Linkflows: enabling a web of linked semantic publishing workflows

Cristina-Iulia Bucur[1]

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. However, 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 artifacts 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 developed a model that is able to support Linked Data Notifications to demonstrate the feasibility of our approach.

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

1 Introduction/Motivation

Publishing is an important practice, not only in the world of science, but also in everyday life. 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 [22]. 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 [5]. 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. The Linked Data principles encourage using dereferenceable HTTP URIs for things like datasets, services, tools, etc. and including links to other URIs. Linked Data also 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 [15] and it permits further collaboration and reuse of scientific discoveries. Unfortunately, according to a recent study published in Nature [20], 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 [19] 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 [4] 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.

In this project we will try to address issues like the provenance and reproducibility in scientific publications and the support for a descentralized system of publishing in which scientific workflows can be modeled using Semantic Web technologies. Moreover, we will define alternative ways of assessing the quality of scientific publications.

2 State of the Art

In the last 25 years, scientific publishing has evolved from the form of a classical article to electronic publishing of scholarly journals. Accessing research publications without the restriction of subscriptions is the idea behind Open Access journals. These journals have been growing in number faster than traditional subscription journals [1]. But the debate of whether Open Access system is damaging the peer-review system and puts the quality of scientific journal publishing at risk was done in [21]. In [8] the authors mention that semantic publishing is inevitable and that it will happen in incremental steps as it is already possible to publish data as RDF statements in the Linked Open Data

Cloud [18]. Semantic Web technologies have launched a revolution in the field of scientific publishing and the idea is to create and facilitate an open access ecosystem where both content and metadata of scientific articles is accessible, together with formalized internal structures of the documents and components, enriched and with semantic connections to other related or similar documents.

In the view of the prevalence of the Semantic Web, considerable research was done in enriching the meaning of a traditional article in the digital publishing environment, facilitating its automatic discovery, having access in a semantic way to and within the article and also being able to link to other related articles or other related parts of articles. Especially notable in this sense are the SPAR ontologies [23], the ontologies central to the task of semantic publishing. All these techniques, methods and approaches can facilitate the scientific publishing domain and our research.

As datasets, documents and, in general, knowledge is spread in the web of the Internet, where everything can be shared and reused and linked, decentralization is a key concept. Decentralization implies that there is no control of a central authority anymore, e.g. a publishing house, over the open content that exists on the Web. There is a lot of research in this area of computer science, but we will focus especially on technologies related to the field of digital scientific publishing. In the past, techniques to ensure the functioning of a secure and decentralized global file system over the Internet to entice collaborations have been described in [7]. Then, the BitTorrent communication peer-to-peer file sharing protocol over the Internet to distribute and access data in the digital publishing environment was studied in [16], while peer-to peer networks for RDF data were developed in [13] and a decentralized architecture to support nanopublications, scientific RDF snippets, was built in [25].

In terms of assessing the quality of scientific publications, the most widely used indicator is the Journal Impact Factor (JIF) [10], but this metric has been the subject of multiple debates in the past as it was shown that it can be favourably manipulated [24]. For example, the JIF can be biased towards journals that publish high number of non-research items (e.g. research notes, comments) and have higher publishing numbers [9]. So, new ways of rating the quality of scientific publications is needed. Semantic Web technologies with ontologies like the Dataset Quality Information (daQ) [14] can support better and unbiased measures of quality, while new dimensions of quality that consider these technologies need to be taken into account. To further support provenance and reproducibility in scientific publications, the notion of scientific workflows was introduced.

A scientific workflow is a mechanism to specify and automate repetitive tasks for computational science or in silico science [12]. These scientific workflows are mostly published as digital artifacts that describe experiments and the additional materials needed to understand them [17]. A scientific workflow can be executed and reproduced because it contains a precise and executable description of a scientific procedure [6]. And, as research becomes more data driven, workflows

are used in the specification of experiments that retrieve, integrate and analyze datasets using distributed resources [11].

3 Problem Statement and Contributions

This research PhD project will be guided by a main research question. This will be the main focus of the thesis, but it will be answered using four sub-research questions. These sub-research questions will build on each other and, in the end, altogether provide an answer to the main research question:

How can scientific workflows that produce and consume digital artifacts 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. 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.

Every sub-research question captures a different aspect of the main research question:

1. *How can we model the decentralized execution of workflows by using Linked Data principles and tools?*
   First, we would like to be able to model scientific workflows. For this, we will provide and build the necessary framework based on PROV-Pings and Linked Data Notifications (LDN) [2], to enable notifications across platforms and tracking of provenance of various scientific workflows. We will use Linked Data principles like dereferenceable URIs using open standards like RDF to publish and link workflows. When workflows are modelled using the Linked Data principles and tools, we call them linkflows, the Linked Data version of workflows. To evaluate the model we will use a case study of at least 20 scientific articles together with their scientific workflows including reviewing and authoring. The innovative aspect is that workflow steps that are produced on various platforms are linked and then they are reused and consumed on other platforms.

2. *How can we execute workflows that produce and consume digital artifacts?*
   This research question allows the creation and execution of the workflow steps modelled previously. Through a software prototype, users will be able to create workflows for a selected corpora of digital artifacts, for example to generate a review. This software prototype will connect and enable linked workflows to flow across platforms and as such involve resources without involving the platform that contains a certain workflow step. This means, for example that the review that was created can be accessed by various interested parties, like online journal editors, who can consume it further by including the review in their own submission system. The innovative aspect that this sub-research question addresses is the same as in the first sub-research question, producing and linking workflows across platforms and consuming them on other platforms. The difference here is that the scientific workflows will be created in automated manner with a focus on workflow decentralization and the small granularity of digital artifacts.
3. *How can we automatically analyze digital artifacts and assess their quality and impact based on the linked workflows that produce and consume them?*
   In this research question we want to analyze the workflows in which digital artifacts were produced and consumed. The goal is to evaluate the quality and impact that these digital artifacts have based on the workflows they are part of. For this, we will build a prototype of a user interface that makes visible the connection between the digital artifact and the workflow that generated it. Next, we will develop metrics to measure the quality of a digital artifact based on the linked workflows that produced it. Moreover, analyzing the workflows that consume the digital artifact, we would be able to measure the impact this artifact has. The innovative aspects here are two-fold: first, tracking and visualizing the workflow steps and the digital artifacts that contain them and second, enabling new ways of measuring quality and impact of digital artifacts that do not rely only on the analysis of their provenance, but how they participate in the flow of dynamic processes, thus in linked workflows across platforms.
4. *Can we use digital artifacts and the linked workflows that contain them to support inquires from users?*
   This last sub-research question will bridge and blend all previous aspects together. An inquiry consists of searches of digital artifacts. The search results will contain not only the static object, the digital artifact, that is relevant for the inquiry, but also the workflow(s) that produce and consume that digital artifact, together with metrics like quality and impact. Moreover, users would be able to generate workflows for digital artifacts at the same time. So, a comment or a review could be added for a digital artifact, opening the execution of workflow steps for users. The innovative aspect would be two-fold: first, the inquiry responses will contain not only the corresponding digital artifact(s) of interest, but the linked workflows that contain them as well and second, users would be able to produce, consume and execute workflows for digital artifacts on the fly.

4 Research Methodology and Approach

In order to answer the research questions from Section 3, we will move away from the idea of a traditional scientific article. We will consider digital artifacts as "universal entities" or objects that can be in the form of text, figures, datasets, code, presentation slides, multimedia objects, etc. as represented in Figure 1. Each of these digital artifacts can be represented in the form of a node in a network. As such, a classical scientific PDF article is comprised of various digital artifacts like text, figures, datasets, code, etc. that are inter-connected. These digital artifacts are considered first class citizens and all bear the same importance. Connections between the nodes of this network of digital artifacts are links, as in Web links.

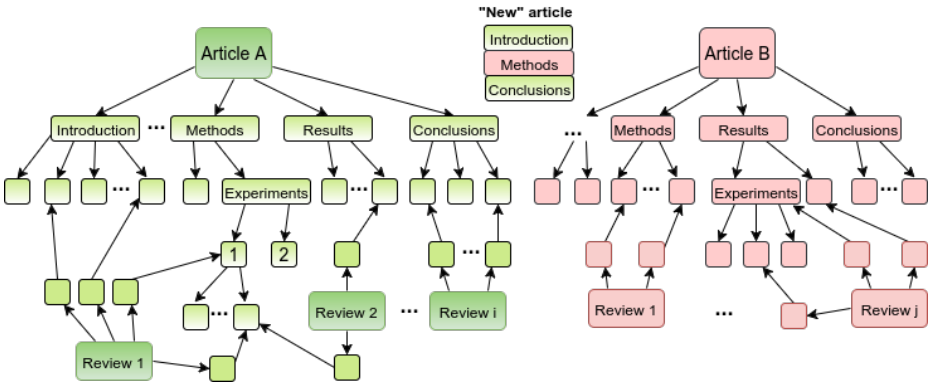

Fig. 1. Traditional scientific article represented as interconnected digital artifacts.

The digital artifacts that connect together to form a scientific contribution undergo specific scientific workflows for their production and even when they are consumed by various systems, platforms and users. A scientific workflow involving a digital artifact represents the processes and actions that the digital artifacts undergo in science, like reviewing, commenting, annotating, etc. The linked workflows of a digital artifact are called linkflows. As such, either small steps of these workflows or entire workflows on a digital artifact are open and distributed in the sense that they can be produced on one platform and consumed on another, thus allowing the smooth flow of a workflow across platforms. This means, for example, that a review of a digital artifact can be produced on platform A, but platforms B and C can get notified of this review and consume or use it on these platforms. In the same way, a scientific workflow can contain multiple steps and these can be produced on different platforms, but linked together to form a coherent flow.

Considering this networked structure, when modifications are made in the network of a scientific contribution, notifications can be sent around about the

changes as these propagate through the network. Here PROV-pings and Linked Data Notifications technologies will help in sending notifications 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 and the use of hashes can help in establishing and tracking the changes that a digital artifact goes through.

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).

Use case 1: The Netherlands Institute of Sound and Vision The Netherlands Institute of Sound and Vision is the biggest audio-visual cultural archive in the Netherlands having almost 70% of the Dutch audio-visual heritage under its tutelage. For this project, we will work closely with the two open access media historical e-journals: VIEW Journal of European Television History and Culture and Tijdschrift voor Mediageschiedenis.

Use case 2: IOS Press IOS Press is an independent publishing house that has about 100 journals (mainly focused on medicine, but also from scientific and technical domains) and around 130 books published annually. For this project we will consider two of the open access journals that are printed by IOS Press: one in the computer science domain, the Data Science Journal, and a journal from the medical domain, the Journal of Alzheimer's Disease Reports (JAD Reports).

Throughout this project, we will implement, test and evaluate the linked workflows (linkflows) on the journals or datasets described in the use cases above. The diversity of the domains, ranging from audio-visual and multimedia to computer science and medicine would ensure a complex coverage of our project.

5 Preliminary Results

For answering the first sub-research question, we model the workflow steps of a journal submission using the Linked Data Notifications (LDNs) protocol [2]. For example, let's consider the scientific workflow of reviewing: 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. 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 and 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. To support this, we use the LDN protocol, but we might also consider OWL-S ontology [3] in the future to describe in more detail the semantic web services that are used.

In the LDN protocol, the reviewers and the digital artifacts that they assess need to be seaparately identified. When a reviewer generates a review for a digital artifact this review is stored on the Web, in a location that is publicly accessible by an URI. Next, platforms or systems that are interested to use or check any actions taken on that digital artifact, can subscribe to receive notifications regarding it when these arrive. And, not only can they receive notifications concerning an action taken regarding a digital artifact, but they can also use the data generated after a certain action is performed. In this case, use the review that one reviewer produced about a certain digital artifact.

6 Evaluation Plan

Throughout the various stages of our research methodology we will perform various evaluations to assess the validity of our results. For this we will mostly consider the use cases provided by the Netherlands Sound and Vision and IOS Press.

For the first sub-research question, where we want to model the decentralized execution of workflows, for a more complex evaluation we will use two different use cases: (i) we will manually model around 20 published scientific articles together with their scientific workflows; (ii) we will use the automated extraction of information from a bioinformatics repository already curated by experts, e.g. DisGeNET, "one of the largest and comprehensive repositories of human gene-disease associations currently available". For the evaluation we will conduct a qualitative analysis on the corpus of selected scientific papers and on the bioinformatics repository. We will consider reviews content, comments and annotations and how these relate to the structured parts of the article or of the bioinformatics repository.

For the second sub-research of how we can execute workflows that produce and consume digital artifacts, we will use the manually created model from the first step. The prototype that we will build will provide the software means to create a fully decentralized reviewing workflow. For evaluation purposes, we will conduct a controlled user experiment where we will ask participants to evaluate both the user interfaces, as well as the new way the reviews are conducted and how these reviewing workflows are generated by comparing it with how this process was carried on previously.

The third sub-research question addresses new ways of evaluating the quality and impact of a digital artifact based on the analysis of the workflows that produced and consumed the respective artifact. Here, we will develop new metrics for quality assessment of digital artifacts based on the scientific workflows that produce them, like reviewing. We will also develop a metrics for the impact that a digital artifact has based on the workflows that consume it. To evaluate these two metrics, we will first use nichesourcing (crowdsourcing with experts). At the same time, we will use existing data to import workflows and use it as a ground truth for generating the workflows.

For the fourth sub-research question, where we want to be able to provide answers to inquiries made by users, we will consider the development and implementation of a prototype that can return digital artifacts in response to user queries, together with the linked workflows that contain them. Furthermore, users would be able to create on the fly workflows for digital artifacts. For evaluation, we will use crowdsourcing to evaluate the software prototype in terms of the relevancy and the results that are returned as answers to user inquiries and also for rating the creation and execution of workflows on digital artifacts.

7 Conclusions

In this project we want to investigate new approaches in the digital environment of scientific publishing by combining Linked Data principles 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. Moreover, these scientific workflows will be used to create new quality dimensions of digital artifacts that take into consideration the dynamic processes that produce and consume them. Thus, by using new and existing Semantic Web technologies we will support the reproducibility of scientific research, the exchange, reuse and linking of all digital artifacts involved in scientific workflows.

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

References

1. Budapest Open Access Initiative. Retrieved from http://www.budapestopenaccessinitiative.org/
2. Linked Data Notifications. Retrieved from https://www.w3.org/TR/ldn/
3. OWL Web Ontology Language for Services (OWL-S). Retrieved from https://www.w3.org/Submission/2004/07/
4. Barend Mons. Which gene did you mean?. In: BMC Bioinformatics 6.142 (2005). doi: 10.1186/1471-2105-6-142.
5. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - the story so far. In: International Journal on Semantic Web and Information 5.3 (2009), pp. 1-22. doi: 10.4018/jswis.2009081901.
6. David de Roure et al. Towards the preservation of scientific workflows. In: Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011)(2011).

7. David Mazieres, M. Frans Kaashoek. Escaping the evils of centralized control with self-certifying pathnames. In: Proceedings of the 8th ACM SIGOPS European workshop on Support for composing distributed applications (1998), pp. 118-125.
8. David Shotton. Semantic publishing: the coming revolution in scientific journal publishing. In: Learned Publishing 22.2 (2009), pp. 85-94. doi: 10.1087/2009202.
9. Dissecting our impact factor. In: Nature Materials (2011). url: https://www.nature.com/articles/nmat3114.
10. Eugene Garfield. The History and Meaning of the Journal Impact Factor. In: JAMA 295.1 (2006), pp. 90-93. doi: 10.1001/jama.295.1.90.
11. Ewa Deelman, Dennis Gannon, Matthew Shields, Ian Taylor. Workflows and e-Science: An overview of workflow system features and capabilities. In: Future Generation Computer Systems 25.5 (2009), pp. 528-540. doi: 10.1016/j.future.2008.06.012.
12. Ian J. Taylor, Ewa Deelman, Dennis B. Gannon, Matthew Shields. Workflows for e-Science: Scientific Workflows for Grids. Springer Publishing Company (2014). ISBN: 1849966192.
13. Imen Filali et al. A survey of structured P2P systems for RDF data storage and retrieval. In: Transactions on large-scale data- and knowledge-centered systems III (2009), pp. 20-55. ISBN: 9783642230738.
14. Jeremy Debattista, Christoph Lange, Soren Auer. daQ, an ontology for dataset quality information. In: Proceedings of the Workshop on Linked Data on the Web 1184 (2014).
15. Jill P. Mesirov. Accessible Reproducible Research. In: International Journal on Semantic Web and Information 327.5964 (2010), pp. 415-416. doi: 10.1126/science.1179653.
16. Joseph Paul Cohen, Henry Z. Lo. Academic Torrents: A Community-Maintained Distributed Repository. In: Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment (2014). doi: 10.1145/2616498.2616528.
17. Khalid Belhajjame et al. Workflow-Centric Research Objects: First Class Citizens in Scholarly Discourse. In: Proceedings of Workshop on the Semantic Publishing (SePublica 2012), 9th Extended Semantic Web Conference, Crete, Greece (2012).
18. Krzysztof Janowicz, Pascal Hitzler. Open and transparent: the review process of the Semantic Web journal. In: Learned Publishing 25 (2012), pp. 48-55. doi: 10.1087/20120107.
19. Mark Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific Data 3 (2016). doi: 10.1038/sdata.2016.18.
20. Monya Baker. 1,500 scientists lift the lid on reproducibility. In: Nature 533.7604 (2016). doi: 10.1038/533452a.
21. Peter Murray-Rust. Open Data in Science. In: Elsevier Serials Review 34.1 (2008), pp. 52-64. doi: 10.1016/j.serrev.2008.01.001.
22. Sean Bechhofer et al. Why linked data is not enough for scientists. In: Future Generation Computer Systems 29.2 (Feb. 2013), pp. 599-611. doi: 10.1016/j.future.2011.08.004.
23. Silvio Peroni. The Semantic Publishing and Referencing Ontologies. In: Semantic Web Technologies and Legal Scholarly Publishing (2014), pp. 121-193. doi: 10.1007/978-3-319-04777-5_5.
24. Tobias Kiesslich, Silke B. Weineck, Dorothea Koelblinger. Reasons for Journal Impact Factor Changes: Influence of Changing Source Items. In: PLOS One (2016). doi: 10.1371/journal.pone.0154199.
25. Tobias Kuhn et al. ecentralized provenance-aware publishing with nanopublications.In: PeerJ Computer Science (2016). doi: 10.7717/peerj-cs.78.