Collaborative platforms for streamlining workflows in Open Science

Konrad U. Förstner[1, 2], Gregor Hagedorn[3], Claudia Koltzenburg[4], M Fabiana Kubke[5], and Daniel Mietchen[6]

[1]Institute for Molecular Infection Biology, University of Würzburg, D-97080 Würzburg
[2]Research Centre for Infectious Diseases, University of Würzburg, D-97080 Würzburg, Germany
[3]Julius Kühn-Institute, Federal Research Center for Cultivated Plants, Berlin, Germany
[4]Managing editor of Cellular Therapy and Transplantation (CTT)
[5]Department of Anatomy with Radiology, University of Auckland
[6]Science 3.0

May 9th, 2011

Abstract

Despite the internet's dynamic and collaborative nature, scientists continue to produce grant proposals, lab notebooks, data files, conclusions etc. that stay in static formats or are not published online and therefore not always easily accessible to the interested public. Because of limited adoption of tools that seamlessly integrate all aspects of a research project (conception, data generation, data evaluation, peer-reviewing and publishing of conclusions), much effort is later spent on reproducing or reformatting individual entities before they can be repurposed independently or as parts of articles.

We propose that workflows - performed both individually and collaboratively - could potentially become more efficient if all steps of the research cycle were coherently represented online and the underlying data were formatted, annotated and licensed for reuse. Such a system would accelerate the process of taking projects from conception to publication stages and allow for continuous updating of the data sets and their interpretation as well as their integration into other independent projects.

A major advantage of such workflows is the increased transparency, both with respect to the scientific process as to the contribution of each

participant. The latter point is important from a perspective of motivation, as it enables the allocation of reputation, which creates incentives for scientists to contribute to projects. Such workflow platforms offering possibilities to fine-tune the accessibility of their content could gradually pave the path from the current static mode of research presentation into a more coherent practice of open science.

1 Introduction

Like most areas of today's life, science has dramatically changed since the advent of the internet. However, the transformation that has taken place until now is just the tip of the iceberg. In the following, we want to discuss the mostly underutilized potential of representing all aspect of science in collaboratively used online workflow platforms. Since such platforms could help to realize Open Science, transparency of the funding cycles and access to all data in the research process, we will shed light on this special aspect and make recommendations regarding implementations.

While there are numerous projects developing and applying so called Virtual Research Environments (VRE) - also known as *Collaboratories* - covering selected stages of the scientific process, a platform spanning every phase is missing so far [1]. Technically overcoming such gaps and creating a seamless transition from bench to publication could speed up the research and, with it, the generation, distribution and reuse of knowledge.

2 The scientific workflow in open VREs

2.1 Conception and project planning

Independent of the nature of a research endeavor - hypotheses-driven or data-driven, performed by a single person or a team - a solid conception phase is the crucial basis for every project. Despite today's common practice of limiting this phase to a small group of people, utilizing collective intelligence during the conception phase could help to avoid redundant research and to improve the design of the study. As the complexity and scope of scientific projects are increasing, the application of project management tools can be useful for managing the processes and parties involved.

2.2 Experiments and data generation

Today, data generation in academic research continues to rely strongly on manual labor. While this is mostly due to the relative low cost of labor force resulting from the academic system and the limited interdisciplinary education of science and engineering, the high potential of automation is mostly neglected.

Not only could the efficiency of invested labor be improved by automation, but also reproducibility could be significantly increased. To make this affordable for the broader research community, a shift from siloed proprietary devices to well-documented pieces of standardized, open-source hardware developed by the scientific community itself in cooperation with potential vendors is needed. Open hardware platforms like Arduino [2] could offer starting points for such a development and first example of such tools are available (e.g. OpenPCR [3]). The devices could and should enrich the primary data with further metadata, convert them into semantified formats and directly upload the output into online repositories.

One promising example which visualizes the potential of such automation of otherwise quite labor-intensive research is the robot scientist ADAM [4]. The streamlining of mechanical steps and the evaluation of results would benefit from formal languages that describe the necessary procedures and make the design and exchange of experimental setups easy [5]. As a long term goal, scientists would mostly engage in programming experiments and engineering the system to automate those steps that have been performed manually so far. The motto "work on the system, not in the system" should guide this development.

2.3 Data release

The online release of experimentally generated data should be done shortly after the generation and can potentially happen in real time. Downstream analysis within the research project but also the reuse by other parties should be kept in mind when selecting data formats. These should, as far as possible, be non-proprietary, machine readable (semantically enriched) and common for the respective domain of research. If no format fulfills all these requirements, the conversion into alternative formats should be permitted. Access to the data could take place via a web interface or domain specific clients. Especially for large or highly accessed data sets, the additional distribution via peer-to-peer networks is recommended.

2.4 Data analysis

Since every step in the data analysis should be transparent and easily reproducible, it should take place preferably in the proposed platform, too. Systems like the analysis workflow tool Taverna [6] could be used for such processing. Already today, many research institution offer grid computing infrastructure for such purposes. Analyses using external tools, especially GUI-tools that do not offer any possibility to log the performed actions, should be avoided if possible, as otherwise documentation has to be created manually. For some computationally intensive analyses the use of shared systems is a more economical usage of the needed infrastructure, provided the management overhead does not exceed the computational efficiency gain. As done for the raw experimental data,

the protocols and the result of the data processing should be documented and stored in repositories to be accessible.

2.5 Knowledge generation

The results of analytical processing as well as the raw data can be used by scientists - or machines [7] - to draw conclusions and to generate knowledge out of the available information in a well documented way. The platform should assist to make this happen collaboratively by offering commenting and rating of statements. Discussions - text, audio- and/or video-based - should be recorded to make the path to finding reconstructible.

2.6 Final publication

As documentation of every step is an inherent feature of the workflow, the final publications resulting from a study can be short reports linking to the major outcomes and putting them into the scientific context. The platform should offer functionalities to perform open peer-review of this final report.

3 Implementation

3.1 Technology

As shown above, the many building blocks of a complete scientific workflow already exist and only need to be connected seamlessly. The development of open standards defining the required interfaces of these parts could enable different parties to assemble the pieces into a consistent workflow and to add further needed parts. This would offer the possibility to implement a platform either as one monolithic application or as separate interacting and exchangeable units.

3.2 Funding

Of similar importance as the technical realization is the adaptation of scientific culture and funding policies. While research institutions like the National Institutes of Health (US) or the Welcome Trust (UK) already require open access for final peer-review manuscripts that results from research they funded [8, 9], the regulations are much weaker for the underlying data, and almost nonexistent for proper annotation. However, the first attempts to establish such requirements are on the horizon [10].

3.3 Licensing

The default copyright restrictions in most jurisdictions hamper the reuse of data. It is therefore highly desirable that, with very few exceptions, each entity generated in the research process is explicitly published under a less restrictive license, e.g., the ones offered by Creative Commons [11] or is released into the

public domain. As the latter concept may differ or be missing in some countries, release through the CC0 license [12] is recommended.

3.4 Reputation

The gain of reputation is the most important incentive for scientists. It is currently mostly determined on the basis of publications in scientific journals and the related measure of success in funding applications. As every contribution to a research project can be attributed to a distinct person and could be rated by others, the allocation of reputation is an inherent element of the proposed platform. The connection to research identifiers like ORCID [13] and the analysis of such microcontributions could assemble a precise image of a scientist's skills and achievements.

4 Challenges

As stated above, considering the allocation of reputation and funding in science is crucial when redesigning scientific processes. To bridge a transient phase until the suggested political changes have taken place, fine granular access control in the research workflow platform could permit that the technology is adapted by scientists despite objection regarding the loss of reputation. With such a control in place, the full process could be opened up after the final publication or at any other desired time.

It is very unlikely that there will be one single platform that can fulfill the requirements of all scientific domains. Building and maintaining completely independent platforms for each domains, on the other hand, may not be sustainable. A modular and flexible system, where possible re-using industry standard software is therefore called for.

Projects presently exploring this are, e.g.:

- the eSciDoc platform which builds on the open-source repository software Fedora Commons [14] and is mainly developed for the Max Planck Society has a similar aim and strategy [15].
- the FP7 funded Virtual Biodiversity Research and Access Network for Taxonomy" (ViBRANT) [16, 17, 18]

They span data collection, analysis and publishing (in collaboration with Pensoft Publishers), are based on the established open-source platforms Drupal [19] and Mediawiki [20] and equipped with specific extensions.

References

[1] Annamaria Carusi, Torsten Reimer. Virtual Research Environment - Landscape Collaborative Study. *JISC*. pp. 72-24 2010.

[2] `http://www.arduino.cc/`

[3] `http://openpcr.org/`

[4] Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, Amanda Clare. The automation of science. *Science*. 3 April 2009: Vol. 324 no. 5923 pp. 85-89

[5] Larisa N Soldatova, Ross D King. An ontology of scientific experiments. *J R Soc Interface*. 2006 December 22; 3(11): 795–803.

[6] Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R. Pocock, Peter Li, Tom Oinn. Taverna: a tool for building and running workflows of services *Nucleic Acids Res*. 2006 Jul 1;34(Web Server issue):W729-32.

[7] Michael Schmidt and Hod Lipson. Distilling Free-Form Natural Laws from Experimental Data. *Science*. 3 April 2009: Vol. 324 no. 5923 pp. 81-85

[8] `http://publicaccess.nih.gov/policy.htm`

[9] `http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002766.htm`

[10] Andrew J Vickers. Making raw data more widely available. *BMJ*. 2011; 342:d2323

[11] `http://creativecommons.org/`

[12] `http://creativecommons.org/publicdomain/zero/1.0/`

[13] `http://orcid.org/`

[14] `http://fedora-commons.org/`

[15] Malte Dreyer, Ulla Tschida, Natasa Bulatovic, Matthias Razum. eSciDoc - a Scholarly Information and Communication Platform for the Max Planck Society. *German e-Science Conference, Baden-Baden*. 2007

[16] `http://vbrant.eu`

[17] Dave Roberts, Vince Smith. ViBRANT - Virtual Biodiversity Research and Access Network for Taxonomy. *Tools for identifying biodiversity: progress and problems. Proceedings of the International Congress, Paris*. September 20-22, 2010. Edited by Pier Luigi Nimis and Régine Vignes Lebbe. p. 54

[18] Vladimir Blagoderov, Irina Brake, Teodor Georgiev, Lyubomir Penev, David Roberts, Simon Ryrcroft, Ben Scott, Donat Agosti, Terry Catapano, Vincent S. Smith. Streamlining taxonomic publication: a working example with Scratchpads and ZooKeys. *Zookeys*. 2010; 50: 17–28

[19] http://drupal.org

[20] http://www.mediawiki.org/