Argument graphs: Literature-Data Integration for Robust and Reproducible Science

Tim Clark
Massachusetts General Hospital and Harvard Medical School, Boston MA, USA
twclark@mgh.harvard.edu

ABSTRACT

Two complementary models for biomedical literature-data integration are presented: entity-based and argument-based. We believe the argument-based model is novel and can be exceptionally useful in providing better support than currently exists for robust and reproducible science. We describe both approaches, along with some current models and available tools for scientific literature annotation. We then show how argument graphs, represented as stand-off annotation on research articles, can help improve the robustness of scientific findings over time.

Keywords

literature-data integration, web annotation, reproducibility

1. BACKGROUND

Literature-data integration as discussed in this article means machine-navigable linkage of specific elements of scientific articles on the Web, to closely related data stored in repositories. The purpose may be to provide explanation or support for terms, images, or statements presented in the target article. The novel model we present, based on argument graphs, is particularly concerned with evidential support for statements made by the author as novel scientific findings. It is one application of the Micropublications model [12], which harmonizes the support-based argumentation model of Toulmin and successors in the AI field [30–32] with the argumentation framework approach of P.M. Dung [13], and its derivations such as Cayrol and Lagasquie-Schiex's bimodal argumentation frameworks [7, 8].

There are at least two useful and complementary models of literature-data integration, although only one has been extensively explored to date. We first present the entity-based model in which terms are linked to curated biomedical databases. We then move on to discuss the argumentation-graph-based approach, which we believe is not only complementary to the first, but may offer very important technical support to improve the robustness of scientific findings over time.

2. ENTITY RECOGNITION IN TEXT LINKED TO CURATED BIOMEDICAL DATABASES

Entity recognition in text with ontology terms linked to curated bioinformatics databases relies on mapping textual terms in the article to elements of controlled vocabularies - especially to biomedical ontologies - inferred to have identical meaning to the textual terms. Because such vocabulary elements can often identify biomedical database entries, such as those for proteins (in UniProt [3], PDB [5], Protein ontology [22–24] , etc.) or gene functions (in GO [1]), the associated data such as protein structures, or gene functions, can be directly mapped to parts of the article text.

Several lines of research and competing approaches have been used to develop such entity recognizers, and well-organized competitions such as BioCreative [19] have undertaken to test out the various algorithms against one another. Not just academic software, but also commercial software exists and is actively marketed to pharmaceutical and biotech companies for this kind of entity extraction (e.g. the Linguamatics textmining suite [4]. Once the text-to-vocabulary mappings have been achieved, they may then serve as the basis for popups and visualizations [2, 25], and/or alerting systems based on researcher or industrial interest. It is clearly essential in this approach, shown conceptually in Figure 1, to employ robust entity recognition algorithms based on sound ontologies.

Entity recognition is typically an ex post facto approach. That is, it enhances the scientific article via post-processing. European PubMed Central, for example, does extensive entity recognition and other text processing on its open access corpus, using tools based on the original WhatIzIt algorithm.

3. THE ARGUMENT-GRAPH MODEL

3.1 Argument Graph Approach

Argument graphs as a model of scientific discourse (or any discourse) are based on the notion - readily verifiable by observation - that scientific articles contain assertions based on cited literature and (for original research publications) direct observations; and that nothing in any scientific article can be considered the last word on the subject. Scientific articles are, as noted by Toulmin, arguments [30]. For any

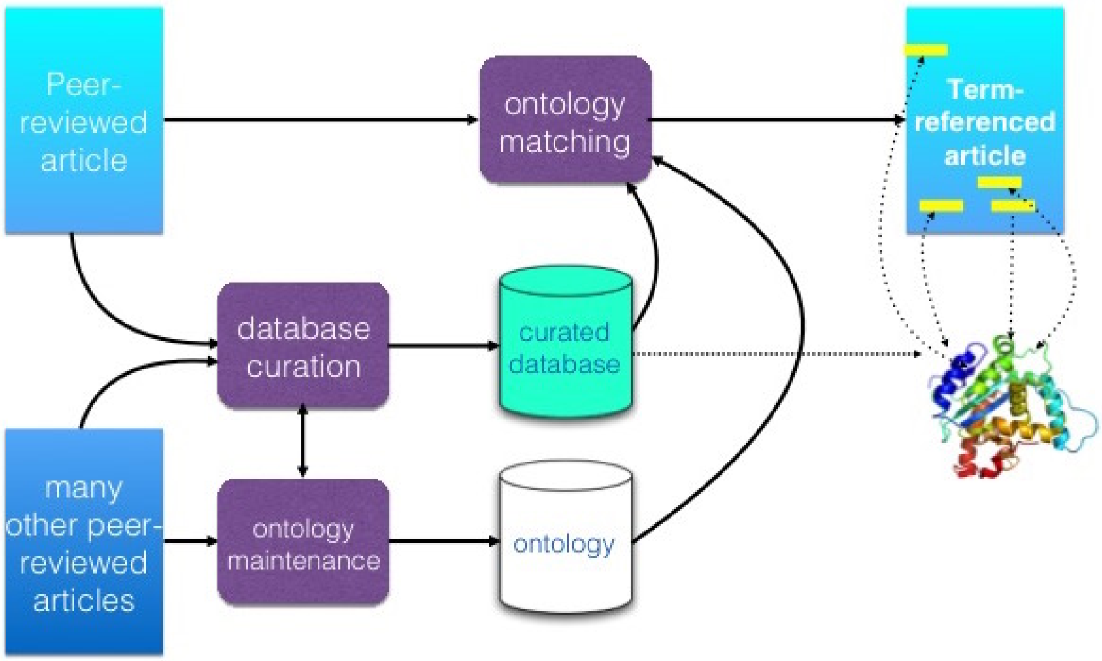

Figure 1: Entity-based Integration

article, its findings may be opposed by counterarguments. Argumentative discourse dates back to the dawn of scholarship (see for example Plato's Symposium [26]). However the model in which arguments are only considered valid if supported by exhibited observational data along with a clear description of how the data were obtained - clear enough to be reproduced in fact or in imagination (virtually) by the reader - is more recent, dating to the 17th century [16]. It was the initiating information model (or "literary technology") of the Scientific Revolution [28].

Definition: A scientific article is a (1) defeasible argument for (2) a set of claims (3) based on a narrative of observations, supported by (4) exhibited data and (5) reusable and/or reproducible methods and materials; which is (6) contextualized in the (7) domain of discourse.

(1) A defeasible argument is an argument that may be contradicted and disproven. (2) A set of claims is a set of assertions, i.e. *truth-bearing* statements. A *truth-bearing* statement is falsifiable. (3) A narrative of observations is a factual description of how a set of data were obtained. (4) Exhibited data is data that is clearly shown. (5) (a) Reusable methods are those which are described in sufficient detail so as to be, at least in principle, reproducible by a sufficiently skilled person. (b) Reproducible materials are tools and/or reagents which may be readily recreated or obtained, again by a person sufficietly skilled in the relevant domain. (6) A *contextualized* argument is one which refers in sufficient depth and detail, through shared technical vocabulary and/or citation of commonly referenced background material, to be readily related by the informed reader to a particluar domain of discourse.

Because scientific arguments are defeasible, accepted scientific truth in any domain evolves by "successive approximation" through a collective process of experiment, theorization, and discourse (argumentation). Not everything you read in a scientific article will be correct, whether today, or in thirty years. What the argument-graph-based model attempts to do, is to create a data structure which reveals these relationships and is tractable for computation and navigation on the Web (Figure 2).

Argument is defined in the Internet Encyclopedia of Philosophy [21] as (a) "a dispute or a fight, or ... " (b) "a collection of truth-bearers some of which are offered as reasons for one of them, the conclusion." It should be clear that (a) is about a scenario of mutually-opposing contradictory statements, and (b) is about mutually-supporting affirming statements. In fact, in any domain of discourse, both aspects are on view. *Within* any particular argument, where argument is considered as a unitary discourse, the statements and evidence (representations which may or may not be statements, but which support or contradict statements) presented are in general mutually supportive. Where this is not the case, contradictory or challenging evidence is generally brought into play as a rhetorical device and ultimately defeated by counterarguments. *Between* arguments (typically those made by different individuals), either mutual support or challenge may prevail. The work of Toulmin [30], who essentially invented modern argumentation theory, deals primarily with the first case, and is focused mainly on intra-argument structure. The work of the AI researcher P.M. Dung [13] and those influenced by him such as Cayrol [7], deals primarily with inter-argument relationships.

Dung's model is formalized as the framework <AF,AR >, where AF is a set of arguments A, and AR is a set of challenge or "attack" relations $R \in AXA$. However in Dung's model, the "arguments" represented by A are "black boxes", with no internal structure, and reduce essentially to assertions. It should be clear that other kinds of relations such as "support" or "noncommittal" could be included in a Dung-

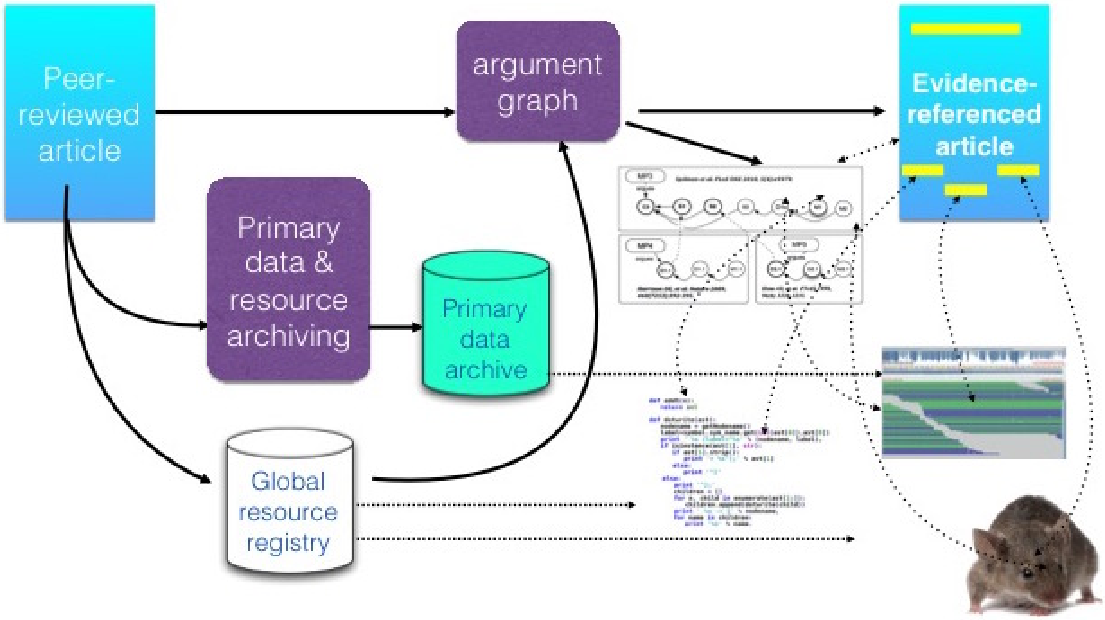

Figure 2: Argument-based Integration

style framework by extension to multimodal frameworks.

3.2 Argument Graphs as Micropublications

The Micropublications model [12] is a formal OWl2 ontology, which can represent any scientific argument as a DAG. The root of this DAG is the argument's central claim. The micropublications vocabulary [1] (a) integrates the Dung and Toulmin approaches in a multimodal (+,-,neutral) system and (b) includes not just statements as support for arguments, but also Representations, a superclass of statements including typical scientific discourse material such as images and data tables. Statements are defined as *declarative sentences*, and sentences are *linear symbolic representations*. These features allow a micropublication to represent both the support and attack or challenge aspects of arguments, as well as both intra- and inter-argument discourse relationships. They also allow micropublications to represent the typical forms of scientific evidence, as well as mathematical expressions. Micropublications are argument graphs with provenance and are expressed using an OWL2 [18] vocabulary, `//http://purl.org/mp`, using RDF [6]. The evidence supporting (or challenging) micropublications statements consists of (1) primary data, methods and materials and/or (2) referenced literature. Evidence may be expressed as statements or more generally as sentences (e.g. citations) or representations (e.g. figures). Evidence and statements, or statements and statements, are connected by *supports*, *challenges* and *discusses* relationships (+, -, neutral). Micropublications in their simplest, most general form take the structure shown in Figure 3. They can express the internal structure of argument in a single biomedical research article, as in Figure 4. Or they may express connected arguments both within and between articles. Or they may combine both supporting and challenging evidence from multiple sources, as in Figure 5.

[1] `http://purl.org/mp`

3.3 Challenge and support

As defeasible arguments, the claims of scientific articles or findings may and often are, challenged. Figure 6 shows how we have modeled this feature of argument graphs in a knowledgebase of drug-drug interactions. What is notable about the current state of information in this domain, despite what you might assume from interactions with your local pharmacist, is that multiple databases of such interactions exist, with contradictory claims and differing evidence supporting these claims. In Figure 6 we illustrate the very simple claim that "Escitalopram does not inhibit CYP2D6", which has contradcitory information. The outer frame in Figure 6 is the micropublication, the inner frame is the nanopublication.

3.4 Logical formalization of claims

For various purposes one may wish to make a logical formalization of a scientific claim. This allows us to do some further reasoning directly on the claims, at the cost perhaps of some loss of fidelity and epistemic or other qualification from its original presentation. We always therefore wish to preserve the original claim as text - what we model as the claim is what its author said. The formalization is a derived construct whose author is the person or algorithm who built it. The derived construct is *supportedBy* the original claim. The relationship *formalizes* is a subproperty of *supportedBy*. Figure 6 shows how a claim may be formalized by the *nanopublication* model [17]. An alternative formalization approach widely used in pharmaceutical companies is BEL, the Biological Expression Language [27].

3.5 How may argumentation be curated?

The first practical questions that may come to mind might be, since this model is not what scientists ordinarily construct for publication and professional reward, how will it be curated? How much time is required to curate each model?

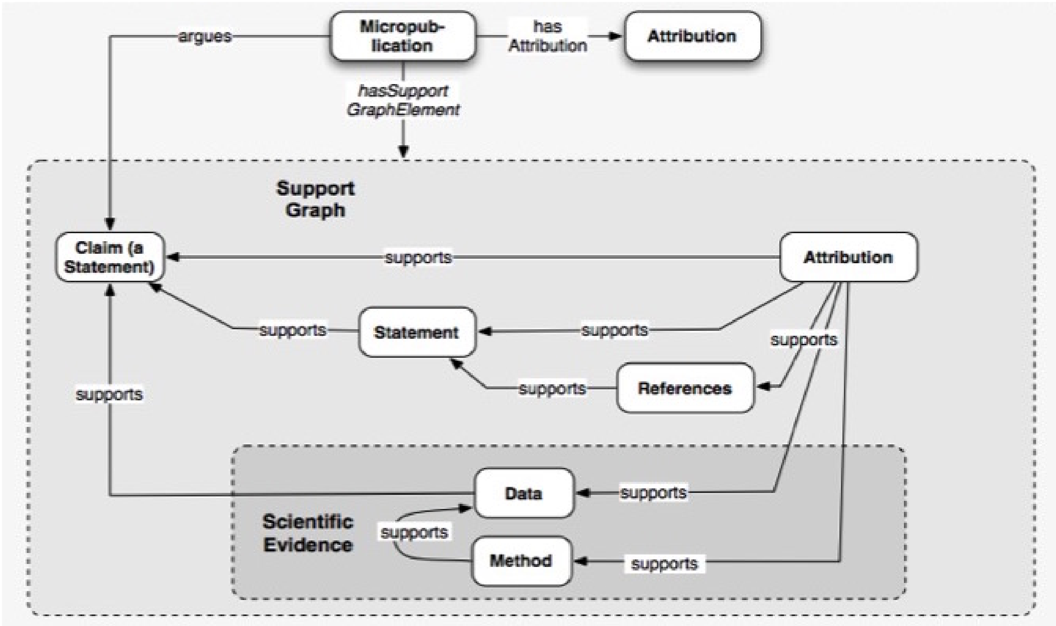

Figure 3: General structure of a micropublication

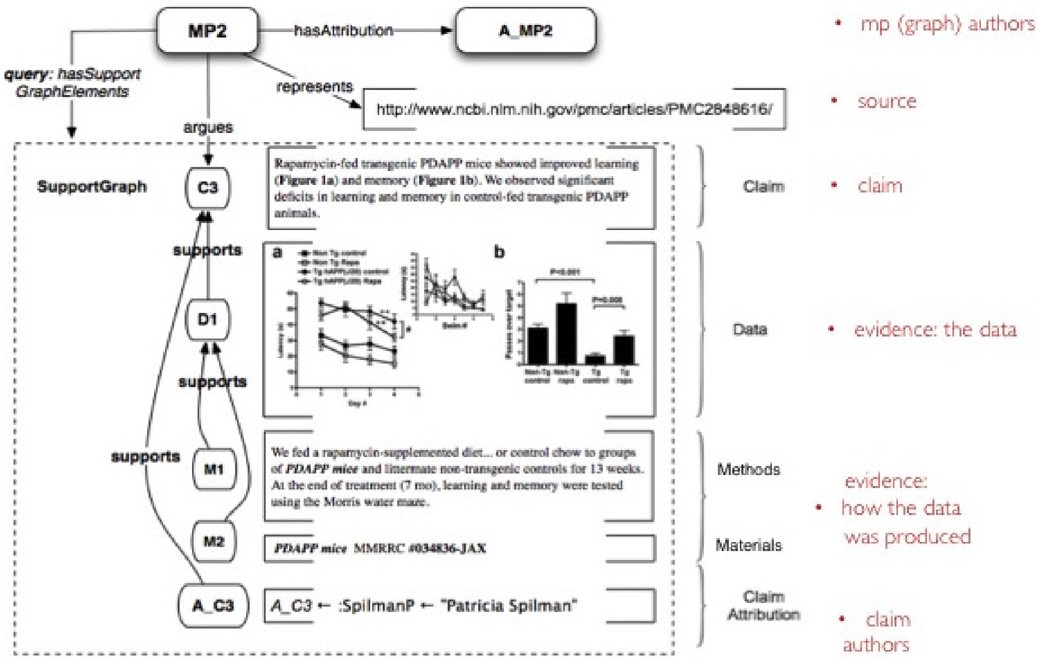

Figure 4: Representing a single research result as a micropublication

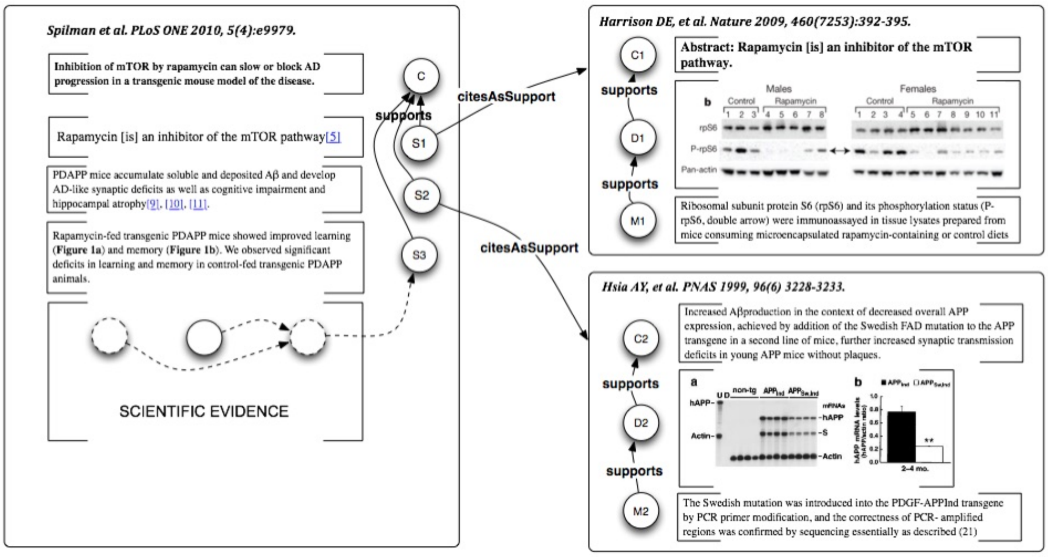

Figure 5: Argument graph across three micropublications

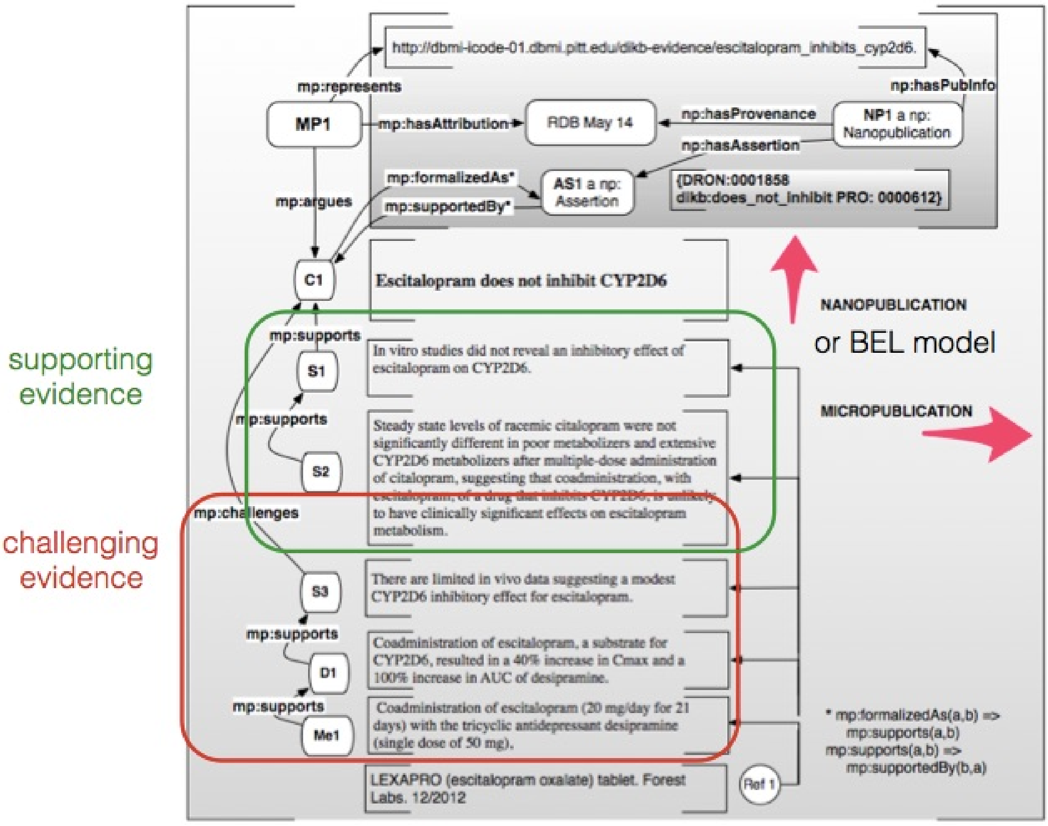

Figure 6: Argument graph with Contradiction

And would this activity scale? We suggest a system in which curation is not separated from the normal practices of day-to-day science. For example: normally in writing articles we make use of bibliographic reference management tools and systems. What do these do? They allow us to encapsulate entire arguments and systems of argument with a set of metadata - the reference. A primary weakness which many have observed, is that when I store an article in one of these systems, the reason for citing it is not saved - and the reason is typically a specific claim. Some systems allow comments to be added to the references, and these can be helpful in documenting the claims for which an article is referenced. If we have the specific claim noted, that is already a step toward our proposals. Specific claims in scientific papers are justified by citing the literature where supporting claims are documented; or by citing evidence presented in the article itself. Internal evidence is often cited similarly to literature references - "see Figure X" - and these references of both types can be parsed into the graph. We have built a pilot system that does this, and stores the graphs in a triple store using the Annotopia server [10] [9] [11]. If a bibliographic reference manager had access to an article's full text, we suggest that the scientist tracking the reference and its associated claim, highlight that claim, and let the software figure out where the claim's support comes from. In this way bit by bit, complex argument graphs may be built up. Publishers may wish to attack this problem by asking authors to identify the major assertions they are making and the support for these assertions. Drug companies doing deep target validation (or better, *continuous* target validation), may wish to construct extensive support and challenge graphs using this model, for the validation hypotheses they must demonstrate to progress a target. In other words, there are a number of points at which to attach this problem and by having a common model with which to represent arguments in biomedical communications, the required information should be able to be built up piecemeal from many contributions as part of an ecosystem. We have made experiments in curation independently of our software while using this general model, and the simplest answer regarding time is, "it depends". The major factor is what depth of modeling is attempted. We have seen experienced neuroscience research associates model an article in their domain, in ten to twenty minutes. However a particularly diligent undergraduate with a strong modeling bent spent four hours on one article. That, we consider excessive. But depending upon who is modeling and why, it may be appropriate. In the case just described, the undergraduate's modeling time incorporated a very large component of effort simply to understand what the article was saying, and what evidence it presented, in great depth. Lastly the question of scaling: this model is intended as a common representation to be used in an ecosystem. It is to be populated, passed along, and its information enhanced, in many cases as essentially a side-effect of doing science. If it is part of a useful set of interoperating tools, it will scale. This leads us to the question of tools, which will be brief because it partially recapitulates what has just been said.

4. INITIAL TOOLS

The micropublications model can be used standalone, to organize a knowledgebase. Or it can be applied as standoff annotation to any existing web documents, or to any information with an argument (claim-evidence) structure having a URI, using the W3C Web Annotation Data Model. The Annotopia Open Annotation Server [11] can read and write micropublications serialized as JSON-LD, and has a restful api. The Domeo web annotation tool [9, 10], which is now integrated with Annotopia, provides a useful set of interactive web forms with which users may annotate HTML documents using micropublications.

5. SOME APPLICATIONS

Argument graphs have several applications which are immediately apparent.

5.1 Claim networks

Argument graphs can be applied to construct large claim networks across entire topics, clarifying what is known and what is merely rhetoric or supposition. Greenberg's 2009 and 2011 articles on citation distortion [14,15] makes it clear that successive levels of distortion may be introduced as authors cite works in support of their own claims. At the least, epistemic qualification tends to be weakened where it casts doubt on a claim, and strengthened where it supports it. In micropublication-based argument graphs, the supporting evidence can be resolved to a single claim in a cited paper; and with systematic use of direct data citation [29], as this becomes more prevalent, ultimately resolved directly to supporting primary data.

5.2 Post-publication peer review

Post publication peer review takes place informally through discussion at conferences, in journal clubs, via emails and on blogs and discussion forums. However, the comments (supporting and challenging) on important research articles are spread out all over the web without any central organizing nexus. Argument graphs based on micropublications, if organized as stand-off W3C Web Annotation, could readily be aggregated across (known) sites.

5.3 Bibliographic reference management

As noted, common bibliographic reference managers are extremely useful for organizing and applying citations to a known corpus of literature. But a common challenge is to recall and to clearly state, what actual assertion or comment from a text is being cited. Today many reference managers allow comments to be associated with a bibliographic entry. There appears to be no reason why one or more very simple argument graphs might not replace the comment(s).

5.4 Target validation

Target validation is the first stage in pharmaceutical drug discovery, and in some firms it is now seen as a continuous process across the life of an entire drug development process. It consists of proving a set of hypotheses considered essential to be demonstrated before investing in a subsequent stage - typically the next stage being high throughput screening. It would be entirely possible to organize these "to be validated" hypotheses as a large argument graph, with data and its interpretation (as support or challenge) attached to the hypotheses as it is generated.

5.5 Drug-drug interactions and other databases of conflicting evidence

One might think that a canonical and well-established set of results would exist that could be readily queried about potentially adverse interactions possibly affecting people taking various drugs. But this would be incorrect. There are several resources and they disagree. Our colleagues at the University of Pittsburg have developed a knowledgebase of drug-drug interactions using micropublications as a model of claims and evidence, with both support and challenge incorporated into the model [20].

6. CONCLUSIONS

In this article we summarized current theory, tools and applications of argument graphs in supporting biomedical knowledge management and in sharply clarifying contradictory and variously supported arguments in biomedical discourse on the Web.

- We reviewed two complementary approaches to literature-data integration, based on (a) entity-tagging and (b) argument graph construction.
- We discussed the "classical" theory of argumentation and argumentation frameworks as presented in the work of Toulmin, Dung and their followers and interpreters.
- We explained how the micropublications model harmonizes the Toulmin (intra-argument, support-based) and Dung (inter-argument, challenge-based) models into a single multimodal framework optimized for representing biomedical knowledge.
- We described several current tools useful in dealing with and instantiating this model.
- We also showed how micropublications can represent contradiction and support logical formalization of claims while preserving their "chain of evidence" intact.
- Finally, we briefly outlined several applications of this model for improving the robustness of scientific findings and their transferability over time.

We believe argument graphs using micropublications and related approaches will prove to be an exceptionally useful informatics technique in exchanging, managing, and developing biomedical knowledge on the Web.

7. REFERENCES

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, *Gene ontology: tool for the unification of biology. the gene ontology consortium*, Nat Genet **25** (2000), no. 1, 25–9.

[2] T. K. Attwood, D. B. Kell, P. McDermott, J. Marsh, S. R. Pettifer, and D. Thorne, *Utopia documents: linking scholarly literature with research data*, Bioinformatics **26** (2010), no. 18, i568–i574.

[3] Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L. Yeh, *The universal protein resource (uniprot)*, Nucleic Acids Research **35** (2007), no. Jan, D193 – D197.

[4] J. Bandy, D. Milward, and S. McQuay, *Mining protein-protein interactions from published literature using linguamatics i2e*, Methods Mol Biol **563** (2009), 3–13.

[5] H. Berman, K. Henrick, and H. Nakamura, *Announcing the worldwide protein data bank*, Nat Struct Biol **10** (2003), no. 12, 980, Berman, Helen Henrick, Kim Nakamura, Haruki eng Letter 2003/11/25 05:00 Nat Struct Biol. 2003 Dec;10(12):980.

[6] Dan Brickley and R.V. Guha, *Rdf vocabulary description language 1.0: Rdf schema*, Tech. report, World Wide Web Consortium, 2004.

[7] Claudette Cayrol and Marie-Christine Lagasquie-Schiex, *Bipolar abstract argumentation systems*, Springer, Dordrecht, 2009.

[8] ———, *Coalitions of arguments: A tool for handling bipolar argumentation frameworks*, International Journal of Intelligent Systems **25** (2010), no. 1, 83–109.

[9] P Ciccarese, M. Ocana, and T. Clark, *Open semantic annotation of scientific publications with domeo*, Journal of Biomedical Semantics **3** (2012), no. Suppl 1, S1.

[10] P Ciccarese, Marco Ocana, and Tim Clark, *Domeo: A web-based tool for semantic annotation of online documents*, 2011.

[11] Paolo Ciccarese and Tim Clark, *Annotopia: An open source universal annotation server for biomedical research*, 7th International Workshop on Semantic Web Applications and Tools for Life Sciences (Adrian Paschke, Albert Burger, Paolo Romano, M. Scott Marshall, and Andrea Splendiani, eds.), vol. 1320, CEUR.

[12] Tim Clark, Paolo Ciccarese, and Carole Goble, *Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications*, Journal of Biomedical Semantics **5** (2014), no. 1.

[13] Phan Minh Dung, *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*, Artificial Intelligence **77** (1995), no. 2, 321–357.

[14] Steven A. Greenberg, *How citation distortions create unfounded authority: analysis of a citation network*, British Medical Journal **339** (2009), b2680, Greenberg, Steven A R01NS43471/NS/NINDS NIH HHS/ R21NS057225/NS/NINDS NIH HHS/ England Clinical research ed. BMJ. 2009 Jul 20;339:b2680. doi: 10.1136/bmj.b2680.

[15] ———, *Understanding belief using citation networks*, Journal of Evaluation in Clinical Practice **17** (2011), no. 2, 389–393.

[16] Alan G. Gross, Joseph E. Harmon, and Michael S. Reidy, *Communicating science: The scientific article from the 17th century to the present*, Oxford University Press, Oxford UK, 2002.

[17] Paul Groth, Andrew Gibson, and Johannes Velterop,

The anatomy of a nano-publication, Information Services and Use **30** (2010), no. 1, 51–56.

[18] OWL2 Working Group, *Owl 2 web ontology language document overview (second edition)*, Tech. report, World Wide Web Consortium, 2012.

[19] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, *Overview of biocreative: critical assessment of information extraction for biology*, BMC Bioinformatics **6 Suppl 1** (2005), S1.

[20] Schneider J, Brochhausen M, Rosko S, Ciccarese P, Hogan W, Malone D, Ning Y, Clark T, and Boyce RD, *Using the micropublications model to organize conflicting evidence and assertions in a database of drug-drug interactions*, Proceedings of the 4th Workshop on Linked Science 2014 (LISC2014), October 19, 2014, Riva Del Garda, Trentino, Italy.

[21] Matthew McKeon, *Argument*, Internet Encyclopedia of Philosophy (James Fieser and Bradley Dowden, eds.), 2015, http://www.iep.utm.edu/argument/.

[22] D. A. Natale, C. N. Arighi, W. C. Barker, J. A. Blake, C. J. Bult, M. Caudy, H. J. Drabkin, P. D'Eustachio, A. V. Evsikov, H. Huang, J. Nchoutmboube, N. V. Roberts, B. Smith, J. Zhang, and C. H. Wu, *The protein ontology: a structured representation of protein forms and complexes*, Nucleic Acids Res **39** (2011), no. Database issue, D539–45.

[23] Darren Natale, Cecilia Arighi, Winona Barker, Judith Blake, Ti-Cheng Chang, Zhangzhi Hu, Hongfang Liu, Barry Smith, and Cathy Wu, *Framework for a protein ontology*, BMC Bioinformatics **8** (2007), no. Suppl 9, S1.

[24] Darren A. Natale, Cecilia N. Arighi, Judith A. Blake, Carol J. Bult, Karen R. Christie, Julie Cowart, Peter DâĂŹEustachio, Alexander D. Diehl, Harold J. Drabkin, Olivia Helfer, Hongzhan Huang, Anna Maria Masci, Jia Ren, Natalia V. Roberts, Karen Ross, Alan Ruttenberg, Veronica Shamovsky, Barry Smith, Meher Shruti Yerramalla, Jian Zhang, Aisha AlJanahi, Irem ÃĞelen, Cynthia Gan, Mengxi Lv, Emily Schuster-Lezell, and Cathy H. Wu, *Protein ontology: a controlled structured network of protein entities*, Nucleic Acids Research (2013).

[25] Sean I. O'Donoghue, Heiko Horn, Evangelos Pafilis, Sven Haag, Michael Kuhn, Venkata P. Satagopam, Reinhard Schneider, and Lars J. Jensen, *Reflect: A practical approach to web semantics*, Web Semantics: Science, Services and Agents on the World Wide Web **8** (2010), no. 2-3, 182–189.

[26] Plato, *Symposium*, Oxford University Press, 2004.

[27] Selventa, *Biological expression language v1.0 overview*, 2011, http://bit.ly/1KzL5Gi.

[28] Steven Shapin, *Pump and circumstance: Robert boyle's literary technology*, Social Studies of Science **14** (1984), no. 4, 481–520, 10.1177/030631284014004001.

[29] Joan Starr, Eleni Castro, MercÃÍ Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, Melissa Haendel, Ivan Herman, Simon Hodson, Joe HourclÃl', John Ernest Kratz, Jennifer Lin, Lars Holm Nielsen, Amy Nurnberger, Stefan Proell, Andreas Rauber, Simone Sacchi, Arthur Smith, Mike Taylor, and Tim Clark, *Achieving human and machine accessibility of cited data in scholarly publications*, PeerJ Computer Science **1** (2015), e1.

[30] Stephen Edelston Toulmin, *The uses of argument*, Cambridge University Press, Cambridge UK, 2003.

[31] Bart Verheij, *Evaluating arguments based on toulminâĂŹs scheme*, Argumentation **19** (2005), no. 3, 347–371.

[32] ———, *The toulmin argument model in artificial intelligence*, Springer, Dordrecht, 2009.