ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

Towards Desiderata for Provenance Ontologies in Biomedicine

Satya S. Sahoo

Division of Medical Informatics, School of Medicine, Case Western Reserve University, Cleveland, OH, USA

Abstract. Provenance is essential metadata in biomedicine to verify data quality, ensure reproducibility of published data, validate experiment protocols, and compute trust of scientific results. Using the requirements identified by the W3C Provenance Incubator Group, seven desired attributes are defined to create an evaluation framework for ten provenance ontologies in biomedicine. Three ontologies, the Experimental Factor Ontology, the Parasite Experiment Ontology, and Ontology for Clinical Research are found to be fully compliant with the desiderata.

1 Introduction

Provenance, derived from the French word *provenir* meaning "to come from" is critical contextual metadata in biomedicine to validate data quality, verify integrity of experiment processes, and compute trust [1] [2]. Provenance metadata is also required for ensuring reproducibility of published scientific results, objective comparison of datasets produced by different research groups, effective biomedical data integration (in form of contextual metadata) [2]. The objective of this study is to propose a set of desired attributes for a provenance ontology in biomedicine that addresses some of the requirements identified by the W3C Provenance Incubator Group and also some of the OBO Foundry principles, and review a list of existing provenance ontologies using the desired attributes as a frame of reference.

Existing work in identifying desired qualities of biomedical ontologies include, evaluation framework for controlled medical vocabularies [3], disease ontologies [4], and the OBO foundry principles (http://www.obofoundry.org/wiki/index.php/OBO_Foundry_Principles).

We derive some of the attributes used in this study from the existing work, but in addition use the "requirements for provenance" identified by the W3C Provenance Incubator Group [5], which are desired attributes relevant for provenance management in biomedicine.

2 Methods

First, a set of existing biomedical ontologies suitable for modeling provenance information are selected. Next, we define the comparison framework by identifying the characteristics that will facilitate provenance management in biomedicine. The selected ontologies are analyzed with respect to the set of desired attributes and the results are represented in an attribute versus provenance ontology matrix.

3 Candidate Ontologies

The National Center for Biomedical Ontologies (NCBO) was the primary source of candidate ontologies, where the "Experimental Conditions" category was used to identify relevant ontologies. We also used our knowledge of other ontologies modeling biomedical provenance to identify additional ontologies. The ten selected ontologies are briefly discussed below:

1. **ProPreO Ontology**. Ontology for modeling the proteomics analysis pipeline as part of the biomedical glycoproteomics project at the University of Georgia.
2. **Ontology for Biomedical Investigations (OBI)**. One of the largest and most comprehensive provenance ontologies covering more than 18 communities, including proteomics, transcriptomics, imaging, and toxigenomics.

3. **Experiment Factor Ontology (EFO).** Ontology developed by the European Bioinformatics Institute (EBI) to model the experimental factors associated with the ArrayExpress database of gene expression and related microarray datasets.
4. **Experiment Conditions Ontology (XCO).** This ontology is one of the three ontologies created for phenotype measurement data.
5. **Biological Imaging Methods (FBbi)**. The ontology models information about the sample preparation methods, imaging process, and visualization techniques used in biomedical imaging that influence the quality and subsequent interpretation of the images.
6. **Parasite Experiment Ontology (PEO).** PEO extends the Provenir top domain ontology for provenance (which in turn uses some Basic Formal Ontology (BFO) classes) to model provenance information of bench biological processes used in human pathogen research.
7. **Ontology for Clinical Research (OCRe).** The OCRe ontology models the provenance associated with human studies, both interventional and observational, which span the design phase, study execution phase, and analysis phases.
8. **Cardiac Electrophysiology Ontology (EP)**: The Cardio Vascular Research Grid (CVRG) has developed the Cardiac Electrophysiology Ontology to represent metadata describing the experimental conditions for cardio vascular research.
9. **Neural ElectroMagnetic Ontology (NEMO).** The ontology aims to represent the provenance information associated with Electro-encephalography (EEG) and Magnetoencephalography (MEG) data to facilitate collection, sharing, and mining of brain electromagnetic data.
10. **SWAN Provenance Authoring and Versioning (PAV) Ontology**. The ontology was developed as part of the Semantic Web Applications in Neuromedicine (SWAN) project and represents the derivation, authoring, and versioning information of biological resources.

4 Desirable Features

We identify seven desirable features for provenance ontologies in biomedicine based on both the requirements identified by the W3C Provenance Incubator Group [5] and the ten OBO foundry principles. The W3C Provenance XG identified a number of requirements for provenance along three dimensions, namely (a) content, (b) usage, and (c) management [5]. In the context of provenance ontologies in biomedicine, we believe provenance ***interoperability***, ***accessibility***, ***entailment***, ***versioning***, and ***understanding*** for end users (to enable use of provenance in applications) are the essential desired attributes [5].

1. **Open source without intellectual property restrictions.** Both the W3C Provenance XG accessibility dimension and the OBO foundry principle#1 recommend that the provenance ontology should be freely available without usage restriction or subject to payment of fee.
2. **Facilitating provenance interoperability** by extending upper level ontologies for creation of domain-specific provenance ontologies.
3. **Well-defined representation format**. Corresponding to the Provenance XG requirement for supporting entailment and OBO foundry principle #2, the provenance ontologies need to be available in a standard representation format to support entailments by available reasoning tools.
4. **Usability in real world applications**. This requirement reflects the "understanding" category of the provenance dimensions defined by the Provenance XG, which facilitates the use of provenance in end-user applications.
5. **Continued development and maintenance**. An important challenge for the biomedical ontology community is ensuring the continued development and maintenance of ontologies, as reflected in the OBO foundry principle#4. Hence, provenance ontologies should continue to

be developed and modified as the requirements of provenance users evolves.

6. **Re-use of existing ontologies**. Corresponding to the OBO foundry principle#5, provenance ontologies in biomedicine need to re-use terms from the large number of biomedical ontologies already created by the community.
7. **Explicit support for versioning**. Versioning information is an important aspect of provenance management as identified by the Provenance XG, hence provenance ontologies themselves need to include explicit support for versioning information.

A framework for evaluating provenance ontologies in biomedicine

Similar to the framework for comparing disease ontologies [4], the seven desired characteristics are not given equal weights. We identify some of the attributes reflecting the requirements of the Provenance XG and some OBO foundry principles to have higher importance as compared to others. We give a maximum weight of 5 to attributes numbered (2), (3), (4), and (7); followed by the weight of 3 to attribute (1); and finally weight of 1 to attributes (5) and (6). The ten provenance ontologies reviewed in the paper are assigned a score of 1 (for full support to a given desired attribute), 0 (for no support for the desired attribute), and a discrete value between 0 and 1 depending on the level of support for the desired attribute.

5 Results

Table 1 represents the results of our evaluation. The findings demonstrate that many existing biomedical provenance ontologies, EFO, PEO, and OCRe, fully support the desired properties identified in the evaluation framework. This is an encouraging trend for the biomedical provenance community and needs to be incorporated in other ontologies, such as the SWAN PAV and XCO, which scored less than 50%.

6 Discussion

The primary areas of concern for provenance ontologies are the support for interoperability, which is either partially (for ProPreO) or not supported at all (EP, PAV, XCO, and FBbi). The use of upper-level ontologies, such as BFO or the Provenir top domain provenance ontology [1] are needed to support consistent modeling, use of ontology design patterns and best practices. The re-use of existing ontologies in the creation of new ontologies has been a focus of continued concern for the OBO Foundry. But, five provenance ontologies are found to have no support for re-use of existing ontologies (ProPreO, XCO, FBbi, NEMO, and SWAN PAV). Hence, it is essential for the provenance ontologies community to ensure maximum re-use of existing ontology terms in development of new ontologies.

	Wt.	ProPreO	OBI	EFO	XCO	FBbi	PEO	OCRe	EP	NEMO	PAV
Open source	*3*	1	1	1	1	1	1	1	1	1	1
Inter-operability	*5*	0.5	1	1	0	0	1	1	0	1	0
Standard format	*5*	1	1	1	1	1	1	1	1	1	1
Understanding	*5*	1	0.5	1	0	0	1	1	0	1	0.5
Continued development	*1*	0	1	1	1	1	1	1	0	1	0
Re-use ontologies	*1*	0	1	1	0	0	1	1	1	0	0
Versioning	*5*	1	1	1	0	1	1	1	1	1	0
Total Score	***25***	**82%**	**90%**	**100%**	**36%**	**56%**	**100%**	**100%**	**56%**	**96%**	**34%**

Table 1. The desiderata applied to provenance ontologies in biomedicine

ProPreO	Proteomics data and process provenance, http://bioportal.bioontology.org/ontologies/13386
OBI	Ontology for Biomedical Investigations, http://bioportal.bioontology.org/ontologies/44899
EFO	Experimental Factor Ontology, http://bioportal.bioontology.org/ontologies/39885
XCO	Experimental Conditions Ontology, http://bioportal.bioontology.org/ontologies/45362
FBbi	Biological imaging methods, http://bioportal.bioontology.org/ontologies/45253
PEO	Parasite Experiment Ontology, http://bioportal.bioontology.org/ontologies/42093
OCRe	Ontology for Clinical Research, http://bioportal.bioontology.org/ontologies/44778
EP	Cardiac Electrophysiology Ontology, http://bioportal.bioontology.org/ontologies/39038
NEMO	Neural ElectroMagnetic Ontologies, http://bioportal.bioontology.org/ontologies/45141
PAV	Provenance, Authoring and Versioning Ontology, http://swan.mindinformatics.org/ontologies/1.2/pav.owl

Table 2. List of biomedical provenance ontologies reviewed in this work

7 Conclusions

We use the W3C Provenance Incubator Group recommendations and OBO foundry principles to define a framework of desired attributes for biomedical provenance ontologies. We use the framework to evaluate ten ontologies and find that a majority of ontologies are compliant.

References

1. S. S. Sahoo, "Semantic Provenance: Modeling, Querying, and Application in Scientific Discovery," Ph.D., Computer Science and Engineering Department, Wright State University, 2010.
2. C. Goble, "Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics," in *Workshop on Data Derivation and Provenance*, Chicago, 2002.
3. J. J. Cimino, "Desiderata for controlled medical vo- cabularies in the twenty-first century.," *Methods Inf Med,* vol. 37, pp. 394-403, 1998.
4. Bodenreider O and Burgun, "Desiderata for an ontology of diseases for the annotation of biological datasets.," in *First International Conference on Biomedical Ontology (ICBO 2009)*, NY, USA, 2009, pp. 39-42.
5. *W3C Provenance Incubator Group Wiki*. Available: http://www.w3.org/2005/Incubator/prov/wiki/Main_Page