Anchors in Shifting Sand: the Primacy of Method in the Web of Data

David De Roure
School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK
+44 23 8059 2418
dder@ecs.soton.ac.uk

Carole Goble
School of Computer Science
The University of Manchester
Manchester M13 9PL, UK
+44 161 275 6195
carole.goble@manchester.ac.uk

ABSTRACT

The wealth of new government and scientific data appearing on the Web is to be welcomed and makes it possible for citizens and scientists to interpret evidence and obtain new insights. But how will they do this, and how will people trust the results? We suggest the Linked Data Web must embrace the "methods" by which results are obtained as well as the results themselves. By making methods first class citizens, results can be explained, interpreted and assessed, and the methods themselves can be shared, discussed, reused and repurposed. We present the myExperiment.org website, a social network of people sharing reusable methods for processing research data, and make some observations on the nature of first class methods in the Web of Data.

Keywords

Reproducible research, linked data, myExperiment, scientific workflows.

1. INTRODUCTION

There is excitement at the prospect of the increasing variety and detail of datasets available on the Web, published as Linked Data amongst other formats. Government initiatives to encourage the re-use of public data are set to enable researchers and citizens alike to gain new insights and understanding. There is a wealth of scientific data which could also be made available in this way, some historical but some real-time, like the streams of data from our instrumented environment which many may wish to interpret, integrate and visualise. With this rich supply of data perhaps we can better debate the evidence for decisions of societal impact from social policy to climate change.

The fact that people can find the very latest information on the Web is one of its great benefits. But it also makes it a place of flux: we can announce a result – be it a statement about crime or flood risk, analysis of a musical similarity or a genome annotation – and point to the sources that led us there, but on the Web they may have changed a nanosecond later. In science every result is tentative but reinforced by reproducibility; unless we address this on the Web, everything is tentative. Meanwhile the established scholarly lifecycle persists in the interests of reproducible research: papers published, immutable and citable, citing data sources which may be curated.

In the interest of reproducibility we must *primae facie* focus on making explicit the *method* by which results are generated. Methods can then be first-class Web citizens in our emerging scientific practice. For example we can create a "pipeline", "script", "mashup", "workflow", "query" or "business process" to generate a result based on data sources, and this provides our route to repeatability, reproducibility and reuse. We get the latest results, and we better understand the provenance of our results so that they can be explained, interpreted, trusted and reused by others.

Crucially, those working with the data also benefit from shareable methods. As a first class citizen on the Web, the method can be discussed and reviewed, reused and repurposed. This enables expertise to grow and methods to become the substance and subject of social networks. As long as we focus only on 'freeing data' the methods remain intangible and even ephemeral.

In this model of "first-class methods" we naturally need URIs which refer to them and mechanisms for discovering, sharing, enacting and curating them – like we do Web pages or datasets – and versioning too. This is exactly what is supported by the myExperiment website (www.myexperiment.org), a social network of people sharing reusable methods for processing research data, in various research communities from bioinformatics and chemistry to climate change and digital humanities [2].

myExperiment embraces the world of changing data by providing shareable methods to process it and mechanisms for validation and reproducibility. Co-evolved with research users, it is an experiment in our notion of primacy of method for emerging science on the highly dynamic and data-intensive Web of Data. This paper introduces myExperiment and reflects on some of the lessons learned from a Web Science perspective.

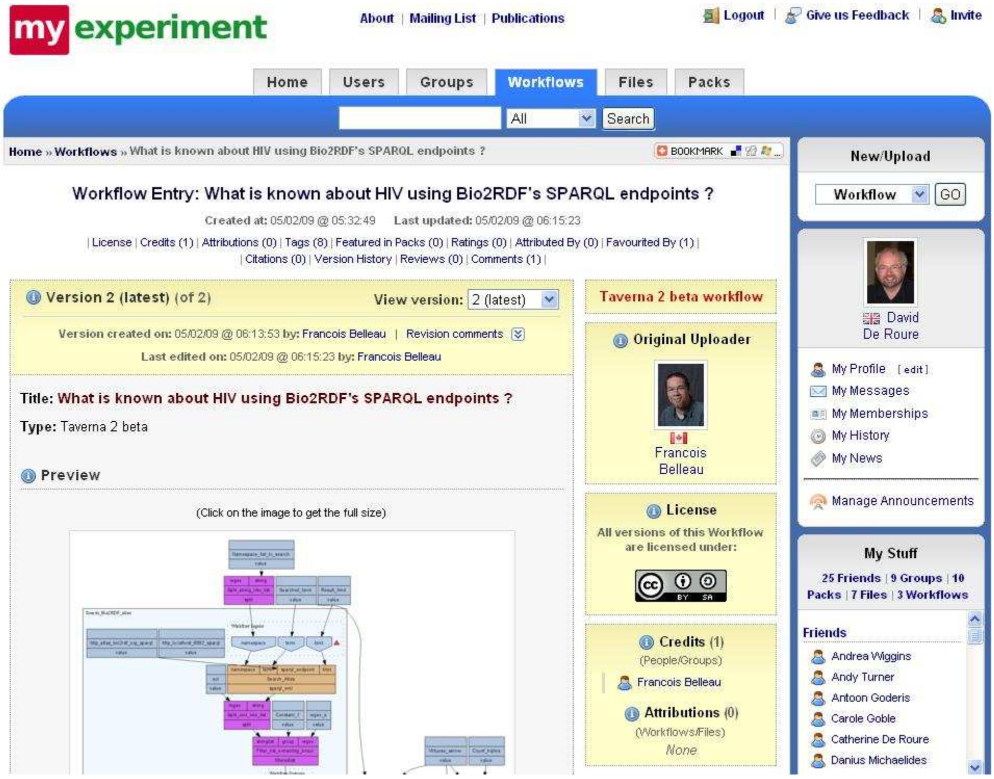
2. MYEXPERIMENT

2.1 Conception

The myExperiment social website was conceived to encourage researchers to share experimental methods and protocols, and thus expertise and know-how. As an "intellectual access ramp" for data intensive researchers it was also driven by the desire to create a site that is highly useable by being immediately familiar, hence we adopted a Web 2.0 approach.

Our focus on methods was partly a reaction to the data-centricity of many repositories: if there is a data deluge then there is also a deluge in the methods used to process it, and we conjectured that these are at least as important and often neglected – methods are candidates for network effects too.

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.



Workflow Entry: What is known about HIV using Bio2RDF's SPARQL endpoints ?

Created at: 05/02/09 @ 05:32:49 **Last updated:** 05/02/09 @ 06:15:23

| License | Credits (1) | Attributions (0) | Tags (8) | Featured in Packs (0) | Ratings (0) | Attributed By (0) | Favourited By (1) |
| Citations (0) | Version History | Reviews (0) | Comments (1) |

Version 2 (latest) (of 2) **View version:** 2 (latest)

Version created on: 05/02/09 @ 06:13:53 **by:** Francois Belleau | Revision comments

Last edited on: 05/02/09 @ 06:15:23 **by:** Francois Belleau

Title: What is known about HIV using Bio2RDF's SPARQL endpoints ?

Type: Taverna 2 beta

Preview

(Click on the image to get the full size)

Taverna 2 beta workflow

Original Uploader

Francois Belleau

License

All versions of this Workflow are licensed under:

BY SA

Credits (1)
(People/Groups)
Francois Belleau

Attributions (0)
(Workflows/Files)
None

Figure 1: A workflow in myExperiment

The Web 2.0 design patterns [10] tell us “Data is the next Intel Inside”: applications seek to own a unique source of data. We followed this principle in myExperiment by focusing first on scientific workflows – descriptions of data analysis pipelines that enable systematic processing of large volumes of data [3]. Workflows are crucial to the automation and reproducibility of data-intensive research [5], however they require expertise to create them, so there is a clear incentive to share them in order to learn, reuse and repurpose.

myExperiment has successfully adopted a Web 2.0 approach in delivering a social website where scientists can discover, publish and curate scientific workflows and other artefacts. While it shares many characteristics with other Web 2.0 sites, myExperiment’s distinctive features to meet the needs of its research user base include support for credit, attributions and licensing, fine control over privacy, a federation model and the ability to execute workflows. Figure 1 shows a workflow in myExperiment, with some of its associated “social metadata”.

myExperiment has an important ‘community curation’ aspect. A fundamental challenge of workflows is that the environment in which they are ‘enacted’ does not remain constant over time. This is a familiar challenge in distributed computing where the researcher is at the mercy of multiple third party services. Consequently, a workflow left alone will effectively ‘decay’. This is hard to fix by automatic means, though it can be detected and repair can be assisted. However, myExperiment enables users of a workflow to maintain it, and provides assistance to curators who proactively validate myExperiment content.

While myExperiment is designed for ease of use interactively, one of our design principles is that the myExperiment functionality can be brought through to other interfaces in the user’s familiar environment; this is also in line with the Web 2.0 principle of “co-operate don’t control”. A variety of interfaces have been built over the RESTful programmatic interface, including Google Gadgets, Facebook applications, and plugins for the Taverna workflow system [9]. There is also a SPARQL endpoint (rdf.myexperiment.org) and a Linked Data compliant service is available. Developers also use the myExperiment source code to create other instances of myExperiment and as a basis for other projects.

Launched at the end of 2007, myExperiment now has over 3300 registered users, with thousands more downloading public content, and is the largest public collection of workflows for systems which include Taverna, Microsoft’s Project Trident [12] and SEASR/Meandre [6]. The diversity of shared artefacts now includes experimental plans, spreadsheets and SPARQL queries.

2.2 Co-Evolution

myExperiment is itself an experiment; for example, it is investigating whether the ‘selfish scientist’ shares workflows enough to benefit from a social website and network effects. By design, myExperiment does not mandate particular ways of working, but rather provides a set of simple mechanisms which users may use however they choose. The capabilities of the site are created in close collaboration with its research users [1]. As such it becomes an interesting probe into emerging research practice in multiple communities.

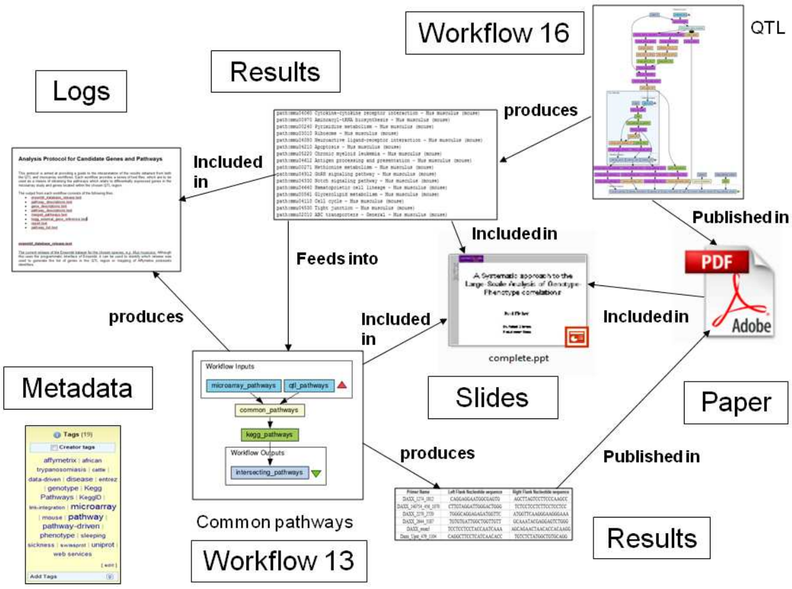


Figure 2: A Pack in myExperiment

Initially these mechanisms included groups, friendships and the various shared contributions. Users were quick to recognise that a workflow can be enriched as a shareable item by bundling it with some other pieces which make up the "experiment". We also observe that in practice researchers do not work with just one content type and moreover that their data is not in just one place – it is distributed, and sometimes disorganised too.

Hence we also developed support for "packs" – collections of items, both inside and outside myExperiment, which can be shared as one bundle. For example, a pack might contain workflows, example input and output data, results, logs, PDFs of papers and slides (see Figure 2) – such a pack captures an experiment and is reusable and repurposeable. Packs are created using the shopping basket (or wishlist) metaphor and can be exported using the Object Reuse and Exchange representation [8] which is gaining increasing adoption in the open repositories community. Approximately 10% of the contributions to myExperiment are packs. This move from artifacts to aggregation is in line with Pepe *et al* [11] who articulate the importance of principled aggregation in scholarly discourse.

Since packs contain references to data outside myExperiment we are again at the mercy of change. Although we do not guarantee that the same versions of external content will be retrieved (that would be the role of a backup or sync service), we can make a weaker guarantee: we can inform users whether the external content may have changed – for example, by the change of version in a repository, by a change in meta-information reported by HTTP (e.g. content length, last modified date, hash) or by comparison with a cached copy. Packs also support alternative URLs for external content, providing a degree of redundancy which aids stability of pack references.

Although a variety of different packs have been created, the archetypal example that has emerged provides a useful insight into the notion of method for the users of myExperiment. A pack contains:

- The workflow
- Example input data and corresponding output data
- The input data
- The output data
- A service invocation log

In addition packs contain information for specific uses and users. The example data serves a crucial role, because it enables the workflow to be validated. These packs, and the myExperiment "virtual research environment" itself, go some way to providing reproducibility in *in silico* research [7].

2.3 Abstractions

One outcome of this exercise is that it is essential to be able to refer to a workflow by a (versioned) URI but, in practice, the more useful shareable artefact is the workflow combined with information about its use. We would argue in fact that this *is* the method, and it corresponds for example to the methods section of papers. These aggregations representing methods are in many senses a new form of scholarly publication. Unlike papers they are machine readable, they drop into the tooling of digital research, and they yield to automated processing as well as human use. It is interesting to consider whether this co-evolution is leading to a replacement for the paper, for recording and reproducing methods.

Abstractly, we might say a method is an aggregation that contains a description of a process and a set of data resources used by that process. However, for research with an *in silico* dimension, the application of the method to data also involves some computation [4]. We might capture information about the computational resources in the method too. While repeatable research using specialist facilities could lock the method into particular resources, a more generically reproducible method should work elsewhere. In fact, working digitally makes it easier to confirm that results are independent of the particular resources provisioned to enact the experiment. Increasingly such provisioning is flexible and dynamic.

In the computer science context, our discussions of this abstraction frequently refer to the notion of a *closure*. A term from programming languages, closure refers to an object which has a *control part* and an *environment part*. Here the workflow is the control – it describes the processing – and the aggregation also includes environment information. In our abstraction, a closure can be applied to data by running it on an *enactor*. This model is particularly powerful when we describe experiments that in turn call other experiments, such that the context of the sub-experiment is inherited from the 'parent' experiment.

An item of data on the Web can be better interpreted and better trusted if we know where it came from – provenance helps with this. In the method-rich world we are articulating here, we can also re-compute that information. It is easy to assume that data is being computed and placed on Web pages where it is consumed by humans and decays as the landscape changes – but note that the computation can happen on-demand (demand driven), or it can happen automatically in response to changing data (data driven). Where appropriate this 'lazy' approach of data-on-demand or even experiment-on-demand should be considered alongside the current data capture practices in archives caricatured as WORN (Write-Once Read-Never).

3. DISCUSSION

We opened by describing a Web of Data in flux. Linked Data is in its infancy and works well just now for a persistent, slowly evolving Web of read-only datasets. To achieve ambitions like Open Government data and scientific datasets on the Web, Linked Data is beginning to address the crucial areas of versioning and discovery.

Here we have proposed another perspective. Fresh data flows through the Web all the time, and those flows are enacted by programmed automation or human workflows. Perhaps we should accept this spectacular dynamism of the Web and ask, what are the fixed pieces? A single method, processing streams of data, is inevitably more persistent than the data itself. We have argued that there is value in reifying these methods. This observation is not limited to the conduct of research on the Web but applies more pervasively.

myExperiment is one approach, and it has revealed the practical realisation of method as an aggregation of environment with process description (closures). Another system of interest is Meandre, which provides publishing schemes to create a distributed repository of shareable components [6]. As computational thinking and scientific method come together we see new artefacts which afford automation and dynamic provisioning of resources, and we have glimpsed the power of this in reproducible data-intensive research.

While many Web Science studies have taken a data-centric perspective or link-centric perspective, we are suggesting there is value in a method-centric perspective. This has both an infrastructural and a social aspect: methods underpin the pipelines of information processing, and by looking at the usage and evolution of methods on the Web we are witnessing the diffusion of knowledge. Our methods become boundary objects [13] shared between the multidisciplinary teams of data-intensive research and indeed shared with the computer itself.

For those studying the scholarly knowledge cycle, two trajectories of work are converging to inform our notion of aggregation:

1. We are recognising that, although arbitrary aggregations can be supported by the technology, there is a principled approach to aggregation arising from established scientific method – and this may itself be evolving as practice changes, for example working with streams of sensor data [11].
2. By providing some basic communication, sharing and aggregation capabilities in myExperiment we are capturing evolving practice in this 'Science 2.0' environment, and this is suggesting a new shareable digital artefact which explicitly contains reusable methods.

Outside the scholarly knowledge cycle, the principle that methods should be first class citizens on the Web transcends scholarly communication. How can we refer to the pipeline that processes data with a unique URI? How is this handled as Linked Data – and in particular, how do we achieve linkage and aggregation to maximise the reproducibility, reusability and repurposeability of methods?

The crucial thing about methods as first class objects is that they empower users. There is talk of citizen scientists and citizen sensors. If there are also to be citizen analysts of data then we must provide re-usable data, methods and mechanisms to build know-how, not just know-what.

4. ACKNOWLEDGMENTS

Our thanks to the myExperiment team and our "friends and family" including Paul Fisher, Matthew Gamble and Marco Roos. myExperiment is supported by JISC under the Information Environment Programme, by EPSRC award EP/G026238/1 (myGrid: A Platform for e-Biology) and by Microsoft External Research.

5. REFERENCES

[1] De Roure, D., and Goble, C.: Software Design for Empowering Scientists, IEEE Software, 2009, 26, pp. 88-95.

[2] De Roure, D., Goble, C., and Stevens, R.: The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows, Future Generation Computer Systems, 2009, 25, pp. 561-567.

[3] Deelman, E., Gannon, D., Shields, M., and Taylor, I.: Workflows and e-Science: An overview of workflow system features and capabilities, Future Generation Computer Systems, 2009, 25 (5), pp. 528-540.

[4] Goble, C., and De Roure, D.: Curating Scientific Web Services and Workflows, Educause Review, 2008, 43 (5)

[5] Goble, C., and De Roure, D.: The impact of workflow tools on data-centric research, in Hey, T., Tansley, S., and Tolle, K. (Eds.): 'Data Intensive Computing: The Fourth Paradigm of Scientific Discovery' (2009), pp. 137-145.

[6] Llorà, X., Ács, B., Auvil, L., Capitanu, B., Welge, M., and Goldberg, D.: Meandre: Semantic-Driven Data-Intensive Flows in the Clouds. Proc. Fourth International Conference on eScience2008 pp. 238-245.

[7] Mesirov, J.P.: Accessible Reproducible Research, Science, 2010, 327 (5964), pp. 415-416.

[8] http://www.openarchives.org/ore/, accessed 29 March 2010.

[9] Oinn, T., Greenwood, M., Addis, M., Alpdemir, N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M., Senger, M., Stevens, R., Wipat, A., and Wroe, C.: Taverna: lessons in creating a workflow environment for the life sciences, Concurrency and Computation: Practice and Experience, 2006, 18 (10), pp. 1067-1100.

[10] Oreilly, T.: What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, Communications & Strategies, 2007, 65, pp. 17-37.

[11] Pepe, A., Mayernik, M., Borgman, C.L., and Van de Sompel, H.: From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web, Journal of the American Society for Information Science and Technology, 2009, 61 (3), pp. 567-582.

[12] Simmhan, Y., Barga, R., van Ingen, C., Lazowska, E., and Szalay, A.: Building the Trident Scientific Workflow Workbench for Data Management in the Cloud. Proc. Third International Conference on Advanced Engineering Computing and Applications in Sciences, Sliema, Malta, 2009 pp. 41-50.

[13] Star, S.L., and Griesemer, J.R.: Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39, Social Studies of Science, 1989, 19 (4), pp. 387-420.