HHS Public Access

Author manuscript

Transform Digit Worlds (2018). Author manuscript; available in PMC 2019 March 01.

Published in final edited form as:

Transform Digit Worlds (2018). 2018 March ; 10766: 620–625. doi:10.1007/978-3-319-78105-1_70.

Semantic Mediation to Improve Reproducibility for Biomolecular NMR Analysis

Michael R. Gryk[1,2] and **Bertram Ludäscher**[1]

[1]University of Illinois, Urbana-Champaign, Champaign IL 61820, USA

[2]UCONN Health, Farmington, CT 06030, USA

Abstract

Two barriers to computational reproducibility are the ability to record the critical metadata required for rerunning a computation, as well as translating the semantics of the metadata so that alternate approaches can easily be configured for verifying computational reproducibility. We are addressing this problem in the context of biomolecular NMR computational analysis by developing a series of linked ontologies which define the semantics of the various software tools used by researchers for data transformation and analysis. Building from a core ontology representing the primary observational data of NMR, the linked data approach allows for the translation of metadata in order to configure alternate software approaches for given computational tasks. In this paper we illustrate the utility of this with a small sample of the core ontology as well as tool-specific semantics for two third-party software tools. This approach to semantic mediation will help support an automated approach to validating the reliability of computation in which the same processing workflow is implemented with different software tools. In addition, the detailed semantics of both the data and the processing functionalities will provide a method for software tool classification.

Keywords

Ontology; Computational Reproducibility; Provenance

1 Introduction

1.1 Computational Reproducibility

Researchers in the natural sciences are becoming increasingly concerned about the repeatability and reproducibility[1] of their studies [1–4]. Biomolecular NMR (bioNMR) is a well-established spectroscopic technique which uses the same principles as MRI in order to observe biomolecules at atomic resolution. It has been common practice to deposit the completed, derived bioNMR datasets with national repositories (e.g., the Protein Data Bank (www.pdb.org) or the BioMagResBank (www.bmrb.org)) while embedding natural language descriptions of the computational analysis within the publications which report the findings.

[1]In this paper, we use the definitions of Vitek & Kalibera [3] that repeatability is the ability for the same researcher to get the same results with the same computational environment, while reproducibility is the ability for others to get similar results with similar computational tools.

This traditional approach has its limitations when considering computational reproducibility and repeatability for bioNMR. First, with the continuing increase in the complexity of the computational pipeline [5, 6], it is nearly impossible to document the process in enough detail for it to be reproduced. Second, the computation itself (in the case of bioNMR) requires dozens of specialized software tools built by academic labs. While typically free to use, these software tools are usually poorly documented, difficult to install and often minimally maintained and/or abandoned.

These issues are being addressed by the new National Center for Biomolecular NMR Data Processing and Analysis (www.nmrbox.org) – a joint NIH-funded project at UCONN Health, the University of Wisconsin and the University of Illinois. The Center has provisioned Ubuntu Virtual Machines (VMs) with more than sixty of the software tools used by bioNMR spectroscopists [7]. (Virtual machines were chosen over container technology as the various software tools are often graphical and used in concert.) These VMs (referred to as NMRbox) are hosted on a private cloud platform at UCONN Health and are available for access through the internet to spectroscopists in non-profit institutions. Since the initial release in the summer of 2016, NMRbox has over 700 registered users. Images of the VMs are also available for download and local installation. Importantly, images of the VMs will be stored indefinitely, giving future researchers the computational infrastructure with which to reproduce prior bioNMR studies, one of the recommendations of Piccolo and Frampton [8]. Of course, this long-term reproducibility requires maintenance of hypervisors capable of running the VM.

Software persistence is only one barrier to reproducibility, however. The other barrier is documenting the analysis workflow in sufficient detail that others can independently reproduce it [2, 8–9]. This requires the capture of the various pieces of metadata regarding software configuration and data manipulation which are necessary to track the process from raw, observational data to the final, derived datasets. Referred to as *provenance* [2], this metadata would ideally be stored in a neutral format which is both human and machine-readable. Importantly, provenance metadata should be readily translatable to any of the software tools capable of performing computations on the data.

1.2 Software Ontologies

In this paper we report preliminary work in developing the infrastructure within NMRbox to gather this important metadata. An important challenge in this endeavor is the diversity of software tools itself. There exist multiple software tools capable of performing each computational step. The choice of one tool over another can be simply a matter of personal preference (as in the choice of Chrome over Firefox) or it can be due to some subtle differences in the software whereby one tool performs better for a specific set of use cases than another tool. Regardless of the rationale, it is doubtful that the bioNMR field will ever unite in support of a single software tool for all computation. Thus, bioNMR workflows more closely resemble those of analyses which combine and compare data from disparate sources, such as genomic bioinformatics [2].

The multiplicity of software tools has consequences for metadata capture and data curation. Considering that multiple tools are capable of performing the same general task, it would be

expected that much of the metadata required to be recorded is conceptually similar. However, since the various tools were developed by different labs at different times, the tools do not use precisely the same vocabulary or nomenclature when referring to parameters or configuration settings. Along the same lines, the parameterization of computational steps often use different units of measure or parametric weights such that the metadata from one tool cannot be used with another without recalibration. Finally, recognizing that some tools are better at some computations than other – this suggests that there may be subtle but profound sematic differences between seemingly similar functional tasks.

Our current approach to mapping this diverse landscape of metadata is inspired by semantic mediation [9, 11, 13]. The general idea is to model the function or functionality of each software tool in order to identify the key metadata necessary to recapitulate the computation. A separate, tool-specific ontology will be created representing the metadata model for each individual software tool. These ontologies will be conceptually linked within the NMRbox VMs – providing a kind of "internal semantic web" for facilitating tool integration and mediation. This will allow for the simple cases of identifying when two different software terms refer to the same thing (the *sameAs* relation) as well as when the same term is used by two software tools to refer to different things (the *differentFrom* relation). Most importantly, it will allow complex mappings when terms from different software tools are similar but not identical. For instance, it is often the case in bioNMR that two implementations of a mathematical operation will use different sign conventions. This can easily be modeled as a *negativeOf* relation. More complex mathematical relationships can be modeled using the terminology defined by MathML and OpenMath.

2 Research Model

Our general strategy for semantic mediation is to construct separate ontologies expressing the semantics of each software tool supported within the NMRbox VM's. These ontologies are linked with each other to support semantic conversion between the various data and process elements for computation. As there is not a complete overlap between the various tools, a core ontology representing the basic data relationships inherent with bioNMR data is used as a foundation for metadata interchange. Ontologies by their nature are always "under-development"; the ontologies described here are accessible through GitHub (https://github.com/CONNJUR/Ontology_Development).

2.1 Core Ontology

The core ontology attempts to model the fundamental concepts of bioNMR experiments and their supporting data/metadata which are used by the various supported software tools. Where possible, we have attempted to use existing, established ontologies for concepts which are not bioNMR specific; for instance, the Friend of a Friend (FOAF) ontology for referring to the various data collectors and curators along the computational workflow; the Event Ontology for referring to data collection events; and the Prov-O model for referring to provenance information along the computational workflow.

As shown in Figure 1, the core ontology deals with both time domain and frequency domain representations of bioNMR spectra, which are mathematical duals of each other. There are many mathematical methods for interconverting between time and frequency, but by far the most common is the Fourier Transform. The core ontology uses the Prov-O vocabulary for defining the provenance of the generation of a frequency spectrum from the time domain recording. The details of the implementation of the Fourier Transform are defined within the tool-specific ontologies.

2.2 Software-based Ontologies

There are many software tools available for converting time domain bioNMR data to frequency plots [7, 12]. This is a multistep process involving several data cleaning steps in which mathematical operations are applied to the data in order to enhance the signal and suppress the noise [5]. The key operations are the Fourier Transform and concomitant phasing of the spectrum to provide so-called absorptive spectral peaks. Both the transform and the phasing operations can be done with either of two sign conventions associated with the integration along the time axis. Of the four major tools supported by Nowling *et al.* [12], two choose a default of a positive sign convention and two with a negative sign convention.

A consequence of the differing sign conventions is that if the primary data are fed naively into each of the four tools, the resulting frequency plots will be reversed and the process would appear irreproducible. However, by correcting for the sign convention during the parameterizing of the Fourier Transform, reproducible results are achieved. The tool-specific ontologies assist in defining these important semantic distinctions by relating the software conventions to those of the core ontology as shown in Figure 1.

3 Conclusions and Expected Contributions

In summary, this approach to metadata curation will help us to expand from the simple level of *repeatability* inherent in archiving the static VMs to the more informative goal of *reproducibility*, in which one can swap different computational tools along the processing workflow [13] in order to validate the final results. Different software tools use alternate sign conventions, units of measure and arbitrary scaling factors in their parameterization of similar computational tasks. Defining the semantics of the computational tasks as well as the mathematical inter-relationships within linked ontologies will assist in the metadata translation necessary to configure alternate tools to execute the workflow in equivalent manners. This is complimentary to defining the overall dataflow of the workflow, as is done with tools such as YesWorkflow [10]. As an extra benefit, this process will allow for the identification of related and/or equivalent tasks as a method of software tool classification. It is also anticipated that by modelling the variants and variations, these linked ontologies will suggest alternative implementations of equivalent workflows (as shown by Bowers & Ludäscher [13]) – also assisting in validating that bioNMR computational results are reproducible.

The software-specific ontologies described in this paper are a similar approach to data integration and reproducibility as proposed by Rijgersbert, *et al.* with the ontology of units of measure [14]. A major difference for bioNMR computation is that many of the scaling

factor differences between software tools are either unit-less or tend to be arbitrary conversions to non-standard units of measure done for the ease of computation, not as a standardize method of reporting findings. Thus a more detailed level of parametric definitions is required than would be supported simply by defining units of measure.

Future work will be the continued development of this ontological framework by expanding the core ontology, adding additional software-specific ontologies, and continued inclusion of other controlled vocabularies such as MathML and OpenMath. In conjunction with the developers of NMRbox, these ontologies will be used for semantic data management within their supported VM's to help support more detailed data curation, assist with workflow management and reuse, validate workflow reproducibility, and eventually enhance data depositions to the BioMagResBank public repository. These ontologies will also be used within the CONNJUR Workflow Builder workflow management system [15] in order to provide a broader abstraction of the individual processing actors which does not rely on the underlying software tool implementation.

Acknowledgments

This work was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number GM-111135.

References

1. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis JPA. A manifesto for reproducible science. Nature Human Behaviour. 2017; 1:1–9.
2. Kanwal S, Khan FZ, Lonie A, Sinnot RO. Investigating reproducibility and tracking provenance – A genomic workflow case study. BMC Bioinformatics. 2017; 18:337. [PubMed: 28701218]
3. Vitek J, Kalibera T. Repeatability reproducibility and rigor in systems research. Proceedings of the ninth ACM international conference on Embedded software (EMSOFT '11); 2011. 33–38.
4. Stodden V, Miguez S. Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. Journal of Open Research Software. 2014; 2(1):e21.
5. Verdi KK, Ellis HJ, Gryk MR. Conceptual-level workflow modeling of scientific experiments using NMR as a case study. BMC Bioinformatics. 2007; 8:31. [PubMed: 17263870]
6. Ellis HJC, Nowling RJ, Vyas J, Martyn TO, Gryk MR. Iterative Development of an Application to Support Nuclear Magnetic Resonance Data Analysis of Proteins. Proc. of the International. Conf. on Inform. Techn: New Generations; 2011; 2011. 1014–1020.
7. Maciejewski MW, Schuyler AD, Gryk MR, Moraru II, Romero PR, Ulrich EL, Eghbalnia HR, Livny M, Delaglio F, Hoch JC. NMRbox: A Resource for Biomolecular NMR Computation. Biophysical Journal. 2017; 112(8):1529–1534. [PubMed: 28445744]
8. Piccolo SR, Frampton MB. Tools and techniques for computational reproducibility. Giga Science. 2016; 5(1):30. [PubMed: 27401684]
9. Bowers S, Ludäscher B. Semantic Web Technologies for Searching and Retrieving Scientific Data (SCISW). 2003. Towards a Generic Framework for Semantic Registration of Scientific Data.
10. McPhillips T, Song T, Kolisnik T, Aulenbach S, Belhajjame K, Bocinsky K, Cao Y, Chirigati F, Dey S, Freire J, Huntzinger D, Jones C, Koop D, Missier P, Schildhauer M, Schwalm C, Wei Y, Cheney J, Bieda M, Ludäscher B. YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. International Journal of Digital Curation. 2015; 10(1):298–313.

11. Madin JS, Bowers S, Schildhauer M, Jones M. Advancing Ecological Research with Ontologies. Trends in Ecology & Evolution. 2008; 23:159–68. [PubMed: 18289717]
12. Nowling RJ, Vyas J, Weatherby G, Fenwick MW, Ellis HJC, Gryk MR. CONNJUR spectrum translator: an open source application for reformatting NMR spectral data. Journal of Biomolecular NMR. 2011; 50(1):83–89. [PubMed: 21409563]
13. Bowers S, Ludäscher B. Actor-Oriented Design of Scientific Workflows; 24st Intl. Conference on Conceptual Modeling (ER); LNCS. Springer; 2005. 369–384.
14. Rijgersberg H, van Assem M, Top J. Ontology of units of measure and related concepts. Semantic Web – Interoperability. Usability, Applicability. 2011; (1)
15. Fenwick M, Weatherby G, Vyas J, Sesanker C, Martyn TO, Ellis HJC, Gryk MR. CONNJUR Workflow Builder: A Software Integration Environment for Spectral Reconstruction. J Biomol NMR. 2015; 62:313–326. [PubMed: 26066803]

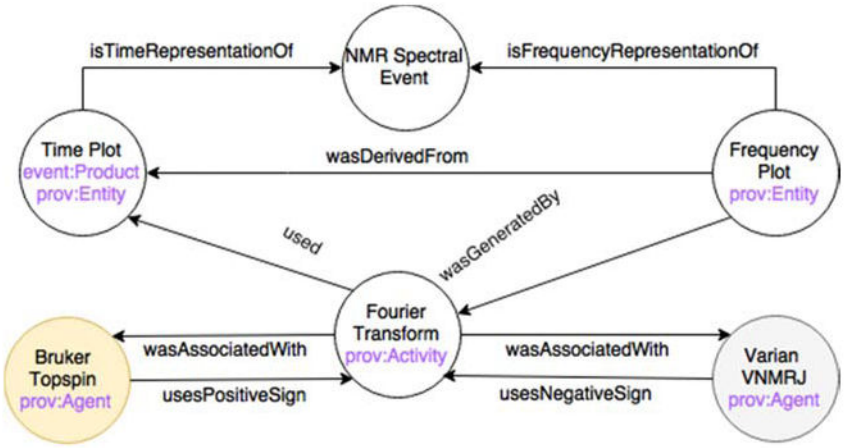

Fig. 1.

Schematic showing a portion of the core ontology for bioNMR spectra (white circles) along with tool-specific elements (beige and gray circles). Where appropriate, entities are mapped to other top ontologies such as the Event and Prov-O ontologies (purple text). As illustrated above, the “Bruker” and “Varian” implementations of the Fourier Transform differ in the sign convention used for the integration.