Matryoshka Modeling: Building one conceptual model within another

Michael R. Gryk; School of Information Sciences, University of Illinois, Urbana-Champaign, USA

Abstract

This paper describes a nested modeling approach for recording the provenance of scientific workflows in the domain of bioNMR. Nested modeling can promote wider adoption by using robust, well-established conceptual models from other domains. However, it draws into contrast the usefulness of the model versus its correctness.

Introduction

George Box famously wrote that "all models are wrong but some are useful." (Box 1979) Implicit within this provocative statement is the admission that models are built to satisfy two distinct goals. One, a model is created to be useful for a specific purpose: for instance, to implement a data store where efficiency of archival and retrieval is a key design specification for the underlying model. Two, a model is created to correctly represent the inherent (and sometimes enigmatic) structure and semantics of the information being modeled. Box's adage challenges the absolute correctness of any given model, of a model's ability to perfectly represent the truth. This emphasis on the importance of correctness in modeling has spawned a variation of Box's statement, "All models are wrong; some are wronger than others." (Carr, Zhu 2018)

The two goals of correctness and usefulness are by no means aligned. The correct botanical model for a tomato is to classify it as a berry. However, from a retrieval point-of-view, it is more appropriate to categorize tomatoes as vegetables as that is where customers expect to find them both in the produce section of the grocery store as well as on the screens of self-check-out scanners[1]. Maximizing both correctness and usefulness poses a significant modeling challenge in many application domains (Moody 2003).

This paper describes conceptual modeling efforts for recording workflow provenance in the scientific domain of biomolecular nuclear magnetic resonance spectroscopy (bioNMR), an example of the balance between correctness and usefulness. The larger context of this work is the NIH-supported Biomedical Technology Research Resource, NMRbox (Maciejewski, Schuyler et al. 2017), whose mission is to foster the reproducibility of bioNMR studies by maintaining a persistent archive of executable software as well as assisting in enhanced metadata capture and data curation for richer depositions to the international repository, BioMagResBank (Ulrich, Akutsu et al. 2008). In the context of this latter goal of data curation, the NMRbox/CONNJUR teams have been developing a scientific workflow system for bioNMR, called CONNJUR Workflow Builder (CWB). At the time of its publication, this workflow system had a custom data model for defining workflows within the CONNJUR application, which could be either stored in a relational database or exported as XML for sharing between researchers (Fenwick, Weatherby et al. 2015). However, being a custom data model, the reusability and transparency was limited to users of CWB (Heintz, Gryk 2018).

[1]The Supreme Court of the United States declared tomatoes to be vegetables (for taxation purposes) in 1893 since they were grown, prepared and consumed like vegetables. Nix v. Hedden, 1893. http://openjurist.org/149/us/304

During the past two years, the data model for CWB has been refactored to foster transparency, data reuse and broader scholarly communication (Heintz, Gryk 2018). The design principles for the refactor were three-fold: (a) to adopt a widely used conceptual model for workflows and provenance description, (b) to include domain-specific models for bioNMR data, and (c) to support data files generated by software tools at various points along the processing workflow. These three requirements pitted correctness against usefulness.

Design Strategy

It was decided to adopt the PREMIS conceptual model as the general framework for defining the provenance of the computation workflows (https://www.loc.gov/standards/premis/). PREMIS is the international standard for digital preservation workflows and provenance. Similar to the PROV standard (http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/), PREMIS relies on three central entities: those of intellectual objects (e.g. files, bit-streams), agents (e.g. software, people), and events (e.g. data collection, transformation)[2]. One benefit of PREMIS is that its usage in the digital preservation community is mature, constrained, well-documented and transparent, thereby passing on these attributes to CWB provenance. A second benefit is that as of version 3, the top level PREMIS entities have extensions for embedded, domain-specific metadata. An additional perk, PREMIS can be expressed as either XML or RDF.

The metadata extensions within PREMIS satisfy the second design principle, allowing custom bioNMR XML to be embedded within a PREMIS record. However, that capability has consequences for the modeling process, as the bioNMR model must be fragmented to adhere to the higher level PREMIS entities of Object, Agent and Event. As will be discussed, this causes further modeling complications when supporting third-party software tools, in which the underlying metadata models do not adhere to these top level PREMIS constructs.

Role of Provenance Capture

Provenance capture during workflow execution is important for several reasons. First, the provenance record provides a report of what transformations and operations the scientist used in cleaning and processing the dataset during a given study. Prior to CWB, the only established mechanism for reporting provenance was the sharing of the UNIX scripts which were used for data processing. While somewhat useful, there is a language burden on anyone wishing to explore the provenance – knowing the language of both UNIX shells (csh, bash, etc.) as well as the syntax and semantics of the underlying processing tools. PREMIS provides a more universal structure for that discourse. The use of controlled vocabularies such as DPCVocab (Chao, Cragin et al. 2015) is also being explored to provide a more universal language for discourse.

Another benefit of provenance capture is that significant properties of the intermediate dataset can be stored within the PREMIS record without having to store the much larger, intermediate files themselves. This led to another modeling consideration: providing metadata for reporting data characteristics which are subject to manipulation throughout a processing workflow. Finally, generating a provenance record provides a useful vehicle for soliciting curation from the data curators themselves (Heintz, Gryk 2018).

[2] A fourth entity for Rights is not applicable to this application of PREMIS for computational workflows.

Results and Conclusions

The provenance metadata for bioNMR processing workflows fit nicely within the three top level PREMIS entities of Object, Event and Agent. Every bioNMR experiment results in a binary file (Object) which is further manipulated through data cleaning and other mathematical transformations. The data collection itself is an Event, as is each processing step. And each of these Events has associated Agents: for data collection there is the scientific instrument and human operator, for processing there are the associated software packages. Metadata and schema definitions for metadata relevant to each of these top level entities is documented on GitHub (CONNJUR_ML).

This form of matryoshka modeling revealed interesting aspects of bioNMR data as well as imposing constraints. For instance, the datasets stored by the spectrometer software lump all metadata together. Yet while modeling the bioNMR experiment collection Event, it was obvious that there were three Events occurring at the time of data collection. The first Event is the sample preparation Event, in which the experimental apparatus provides both the mechanism of data collection as well as limits the universe of possible observations (Coombs 1964). The second Event is the actual observation Event, in which the magnetization of the sample is measured within the apparatus. The third Event is the digitization and recording of the measurement.

All three of these Events put constraints on the measurement, and each is recorded in separate metadata fields which are lumped together. For instance, the first Event constrains the observable frequencies as the reciprocal of the length of the read-out pulse, recorded as “pw” or “p1”. The second Event further constrains the observation by applying bandpass filters, recorded as “fb” for filter bandwidth. Finally, the third Event constrains the observation by imparting a Nyquist grid during digitization, recorded as “sweep width”.

While the author argues that these three Events are a more correct model than assuming a single data collection Event, the model may be more useful to its audience if referred to a single Event. This compromise is elegantly handled by the Event ontology, in which a single Event can be composed of sub-Events. However, PREMIS does not provide the ability of linking Events to Events, and so in order to maintain this richness of expression, the Object must be linked to all of the sub-Events. This is an important consideration of nested modeling – limitations of one model can either be ignored or accommodated within the nested model.

The above example describes multiple Events being conflated within the metadata of one record. Many software tools ignore the concept of Events and record all metadata as properties of the Object. This also has serious modeling implications. The PREMIS record must be capable of partitioning metadata into Event entities when these Events are sufficiently documented, but must also be capable of storing the same metadata within Objects if the metadata record is insufficiently detailed. This is a serious challenge for nested modeling, as the Event metadata and Object metadata are designed to be both redundant and unrelated.

In summary, nested modeling has benefits in wider adoption by amalgamating robust, well-established conceptual models from other domains. However, it exacerbates Fox’s problem of pitting the usefulness of the model against its correctness.

Literature Cited

BOX, G., 1979. *Robustness in the Strategy of Scientific Model Building.* New York : .

CARR, P. and ZHU, Q.J., 2018. *Convex Duality and Financial Mathematics.* Springer International Publishing.

CHAO, T.C., CRAGIN, M.H. and PALMER, C., L., 2015. Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology/,* **66**(3), pp. 616-633.

COOMBS, C.H., 1964. *A Theory of Data.* New York: John Wiley & Sons.

FENWICK, M., WEATHERBY, G., VYAS, J., SESANKER, C., MARTYN, T.O., ELLIS, H.J. and GRYK, M.R., 2015. CONNJUR Workflow Builder: a software integration environment for spectral reconstruction. *Journal of Biomolecular NMR,* **62**(3), pp. 313-326.

HEINTZ, D. and GRYK, M.R., 2018. Curating Scientific Workflows for Biomolecular Nuclear Magnetic Resonance Spectroscopy. *International Journal of Digital Curation,* **13**(1), pp. 286-293.

MACIEJEWSKI, M.W., SCHUYLER, A.D., GRYK, M.R., MORARU, I.I., ROMERO, P.R., ULRICH, E.L., EGHBALNIA, H.R., LIVNY, M., DELAGLIO, F. and HOCH, J.C., 2017. NMRbox: A Resource for Biomolecular NMR Computation. *Biophysical journal,* **112**(8), pp. 1529-1534.

MOODY, D.L., 2003. Measuring the Quality of Data Models: An Empirical Evaluation of the Use of Quality Metrics in Practice. *ECIS 2003 Proceedings,* **78**.

ULRICH, E.L., AKUTSU, H., DORELEIJERS, J.F., HARANO, Y., IOANNIDIS, Y.E., LIN, J., LIVNY, M., MADING, S., MAZIUK, D., MILLER, Z., NAKATANI, E., SCHULTE, C.F., TOLMIE, D.E., KENT WENGER, R., YAO, H. and MARKLEY, J.L., 2008. BioMagResBank. *Nucleic acids research,* **36**(Database issue), pp. D402-8.