Where did you hear that? Information and the Sources They Come From

James P. McCusker[1], Timothy Lebo[1], Li Ding[1], Cynthia Chang[1], Paulo Pinheiro da Silva[2], and Deborah L. McGuinness[1]

[1]Tetherless World Constellation
Rensselaer Polytechnic Institute
110 8th St., Troy, NY 12180, USA
[2]CyberShARE Center, University of Texas at El Paso
500 W University Ave, El Paso TX 79968, USA
{mccusj,lebot}@rpi.edu,{dingl,csc}@cs.rpi.edu,paulo@utep.edu,dlm@rpi.edu
http://tw.rpi.edu

Abstract. One current challenge in linked science is to adequately describe where a piece of information in the linked science cloud came from. Provenance models, such as Proof Markup Language (PML), have developed methods for expressing simple relationships between information and the sources of information. We argue that the representation of where information comes from is central to trusting linked data in scientific applications. We introduce the notion of a model of information source and the usage of the source to obtain information by describing the Proof Markup Languages notion of source usage and show how this relationship can be modeled in a library science schema, Functional Requirements for Bibliographic Resources (FRBR). We discuss how these kinds of representations are critical to provenance models.

1 Introduction

Before publishing a result, scientists need to check their facts. We stand on the shoulders of giants, but as we push forward in science, we need to make sure that we aren't standing on a giant house of cards. Knowing how, when and where your data comes from is critical for good science, and it's even more critical for linked science, where it isn't immediately clear where a database record or knowledge assertion came from. Sources of information become critical to evaluate information quality. It is difficult if not impossible to assess the trust of information, or to encode it as knowledge, without having a link between information and their sources. For example, one may want to know if the information came from a source such as the New York Times, and further, it may be useful to know the date, edition, page, and exact text fragment where the information was asserted.

There are many challenges in the task of assigning a source to a piece of information. First, it may not be easy to characterize the piece of information in larger information containers (databases, printed documents, web documents, documents that require parameters from a system to be retrieved, etc). Second,

the source of a piece of information is often a source of other pieces of information and should be referenced by an identifier and characterized elsewhere. Third, the assertion of the piece of information is a point-time event that occurs during the life-time of the information source. Thus, not any assertion event is a valid assertion: it needs to occur during the lifespan of its source(s) and in places where the sources are located. Those are all critical conditions that need to be properly captured in provenance languages if one wants to make proper use of source information in combination with linked data.

2 Current Implementation: Proof Markup Language

The Proof Markup Language (PML) [1][2] evolved to include language constructs to handle use cases such as those encountered in many text analytic settings [3], where components, such as entity extractors, review natural language text and infer structured assertions from the text. In order to maintain provenance, it needed to capture the source that was accessed, in this case by the text analytic component, and the information that was obtained or inferred. Further, in many cases it was important to be able to encode the particular fragment of the source that was used. The notion of using a source to obtain information is captured in PMLs SourceUsage class, which serves to record the event of information assertion that can also be a general case for the event of information retrieval and information extraction. Sources can be as fine-grained as particular regions of text or data files, text fragments, or as broad as entire online databases. The raw data that was received from the Source is attached to Information using the property hasRawString. Using this, it is possible to determine if two pieces of Information came from the same Source, and if the Information has been derived from the same data fragment.

An example of this representation can be seen in Figure 1[1]. The top level concept of a NodeSet supports the encoding of support for a particular piece of information that can be viewed as a conclusion of some inference step. That inference step can be as simple as a told assertion or could be an inference using some antecdents and resulting in a conclusion. One type of inference includes the usage of a document source to obtain a piece of information. Figure 1 shows a particular usage of a document (the PML primer) and includes an encoding of the time it was used and the fragment of the text used. The inference steps use of a SourceUsage identifies the following: the date, time, location and source of the assertion. The Source is a pmlp:Source concept that can be used to represent things like publications, documents, websites, datasets, person, organizations, etc. Additional examples are available in Murdock *et al.* [4] and Welty, *et al.*, [3].

The PML classes for Source, SourceUsage, and Information are empirically derived, that is, they are responses to a set of use cases that required tracking where information came from and how it was used. This model was effective

[1] The RDF can be downloaded from http://inference-web.org/proofs/csctest/iwp-pml-2.rdf

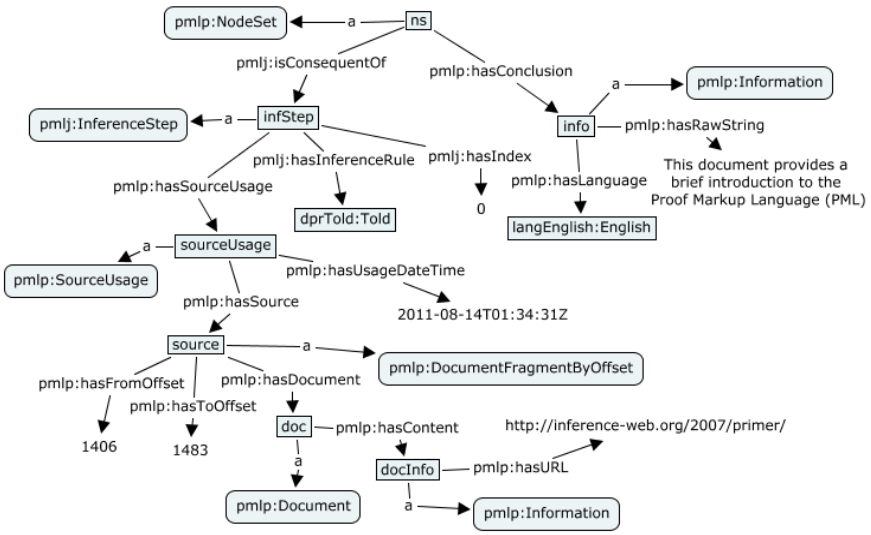

Fig. 1. A description of downloading the PML Primer from a web site and selecting a quote from it using PML.

at capturing the text analytic requirements from the Unstructured Information Management Architecture components, however a more general representation may be beneficial to support additional use cases such as copying files, transforming data from one format to another, and so on. Files on disk are considered to be pmlp:Sources, which limit the ability to describe mechanical duplication, as derivational provenance in PML is limited to pmlp:Information. Similarly, transformation of data from one file format to another results in, in one perspective, two pmlp:Information instances that have the same information (they have the same information), but in one perspective have completely different file content, since the information is being represented using a different file encoding. Generalization of these sorts of relationships can allow faithful representation of these operations and allow for extension and decomposition of concepts like "the source of a piece of information".

Library Science has spent significant time dealing with some kinds of provenance in the realm of bibliographic resources. Functional Requirements for Bibliographic References (FRBR) [5] is designed to address issues of abstraction in bibliographic resources. For instance, when we mention The Art of Computer Programming, we could be referring to the work as a whole, a particular edition, a particular rendering of that edition (electronic versus paper, for instance), or a particular copy. FRBR separates these different levels into, respectively, Work, Expression, Manifestation, and Item. Electronic information resources can be similarly distinguished by using Item to refer to a particular copy, Manifestation

to refer to a specific bit image, Expression to refer to a fixed set of information, and a Work to refer to all versions of that information. Here we refer to sets of Work, Expression, Manifestation, Item that are interlinked as a FRBR stack, in that it is a complete stack of instances representing a particular piece of information at all abstractive levels.

3 Mapping Source and Information into FRBR

Describing how information is derived from a source is a little more complicated in FRBR, but is more generalizable and allows for greater shades of distinction. We propose that, since pmlp:Source is considered to be an opaque, specific resource, it should be a subclass of frbr:Item. pmlp:Information is actually a role of a frbr:Work, frbr:Expression, or frbr:Manifestation. For instance, an expression may not be information but may play the role of being information in the context of an assertion. In Figure 2 two FRBR stacks show how a quote "This document provides a brief introduction to the Proof Markup Language (PML)" from the PML Primer[2] is derived from a downloaded copy of the primer. In the representation, we use the abstractive perspective to allow for description of physical movement of data and transformation of information using the same derivational ontology. Conversely, it allows description of what happened encoded directly in the relationships. For instance, the fact that the copyEvent produced an identical copy is stated by the fact that the server copy and client copy are exemplars of the same Manifestation, while the subset event produces a quote simply because that stack is declared to be partOf the PML Primer stack. The derivational ontology is not named, but PML is an adequate candidate for this task. Its information source construction can be replaced by FRBR.

4 Conclusion

We believe that any serious model of provenance that supports linked science must provide a mechanism for describing information sources and their usage. This can be and has been achieved using the modeling primitives provided in PML. By using the mapping we describe using FRBR, we can also model additional nuanced explanations of data access, transformation, and analysis. Generalized models of abstractive provenance also provide opportunities to express nuanced explanations of data access, transformation, and analysis. We show how the link between information and source can be modeled using a combination of FRBR and a derivational provenance model. This combination is powerful, and allows for unambiguous descriptions of data and information access and transformation. Finally, we argue that the abstractive dimension should be a key component of any provenance model that attempts to deal with artifacts of information or data.

[2] http://inference-web.org/2007/primer/

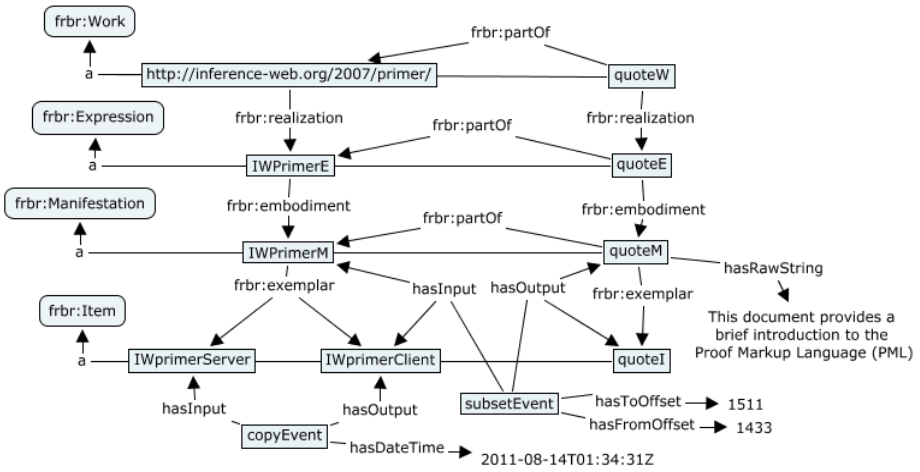

Fig. 2. A description of downloading the PML Primer from a web site and selecting a quote from it using FRBR and derivational events. Each of these levels represents a different aspect of the primer and the quote from it.

5 Acknowledgements

The Tetherless World Constellation is partially funded by DARPA, U.S. Department of Energy, Fujitsu, LGS, Lockheed Martin, Microsoft Research, NASA, National Ecological Observatory Network (NEON), the National Science Foundation, Qualcomm, and the Woods Hole Oceonographic Institution (WHOI). This research was partially funded by the National Science Foundation under CREST Grant No. HRD-0734825.

References

1. McGuinness, D.L., Ding, L., Pinheiro Da Silva, P., Chang, C.: PML 2: A modular explanation interlingua. In: Proceedings of AAAI. Volume 7. (2007)
2. Pinheiro da Silva, P., McGuinness, D.L., Fikes, R.: A Proof Markup Language for Semantic Web Services. Information Systems **31**(4-5) (2006) 381395
3. Welty, C., Murdock, J.W., Silva, P.P.D., McGuinness, D.L., Ferrucci, D., Fikes, R.: Tracking information extraction from intelligence documents. In: Proceedings of the 2005 International Conference on Intelligence Analysis (IA 2005). (2005)
4. Murdock, J., McGuinness, D.L., Pinheiro da Silva, P., Welty, C., Ferrucci, D.: Explaining conclusions from diverse knowledge sources. The Semantic Web-ISWC 2006 (2006) 861–872
5. O'Neill, E.: FRBR: Functional Requirements for Bibliographic Records. Library resources & technical services **46**(4) (2002) 150–159
6. Wilkinson, M., Vandervalk, B., McCarthy, L.: SADI Semantic Web Services-cause you can't always GET what you want! In: Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific, IEEE (2009) 13–18