Web Science Studies into Semantic Web Service-Based Research Environments

Mark D. Wilkinson
Centro de Biotecnología y Genómica de Plantas
Universidad Politécnica de Madrid (UPM)
Madrid, Spain
mark.wilkinson@upm.es

***Abstract*— The emergent domain of Web Science has a number of as-yet unrealized goals. Among these are: to facilitate scientific discourse by supporting the explicit comparison and evaluation of hypotheses; to simplify *in silico* experiments by providing an ecosystem of expert analytical strategies that can be automatically assembled; to enhance scientific rigor by reducing bias, and improving reproducibility; and to integrate the knowledge gained from the experiment back into the Web. SHARE is a novel orchestration system that automatically chains-together Semantic Automated Discovery and Integration (SADI)–style Semantic Web Services. During development of SHARE, we noted that many requirements of such an end-to-end Web Science research environment were being realized. These include formally-defined, machine-readable, Web-embedded research hypotheses; an explicit, transparent, rigorous, and reproducible research methodology utilizing the most up-to-date data and expert-knowledge from the community; immediate dissemination and re-use of the resulting data and knowledge; and enhanced support for peer-review. This manuscript describes how SHARE is now being tested as a prototype Web Science framework.**

Keywords – SADI, SHARE, Semantic Web Services, Workflow Orchestration, Reproducibility, Transparency, Personalized Web

I. MOTIVATION AND REQUIREMENTS-GATHERING

Web Science is a recently-established, cross-disciplinary research domain spanning technology, sociology, psychology, and policy. Web Science research considers the Web as both a subject of, and/or a platform for, scientific investigation. It is the latter perspective that is the focus of this work. This domain of Web Science includes investigations into how Web environments might augment many aspects of scientific research, from scientific discourse through experimentation to peer-review and publication. In this manuscript, we are interested in how Semantic Web technologies might improve the execution of, and reproducibility of, high-throughput biomedical research.

Arguably, the Web is the most important and ubiquitous tool in modern biological research; used to collect data, submit data for processing, research prior art, and publish research results, there are few moments in the day when a biological researcher does not have their Web browser open. One could therefore argue that we are already engaged in "Web Science". However, the Web, as used by researchers today, is merely a conduit through which ideas and results are transmitted, usually in a form that the Web itself cannot "understand" or take advantage of (i.e. PDF-formatted discussions); like a carrier pigeon delivering paper notes,

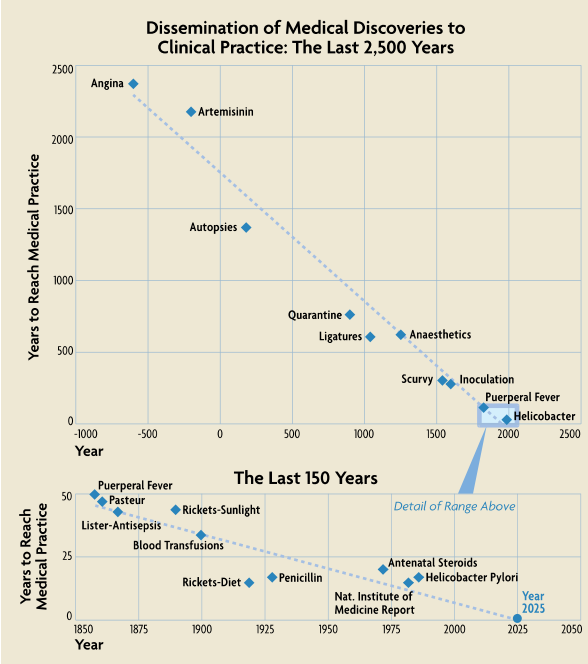

Figure 1. The time from discovery to implementation in Medicine. Over time, the delay between discovery and implementation has decreased in a near-linear manner. This pattern converges on the X-axis in approximately 2025 taken from [1] under c.c. License)

current Web Science is done *over* the web, not *within* it. This significantly under-utilizes the power of modern Web technologies – Semantic Web technologies in particular – to automatically discover and integrate data, knowledge, and analytical resources on a global scale. Such technologies are, therefore, the obvious choice for investigating novel Web Science infrastructures to support both the scientist and their science. The desire to understand if and how Web Science technologies can be applied is being driven by a number of intersecting trends, technical problems, and changing requirements in light of increasingly large datasets.

In Figure 1, Gilliam et al. suggest that the length of time between discovery, and the implementation of that discovery in-practice, is shrinking rapidly [1]. The rate at which discovery and implementation are converging appears constant, and the two are predicted to meet in approximately 2025. At this point, according to Gilliam, the moment of discovery, dissemination, and utilization merge into a single event. While the figure relates to healthcare, the same phenomenon likely holds in all areas of life science.

 ISBN: 978-1-61208-270-7

In practice, to achieve this end-point, research would be conducted in a medium that immediately interpreted and disseminated the results of an experiment; disseminated these results in a form that immediately (and actively) affected the results of other studies; and affected those studies without requiring those investigators to be aware of the new results or knowledge (since researchers are unable to stay-abreast of the literature, even in their own specialized field [1]). Moreover, it would be desirable for the experiment to be thoroughly documented, in a machine-readable manner, with a full provenance record including purpose, hypothesis, source, the data and algorithms used, versions, etc. This would allow both machines and humans to better assess the reliability, applicability, and validity of these results prior to using them in subsequent experiments. Despite being projected to be only a decade away, such knowledge-generating and knowledge-disseminating technologies and frameworks do not yet exist.

In parallel with the growing requirement for speed in knowledge-dissemination, there are increasingly worrisome observations of human limitations with respect to managing and manipulating these massive and complex datasets, and the resulting ease of making errors during data analysis - "the most common errors are simple... the most simple errors are common" [2]. Many researchers lack the skills to programmatically manipulate large datasets, and continue to use inappropriate tools to manage 'big data'. Serious errors introduced during data manipulation are difficult to detect by the researcher and, because they go un-recorded, are nearly impossible to trace during peer-review. In addition, the statistical expertise required to correctly analyze high-throughput data is rare, and biological researchers are seldom adequately trained in appropriate statistical analyses of high-throughput datasets. As such, inappropriate approaches, including trial-and-error, may be applied until a "sensible" answer is found [3]. Finally, because manually-driven analyses of high-throughput data can be extremely time-consuming and monotonous, researchers will sometimes inappropriately use a hypothesis-driven approach – examining only possibilities that they already believe are likely, based on their interpretation of prior biological knowledge, or personal bias towards where they believe the "sensible" answer would be found [4]. Thus, the scientific literature becomes contaminated with errors resulting from "fishing for significance", from research bias, and even from outright mistakes. These problems are becoming pervasive in omics-scale science - the affordability and accessibility of high-throughput technologies is such that now even small groups and individual laboratories can generate datasets that far exceed their capacity, both curatorially and statistically, to accurately manipulate and evaluate.

Even more troubling is that peer-review is failing to catch serious errors. While the Baggerly study into high-throughput publication quality [2] triggered retractions and a scientific misconduct investigation [5], the Ioannidis study reveals that, even in the prestigious Nature Genetics, more than half of the peer-reviewed, high-throughput studies cannot be replicated [6]. The failure of peer-review to detect non-reproducible research is, at least in part, because the analytical methodology is not adequately described [6], but perhaps equally because a proper evaluation of an experiment that controlled for errors would necessitate a re-execution of the experiment itself – something that is not reasonable to expect from reviewers. Thus, in the "big data" world, traditional peer-review is demonstrably ineffective.

In recognition of these limitations, the Institute of Medicine in 2012 published several recommendations relating to proper conduct of high-throughput analyses [7]. These include: rigorously-described, annotated, and followed data management procedures; "locking down" the computational analysis pipeline once it has been selected; and publishing the workflow of this analytical pipeline in a formal manner, together with the full starting and result datasets. These recommendations help ensure that (a) errors are not introduced through manual data manipulation, (b) there can be no human-intervention in the data as it passes through the analytical process, and (c) that third-parties can properly evaluate the data, the analytical methodology, and the result at the time of peer-review. While formal workflow technologies have proven effective at resolving some of these issues [8], integration of workflows into the overall scientific process continues to be *ad hoc*, and workflows themselves tend to be sparsely documented, difficult to review, difficult to re-use and re-purpose, and are not integrated with other forms of Web knowledge and expertise [9].

We lack the frameworks, standards, and infrastructures required to meet most of the intersecting trends, requirements and recommendations described above; moreover, these requirements appear to necessitate mechanization of much of the scientific process. As such, there is now some urgency around the necessary and inevitable creation of next-generation Web Science technologies, frameworks, and infrastructures to support the activities of high-throughput researchers.

In the remainder of this work-in-progress manuscript, we first describe, in Section II, the results of our examination of a prototype Semantic Web Service-based Web Science platform. We then discuss, in Section III, the underlying technologies that were used in that prototype, and how we propose to further examine these technologies. We then conclude in Section IV with a brief discussion of related projects, and the potential impact such platforms might have on the scientific process.

II. A PROTOTYPE WEB SCIENCE RESEARCH PLATFORM

Almost ubiquitously, scientific Web Services exhibit a set of features/behaviors that make them easier to connect into workflows compared to business-oriented Services [10]. We leveraged this to create a prototype Semantic Web Service workflow orchestration engine, and the results of these studies were recently published [9]. We demonstrated that, by constructing and publishing in the Web a semantic model – an ontology – describing a hypothetical biological phenomenon (Figure 2, left), we were able to automatically synthesize and execute an integrative analytical workflow (Figure 2, right) that discovered and/or synthesized data matching that model. Put concisely, by building a formal

 ISBN: 978-1-61208-270-7

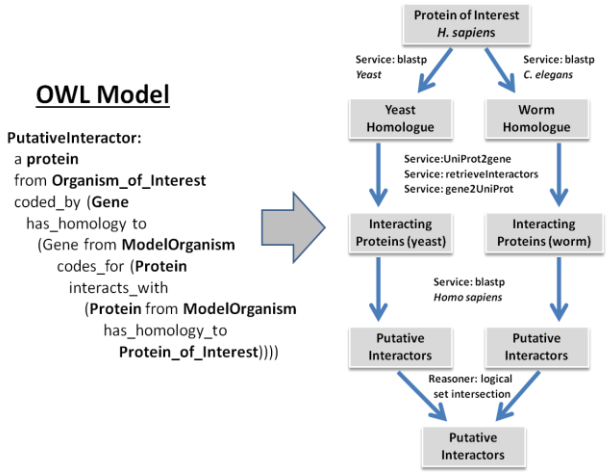

Figure 2. Conversion of models into analytical workflows: A biological model, in this case describing the chain of properties that would be expected of proteins that interact in a particular organism, is constructed in OWL (left). The SHARE software then analyses the knowledge in this model, and constructs an analytical workflow of inter-connected SADI Semantic Web Services (right) which is capable of generating data that conforms to that model.

model representing a hypothetical biological scenario, we were able to automatically find data compatible with that scenario without introducing (manual) bias. Moreover, the hypothetical model, the analytical workflow, and the result, were: (a) explicit and machine-readable; (b) inherently connected into local and remote data-sets on the Web, (c) a merger of explicit local and remote biological data, knowledge and analytical expertise, and (d) automatically published on the Web for peer-review and re-use.

III. TECHNOLOGIES AND APPROACH

The Web Science research platform above utilized both standard Semantic Web technologies, as well as novel tools developed in our group. The core technologies were:

Resource Description Framework (RDF) - a data syntax consisting of statements in the form of "triples" of subject, predicate, and object, where each component of the triple is a Uniform Resource Identifier (URI).

Web Ontology Language (OWL) - a description logic used to create machine-readable assertions that direct the automated interpretation of sets of RDF triples. It is worth noting that this use of OWL – for *ad hoc* data interpretation – is not typical within the Life Sciences community, where OWL is more often used to create concept hierarchies, or as a data template or schema. A Web Science research platform, however, requires discovery of data matching a hypothetical model from a 'mashup' of multiple disparate external data resources that will not have a predictable structure. As such, biological knowledge representing hypotheses is modeled in OWL, and this model is used to both formulate the experimental workflow (see "SHARE" below) as well as to automatically discover matching data hidden within vast integrated datasets.

Semantic Automated Discovery and Integration (SADI) [11] is our set of design patterns for scientific Semantic Web Service publishing. SADI Services are distinct in that they require Web Service publishers to (a) consume and produce RDF natively; (b) model their input and output as OWL classes; and (c) explicitly model the semantic relationship between input and output data through properties in the output class. The SADI design patterns dramatically improve the ability of software to automatically discover appropriate data retrieval and analysis services, and chain these into complex analytical workflows [11].

Semantic Health and Research Environment (SHARE) [12] is a specialized SPARQL-DL query engine that (a) responds to SPARQL queries by mapping query clauses to SADI Semantic Web Services, and (b) finds instances of an OWL ontological class by recursively mapping the class-defining property restrictions to SADI service invocations, then pipelining those services into an automatically-executed workflow. Succinctly, given the OWL model of a biological concept, it will attempt to find data on the Web consistent with that model. What is most important to note about SHARE is that is capable of automatically mapping a biological model onto a computational workflow made-up of SADI services. Since our belief is that a Web Science research platform should automatically evaluate arbitrary biological hypotheses, it is this separation of, and automated mapping between, the formal biological question, and the formal computational solution, that led us investigate SHARE's utility and behaviors further. In particular, the Web Science-like features of the SHARE *in silico* research platform include:

1) That the research process is entirely Web-embedded
2) Distributed expert-knowledge encoded in OWL is used, through OWL imports, both to construct the hypothesis as well as to drive the formulation of the solution; thus, a researcher need not possess all of the knowledge the experiment requires.
3) The experimental workflow is explicit, and no manual intervention occurs during the execution of the experiment.
4) The experimental workflow can be re-generated and re-run over the same dataset (reproducibility) and more importantly, will automatically adapt to a new dataset (re-usability) due to SHARE's ability to dynamically discover appropriate Services based on both the specific dataset and the OWL model [9].
5) The experimental workflow is self-annotating, as a result of being derived from a biological model; thus, review of the experiment is dramatically simplified with no additional human curation.
6) The starting datasets, and result datasets, are encoded in RDF, and thus are by nature a part of the Web. These, together with the initiating OWL model and workflow, aid third-party evaluation.

To investigate this Web Science platform further, we are now attempting to determine the extent to which common high-throughput life science problems can be modeled in OWL, and test the resiliency of the SHARE OWL-to-

 ISBN: 978-1-61208-270-7

workflow transformations. In particular, the limitations of the OWL/OWL-2 languages are well-known, and we anticipate that certain hypothetical constructs will require other types of semantics or logical filters, such as rules and/or the filters available within the SPARQL language itself. To explore the boundaries of what types of *in silico* hypotheses can be modeled using available logics and rules languages, the following experiments are being conducted:

1) Select 15-20 peer-reviewed papers spanning a broad range of high-throughput experimental scenarios, ensuring that the datasets and Services required to execute the experiment are available.
2) Empirically attempt to model these in OWL.
3) For any hypothetical construct asserted in the publication that cannot be modeled in OWL, attempt to model it using an alternative tool such as rules or using query filters.
4) For any construct that cannot be modeled using any available tool or language, construct an appropriate Web Service that will generate data conforming to this construct (while we acknowledge that building a Web Service to resolve every difficult case is not a scalable solution, it ensures that we can progress to the end of the investigation).
5) Manually construct a workflow, in collaboration with biological researchers, that represents their expert solution to the hypothesis.
6) Provide the OWL/rules model to SHARE; examine and compare the workflow that is automatically constructed.
7) Similarly, compare the outputs of the two workflows to determine if any methodological differences were consequential. Biologists will also evaluate the automatically-generated output.

IV. Discussion and Conclusion

Approaches to research methodology and scientific publishing have changed very little with the advent of the Web, and remain largely unaffected by the emergence of powerful Semantic Web standards and tools. We believe that, by continuing to develop the technologies described here, we can dramatically change the way *in silico* research is conducted and disseminated by establishing the fundamental concepts of "Web Science" – a novel approach to scientific research where hypotheses and experiments are explicitly described, publicly shared, intimately linked into existing data and knowledge, dynamically executed in an unbiased manner using the encoded expertise of the global community, and automatically published with a full provenance record, enabling rigorous peer-review. We propose, further, that formal semantic models can (and should) be used as the mediators of scientific discourse and disagreement.

The objective with this report is to present these ideas to the community in order to: raise awareness of the project; foster debate about its plausibility and utility; and encourage both collaborative and independent pursuit of similar research problems with the goal of rapidly bringing Web Science to fruition. In this regard, we are currently collaborating with the HyQue project [13] which utilizes RDF/OWL to model data and knowledge, and evaluates hypotheses formulated as SPARQL queries using a novel ontology. The significant differences are that SHARE uses distributed resources rather than a warehouse; utilizes Web Services, thus can execute analyses in addition to static data retrievals; and does not utilize any project-specific ontologies. Nevertheless, these independently-derived, yet highly similar, Web Science research environments suggest that the potential "solution space" for Web Science infrastructures may be very small.

Acknowledgment

Funding from the Marie Curie Co-fund, the Heart and Stroke Foundation of Canada, Microsoft Research, the Canadian Institutes for Health Research, and CANARIE.

References

[1] M. Gillam, et al. "The healthcare singularity and the age of semantic medicine," in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, T. Hey, S. Tansley, and K. Tolle, Eds. Microsoft Research, 2009, pp. 57–64.

[2] K. A. Baggerly and K. R. Coombes, "Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology," *Ann. Appl. Stat.*, vol. 3, no. 4, Dec. 2009, pp. 1309–1334.

[3] A.-L. Boulesteix, "Over-optimism in bioinformatics research" *Bioinformatics*, vol. 26, no. 3, Feb. 2010, pp. 437–439.

[4] P. Fisher, et al., "A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis" *Nucleic Acids Res.*, vol. 35, no. 16, Jan. 2007, pp. 5625–33.

[5] V. Gewin, "Research: Uncovering misconduct," *Nature*, vol. 485, no. 7396, May 2012, pp. 137–139.

[6] J. P. A. Ioannidis, et al., "Repeatability of published microarray gene expression analyses" *Nat. Genetics*, vol. 41, no. 2, Feb. 2009, pp. 149–155.

[7] C. M. Micheel, S. J. Nass, and G. S. Omenn, Eds., *Evolution of Translational Omics Lessons Learned and the Path Forward*. The Institute of Medicine of the National Academies, 2012, p. 354.

[8] M. Hauder, Y. Gil, R. Sethi, Y. Liu, and H. Jo, "Making data analysis expertise broadly accessible through workflows" in *Proceedings of the 6th workshop on Workflows in support of large-scale science - WORKS '11,* 2011, p. 77.

[9] I. Wood, B. Vandervalk, L. McCarthy, and M. Wilkinson, "OWL-DL Domain-Models as Abstract Workflows" in *Leveraging Applications of Formal Methods, Verification and Validation. Applications and Case Studies*, T. Margaria and B. Steffen, Eds. Springer Berlin/Heidelberg, 2012, pp. 56–66.

[10] P. Lord, et al., "Applying Semantic Web Services to Bioinformatics: Experiences Gained, Lessons Learnt" *LNCS*, vol. 3298, Jan. 2004, pp. 350–364.

[11] M. D. Wilkinson, B. Vandervalk, and L. McCarthy, "The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation" *J. Biomed. Semant.*, vol. 2, no. 1, Oct. 2011, p. 8.

[12] B. Vandervalk, L. McCarthy, and M. Wilkinson, "SHARE: A Semantic Web Query Engine for Bioinformatics" in The Semantic Web, *LNCS Proc. ASWC 2009*, vol. 5926, pp. 367–369.

[13] A. Callahan, M. Dumontier, N.H. Shah, "HyQue: Evaluating hypotheses Using Semantic Web Technologies," *J. Biomed. Semant.* vol 2(Suppl.2): S3.

 ISBN: 978-1-61208-270-7