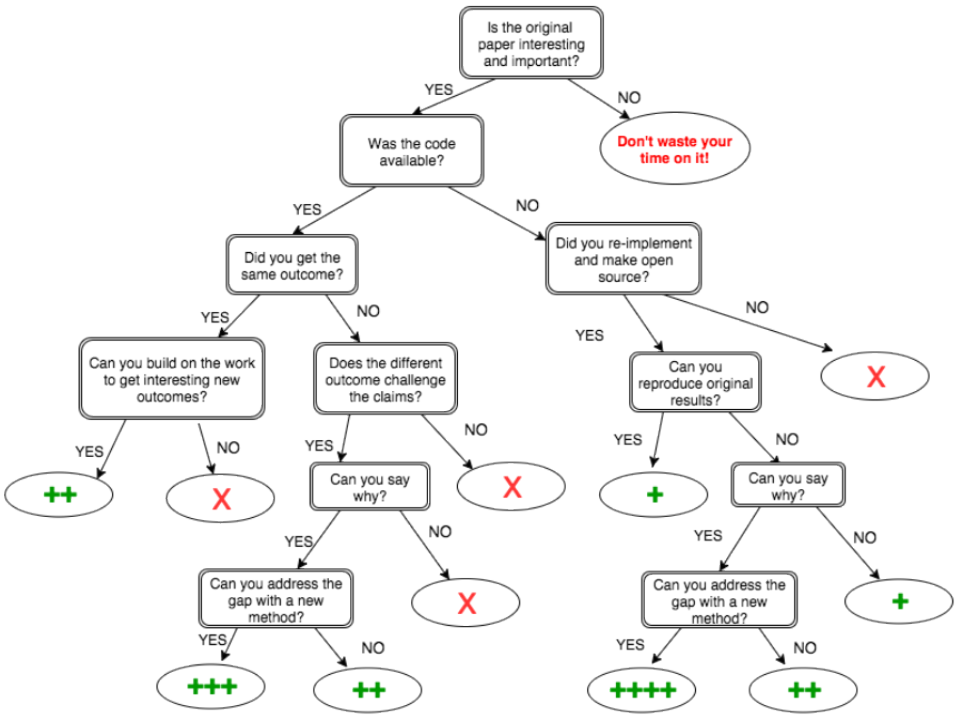

Figure 3 How publishable is a paper attempting to independently replicate/reproduce an earlier paper? The decision tree is intended to give guidance to someone considering writing a paper that reports on an independent study of a previouslu paper (or an editor/reviewer considering such a paper). In some cases "code" could be replaced by "code and data". Data considerations may be different, but similar principles apply.

limiting sense, crystallizing exceptions. With this in mind, the leaves in the diagram that are marked as "X" represent outcomes that are less informative, or unsurprising, outcomes. Conversely, the leaves marked with "+" represent outcomes where ongoing research activities will benefit from publication of this new work.

The number of "+" marks in each leaf is intended to provide some level of guidance as to the value of such a paper, but should not be regarded as being prescriptive. The single node marked as "++++" represents a clear and unambiguous contribution, and in some ways is the ideal. But note that the pathway to that node commenced with a careful review of some previous work – a baseline or starting point – and must still be regarded as having been initiated by a reproduction of that prior activity. As already noted, that is the nature of science: to demonstrate in deeds that previous thought and understandings of the subject can be enhanced or improved. From the point of view of the authors of the prior paper, having that work selected for an in-depth re-evaluation should be regarded as being a sign of respect, and of recognition of contribution. Einstein improved upon Newton, and yet Newton was by no means dumb.

Because of its simplicity, the diagram fails in some cases. One might choose to independently reimplement a method even if code exists; in this case, the results are strengthened if the new code is validated relative to the old experiments before any new experiments are undertaken. This would add to the new work, rather than weaken it.

In summary, papers should always be evaluated on their merits, rather than formulaic requirements. Editors and reviewers need to be aware of the benefits of work that reproduces,

extends, or otherwise refines the work of others, and be encouraging and supportive of authors who pursue these goals. The work that gets encapsulated in such publications should not be regarded as being second-class, or be dismissed as being insufficiently novel.

6.7 Incentives and barriers to reproducibility: investments and returns

Paul Rosenthal (TU Chemnitz, DE), Rudolf Mayer (SBA Research – Wien, AT), Kevin Page (University of Oxford, GB), Rainer Stotzka (KIT – Karlsruher Institut für Technologie, DE), and Evelyne Viegas (Microsoft Research – Redmond, US)

There have been many studies on what motivates people to change their behavior in their personal or professional lives [1]. Simply put, motivation is driven by need: the greater the need the greater the motivation. Other studies have focused on incentives as a means to change behaviours [2] while emphasizing the necessity to engineer the incentive to align with the need and goal to achieve and avoid the incentive to backfire [3]. For instance, as part of reaching better reproducibility, one should not build an incentive towards just making (any) code available, but rather the incentive should be built to make good quality code.

6.7.1 Investments of value

Table 5 lists the different investments of value from the different actors to achieve good reproducibility. The definition of actors is inspired by Section 6.4, with minor differences:

- **Creators** are persons responsible for the creation of possibly reproducible scientific artifacts, i.e. students, authors, academic leaders/lab directors, research software engineers, thesis supervisors, industrial researchers
- **Enablers** are persons and institutions enabling research to be conducted and published, e.g. funders, publishers, editors, institutions, academic leaders/lab directors, data providers, thesis supervisors, digital archives, professional societies, industry, research software engineers
- **Consumers** are persons and institutions consuming and utilizing scientific artifacts, e.g. readers, authors, students, policy makers, educators, adopters, technical communities, IT services, industrial researchers, users, research software engineers, PhD students
- **Examiners** are persons examining the quality of scientific artifacts, e.g. reviewers, thesis examiners, research evaluation committees, funders, policy makers, institutions, professional societies

Table 5 indicates a set of examples for each type of investment and which actors have to invest them.

6.7.2 Returns of Value

Table 6 indicates the returns of value from good reproducibility broken down with respect to the different actors.

6.7.3 Incentives

Having documented the investments, returns of value and the needs for reproducibility per actor (see Section 6.4), in the following we look at incentives required to transition from

Table 5 Investments of value for creators (creat.), enablers (enabl.), consumers (consum.) and examiners (exam.)

Investments	Creat.	Enabl.	Consum.	Exam.
Artifact preparation (clean code, negotiate rights of code and data, annotate, document code and data, support incentives to promote reproducibility)	x			
Research documentation (carefully and reproducibly document all research steps, enable access to documentation)	x			
Education (training good reproducibility structures and methods, examine reproducibility knowledge)	x			x
Infrastructure (establish systems for reproducing computations, to publish reproducible research, and to review reproducibility)	x	x		
Citation (careful creating/citing literature, software, and data)	x		x	
project resources (enable researchers to make investments into reproducibility with respect to time and work force)		x		x
publication guidelines (prepare guidelines for authors and reviewers of publications with focus on reproducibility)		x		x
time for reviewers (give reviewers time to assess the reproducibility documentation of publications)		x		x
principles (create software citation, data management, and reproducibility plan principles)		x		x
requirements validation (establish methods to validate proposal with respect to established principles, adapt panel behaviours to follow principles)		x		x
credit mechanisms (establish community and institutional mechanisms to credit reproducible research)	x	x		

current behaviours to the desired ones to reach reproducibility. We do so in the context of four categories (natural, moral, financial and coercive), and their relationship to the actors. We adapted the categories of incentives from McClelland [3] and Dalkir [4] to address reproducibility as follows:

- **Natural Incentives**: an actor applies her/his curiosity towards PRIMAD, searches for the pursuit of true science, or wants to participate to accelerating research and innovation for the benefit of the social good in the world.
- **Moral incentives**: the choice made by the actor of embracing PRIMAD (see Section 6.1) to make her/his work reproducibility is widely regarded as the right thing to do, or as admirable and the actor can expect a sense of self-esteem, while the failure to adopt PRIMAD is condemned as the wrong thing to do, or as condemnable, and the actor can expect a sense of guilt.
- **Financial incentives**: the actor can expect some form of material reward (e.g. prize, grant, and more generally money) – in exchange for making her/his work reproducible.
- **Coercive incentives**: the actor failing to embrace PRIMAD will see her/his reputation shaken, portfolio of opportunities (e.g. grants, government budget) diminished.

Table 6 Possible returns of value for creators (creat.), enablers (enabl.), consumers (consum.) and examiners (exam.)

Returns of Value	Creat.	Enabl.	Consum.	Exam.
Publicity (more citations and promotion for papers, code, and data, awareness of own and other communities, visibility for possible industry partners)	x	x		
Insight (better estimation of costs for reproducibility, easy incorporation into future proposal and plans)	x			
Impact in industry (commercialisation of research, recognition of results in industry, impact of research in industry)	x	x		
entry into industry (knowledge entry into industry eased, acceleration of transfer, wider economic value and relevance of research, easier reuse for industry)		x	x	
personal satisfaction (providing good reproducibility can give good conscience and satisfaction due to the good cause)	x	x		
incorporation in teaching (reduced preparation time and costs for education by using reproducible research artifacts)			x	
research reuse (easier, quicker, and more reliable research, building on reproducible results, code, data, and methods)	x	x	x	
innovation (more innovation through saving time to reproduce)			x	
ease of reproducibility (well-established mechanisms of reproducibility and accountability through introduction of common culture)	x	x	x	
funding effectivity (funding agencies get reproducible and reusable research, ineffective duplicated investigation and implementation is avoided, greater funds are conserved for novel research)		x	x	
interdisciplinarity (research between agencies, institutions, and labs becomes easier through a common ground of reproducibility)	x	x	x	
comparability (easier comparison with state of the art methods)			x	x

The principles underlined in the 4 reproducibility incentives categories are proposed to help design incentives that meet the reproducibility needs of each community and we expect that they will vary across communities, cultures and actors.

We propose below some examples of incentives per actor:

- The researcher who embraces PRIMAD creates a financial incentive (e.g. app research store) to get more investment from the funders, from industry into her/his research
- The researcher/community who embraces PRIMAD creates a coercive incentive (e.g. "no PRIMAD" stamp) for funders who ignore PRIMAD cost in research
- The community creates a natural incentive (e.g. best reproducibility award) for the researcher to make her/his research reproducible.
- The community creates a moral incentive (e.g. hall of fame) for the researcher to make her/his research reproducible
- The funders create a natural incentive (e.g. interdisciplinary badge) for the researcher to make her/his research reproducible where research is reused across scientific areas
- The funders create a coercive incentive (e.g. grant application section on reproducibility) for the researcher to make her/his research reproducible
- The funders create a financial incentive (e.g. grant, in kind resources) for the researcher to make her/his research reproducible

In structuring these incentives we also note the potential for deferred returns of value to act as a barrier for adoption and implementation of reproducibility. Where an actor must make an investment of value (Table 5), frequently as an individual, a significant period of time before reaping an equivalent or greater return of value (Table 6), often through membership of a community, the interim "debt" may become a disincentive to make that investment; i.e. beyond principled or altruistic motivations it may be difficult to justify that investment above the many other demands for priority faced by researchers and their organisations. As such, despite the long-term sustainability of reproducibility as an economic system through a beneficial cycle of investment and returns, it may be desirable – perhaps necessary – for enabling organisations to provide an initial pump priming investment of value to provide a "bridging loan" to creators until the system is self-sustained.

References

1 The World Bank Group. Theories of Behavior Change, Communication for Governance and Accountability Program.

2 Rothman AJ. Initiatives to Motivate Change: A Review of Theory and Practice and Their Implications for Older Adults. In: National Research Council (US) Committee on Aging Frontiers in Social Psychology, Personality, and Adult Developmental Psychology; Carstensen LL, Hartel CR, editors. When I'm 64. Washington (DC): National Academies Press (US); 2006.

3 McClelland, David C. (1987). Human Motivation. CUP Archive.

4 Dalkir, Kimiz (2013). Knowledge management in theory and practice. Routledge. McClelland, David C. (1987). Human Motivation. CUP Archive.

6.8 User Studies in IR

Nicola Ferro (University of Padova, IT), Norbert Fuhr (Universität Duisburg-Essen, DE), Kalervo Järvelin (University of Tampere, FI), Noriko Kando (National Institute of Informatics – Tokyo, JP), and Matthias Lippold (Universität Duisburg-Essen, DE)

The goal of information retrieval (IR) is to best serve a user information need by presenting him/her with a list of documents (information objects) potentially relevant to this need. This calls for specific evaluation methodologies which take into account the user, since determining the quality of a produced ranking, i.e. the effectiveness of a system, is directly depending on the user notion of what is satisfactory for his/her information need.

This setting is quite different from what we have, for example, in databases, where queries are exact and the correctness of results is not an issue, putting the emphasis on efficiency rather than effectiveness.

Therefore, it becomes central to understand what reproducibility is and how it can be achieved when users are in the loop.

6.8.1 Methodological Background

6.8.1.1 Experiments in psychology

The knowledge acquired in psychology is based on empirical results of experiments. An experiment is a research method in which one or more independent variables (IV) are manipulated to determine the effect(s) on a dependent variable. Other relevant factors need to be controlled in this setting. For instance, in the case of a user experiment in information retrieval, the independent variable could be a different search algorithm and the dependent variable could be the time to finish the search.

Psychological experiments needs to fulfil three criteria: **validity**, **objectivity** and **reliability**.

6.8.1.2 Validity

They need to be valid, which it is when the measures what it claims to measure is really measured. A problem could be that some participants might not be paying attention during the experiment, because of a lack of motivation. In some cases a manipulations check, which tests the attention of the user can be useful.

6.8.1.3 Objectivity

Objectivity is also important. An experiment has to be objective in two ways, the result of the experiment should not be influenced by the experimenter and that the interpretation of the data should not depend on the examiner.

6.8.1.4 Reliability

An experiment has to be reliable. When you repeat your experiment or another person repeats your experiment should come to a similar result. To ensure reliability, scientists have to specify their experimental design, they have to describe the conditions, under which the experiment is conducted and share information about the participants. The material and the

raw data of the experiments needs to be stored and shared on demand by the corresponding author.

6.8.1.5 Reproducibility crisis in psychology

In a recent study (Open Science Collaboration, 2015) the results from 100 experiments from four top journals could just be partially replicated. That started a big discussion about the reasons.

6.8.1.6 Reasons for failed replication

Theoretical reason can be in the theories selection itself. If you have an **ill-defined theory**, which does not specify the outcome of the experiment and you use the result of the experiment as evidence for you theory, then the result did not matter and most likely can not be reproduced. For the IR experiments it might be necessary to define for which population the tools are produced and if the result can be generalized for all possible users. Older people might use the search engines in a different way than students do, which usually are the participants of the experiments.

Another theoretical threat are post theories and **post hypotheses or predictions**. If the hypotheses and the theoretical background are selected after the result of the experiment is known, you can not claim that you knew before. When this is happening the probabilities and the p-values are wrong.

Concerning the methodology, this is also a problem in psychology. Researchers rely almost exclusively on the p-value and **do not consider the effect sizes**, which are more important. The question in IR should not primarily be, is there a difference, but how big is the difference and would the user actually notice this difference. Furthermore, a lot of experiments are conducted with **low statistical power**, so the effect in this kind of experiment might not be the real effect and a replication can not find this result.

6.8.2 Context of User-oriented IR Evaluation

In IR, we have different kinds of user studies:

- laboratory experiments, where users are observed in the lab
- in situ observation of users at their workplace
- living labs, where the researcher analyses the system logs and possibly also manipulates the system employed by the users for their daily work.

Besides these types of experiments, there are studies that focus mainly on data collection methods, for which the discussion below only partially applies:

- exploratory user studies,
- focus groups, where researchers interview users
- longitudinal studies of users.

For discussing the reproducibility issues for the specific case of user studies, we follow the PRIMAD model (see Section 6.1) described above:

- Research objective is the research question to be addressed. In most cases, this part should also include the hypotheses to be tested with the experiment described in the remainder of the research paper.
- Model relates here to the experimental settings, which are used for testing the hypotheses specified before, So, besides the type of study, also the relevant aspects of the settings that refer to the research objectives are part of the model

- Implementation and Platform correspond here to the environment in which the study was carried out. Besides the system used for the study, also the group of users participating in the study as well as the exact conditions under which they participated belong to this aspect.
- Actor is the experimenter. In cases where the experimenter has direct contact with the users, the actor might have influence on the results of the study. Thus the actor should be kept constant throughout the study
- Data has a twofold meaning in user studies. First, there is the data that comprises the so-called testbed, like the document collection, the tasks carried out by the users, etc.. Second, there is the observation data collected throughout the study (thus, the user is regarded here as a data generator))

For enabling reproducibility, a researcher should share this context with other users to the maximum extent possible. Research objective and method are usually described in the research paper. In the past, the main research objective was the effectiveness of the methods investigated. Nowadays, also other aspects are considered, which are either more closely related to the actual user task, or to more subjective factors such as user satisfaction or engagement (which, in turn, can be measured via different variables). The more factors are considered, the more it becomes important to state the research hypotheses before actually carrying out the study, in order to achieve statistically valid results.

The environment usually can only be shared partially (mainly the system), while most other aspects (e.g. the users, the hardware, etc.) should be described at a reasonable level of detail in order to ease reproducibility. The same holds for the actor.

For the data, sharing testbeds is widely accepted nowadays, since the state of the art does not allow yet to characterize testbeds to such an extent that an independent researcher would be able to create a comparable testbed that could be expected to give the same results. The observation data, on the other hand, is essential for verifying the claims of a research paper. To a limited extent, it also can be used for simulation studies, depending on the degree of interactivity involved in the study (in classical IR experiments, the only data of this kind are relevance judgments).

6.8.3 Barriers/Obstacles

The research on Information Retrieval (IR) using computer started in 1950s and is said that IR is the first area in computer science using the human judgement as a success criteria of the technology [1]. This makes IR interesting and complex, and therefore the IR community has a strong tradition on evaluation to cope with how users incorporated in the experiment and testing, and make the reasonable comparison across the systems and the algorithms in the same system. Moreover, the commercial online search services started in 1960s and then the issue of working with real users in an interactive environment came up.

Since the Cranfield project in early 1960 [2], researchers constructed and shared testbeds called “test collections” which consist of the three types of data: document collections, the set of search requests, and the static set of human relevance judgment for each search request on the document collection. Such re-usable static testbeds were shared by the community as an infrastructure for the comparative evaluation and as one of the major elements for reproducibility of the experiments.

However, the technology and the society evolved tremendously: interactive online search for various purposes by ordinary persons became pervasive in everyday life, the various data collections including various web services and the social media are enhanced, search

on multiple devices for multi-tasking become more common. The traditional evaluation paradigm (based on the batch-style one-time judgments)can not cover all the problems in the IR research and we are facing various new challenges and obstacles to make the research reproducible:

For studying users in interactive IR, there are various barriers and obstacles for reproducibility:

- Privacy/limitations of anonymizing
- Confidentiality
- Volatility of data (live streams, when the same situation never happens again, etc)
- Validity of the data: the data is so multidimensional that it is difficult to ensure the external validity of the experiment. This complexity is present also in IR test collection, and even more if we consider dynamic test collections.

Online web services are generally based on algorithms using user behaviour data in some way. This data is intrinsically rich in privacy and often includes confidential information. With interactive research IR systems, the situation is similar. Although various research efforts have targeted anonymization, there are still limitations, and these make it difficult to release the user behaviour data for external research groups, which, in turn, hampers reproducibility.

Large-scale users logs are generated in commercial search services and substantial studies on modeling and predicting users behaviour have been conducted based on these data, but again, the underlying data is not accessible for other researchers. Not only user modeling studies, but also various operational search mechanisms exploit user behavior data in the search and ranking algorithms, thus making it difficult to reproduce these methods.

To tackle the problems of the document data with privacy information and/or copyright problems, various evaluation-as-a-service approaches have been proposed and some of them were implemented successfully. However, these are still not sufficient for all the data produced by the users in-situ and lab environments.

For volatility, IR experiments can be conducted on live streaming data or commercial search services, in which the data and algorithms are continuously changing and the same data will never obtained again. Also, user experiments can not be "re-run" with the same users as the users learn from the previous experience.

In IR, interactivity and user behaviour or search experience through whole search sessions (or sometimes even a task involving multiple sessions) become more important, in order to consider real-world contexts. Various algorithms and softwares to support such interactions have been studied and proposed. The data obtained from the users in such task-based or whole session-based studies are highly complex, comprising e.g. the nature of the tasks conducted as well as the characteristics of each user. More research is needed for developing a framework that is able to describe such complex, multi-dimensional data as well as for devising methods for proper scientific analysis of the data collected.

6.8.4 Actions to Improve reproducibility

Actions to improve reproducibility of user-oriented experiments include checklists for authors (and reviewers, editors, chairs, etc.), sample exemplary papers, method inventories, extended methodological sections in papers, and critical discussions on the components/tools/other data used. These are considered briefly below:

Checklists should be provided on the methodology applied in the study. Kelly [3] is a useful source for constructing a checklist for user-oriented IR studies. Examples of items to check are:

- Research questions and experimental design (latin square, intra/inter subject, etc.)
- Participant characteristics and the population they are claimed to represent
- Methods of data collection, including the experimental protocol, environmental conditions, and variables used in the study (how to describe, how to measure, operational definition, observables)
- Experimenter
- Retrieval systems and their interfaces
- Methods for data analysis, including assumptions of statistical analyses (and adjustments if assumptions are not met)
- Degree of control on the system by the experimenter

Exemplary papers representing various types of user experiments could be offered in some community-based repository and annotated for their strong features, see also next section.

Inventories of typical variables various types of user experiments and standard ways to operationalize and measure them in different (sample) study settings could be provided by the community), as further discussed below.

Methodological sections could be emphasized in document templates, author guidelines and review guidelines. More space might be allocated to these sections and or authors encouraged to provide methodological appendixes or technical reports.

Finally, authors could be encouraged to *critically discuss* how suitable the set of tools and collections is to answer the research questions, what claims can the tools/collection support, describing the generalizability of the findings on the basis of the tools/collection that have been used.

6.8.5 Community Support to Reproducibility

In order to embody the vision described above and foster reproducibility in user-oriented studies, the involvement of the research community is crucial and it should consist of two complementary actions:

1. Support to the creation of shared resources;
2. Taking up and implementation of shared practices.

When it comes to *shared resources*, we can foresee several examples of them:

- **Inventories:** in order to streamline the reproducibility process, there is a need for catalogues accounting for the most appropriate experimental designs, the kind of independent and dependent variables you typically encounter in these settings, how to describe and measure such variables, the proper data analysis methodologies and statistical validation methods to apply to these variables in the different experimental designs, and so on;
- **Do's and don'ts:** in order to facilitate the understanding and adoption of the above facilitators of reproducibility, real and hands-on examples of appropriate and inappropriate ways to carry out user-oriented experiments are needed to clearly explain why a seemingly appropriate experimental setup is or is not working as expected. This could be partnered with a selection of exemplary and well-known papers, which should be annotated and enriched with links and explanations related to the above inventories, in order to clarify the researcher how and when to apply a given approach by means of concrete and remarkable case studies;
- **Repositories:** the adoption of shared repositories to gather collections of documents, interaction data, tasks and topics, and more is a key step to extend the reach of reproducibility in user-oriented experimentation;

- **Data formats:** the development of commonly understood and well-documented data formats, which can be extended to specific needs, as well as the introduction of proper metadata (descriptive, administrative, copyright, etc.) to model, describe, and annotate the data and the experimental outcomes is a crucial factor in lowering the barriers to reproducibility in user-oriented experimentation.

The methodological instruments, the checklists, the critical discussions, the different kinds of shared resources previously described are all key "ingredients" for successfully reproducing user-oriented experiments but the actual catalyst is the systematic and wide adoption by the community of *shared practices*, effectively exploiting all of these "ingredients", as also discussed in Sections 6.4 and 6.7.

References

1 W. B. Croft. Information retrieval and computer science: an evolving relationship. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2–3. July 28 – August 1, 2003, Toronto, Canada.
2 C. Cleverdon. The Cranfield tests on index language devices. Aslib Proceedings 16, 6:173–194. 1967.
3 D. Kelly. Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval, 3(1–2), pp. 1–224. 2009.

7 Open problems

7.1 Open Research Problems in Reproducibility

Carole Goble (University of Manchester, GB) and Daniel Garijo (Technical University of Madrid, ES)

When referring to reproducibility, we can distinguish two main types of research agendas, each with their scope and social implications. There is a **macro research agenda**, which consists of the topics of interest of the main funding agencies, and a **micro research agenda**, which would consist of the particular topics for new PhD students. While the macro agenda is influenced by the political tendencies of the moment, the micro agenda is influenced by the particular interests of researchers. Reproducibility initiatives may work fine for specific domains, but they may collapse when applying them at a macro level. Since most of the people in the group did not belong to funding agencies, the discussion focused on the micro agenda.

Regarding the social implications of reproducibility, an agenda should be issued in terms of productivity. Reproducibility can be seen as an investment for productivity, and part of its agenda should study and make explicit the correlation between these two features. Another challenge is addressing how the quality and quantity of the research work is affected by reproducibility. Currently, when given the opportunity, a researcher will choose to publish two publications rather than a highly reproducible one. **It is important to be able to show the long term value of high quality reproducible work**.

Another important aspect to take into consideration is the analysis of infrastructure, which includes the improvement of record keeping. The best way of holding trusted resources is to convince institutions to get involved. **Labs, companies and people are temporary,**

while institutions tend to last even centuries. In order to achieve this, it is critical to address the intellectual property rights of the resources to be archived. Having company-friendly licenses may help in their adoption.

7.1.1 Open Research Challenges

We summarize the main challenges related to reproducibility in the list below

1. What are the interventions needed to change of behavior of the researchers? Making a paper reproducible is often related to the ways people are used to work within a given community. Knowing which are the necessary changes to change the behaviour of scientists towards adopting reproducibility may help to make the transition in a more effective way.
2. Do reproducibility and replicability translate in long term impact for your work? By showing empirical proof of the impact of reproducible versus non reproducible work on a community, more authors may be convinced on adopting a reproducible approach.
3. How do we set the research environment for enabling reproducibility? If making a paper reproducible takes a lot of time, people will not do it. Instead, working towards the creation of environments for enabling reproducibility seems like a more sensible approach.
4. How can we obtain long term digital archiving? Having a long lasting record of the resources used for a paper is a crucial requirement for reproducibility. Existing institutions (libraries, church) have archived successfully knowledge for centuries, and we should learn their methods and apply it to software as well.
5. How can we track the components that are part of the materials that have been used in a project? A researcher may forget to include data considered trivial in an experiment, but that same data may crucial for another researcher that aims to reproduce it years later. Is it possible to auto-document your research?
6. Is it possible to define roles of contributors for getting the credit for a work? Capturing the finer grain of contributions is crucial to provide appropriate credit to all the contributors of a research work. Some initiatives have started proposing sharing taxonomies[28] and distribute credit [1], which are a first step towards addressing this challenge.
7. Can we measure the cost of reproducibility/repeatability/documentation? What are the difficulties for newcomers? Understanding how to lower the barrier for adopting reproducibility and its costs is another of the key aspects to take into consideration when convincing a community to make their work reproducible.

References

1 D. S. Katz, and A. M.Smith. (2014). Implementing transitive credit with JSON-LD. arXiv preprint arXiv:1407.5117.

[28] http://casrai.org/CRediT

Participants

- Vanessa Braganholo
Fluminense Federal University, BR
- Fernando Chirigati
NYU Tandon School of Engineering, US
- Christian Collberg
University of Arizona – Tucson, US
- Shane Culpepper
RMIT University – Melbourne, AU
- David De Roure
University of Oxford, GB
- Arjen P. de Vries
Radboud University Nijmegen, NL
- Jens Dittrich
Universität des Saarlandes, DE
- Nicola Ferro
University of Padova, IT
- Juliana Freire
New York University, US
- Norbert Fuhr
Universität Duisburg-Essen, DE
- Daniel Garijo
Technical University of Madrid, ES
- Carole Goble
University of Manchester, GB
- Kalervo Järvelin
University of Tampere, FI
- Noriko Kando
National Institute of Informatics – Tokyo, JP
- Randall J. LeVeque
University of Washington – Seattle, US
- Matthias Lippold
Universität Duisburg-Essen, DE
- Bertram Ludäscher
University of Illinois at Urbana-Champaign, US
- Mihai Lupu
TU Wien, AT
- Tanu Malik
University of Chicago, US
- Rudolf Mayer
SBA Research – Wien, AT
- Alistair Moffat
The University of Melbourne, AU
- Kevin Page
University of Oxford, GB
- Raul Antonio Palma de Leon
Poznan Supercomputing and Networking Center, PL
- Martin Potthast
Bauhaus-Universität Weimar, DE
- Andreas Rauber
TU Wien, AT
- Paul Rosenthal
TU Chemnitz, DE
- Claudio T. Silva
New York University, US
- Stian Soiland-Reyes
University of Manchester, GB
- Benno Stein
Bauhaus-Universität Weimar, DE
- Rainer Stotzka
KIT – Karlsruher Institut für Technologie, DE
- Evelyne Viegas
Microsoft Research – Redmond, US
- Stefan Winkler-Nees
DFG – Bonn, DE
- Torsten Zesch
Universität Duisburg-Essen, DE
- Justin Zobel
The University of Melbourne, AU