A Semantic Workflow Approach to Web Science Analytics

Extended Abstract

Spencer C. Norris
Rensselaer Polytechnic Institute
Troy, New York 12180
norris@rpi.edu

John S. Erickson
Rensselaer Polytechnic Institute
Troy, New York 12180
erickj4@rpi.edu

Deborah L. McGuinness
Rensselaer Polytechnic Institute
Troy, New York
dlm@cs.rpi.edu

ABSTRACT

Reproducibility and reuse are rapidly becoming guiding principles in publishing and sharing scientific results[1, 10]. In order to enhance researchers' ability to leverage existing results, many are moving in the direction of semantic workflow systems, which enable users to define and share experimental procedures as linked data on the Web. These workflows provide a powerful mechanism for reproducing *in silico* experiments and thus are well-suited for Web Science tasks. In order to aid users in the process of experiment design and reproduction, we are integrating the Workflow INstance Generation and Specialization (WINGS) system [7] with our existing Semantic Numeric Exploration Technology (SemNExT) framework [9]. This will provide a completely open-source stack for designing *in silico* experiments using a combination of semantic and numeric analyses.

We will explore how this system may be configured to create reproducible Web Science workflows, especially as it pertains to data federation across the Web. We are leveraging our existing tooling as we develop new approaches for automatically generating provenance for interacting with remote endpoints of heterogeneous data sources. This will support not only the aggregation of diverse and geographically-disparate data sources across the Web, but also collaborative science by allowing other users to reproduce and expand on the same results using shared workflows.

CCS CONCEPTS

•**Information Systems** → *Semantic web description languages;* •**Computer Systems Organization** → **Architectures - Other Architectures - Heterogeneous (hybrid) systems;** *Architectures - Distributed architectures;* •**Web interfaces** → Mashups;

KEYWORDS

Semantic Workflows, Reproducibility, Data Federation

1 BACKGROUND

Semantic workflows can be viewed as descriptions of computational procedures encoded using the Resource Description Framework (RDF) [8], the Web Ontology Language (OWL) [3] and other open Web standards. Semantic workflows have been shown to provide a robust approach for reproducing *in silico* experiments, especially in systems biology and bioinformatics domains. Because of this, there is a strong motivation for using them to coordinate and publish reproducible experiments in the domain of Web Science, such as data collection and aggregation methods that rely on the federation of heterogeneous, geographically-separated data sources.

2 RELATED WORK

Semantic workflows have seen early successes concerning reproducibility, e.g., in recreating the Tuberculosis drugome [4], demonstrating their applicability to *in silico* experiments for -omics and systems biology. However, there have not been many examples of semantic workflows being used to explicitly support tasks pertinent to Web Science. The closest analogues have been standalone software solutions, such as BioMart, and experiments designed using the Galaxy workflow system. The BioMart data federation system has been used to support aggregation across geographically-dispersed data stores [11]. While BioMart is a powerful, highly scalable system, it suffers the usual drawbacks of a non-generic interface, as it requires that the databases it wraps to expose their data as BioMart datasets. Additionally, BioMart is only an aggregation tool; it does nothing in terms of experimental workflow design.

The Galaxy workflow system has done similar work with workflows for data aggregation using SADI services [2]. However, a key drawback of Galaxy's approach to workflows is the lack of provenance capture as RDF. Galaxy workflows are encoded as unintuitive JSON representations, notably lacking URIs for output results, module names and other important features. Retrospective provenance is captured as a history log which can be exported and reused in other experiments. However, these encodings only support other instances of Galaxy and fail to capture the semantics of the workflow using either an ontology or controlled vocabulary. This limits Galaxy's usefulness in workflow discovery and recombination, and thus its applicability to collaborative Web Science.

3 WEB SCIENCE AS SEMANTIC WORKFLOWS

The Semantic Numeric Exploration Technology (SemNExT) project [9] aims to address the issue of workflow reproducibility for Web Science by integrating with the WINGS semantic workflow system [7]. Building on top of WINGS confers a number of benefits, most notably automated prospective and retrospective provenance capture for workflows as linked data. This allows workflows not only to be reproduced, but also to be readily discovered and recombined to generate larger experiments. Additionally, WINGS has developed a recommendation system for dataset discovery, assisting users in selecting datasets based on their semantic qualities [5] [6]. These factors make WINGS an attractive candidate for supporting workflow reuse and publication on the Web.

WebSci '17, June 25–28, 2017, Troy, NY, USA
© 2017 Copyright held by the owner/author(s). 978-1-4503-4896-6/17/06. .
DOI: http://dx.doi.org/10.1145/3091478.3098875

SemNExT introduces the components necessary to interface with WINGS and hook it into data repositories across the Web. Specifically, it provides the Python object model required to wrap data sources and expose their results in semantically consistent ways. Datasource objects expose a given service through a collection of developer-defined methods. Datasources are then plugged into Annotator objects, which interface with Datasources and narrowly define how they will be used to annotate an input RDF graph. These components all have analogues in the SemNExT ontology, allowing the different elements to be recombined using semantic reasoning.

Using this approach, we can recombine Web resources while preserving the consistency of assumptions about the processing steps performed on the input data, the types of different entities, and much more. This approach is inspired by SADI services but is generalizable as a method of wrapping any remote data source in reusable objects capable of mapping query results into RDF. Using the associated ontological concepts, WINGS is able to reason about their behavior and use in a given workflow, as well as recommend them to experimenters during the design process.

4 REPRODUCIBILITY AND REUSE

Combining WINGS and SemNExT will drive the creation of reproducible workflows for Web Science tasks in a way that supports discovery, reuse and collaboration. Because all WINGS workflows are encoded using W3C standards, they can be discovered and reasoned over as semantically-enriched linked data. This is an important next step for workflow technology, as defining workflows with open standards supports interoperability, recombination, and future research inquiries, as well as potentially improving the representation of workflows in online data observatories.

Additionally, the segmentation of programmatic interfaces into unique components and the association of those components with ontological concepts allows the discovery and reuse of the code itself supporting the workflow. Because the code would be structured atomically and have minimal external dependencies, it would allow code segments with well-defined interfaces to be dynamically loaded according to the requirements of a given workflow. This has the effect of treating code as data on the Web and allowing it to be used in contexts outside of monolithic software paradigms. Furthermore, embracing Annotator and Datasource entities as atomic operational units means that both transformative and generative procedures can be directly invoked based only on the semantics of their associated concepts rather than relying on human inference. This means that the programmatic components would no longer require coding expertise to invoke.

Lastly, the use of ontologies as organizing utilities in workflow design, generation and execution means that the interactions between these atomized components are also subject to logical constraints; thus, they can be recombined using lightweight reasoning methods making data mash-ups that invoke Web-based data repositories trivial to implement (provided the desired components already exist). This could lead to fully-automated service composition in future work, provided that the semantics are rigorous enough and the code requirements properly encapsulated.

The ability to create semantic workflow descriptions that rely on openly accessible data resources and reusable components for automating transactions with endpoints has the potential to create a new approach to composing services over the Web. By allowing code components to be directly referenced in workflow descriptions and exposing that code as linked data, it may be possible to create fully atomized workflow descriptions in support of Web Science.

5 CONCLUSION

As the amount and diversity of data on the Web continues to grow, it is increasingly important to develop experimentation methods that allow Web Science analyses to be recreated and recombined. By fusing programmatic interfaces for Web resources with semantic workflows, researchers will be able to consistently reproduce results and aggregate information across the Web, as well as redistribute their experiments and recombine them with others in a semantically consistent way. Because of their suitability for this task, we will likely start by ingesting SADI service descriptions and wrapping them in Datasource and Annotator objects to support the SemNExT project with robust annotation methods for open genetics data. Beyond this, we will likely attempt a similar configuration using remote SPARQL endpoints and relational databases.

REFERENCES

[1] On data availability, reproducibility and reuse. 19 (????), 259. Issue 4. https://doi.org/10.1038/ncb3506

[2] Mikel Egaña Aranguren and Mark D Wilkinson. 2015. Enhanced reproducibility of SADI web service workflows with Galaxy and Docker. *GigaScience* 4 (2015), 59. https://doi.org/10.1186/s13742-015-0092-3

[3] Frank van Harmelen Deborah L. McGuinness. 2004. Owl web ontology language overview. *W3C recommendation 10.2004-03* 2004, February (2004), 1–12. https://doi.org/10.1145/1295289.1295290

[4] Daniel Garijo, Sarah Kinnings, Li Xie, Lei Xie, Yinliang Zhang, Philip E. Bourne, and Yolanda Gil. 2013. Quantifying reproducibility in computational biology: The case of the tuberculosis drugome. *PLoS ONE* 8, 11 (2013). https://doi.org/10.1371/journal.pone.0080278

[5] Yolanda Gil, Pedro A Gonzalez-Calero, Jihie Kim, Joshua Moody, and Varun Ratnakar. 2011. A semantic framework for automatic generation of computational workflows using distributed data and component catalogues. *Journal of Experimental & Theoretical Artificial Intelligence* 23, 4 (2011), 389–467.

[6] Yolanda Gil, Varun Ratnakar, and Christian Fritz. 2010. Assisting Scientists with Complex Data Analysis Tasks through Semantic Workflows.. In *AAAI Fall Symposium: Proactive Assistant Agents*. Citeseer.

[7] Yolanda Gil, Varun Ratnakar, Jihie Kim, Pedro González-Calero, Paul Groth, Joshua Moody, and Ewa Deelman. 2011. Wings: Intelligent workflow-based design of computational experiments. *IEEE Intelligent Systems* 26, 1 (2011), 62–72. https://doi.org/10.1109/MIS.2010.9

[8] Graham Klyne and Jeremy J Carroll. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. *W3C Recommendation* 10, October (2004), 1—-20. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[9] D.L. McGuinness and K. Bennett. 2015. Integrating Semantics and Numerics: Case Study on Enhancing Genomic and Disease Data Using Linked Data Technologies. In *Proceedings of SmartData 2015 (August 18-20 2015, San Jose, CA).*

[10] Gaël Varoquaux. 2015. Of software and Science. Reproducible science: what, why, and how. (2015).

[11] Junjun Zhang, Syed Haider, Joachim Baran, Anthony Cros, Jonathan M. Guberman, Jack Hsu, Yong Liang, Long Yao, and Arek Kasprzyk. 2011. BioMart: A data federation framework for large collaborative projects. *Database* 2011 (2011). https://doi.org/10.1093/database/bar038