Linkflows: enabling a web of linked semantic publishing workflows

Cristina-Iulia Bucur

Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands.
c.i.bucur@vu.nl

Abstract. In recent years, the prevalence of the Internet and Semantic Web technologies has shifted the traditional scientific journal publishing framework towards the digital environment. In support of this, ontology suites like SPAR (The Semantic Publishing and Referencing) are built to support digital publishing. Additionally, new ways to represent fine-grained knowledge in the form of nanopublications, for example, have emerged. These fine-grained technologies facilitate the decomposition of traditional science articles in constituent machine-readable parts that are linked not only with one another, but also to other related fine-grained parts of knowledge on the Web following the Linked Data principles. But, these resulting digital artifacts of fine-grained knowledge are static objects that do not take dynamic processes, or *scientific workflows*, into account. And *scientific workflows* are important because they show how digital objects are produced and consumed. In this project, we enable the decentralized execution of scientific workflows of digital artifacts across platforms such that individual steps of single workflows can be distributed. By considering these *scientific workflows*, we can further find new dimensions with respect to the quality and impact of digital artifacts. In our preliminary results we have used Linked Data Notifications to demonstrate the feasibility of our approach.

Keywords: semantic publishing, digital publishing workflows, scientific workflows, semantic web.

1 Motivation

Publishing is an important practice, not only in the world of science, but also in everyday life. Since a few centuries ago, when Gutenberg invented the publishing press, publishing has proved to be an effective means to share information and knowledge, making them more and more accessible to everyone throughout the centuries. In the recent years, with the pervasiveness of technology and the Internet, we changed not only the way we do science, but also how we perform and disseminate science. As such, scientific publishing became a more versatile and multifaceted process changing completely the initial paradigm of publishing towards a digital environment with new methods of electronic publication including scientific workflows, research protocols and standard operating procedures [1]. Still, traditional article publishing remains an important part of the scientific process of one's research, providing a means not only to spread knowledge and gain a certain validation from the scientific community, but also giving a sense of value to one's research by ensuring reproducibility among others. Especially in the sciences, moving towards a more digital environment and generating digital content seems to be more the rule and challenges the classic ways of publishing.

In this rather new digital publishing context, Linked Data is a framework that supports scientific publications by enabling the exchange, reuse and linking of data on the Web [2]. While the Linked Data set of best practices to connect and publish structured data on the Web is not enough to enable the entire scientific publication process, it is an important layer that facilitates it. Linked Data supports provenance (meta-)information about the resources that are linked, thus giving a way to locate various versions of data and access information like ownership and copyright. In turn this would sustain reproducibility.

Reproducibility plays a crucial role in scientific research because it allows others to test, check and verify the validity of one's claims and methods and it permits further collaboration and reuse of scientific discoveries. A scientific publication is a way to announce and describe a result of interest following research work that can benefit the scientific community as a whole. At the same time, it tries to prove that the claims exposed in the publication are true [3]. For this, it is very important to provide the means to publish these scientific contributions in an understandable way that allows for verification and thus facilitates reproducibility. Unfortunately, according to a recent study published in Nature [4], over 70% of the 1500 interrogated scientists admitted to have failed to reproduce the work of other researchers at some point in time. The FAIR principles for scientific information [5] can be key factors in guiding towards reproducible research. According to these, data should be (i) findable both for humans and machines; (ii) accessible on the long term; (iii) interoperable by the use of shared vocabularies, for example; and (iv) reusable for both humans and machines. And, following these guidelines should, in turn, support reproducibility.

As Mons [6] notices, an important problem with traditional articles and another hurdle in the way of reproducibility is the process of "Knowledge Burying". That is all information is written and published in one bulk of text - the article - that contains the scientific hypotheses, arguments, methods and results. So, in order to extract knowledge from an article and have information in a structured form, additional methods like text mining need to be applied, thus resulting in a loss of knowledge. And, this is true for human readers as well: reading a traditional scientific article is an information retrieval task in itself, independent of the domain of science. Readers know how to recognize parts of the research by following standardized section headlines like "Introduction", "Methods", etc. and read or follow the parts they are most interested in.

To further support provenance and reproducibility in scientific publications, scientific workflows were introduced. A scientific workflow is a mechanism to specify and automate repetitive tasks for computational science or *in silico* science [7]. These scientific workflows are mostly published as digital artifacts that describe experiments and the additional materials needed to understand them [8]. A scientific workflow can be executed and reproduced because it contains a precise and executable description of a scientific procedure [9]. And, as research becomes more data driven, workflows are used in the specification of experiments that retrieve, integrate and analyze datasets using distributed resources [10]. Normally, scientific workflows are represented in the form of directed graphs where nodes represent local or remote tasks, actions or operations and edges represent the dependencies between them. We will consider the notion of scientific workflows in the context of scientific publications. Such a scientific publishing workflow in a digital environment can be considered the reviewing process of open access journals.

2 Approach

In this project we are guided by the following research question:

How can scientific workflows be assessed, linked and decentrally executed across platforms, such that individual steps of a single workflow can be distributed?

Digital artifacts can be considered all objects or resources that belong to a scientific publication, such as text, datasets, code, multimedia objects, spreadsheets, reviews, figures, methods, protocols, and results, as represented in **Fig. 1**. The *scientific workflows* refer to processes, actions or operations that produce or consume these *digital artifacts* like authoring, revising, editing, reviewing, commenting and annotating. A single scientific workflow can be composed of multiple steps and we argue that these various steps can be spread on various platforms like repositories, code bases and collaboration platforms. The innovative aspect of the project is that one platform would not be in full control of the complete workflow, but would provide the means to link to a workflow step as it is produced. Thus, a complete scientific workflow of a digital artifact would then be composed of these workflow steps that are distributed on different platforms. As such, the static digital objects will be linked to the dynamic processes that contain them. Another innovative aspect lies in the fact that new quality and impact measures of digital artifacts can be derived by considering the workflows that consume and produce them.

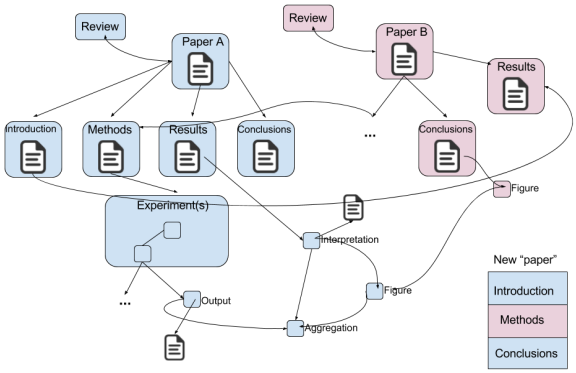

Fig.1. Traditional article represented as interconnected digital artifacts.

The combination of SPAR ontologies [11] will be useful to break down the article into interconnected digital artifacts. PROV-pings [12] and Linked Data Notifications [13] technologies will help in sending notifications about digital artifacts to the interested parties. But, every node in the network, each digital artifact or workflow step, can be considered a fixed and immutable entity. This way provenance and versioning would be possible. Using Trusty URIs [14] and the use of hashes can provide helpful in establishing and tracking the changes that a digital artifact goes through. Plus, various platforms, may refer to different versions of a digital artifact in their workflows.

In this project we will collaborate closely with two organizations: The Netherlands Sound and Vision and IOS Press. They both publish proprietary journals in various domains like the humanities (audio-visual history, science, medicine). Throughout this project, we will implement, test and evaluate the linked workflows (linkflows) on the journals or datasets provided by these two organizations. The diversity of the domains, ranging from audio-visual and multimedia to computer science and medicine would ensure a complex coverage of our project.

3 Preliminary Results

To answer the research question, we will first model the workflow steps of a journal submission using the Linked Data Notifications (LDNs) protocol. For example, let's consider the scientific workflow of reviewing (see **Fig. 2.)**: a digital artifact is submitted for review, then the editor sends it to reviewers that assess it and also determine if other reviewers are needed. Then, the digital artifact can be rejected, accepted or accepted based on modifications. In the end, if all modifications are accepted, the camera-ready version will be ready for publication.

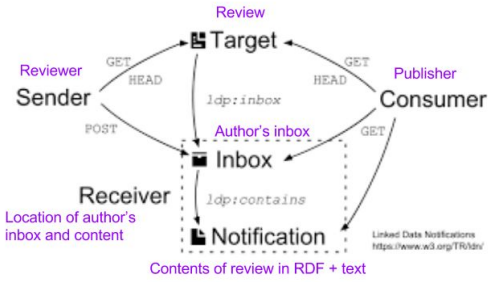

Fig. 2. How LDN maps to a part of the reviewing workflow.

All these steps can be carried out on different platforms and servers, thus produced and consumed in different locations on the Web. For example, both IOS Press and the Netherlands Institute of Sound & Vision can subscribe to get the review notifications concerning a digital artifact and if an acceptance is given for the publishing of the digital artifact, they can include it in their publishing journals as well. As such, reviews can be consumed on any platform or by any system that wants to include it in its publishing workflow.

4 Conclusions

In this project we want to investigate new approaches in the digital environment of scientific publishing by combining Linked Data technologies to address problems like "Knowledge Burying" of traditional articles. Furthermore, we want to provide a framework that supports scientific workflows for digital artifacts. As such, we aim to link and connect the static products of dynamic processes - digital artifacts - to the processes that produce and consume them. The main innovative aspect of this research is the fact that scientific workflows are executed decentrally and linked across platforms, such that individual steps of a single workflow can be distributed.

Acknowledgements. We would like to thank Tobias Kuhn, Davide Ceolin, Lora Aroyo, Johan Oomen, Erwin Verbruggen, Maarten Frohlich and Stephanie Delbecque for their valuable and constant feedback and ideas.

References

[1] Sean Bechhofer et al. "Why linked data is not enough for scientists". In: *Future Generation Computer Systems* 29.2 (Feb. 2013), pp. 599611. doi: 10.1016/j.future.2011.08.004.
[2] Christian Bizer, Tom Heath, and Tim Berners-Lee. "Linked Data - the story so far". In: *International Journal on Semantic Web and Information*. doi: 10.4018/jswis.2009081901.
[3] Jill P. Mesirov. "Accessible reproducible research". In: *Science* 327.5964 (Jan. 2010), pp. 415-416. doi: 10.1126/science.1179653.

[4] Monya Baker. “1500 scientists lift the lid on reproducibility”. In: *Nature* 533.7604 (May 2016). doi: 10.1038/533452a
[5] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3 (Mar. 2016). doi: 10.1038/sdata.2016.18
[6] Barend Mons. “Which gene did you mean?”. In: *BMC Bioinformatics* 6.142 (Jun. 2005). doi: 10.1186/1471-2105-6-142
[7] Ian J. Taylor, Ewa Deelman, Dennis B. Gannon, Matthew Shields. “Workflows for eScience: Scientific Workflows for Grids”. Springer, 2014. ISBN:1849966192 9781849966191.
[8] Khalid Belhajjame et al. “Workflow-centric research objects: first class citizens in scholarly discourse”. In: *Proceedings of Workshop on Semantic Publishing, 9th Extended Semantic Web Conference (ESWC)*. Springer, 2012.
[9] David de Roure et al. “Towards the preservation of scientific workflows”. In: *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES)*, ACM, 2011.
[10] Ewa Deelman et al. “Workflows and eScience: An overview of workflow system features and capabilities”. In: *Future Generation Computer Systems* 25.5 (May 2009), pp. 528-540. doi: 10.1016/j.future.2008.06.012
[11] SPAR ontologies: http://www.sparontologies.net/
[12] PROV-pings: http://git2prov.org:8902/prov-pings/index.html
[13] Sarven Capadisli et al. “Linked Data Notifications: A Resource-Centric Communication Protocol”. In: *Proceedings of the 14th Extended Semantic Web Conference (ESWC)*, Springer, 2017. doi: 10.1007/978-3-319-58068-5_33
[14] Tobias Kuhn and Michel Dumontier. “Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data”. In *Proceedings of the 11th Extended Semantic Web Conference (ESWC)*. Springer, 2014. doi: 10.1007/978-3-319-07443-6_27