Using the Micropublications ontology and the Open Annotation Data Model to represent evidence within a drug-drug interaction knowledge base

Jodi Schneider[1], Paolo Ciccarese[2], Tim Clark[2], and Richard D. Boyce[3]

[1] INRIA Sophia Antipolis France
jodi.schneider@inria.fr
[2] Massachusetts General Hospital and Harvard Medical School
paolo.ciccarese@gmail.com; tim_clark@harvard.edu
[3] University of Pittsburgh
rdb20@pitt.edu

Abstract. Semantic web technologies can support the rapid and transparent validation of scientific claims by interconnecting the assumptions and evidence used to support or challenge assertions. One important application domain is medication safety, where more efficient acquisition, representation, and synthesis of evidence about potential drug-drug interactions is needed. Potential drug-drug interactions (PDDIs), defined as two or more drugs for which an interaction is known to be possible, are a significant source of preventable drug-related harm. The combination of poor quality evidence on PDDIs, and a general lack of PDDI knowledge by prescribers, results in many thousands of preventable medication errors each year. While many sources of PDDI evidence exist to help improve prescriber knowledge, they are not concordant in their coverage, accuracy, and agreement. The goal of this project is to research and develop core components of a new model that supports more efficient acquisition, representation, and synthesis of evidence about potential drug-drug interactions. Two Semantic Web models—the Micropublications Ontology and the Open Annotation Data Model—have great potential to provide linkages from PDDI assertions to their supporting evidence: statements in source documents that mention data, materials, and methods. In this paper, we describe the context and goals of our work, propose competency questions for a dynamic PDDI evidence base, outline our new knowledge representation model for PDDIs, and discuss the challenges and potential of our approach.

Keywords: Linked Data, drug-drug interactions, evidence bases, Micropublications, Open Annotation Data Model, knowledge bases

1 Introduction

Scientific knowledge depends on the verification and integration of large systems of interconnected assertions, assumptions, and evidence. These systems are con-

tinually growing and changing, as new scientific studies are completed and new documents are published. The state of current knowledge in any given domain can be difficult for any one individual to fully grasp, because bits of knowledge are updated at frequent intervals.

In the biosciences, this problem has taken on particular importance, due to an exponential growth in the aggregate publication rate. Manually curated databases are used to record certain types of knowledge. To update and maintain these databases, curators must make knowledge-intensive decisions, identifying the best available evidence in the current scientific literature. Maintaining such databases is challenging because there is limited tracking of the source information.

In an ongoing project, we are experimenting with using the Micropublications Ontology[4] [Clark2014] and the Open Annotation Data Model[5] [W3C2013] to create an audit trail between assertions, evidence, and source documents, so that assertions and evidence can be flagged for update in flexible and intelligent ways. Updates may be needed when the underlying sources change, when a particular method for establishing an assertion is discredited, etc. Our goal is to provide better linkages between an assertion recorded in a knowledge base and its supporting evidence (i.e., data, materials, and methods) found in source documents.

In the remainder of the paper, we describe the competency questions for our evidence base and the new evidence model that we are creating, which combines the Micropublication Ontology and the Open Annotation Data Model, and adapts them to the existing evidence modeling of the Drug Interaction Knowledge Base[6] [Boyce2007,Boyce2009]. We then reflect on how the new model performs for our goal of creating an audit trail between assertions, evidence, and source documents.

2 Context and goals

Our work is in the context of a larger project on organizing and synthesizing scientific evidence from the biomedical literature on potential drug-drug interactions. Potential drug-drug interactions (PDDIs), defined as two or more drugs for which an interaction is known to be possible, are a significant source of preventable drug-related harm (i.e., adverse drug events, or ADEs). The combination of poor quality evidence on PDDIs, and a general lack of PDDI knowledge by prescribers, results in many thousands of preventable medication errors each year. While many sources of PDDI evidence exist to help improve prescriber knowledge, they are not concordant in their coverage [Saverno2011], accuracy [Wang2010], and agreement [Abarca2003]. Difficulties with synthesizing evidence, and gaps in the scientific knowledge of PDDI clinical relevance, underlie such disagreement.

[4] `http://purl.org/mp/`
[5] `http://www.openannotation.org/spec/core/`
[6] `http://purl.net/net/drug-interaction-knowledge-base/`

To address these problems, our research group is studying the potential benefit of applying recent developments from the Semantic Web community on scientific discourse modeling and open annotation. The goal is to develop core components of a new PDDI knowledge representation model that will support a more efficient acquisition, representation, and synthesis of PDDI evidence. The desired knowledge representation will provide better linkages between PDDI assertions and their supporting evidence, by directly connecting to annotated section(s) of relevant source documents.

3 Approach

Our new approach will draw upon the current version (1.2) of the Drug Interaction Knowledge Base [Boyce2007,Boyce2009], the Open Annotation Data Model [W3C2013], and the Micropublications Ontology [Clark2014].

The Drug Interaction Knowledge Base (DIKB) is a static, manually constructed evidence base that indexes assertions and evidence of PDDI for over 60 drugs. Its taxonomy of assertion types and evidence types [Boyce2014] is a starting point for the new knowledge base. The current version of the DIKB implements a version of the SWAN semantic discourse ontology [Ciccarese2008] to represent evidence relations. Specifically, the knowledge base uses *swanco:citesAsSupportingEvidence* and *swanco:citesAsRefutingEvidence* to link to an entire source document as a supporting or refuting citation. At the time the DIKB 1.2 was constructed (2007–2009), annotation methodologies were less well developed. Consequently, version 1.2 of the DIKB stores quotes as textual strings manually copied from source documents. The text has been enriched with metadata about the source section, but it is non-trivial to return to the appropriate segment of the text from this information.

Our use of the Open Annotation Data Model (OA) reflects a change in the state of the art. OA is an "an interoperable framework for creating associations between related resources, annotations, using a methodology that conforms to the Architecture of the World Wide Web"[7]. In particular, OA allows an evidence database to provide explicit connections from quotes to their source documents. For example, as shown in Figure 1, an OA resource can be used to quote a specific part of a drug product label (also known as a summary of product characteristics) to indicate evidence that *escitalopram inhibits CYP2D6*. In general, OA enables queryable links between selections from source documents (as target) to the instances of data, methods, and materials (as body) that we want to model to support drug interaction knowledge base use cases.

Similarly, the Micropublications Ontology improves the depth with which evidence can be represented and queried. The most important feature of the Micropublications model, in our view, is its ability to represent the data, methods, and materials that act as support for a claim, and to transitively close chains

[7] `http://www.openannotation.org/spec/core/`

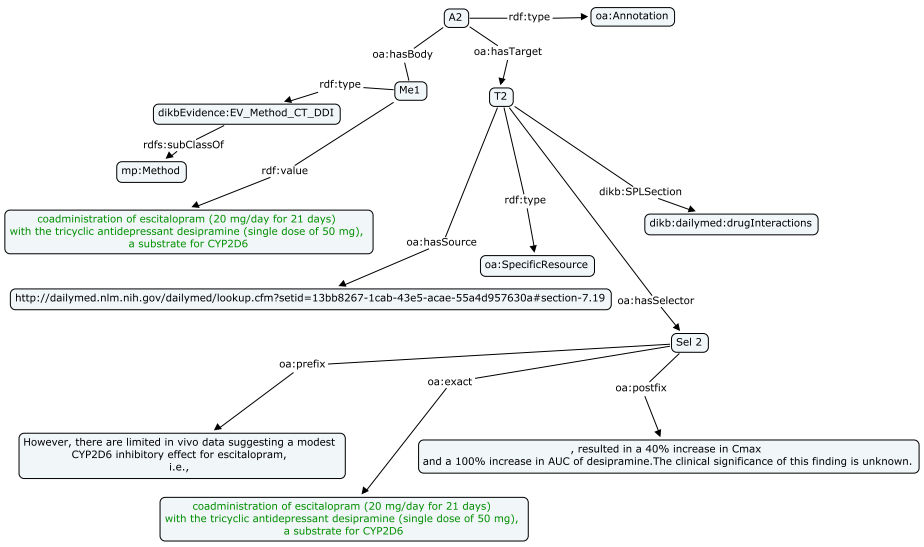

Fig. 1. The Open Annotation ontology (*oa*) can be used to quote the evidence (here a method described in a DailyMed product label) and associate it with an instance of the Micropublication ontology (*mp*). The annotation records quoted text (*oa:exact*) as the target, while the body of the annotation is a *mp:Method* instance, *Me1*, supporting the Claim *Escitalopram inhibits CYP2D6* shown in Figure 2. We use existing terminology from the DIKB ontology to specify the section of the DailyMed product label and to indicate the *dikbEvidence* type.

of claims[8] and citations across the literature to their fundamental supporting evidence. A *mp:Micropublication* *mp:argues* a *mp:Claim* based on connecting any number of *mp:Representation*s. The whole Micropublication is a Representation, as are Data and Methods (including Materials and Procedures), whether textual or pictoral. A *mp:Representation* may *mp:support* or *mp:challenge* any other *mp:Representation*, making the evidence explicit and queryable.

4 Competency Questions

To design an appropriate enhancement of the DIKB model with Micropublications and the Annotation Ontology, we need to understand what sorts of questions experts would like to retrieve about the PDDIs. The competency questions below were elicited from experienced editors of clinically oriented drug compendia during the process of developing DIKB 1.2. Most fall into three categories: finding assertions and evidence; assessing the evidence; and enabling updates. A second area of interest is statistical information about the evidence base which is useful for various analytics related to knowledge base maintainance.

4.1 Finding assertions and evidence

1. **Finding assertions:**
   (a) List all assertions that are not supported by evidence
   (b) Which assertions are supported (or refuted) by just one type of evidence?
   (c) Which assertions have evidence from source X (e.g., product labeling)
   (d) Which assertions have both evidence for and evidence against from a single source X?
2. **Finding evidence:**
   (a) List all evidence for or against assertion X (by evidence type, drug, drug pair, transporter, metabolic enzyme, etc.)
   (b) What is the in vitro evidence for assertion X? the in vivo evidence?
   (c) List all evidence that has been flagged as rejected from entry into the the knowledge base
   (d) Which single evidence items act as support or rebuttal for multiple assertions of type X (e.g., $substrate\_of$ assertions)?

4.2 Assessing the evidence:

1. **Understanding evidence coming from a given study:**
   (a) What data, methods, materials, are reported in evidence item X?
   (b) Which evidence items are related to and follow-up on evidence item X?
   (c) Which research group conducted the study used for evidence item X?
   (d) Are the evidence use assumptions for evidence item X concordant? unique? non-ambiguous?

[8] 'Assertion' in DIKB terminology corresponds to a 'Claim' in the Micropublications model; this variation in terms is because the term 'claim' is used in a different sense in medical billing.

2. **Verifying plausibility of an evidence item:**
   (a) Has evidence item X been rejected for assertion Y? If so, why and by whom?
   (b) Which other assertions are being supported/challenged by this evidence item?
   (c) What are the assumptions required for use of this evidence item to support/refute assertion X?
3. **Checking assertions about pharmacokinetic parameters (i.e., area under the concentration time curve (AUC))**
   (a) How many pharmacokinetic studies used for evidence items in the DIKB could be used to support or refute an assertion about pharmacokinetic paramater X (e.g., 'X increases AUC')?
   (b) How many pharmacokinetic studies in the DIKB used for evidence items for assertion X are based on data from the product label?
   (c) What is the result of averaging (or applying some other statistical operation) to the values for pharmacokinetic parameter X across all relevant studies used for evidence items?
4. **Checking for differences in the product labeling:**
   (a) Are there differences in the evidence items that were identified across different versions of product labeling for the same drug?
   (b) What version of product labeling was used for evidence item X? Original manufacturer or repackager? Most current label or outdated? Is the drug on market in country X or not? American or country X?

4.3 Supporting updates to evidence and assertions

1. **Changing status of redundant and refuted evidence:**
   (a) Remove a older version of a redundant evidence item
   (b) Change the modality of a supporting evidence item to be a refuting evidence item
2. **Updating when key sources change:**
   (a) Get all assertions that are supported by evidence items identified from an FDA guidance or other source document just released as an updated version.

4.4 Understanding the evidence base

1. **Statistical information about the evidence base:**
   (a) Number of assertions in the system
   (b) Number of evidence items for and against each assertion type
   (c) Show the distribution of the levels of evidence for various assertion types (e.g., pharmacokinetic assertions)

5 Modeling evidence about drug-drug interactions

Figure 2 shows how the new Micropublications model of evidence on PDDIs would represent some of the evidence supporting and challenging the assertion *escitalopram does not inhibit CYP2D6*. We created the example by hand using a sample assertion and evidence items from the DIKB version 1.2[9].

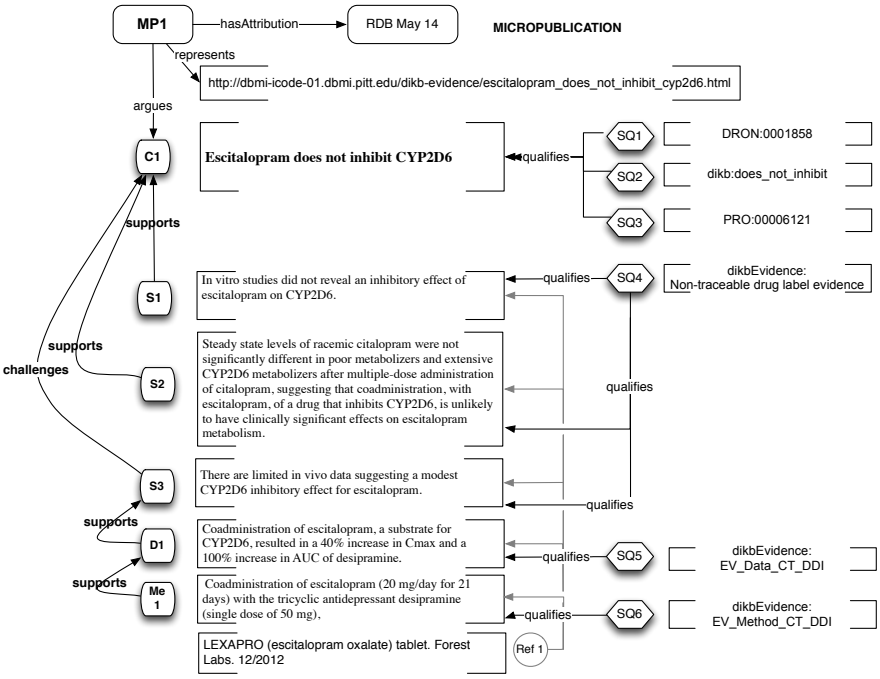

Fig. 2. A model of the evidence for and against the assertion *escitalopram does not inhibit CYP2D6*. This is based on the Micropublications ontology, and reuses the evidence taxonomy (dikbEvidence), terms (dikb), and data from the DIKB. The Drug Ontology (DRON) and Protein Ontology (PRO) are reused in semantic qualifiers. A more detailed view of Method *Me1* is shown in Figure 1.

The Micropublications ontology is used to structure the evidence relating to data, methods, and materials, and the overall indication that evidence *mp:supports* or *mp:challenges* a *mp:Claim*. We qualify Claims (C1 in the figure) by reusing identifiers from DRON[10] [Hanna2013] and the Protein Ontology[11] [Natale2011].

[9] `http://dbmi-icode-01.dbmi.pitt.edu/dikb-evidence/escitalopram_does_not_inhibit_cyp2d6.html`

[10] `http://purl.obolibrary.org/obo/dron.owl`

[11] `http://pir.georgetown.edu/pro/`

The new model reuses the DIKB evidence taxonomy[12] to provide epistemic qualification (SQ2, SQ5, SQ6 in the figure) to statements (S1, S2, and S3 in the figure), data (D1 in the figure), methods (Me1 in the figure), and materials (not shown in this example). The Open Annotation Data Model (previously shown in Figure 1) is used to link quotes taken from source documents back to their originating information artifacts. The approach to modeling other DIKB assertions would be similar to this example.

6 Discussion

6.1 Expected Benefits

Certain benefits accrue from upgrading from the current DIKB. Many of the competency questions (Section 4) are not supported in the DIKB 1.2. The new model is designed to support these and additional questions relevant in the domain. Visual inspection of the model suggests that we will be able to answer some competency questions quite naturally. In particular, finding the assertions that are not supported by evidence already in the evidence base, the evidence that should be checked most thoroughly (e.g. evidence that by itself supports multiple assertions), and the data, methods, and materials associated with a given evidence item as described in source documents.

Further, as a Linked Data resource, our new knowledge base will also enable innovative queries using knowledge from other sources about tagged entities (i.e., drugs and proteins) represented in the evidence base. Unlike the current DIKB, we will be able to render annotations in their original context. We also expect to be able to support distributed community annotation/curation, since MP and OA take account of provenance, and since OA is being increasingly adopted by a variety of annotation tools.

6.2 Modeling challenges

Our project does raise certain modeling challenges. To date, MP has not been used to represent both unstructured claims and the related logical sentences. Figure 1 shows the assertion *escitalopram does not inhibit CYP2D6* as unstructured text. However, the DIKB requires that 1) assertions about PDDIs be formulated by experts prior to collecting evidence, and 2) that the assertions be represented both as unstructured statements and sentences in a logical formalism. Careful thought is being put into how to properly accommodate this use case. Such challenges are to be expected since MP is a relatively new ontology and since this is a new application of it.

Another challenge is to ensure that, as the evidence base scales, competency questions can be answered efficiently. To address this, we building the model using an iterative design-and-test approach. In this process, efficient querying is a key requirement.

[12] `http://bioportal.bioontology.org/ontologies/DIKB`

6.3 Other issues

For enabling synthesis over the PDDI information, the model is not the only concern. Applying this model will require integration work. One challenge is inherent to scholarly documents: the existing evidence items within the DIKB refer to many data, materials, and methods that exist only in PDF documents accessible only through proprietary portals or academic library systems. Consequently, resolving annotations requires a method for pointing to proprietary *oa:target*s.

7 Conclusions & Future Work

We are currently iterating and refining the PDDI evidence and annotation model. Once it is stable, we plan to use the new model to represent as Linked Data evidence collected by an evidence board consisting of drug experts. The evidence collection effort is planned as part of a research project funded by the National Library of Medicine ("Addressing gaps in clinically useful evidence on drug-drug interactions", 1R01LM011838-01) and will focus on PDDI assertions for a number of commonly prescribed drugs (anticoagulants, statins, and psychotropics). We plan to implement a pipeline for extracting PDDI mentions from a variety of publicly available sources, including published journal articles indexed in PubMed or PubMed Central, FDA Guidance Documents, and drug product labels from the National Library of Medicine's DailyMed website[13]. Candidate PDDI assertions will be linked by machine to the Internet-accessible versions of the information artifacts used as evidence.

An existing Micropublication plugin for Domeo [Ciccarese2014] is being modified as part of the project. Our plan is to use the revised plugin to support the evidence board with the collection of the evidence and associated annotation data. It will also enable the broader community to access and view annotations of PDDIs highlighted in a web-based interface. We anticipate that this approach will enable a broader community of experts to review each PDDI recorded in the DIKB and examine the underlying research study to confirm its appropriateness and relevance to the evidence base.

The usability of the annotation plug-in is critically important so that the panel of domain experts will not face barriers to annotating and entering evidence. This will require usability studies of the new PDDI Micropublication plugin. Another issue is that many PDDI evidence items can be found only in PDF documents. Currently, the tool chain for PDF annotation is relatively weak: compared to text and HTML, PDF annotation tools are not as widely available and not as familiar to end-users. Suitable tools will have to be integrated into the revised plugin.

Knowledge representations combining MP and OA have the potential to allow more granular and reusable representation of evidence (data, materials, and methods), which are needed for synthesizing contested knowledge at the state

[13] `http://dailymed.nlm.nih.gov/dailymed/about.cfm`

of the art from scientific documents. The knowledge representations we are now creating will be beneficial for integrating PDDI evidence, and we hope they will inspire an increased use of linked data for evidence synthesis in other domains.

Acknowledgments

This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 246016, and a grant from the National Library of Medicine (1R01LM011838-01). We thank Carol Collins, Lisa Hines, and John R Horn for serving on the Evidence Panel of "Addressing PDDI Evidence Gaps", and for contributing to the competency questions presented here.

References

[Abarca2003] Abarca, Jacob, Daniel C. Malone, Edward P. Armstrong, Amy J. Grizzle, Philip D. Hansten, Robin C. Van Bergen, and Richard B. Lipton. "Concordance of severity ratings provided in four drug interaction compendia." *Journal of the American Pharmacists Association* 44;2 (2003): 136–141.

[Boyce2014] Boyce, R.D. "A Draft Evidence Taxonomy and Inclusion Criteria for the Drug Interaction Knowledge Base." August 9, 2014, url: http://purl.net/net/drug-interaction-knowledge-base/evidence-types-and-inclusion-criteria

[Boyce2007] Boyce, Richard D., Carol Collins, John Horn, and Ira Kalet. "Modeling Drug Mechanism Knowledge Using Evidence and Truth Maintenance." *IEEE Transactions on Information Technology in Biomedicine* 11;4 (2007): 386–397.

[Boyce2009] Boyce, Richard D., Carol Collins, John Horn, and Ira Kalet. "Computing with evidence: Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment." *Journal of Biomedical Informatics* 42;6 (2009): 979–989.

[Ciccarese2008] Ciccarese, Paolo N., Elizabeth Wu, Gwen Wong, Marco Ocana, June Kinoshita, Alan Ruttenberg, and Tim Clark. "The SWAN biomedical discourse ontology." *Journal of Biomedical Informatics* 41;5 (2008): 739–751.

[Ciccarese2014] Ciccarese, Paolo N., Marco Ocana, and Tim Clark. "Open semantic annotation of scientific publications using DOMEO." *Journal of Biomedical Semantics* Apr 24;3 (2012): Suppl 1:S1.

[Clark2014] Clark, Tim, Paolo N. Ciccarese, and Carole A. Goble. "Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications." *Journal of Biomedical Semantics* 5;28 (2014).

[Hanna2013] Hanna, Josh, Eric Joseph, Mathias Brochhausen, and William R. Hogan. "Building a drug ontology based on RxNorm and other sources." *Journal of Biomedical Semantics* 4 (2013): 44–52.

[Natale2011] Natale, Darren A., Cecilia N. Arighi, Winona C. Barker, Judith A. Blake, Carol J. Bult, Michael Caudy, Harold J. Drabkin, Peter D'Eustachio, Alexei V. Evsikov, Hongzhan Huang, Jules Nchoutmboube, Natalia V. Roberts, Barry Smith, Jian Zhang and Cathy H. Wu. "The Protein Ontology: a structured representation of protein forms and complexes." *Nucleic acids research* 39, no. suppl 1 (2011): D539–D545.

[Saverno2011] Saverno, Kim R., Lisa E. Hines, Terri L. Warholak, Amy J. Grizzle, Lauren Babits, Courtney Clark, Ann M. Taylor, and Daniel C. Malone. "Ability of pharmacy clinical decision-support software to alert users about clinically important drug-drug interactions." *Journal of the American Medical Informatics Association* 18;1 (2011): 32–37.

[Wang2010] Wang, Lorraine M., Maple Wong, James M. Lightwood, and Christine M. Cheng. "Black box warning contraindicated comedications: concordance among three major drug interaction screening programs." *Annals of Pharmacotherapy* 44; 1 (2010): 28–34.

[W3C2013] Sanderson, Rob, Paolo N. Ciccarese, and Herbert Van de Sompel (editors). "Open Annotation Data Model", W3C Community Group Draft, 08 February 2013, url: http://www.openannotation.org/spec/core/