Formalising Uncertainty: An Ontology of Reasoning, Certainty and Attribution (ORCA)

Anita de Waard[1] and Jodi Schneider[2]

[1] Elsevier Labs, Jericho, VT A.dewaard@elsevier.com
[2] Digital Enterprise Research Institute, National University of Ireland
jodi.schneider@deri.org

Abstract. To enable better representations of biomedical argumentation over collections of research papers, we propose a model and a lightweight ontology to represent interpersonal, discourse-based, data-driven reasoning. This model is applied to a collection of scientific documents, to show how it can be applied in practice. We present three biomedical applications for this work, and suggest connections with other, existing, ontologies and reasoning tools. Specifically, this model offers a lightweight way to connect nanopublication-like formal representations to scientific papers written in natural language.

Keywords: scholarly communication, ontologies, nanopublications

1 Introduction

Biological understanding is created by scientists collaboratively working on understanding a (part of a) living system. To contribute knowledge to this collective understanding, biologists perform experiments and draw observational and interpretational assertions about these models [19]. In the social and linguistic practice of scientific publishing, the truth value (the confidence in the certainty of a statement), the knowledge source (who stated it) and basis (what was the statement based on) of an assertions are generally indicated through some linguistic expression of certainty or attribution, a 'hedge', of the type 'These results suggest that [A causes B]', or 'Author X implied that [A causes B]', or, in case a proposition is presumed true, the unmodulated '[A causes B]'. In this way, biological papers provide explicit truth evaluations of their own and other authors' propositions and these evaluations and attributions are a core component of shared knowledge constructions.

The goal of the present work is to provide a lightweight, formal model of this knowledge value and attribution, to assist current efforts that offer formal representations of biological knowledge and tie them more directly to the natural language text of scientific papers. Formal knowledge representations (and their corresponding ontologies) generally consist of statements of the system 'A causes B', or 'A is-a B', that do not leave room for doubt, dispute, and disagreement. But if we want to model the process (as opposed to the final consensus of outcome) of science, we need to trace the heritage of claims. Very often, claims can be traced to interpretations of data – so to model claim-evidence networks [6,15] we need to allow for links from claims to non-textual

elements such as figures and provenance trails, to trace the attribution of claims to people, organizations, and data processes [8].

Our model is based on an analysis of scientific argumentation from different fields: linguistics, sentiment analysis and genre studies. We have developed a lightweight ontology, dubbed 'ORCA', the Ontology of Reasoning, Certainty and Attribution, for making RDF representations of the certainty and source of claims. The goal of this model is to assist and augment other efforts in bioinformatics, discourse representation and computational linguistics with a lightweight way of representing truth value, basis and source. In this paper, we present our model and show different scenarios for the practical application of this work. We provide a brief overview of related projects, and sketch our thoughts on possible alignments with complementary ongoing efforts.

Following this introduction, in Section 2 we discuss our proposal for representing the strength and source. Then in Section 3 we discuss related work, followed by some realistic application areas in Section 4. We conclude the paper with a discussion of next steps in Section 5.

2 Our proposal

2.1 Model

In science, the strength and source of claims are important. Attribution is particularly central in science, yet existing models of provenance do not capture some simple, key distinctions: is the work cited or referred to done by the author or by another person? Is that work backed by data or by inference? Further, the strength of claims is of particular concern, especially during the reviewing process. It is common for authors to need to add qualification to their words, in order to get through the publication process. Even titles must appropriately indicate this in order to get a paper published. For instance, the author proposing the title "miRNA-372 and miRNA-373 Are Implicated As Oncogenes in Testicular Germ Cell Tumors" was instructed to softens the claim by saying that data is the source, making the (un)certainty of this result clearer. To get the paper published, it had to be retitled: "A Genetic Screen Implicates miRNA-372 and miRNA-373 As Oncogenes in Testicular Germ Cell Tumors".

Following concepts in linguistics and computational linguistics (for a full overview of literature, see [7]) we identify a hedged or attributed clause as a Proposition P that is modified by an evaluation E that identifies the truth value and attribution of P. Based on work in linguistics, genre studies and computational linguistics, we identify a three-part taxonomy of epistemic evaluation and knowledge attribution which covers the most commonly occurring types of knowledge attribution and evaluation in scientific text, as shown in the table. This taxonomy is summarized in Figure 1.

From our corpus study [7] it appeared that for all biological statements or 'bio-events' [18] a certain value of E (V, B, S) can be found without much difficulty. Linguistic markers for a lack of full certainty (i.e. where Value ¡ 3) include the use of hedging adverbials and adjectives ('possibly', 'potential' etc), the use of modal auxiliary verbs ('might', 'could') and, most frequently, the use of reporting verbs ('suggest', 'imply', 'hypothesize', etc.). As is clear from the examples in Figure 1, the absence or presence of a single linguistic marker identifies a value in each of the three dimensions:

Concept	Values	Example
Value	0 - Lack of knowledge	*The mechanism of action of this system is not known*
	1 – Hypothetical: low certainty	*We can hypothesize that…*
	2 – Dubitative: higher likelihood but short of complete certainty	*These results suggest that…*
	3 – Doxastic: complete certainty, reflecting an accepted, known and/or proven fact.	*REST-FS lacks the C-terminal repressor domain that interacts with CoREST…*
Basis	R – Reasoning	*Therefore, one can argue…*
	D – Data	*These results suggest…*
	0 – Unidentified	*Studies report that…*
Source	A - Author: Explicit mention of author/speaker or current paper as source	*Figure 2a shows that…*
	N - Named external source, either explicitly or as a reference	*….several reports have documented this expression [11-16,42].*
	IA - Implicit attribution to the author	*Electrophoretic mobility shift analysis revealed that…*
	NN – Nameless external source	*…no eosinophil-specific transcription factors have been reported…*
	0 – No source of knowledge	*transcription factors are the final common pathway driving differentiation*

Fig. 1. Taxonomy

- 'These results suggest': Value = 2, Source = Author, Basis = Data;
- 'REST-FS lacks the C-terminal repressor domain that interacts with CoREST': Recommended Value = 3, Source = Not specified; Basis = Not specified.

2.2 Ontology

We then model this in a lightweight ontology, ORCA – the Ontology of Reasoning, Certainty and Attribution. From the taxonomy, we have three core aspects: the source of knowledge, its basis, and its certainty. Our ontology should allow us to associate values for each of these. Thus we model them as classes and add associated Object Properties to add flexibility of expression. Controlled values for the taxonomy (e.g. "Named External Source") are represented as instances. Further, we induce an order on the certainty values, using transitive properties, to make it evident that, e.g. "Hypothetical Knowledge" is less certain than "Dubitative Knowledge". We considered using SKOS[3] to induce this order; however, skos:broaderThan is not appropriate, and skos Collections add an unwanted layer of complexity.

A clear application of this work is to support and help underpin the relations between formal knowledge representation such as nanopublications, and scientific text. A recent paper in Nature Biogenetics argues that

> Some argue that the rhetoric in articles is difficult to mine and to represent in the machine-readable format. Agreed, but frankly, why should we try? All nanopublications will be linked to their supporting article by its DOI. [17]

We think that adding a layer of epistemic validation with knowledge attribution enables a 'good enough' representation of the first level of scientific argumentation: the

[3] http://www.w3.org/TR/2009/REC-skos-reference-20090818/

statement and citation of claims. “Frankly, we should try” to do this, since this creates a superior representation of scientific argumentation, and as a bonus, allows us to connect nanopublications at a much more fine-grained level that merely the DOI of the paper that contains the statement. By adding a markup that contains the triple representation of the bioevent, augmented by the ORCA values, an evaluated and traceable network of knowledge can be created within and between documents, that can be represented and reasoned with using the same tools and utlilising the RDF-based standards that are currently being developed for other semantic representation projects such as OpenBEL[4], OpenPhacts [25], Eagle-I[5], and others.

As an example, in another paper on nanopublications, Clare et al (2011) [5] propose that the route to a scientific nanopublication can be facilitated by enabling the annotation of scientific notes or blogs at multiple levels of detail. Specifically, the authors propose that an author annotates a statement such as ‘isoproterenol binds to the Alpha-2 adrenergic receptor’ with a triple, linking the concepts ‘isoproterenol’, ‘Alpha-2 adrenergic receptor’ and ‘binds’ to the CHeBI, UniProt and NCI ontologies, respectively.

Enriching this model, we propose to add an epistemic evaluation to a similar statement: ‘These data demonstrated that [...] isoproterenol modulated the binding characteristics of alpha 2-adrenergic receptors’, which we would represent as follows:

```
@prefix orca: <http://vocab.deri.ie/orca#> .

"isoproterenol modulates binding characteristics
alpha 2-adrenergic receptors"
orca:hasSource orca:AuthorExplicitly ;
orca:hasBasis orca:Data ;
orca:hasConfidenceLevel orca:DoxasticKnowledge .
```

This provides a formal representation of the scientifically relevant aspects–the source of the statement, its basis, and its confidence level; or ORCA could be combined with annotation ontologies. This opens up new possibilities, beyond existing work, as we now discuss.

3 Related Work

There are a wealth of efforts in various fields that aim to represent the argumentation of (biomedical) scientific text, which our work builds on and which we hope this little ontology can support. We will only briefly mention efforts pertaining to scientific discourse efforts and computational linguistics - for a more detailed overview, see [7]).

Semantic Scientific Discourse Regarding semantic scientific discourse work, seminal efforts by the Knowledge Media Institute led by Buckingham Shum aimed to represent scientific sense making and offered ScholOnto, a scientific argumentation ontology, see e.g. [3,13]. In addition to SWAN, Clark et al. and the Annotation Ontology [4] which

[4] http://www.openbel.org/
[5] https://www.eagle-i.net/

aims to capture networks of hypotheses and evidence; this work is currently being combined with work on the Open Annotation framework and experiencing a lively series of developments to enable the creation of a robustly scalable framework for supporting argumentation modeling on the semantic web. Other efforts include SALT [10,11,9], the ABCDE format [1], and ontologies such as CiTO [21] are meant to create environments for authoring and citing specific portions of papers (see also [6] for a summary of these efforts). We believe our work can complement all of these efforts. ORCA can easily be used, alone or in combination with annotation ontologies, in order to link evidence to its source, basis, and confidence level.

Biomedical Informatics Several biomedical informatics systems categorize evidence; for a review, see [2]. We see the modularity as a key advantage: while the Gene Ontology[6] indicates evidence codes for biological inference, these cannot be used without importing the entire ontology. By contrast, domain ontologies could easily incorporate a lightweight, modular RDF ontology such as ORCA. Compared to the Evidence Code Ontology[7], ORCA is more suitable for annotating discourse: for instance, it handles citations to external sources and explicitly indicates confidence levels.

Computational Linguistics Within computational linguistics a number of efforts have focused on detecting the key components and salient knowledge claims in scientific papers, starting with the seminal work of Teufel [22] who developed a system for describing and set of tools to find 'argumentative zones' in scientific papers. Separate efforts to identify epistemically modulated claims and bioevents started with the work of Light et al [14] among (many) others (e.g. [16,23,24]; for a more complete literature overview, see [7]).

4 Possible applications

Improving the evidence behind drug product labels Drug product labels represent drug-drug interactions to help practitioners appropriately prescribe and avoid adverse drug events [2]. Representing the currently known information from the literature is important, yet the certainty of this knowledge varies considerably. Thus, drug-drug interactions are another important use case for ORCA. Information that two drugs interact should be qualified by an indication of what data backs this finding. The level of certainty is indicated in the literature, and representing this would allow different actions to be taken as appropriate. For frail patients, even suspected, unverified drug-drug interactions, could be avoided, as ongoing research confirms the circumstances and certainty of this information. Experimental treatments might accept suspected bad interactions, up to some higher level of certainty.

Data 2 Semantics Use Case As another potential use case, the Data2Semantics project[8] aims to build a semantic infrastructure to connect (and in future, semi-automatically detect and reconstruct) chains linking clinical recommendations to clinical summaries to

[6] `http://www.geneontology.org/`
[7] `http://www.evidenceontology.org/`
[8] `http://www.data2semantics.org/`

the underlying evidence. Still missing from the lightweight ontology being explored by this project is a formalization of the strength and attribution of these clinical recommendations offering another possible use of ORCA. Specifically, we imagine adding an 'ORCA-layer' (either during authoring, or post hoc), to the recommendations provided in clinical trials, so that these can be assessed and directly cited from clinical guidelines. One can then imagine a semantic representation of the clinical finding itself (augmented with an ORCA-structured clause) that can be automatically mined to prepopulate a proposed set of guideline recommendations, that merely need to be checked off my an editor, and can be constantly updated.

Enriching semantic search As a further application of our work is to enrich semantic search platforms: systems that allow search and retrieval subject-object-relationship triples. For instance, MedIE[9] is a triple-based search platform that can be used to find biomedical correlation sentences. But this search does not distinguish between fact and perhaps-fact – nor between novel information and well-known information. We can imagine a number of questions a user might want to answer:

- Give me all completely verified facts about X
- Tell me who found out what about X
- Show me what X is based on?
- Show me all claims which an author says are true, based on their own data. (Such data-based claim knowledge updates have Value = 2 or 3, Basis = Author, and Source = Data; they can be found with using state-of-the-art semantic parsing [20].)

By representing information with ORCA, semantic search engines such as MedIE could provide a better answer to these questions.

5 Next steps for using ORCA

We envision a mixed-initiative approach for applying ORCA to scientific papers. A text mining system (such as mentioned above) would present an author with a tentative list of ORCA-enhanced claims, i.e. a list of the bioevents or key claims in a paper along with a suggested ORCA assignment of the 'veracity value' for each claim. The author (or editor or curator) would then validate both the claim and its 'veracity value', resulting in a set of claims enhanced with ORCA 'veracity values'. Through concept networks such as those proposed in nanopublications or systems such as BEL[10] articles can then be connected by and enriched with these knowledge networks.

Thus, we believe that future collaborations between efforts in semantic web technologies, bioinformatics and computational linguistics can help develop a future where authors can interact with systems that acknowledge and identify their core claims. Similar mixed-initiative systems have already been used to automatically highlight phrases that could be semantically annotated [12]. In particular, we have taken some preliminary steps to identify 'Claimed Knowledge Updates' - a special case of a bioevent that is claimed by the author and based on data - using state-of-the-art semantic parsing [20].

[9] `http://www.nactem.ac.uk/medie/`
[10] `http://openbel.org`

One potential roadblock to such a system is that, at least at first, the system output will need to be corrected, which means another step in submission and editing for this already beleaguered author. Another serious issue is that the current system of providing slightly vague, hedged claims serves a social purpose: authors prefer to think that they have made many improvements on the state of the art, and as long as they hedge their statements appropriately, reviewers will let them get away with it. If authors have to make the validity/strength of the claim explicit at the authoring stage, this might introduce a precision in applying truth value that makes all parties uncomfortable. In fact, many reviews mainly concern the degree or strength of the claims made, with the addition of hedging being a frequent demand.

Yet, given the fact that scientific knowledge continues to grow at a dizzying pace, it seems inevitable that sooner or later we will need to represent more exact representations of that knowledge across collections of papers. Widespread use of systems for marking the value, basis, and source of the hedge will help to represent the richness of this knowledge. And there is no particular reason why this model would be limited to the life sciences. As a succinct, simple, and interoperable ontology in that can be used in combination with any RDF-based system, we hope that ORCA can contribute a small building block to what will prove, undoubtedly, to be a collective effort.

Acknowledgements

Thanks to Aidan Hogan for collaboration on the ontology. Jodi Schneider's work was supported by Science Foundation Ireland under Grant No. SFI/09/CE/I1380 Líon2.

References

1. Anita de Waard and Gerard Tel. The ABCDE format enabling semantic conference proceedings. In *ESWC 2006*.
2. R. Boyce, C. Collins, J. Horn, and I. Kalet. Computing with evidence: Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment. *Journal of Biomedical Informatics*, 42(6):979 – 989, 2009.
3. S. Buckingham Shum, E. Motta, and J. Domingue. ScholOnto: an ontology-based digital library server for research documents and discourse. *International Journal on Digital Libraries*, 3(3):237–248, Oct. 2000.
4. P. Ciccarese, M. Ocana, S. Das, and T. Clark. AO: an open annotation ontology for science on the web. In *BioOntologies 2010*, May 2010.
5. A. Clare, S. Croset, C. Grabmueller, S. Kafkas, M. Liakata, A. Oellrich, and D. Rebholz-Schuhmann. Exploring the generation and integration of publishable scientic facts using the concept of nano-publications. In *Proc. 1st Workshop on Semantic Publishing 2011 at ESWC2011*.
6. A. de Waard, S. Buckingham Shum, A. Carusi, J. Park, M. Samwald, and Á. Sándor Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims. Proc. of the Workshop on Semantic Web Applications in Scientific Discourse at ISWC-2009.
7. A. de Waard and H. Pander Maat. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proc. of the Workshop on Detecting Structure in Scholarly Discourse*, pages 47–55, 2012. Association for Computational Linguistics.

8. P. Groth, Y. Gil, J. Cheney, and S. Miles. Requirements for provenance on the web. *International Journal of Digital Curation*, 7(1):39 – 56, 2012.
9. T. Groza, S. Handschuh, K. Möller, and S. Decker. KonneX-SALT: First Steps Towards a Semantic Claim Federation Infrastructure. In S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, editors, *The Semantic Web: Research and Applications, Proc. of ESWC 2008*, volume 5021 of *LNCS*, pages 80–94. Springer, 2008.
10. T. Groza, S. Handschuh, K. Mller, and S. Decker. SALT - semantically annotated LaTeX for scientific publications. In *The Semantic Web: Research and Applications, Proc. of ESWC 2007*, LNCS. Springer, 2007.
11. T. Groza, K. Möller, S. Handschuh, D. Trif, and S. Decker. SALT: Weaving the claim web. *The Semantic Web: Research and Applications, Proc. of ISWC+ASWC 2007*, 2007.
12. J.L. Fink, P. Fernicola, R. Chandran, S. Parastatidis, A. Wade A, O. Naim, G. B. Quinn, P. E. Bourne. Word add-in for ontology recognition: semantic enrichment of scientific literature. *BMC Bioinformatics*, 24(11):103, 2010.
13. A. D. Liddo, Á. Sándor, and S. B. Shum. Cohere and XIP: human annotation harnessing machine annotation power. In *CSCW (Companion)*, pages 49–50, 2012.
14. M. Light, X. Qiu, and P. Srinivasan. The language of bioscience: Facts, speculations, and statements in between. In *Proc. of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users*, pages 17–24, 2004.
15. M. S. Marshall, R. Boyce, H. F. Deus, J. Zhao, E. L. Willighagen, M. Samwald, E. Pichler, J. Hajagos, E. Prudhommeaux, and S. Stephens. Emerging practices for mapping and linking life sciences data using RDF a case series. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14:2 – 13, 2012.
16. B. Medlock and T. Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *Association for Computational Linguistics*, volume 45, page 992, 2007.
17. B. Mons, H. van Haagen, C. Chichester, P.-B. t. Hoen, J. T. den Dunnen, G. van Ommen, E. van Mulligen, B. Singh, R. Hooft, M. Roos, J. Hammond, B. Kiesel, B. Giardine, J. Velterop, P. Groth, and E. Schultes. The value of data. *Nature Genetics*, 43(4):281–283, 2011.
18. R. Nawaz, P. Thompson, and S. Ananiadou. Evaluating a meta-knowledge annotation scheme for bio-events. In *Proc. of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pages 69–77, 2010. Association for Computational Linguistics.
19. T. Russ, C. Ramakrishnan, E. H. Hovy, M. Bota, and G. Burns. Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case. *BMC Bioinformatics*, 12:351 – 366, 2011.
20. Á. Sándor and A. de Waard. Identifying claimed knowledge updates in biomedical research articles. In *Proc. of the Workshop on Detecting Structure in Scholarly Discourse*, pages 10–17, 2012. Association for Computational Linguistics.
21. D. Shotton. CiTO, the citation typing ontology. *Journal of Biomedical Semantics*, 1(Suppl 1):S6, 2010.
22. S. Teufel. *Argumentative Zoning: Information Extraction from Scientific Articles*. Ph.D., University of Edinburgh, Edinburgh, Scotland.
23. P. Thompson, G. Venturi, J. McNaught, S. Montemagni, and S. Ananiadou. Categorising modality in biomedical texts. In *Proc. of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 27–34, 2008.
24. V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(Suppl 11):S9, 2008.
25. A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, C. T. Evelo, N. Blomberg, G. Ecker, C. Goble, and B. Mons. Open Phacts: semantic interoperability for drug discovery. *Drug Discovery Today*, 2012.