Proceedings of the First International Workshop on the role of Semantic Web
in Provenance Management (SWPM 2009) 2009:(electronic proceedings: http://ceur-ws.org/Vol-526/).

Provenance information in biomedical knowledge repositories – A use case

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
US National Library of Medicine
Bethesda, Maryland, USA
olivier@nlm.nih.gov

***Abstract*—We present a use case for provenance information in biomedical knowledge repositories designed to support applications including information retrieval and knowledge discovery. We show that information about the knowledge sources from which statements are extracted must be recorded in addition to the statement themselves in order to support these applications. While the storage and processing of statements has been greatly facilitated by the emergence of powerful triple stores and the standardization of query languages (e.g., SPARQL), recording and exploiting provenance information (i.e., statements about statements) remains challenging.**

Keywords-provenance information; use case; biomedical knowledge repository

I. INTRODUCTION

Biomedical knowledge is produced and consumed by biomedical researchers and health care practitioners. The biomedical literature (textbooks and journal articles) represents the main source of unstructured biomedical knowledge. Knowledge bases (e.g., model organism databases annotated to the Gene Ontology) result from the curation of the primary literature in an attempt to make knowledge more accessible and actionable. Finally, ontologies represent the ultimate form of computationable knowledge, but are often limited in scope and tend to focus on definitional, as opposed to assertional knowledge. Rich sets of metadata have been defined and are collected along with the primary data, using standards such as the Dublin Core for the literature [1] and MIAME for gene expression data [2].

Attempts to make knowledge accessible to agents in addition to humans have focused on the extraction of knowledge from unstructured sources, as well as the interoperability of structured knowledge sources. Text mining techniques are used to extract "predications" (i.e., statements) from text, for example in the Semantic Medline project [3]. Metadata are often stored in an *ad hoc* format in order to help associate predications with the articles from which they have been extracted. The Linked Data initiative [4] promotes the use of RDF (the Resource Description Framework) [5] to link biomedical datasets, with a strong emphasis on shared URIs (Uniform Resource Identifiers) in order to relate concepts sharing the same identifiers across datasets. In most cases, however, such repositories of linked data have little metadata, in part because simple RDF representations make it difficult to represent statements about statements. These examples illustrate the difficulty of representing – let along computing with – provenance information in biomedical knowledge repositories.

One such repository is being created as part of a research project at the National Library of Medicine [6]. It includes knowledge extracted from Medline abstracts by text mining tools, structured knowledge derived from existing knowledge bases (e.g., NCBI's Entrez system [7]) and terminological knowledge from the Unified Medical Language System [8]. In this project, we are also interested in recording and processing information about the statements (e.g., location in the information space and time annotations), in order to support applications including enhanced information retrieval, multi-document summarization, question answering and knowledge discovery.

In this paper, we briefly examine the types of metadata required in the context of our biomedical knowledge repository. In other words, we look at provenance information through the use case of this knowledge repository and discuss some of the issues encountered along the way and challenges ahead.

II. PROVENANCE INFORMATION IN TYPICAL APPLICATIONS

The four applications our repository has been designed to support require various types of provenance information [6]. Common to all applications is the requirement that the origin of any statement be identifiable (e.g., from which knowledge sources was it extracted?, using which extraction techniques, if any?) Because biomedical knowledge evolves over time, it is also indispensable that some time annotation be associated with each statement (e.g., date of publication of the article from which the statement was extracted, date when the statement was curated in a given knowledge base, or date when a given ontology was last revised.) When available, the degree of confidence associated with a given statement should also be recorded. Confidence can be indicated by the tools used for the production of the statements (e.g., text mining tools) or approximated through frequency information. In the following discussion, the association between types of applications and types of provenance information is somewhat arbitrary and presented essentially for illustrative purposes.

A. *Information retrieval*

The enhanced information retrieval envisioned goes beyond keyword or concept searches and supports searches based on relations. For example, finding all the documents in which the statement "IL-13 inhibits COX-2" is found. Like with traditional search engines, there is a need for associating a document identifier with a given statement. The list of all document identifiers associated with a given statement forms the basic index in such a system. Conversely, indexing a document consists in associating this document with all the statements extracted from it by the text mining tool.

B. *Multi-document summarization*

In addition to the basic index required for information retrieval, information is needed for the prioritization of statements (among all relevant statements) in multi-document summarization. Statements below a certain threshold of confidence may be hidden as a way of restricting the amount of information provided in the summary. Low confidence can be indicated by a text mining tool, for example, when ambiguity in natural language cannot be resolved by the system.

C. *Question answering*

In question answering applications, answers must be collected from reputable sources. Here, statements from the biomedical knowledge repository are used as potential answers to input questions (e.g., what genes does IL-13 inhibit?) Not only must the origin of the statement be present as for information retrieval and summarization purposes, but additional metatada associated with the document must also be available (e.g., does this document come from a reputable source, such as an article about randomized clinical in the case of clinical effectiveness statements? Does this statement come from a document published/a knowledge base revised recently?) The distinction here is between metadata directly associated with the statement (e.g., document identifiers), and metadata about the documents themselves, indirectly associated with the statement (reputability of the source, publication date).

D. *Knowledge discovery*

Information retrieval, summarization and question answering can be thought of as exploiting a static repository, mostly through look-ups in the repository, with no (or limited) need for inference. In contrast, knowledge discovery processes aim at inferring new knowledge from patterns of statements in the repository. Inference is one major technique for deriving new knowledge from existing knowledge. Production rules provide a simple mechanism for formalizing inference and rule engines are implemented in many systems that store statements. Knowledge discovery systems require not only production rules and rule engines for the production of entailed statements from rules, but also the production of the metadata associated with the entailed statements (i.e., inferred provenance information). Provenance for both asserted and inferred statements is required so that the universe of statements can be restricted to degree of confidence, specific time periods or sources. For example, can a path be found in a graph, directly (asserted links) or indirectly (inferred links), between two nodes (e.g., between a disease and a drug), when links are restricted to a specific source? The issue here is not only to associate provenance information to asserted statements, but also to compute such information for inferred statements as well.

III. Issues and Challenges

Limitation of naïve implementations. RDF provides a simple mechanism for recording statements about statements through "blank nodes" [5]. A blank node can be used as an identifier for the statement, each component of which – subject, predicate and object – is then linked to it through predicates such as *hasSubject*, *hasPredicate* and *hasObject*. Similarly, provenance information can be linked to the statement identifier (e.g., link to the article from which it is extracted through a *hasSource* predicate). This mechanism, called reification, is inefficient as it increases the number of triples required for implementing a statement (at least one for the relation of the blank node to each of the three components of the original statement). Scalability issues are thus likely with large biomedical repositories (typically several hundred million asserted statements). Moreover, by significantly increasing the complexity of queries, reification also puts an unnecessary cognitive burden on the user.

Lack of support for provenance information in mainstream triple stores, There is currently no support for exploiting provenance information in off-the-shelf triple stores. Support is generally provided for named graphs in so-called "quad stores", but named graphs hardly provide the level of granularity needed for provenance information required in biomedical applications. Beside reification, SPARQL does not offer support for seamless processing of provenance information. There is a need for a standardization of emerging models of provenance (e.g., OPM [9]) and their efficient implementation in triple stores.

Limitations for the applications. The emerging paradigm of linked data and mashups had met tremendous enthusiasm in the biomedical community [10, 11]. At this early stage, the possibility of easily integrating disparate datasets still outweighs the lack of fine control over constraints on these data sources. However, when applications mature beyond answering questions such as "is there a path between this and that?" to restricting graph traversal with constraints specific to properties of the links (statements, not simply predicates), the lack of standard models and implementations for provenance information will appear as a serious limitation.

Acknowledgment

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

[1] "The Dublin Core® Metadata Initiative," http://www.dublincore.org/.

[2] A. Brazma, "Minimum Information About a Microarray Experiment (MIAME)--successes, failures, challenges," ScientificWorldJournal, vol. 9, 2009, pp. 420-423.

[3] M. Fiszman, et al., “Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation,” J Biomed Inform, vol. 42, no. 5, 2009, pp. 801-813.

[4] “Linked data,” http://linkeddata.org/.

[5] W3C, “Resource Description Framework (RDF),” http://www.w3.org/RDF/.

[6] O. Bodenreider and T.C. Rindflesch, Advanced library services: Developing a biomedical knowledge repository to support advanced information management applications, Technical report, Lister Hill National Center for Biomedical Communications, National Library of Medicine, 2006.

[7] E.W. Sayers, et al., “Database resources of the National Center for Biotechnology Information,” Nucleic Acids Res, vol. 37, no. Database issue, 2009, pp. D5-15.

[8] O. Bodenreider, “The Unified Medical Language System (UMLS): Integrating biomedical terminology,” Nucleic Acids Res, vol. 32 Database issue, 2004, pp. D267-270.

[9] L. Moreau, et al., “The Open Provenance Model: An Overview,” Provenance and Annotation of Data and Processes, Lecture Notes in Computer Science 5272, Springer, 2008, pp. 323-326.

[10] F. Belleau, et al., “Bio2RDF: towards a mashup to build bioinformatics knowledge systems,” J Biomed Inform, vol. 41, no. 5, 2008, pp. 706-716.

[11] K.H. Cheung, et al., “Semantic mashup of biomedical data,” J Biomed Inform, vol. 41, no. 5, 2008, pp. 683-686.