Next Generation Scientific Publishing and the Web of Data

Editorial

Tim Clark
Harvard Medical School, Boston MA USA; Massachusetts General Hospital, Boston MA USA; School of Computer Science, University of Manchester, Manchester UK

Abstract Next Generation scientific publishing will exploit both the web of documents and the web of data to help resolve many of today's serious problems in reproducibility, citation and citation-claim integrity, and publication volume intractability. Here we briefly review several developments in this field using new semantic models and approaches to help achieve more robust scientific, and particularly biomedical, communications.

Keywords: scientific publishing; biomedical communications; semantic web; argumentation; reproducibility; data integrity; ontologies; scientific data; reusability

The entire topic of what has been called "Next Generation Scientific Publishing" has garnered significant interest lately, due to widely-reported defects in the current ecosystem of biomedical communications affecting among other things, reproducibility, citation integrity, and intractability of the present volumes of publications per field.

Reproducibility is a very real problem, particularly in therapeutic development. Begley and Ellis [1] at Amgen reported in 2012 that of the academic research results they reviewed and tested in hematology-oncology, only 11% were actually reproducible in the laboratory. Researchers at Bayer [2] reported similarly disappointing results. The author has had discussions with colleagues collaborating with another pharmaceutical company, who quoted, again, very similar results.

This issue was recently brought into even sharper focus by a controversy in the regenerative biology community over the findings on so-called STAP (Stimulus Transitioned Acquisition of Pluripotency) cells. Obakata et al. [3,4] reported very surprising and promising results that simply could not be reproduced in many laboratories. An investigation conducted by Riken subsequently found that she had mishandled data by illicit alterations to images. These alterations were first discussed on PubPeer and various blogs (see Figure 1).

The Obakata et al. results were striking and potentially ground-breaking, which was why they were scrutinized so rapidly and carefully. But how many somewhat more mundane articles, that do not receive this scrutiny, suffer from similar if not as radical problems? Begley and Ellis point to more insidious problems: cheery-picking data (experiment works once out of ten tries, publish the one dataset only); and poor description of research reagents, preventing the experiment from actually being reproduced with the identical materials.

Current efforts to deal with reproducibility include direct data citation [5-8] and resource citation [9,10]. Groups working in these areas are making intensive use of semantic technologies to develop solutions. The ELIXIR pilot project FAIRPORT, to develop a common web services interface for biomedical databases, is an example of this kind of approach, and will make the tasks of newly-emergent "data publications" in the scientific literature, such as *Nature Scientific Data*, easier.

Proposals for various forms of semantic “fact” or “assertion” extraction such as that of Groth et al. [11,12] or the article in this issue by Marcondes et al. must generally assume as an underlying base for reasoning, that the extracted material is sound. The questions raised on reproducibility and integrity of data show that in some cases they may not be, and that sorting out ways to improve data integrity in science publications are a very necessary complement to semantics-based reformulation of portions of the material for ease of interoperation or search and retrieval.

Citations and cited claims can also be at issue. Greenberg [13,14] and others [15,16] studying claim and citation networks have found that citations – as is anecdotally well-known – are often corrupted in various ways. Greenberg [13,14] studied the selection pressure on citation-based claim strength “evolving” through a chain from original research to review articles. He found that qualifiers tended to be removed from claims as they were successively cited – in some cases all the way from “we hypothesize” and “it may be that” to “it has been shown that” – without the introduction of any actual confirmatory data along the way. Other researchers [15,16] have found extensive use of copied citations where it is not possible that the citing researchers actually read the relevant articles.

Greenberg’s approach to constructing a claim network is too labor intensive to be brought into general “production” use. However other authors [17] have developed methods for creating entire argument graphs, such as the Micropublication vocabulary (http://purl.org/mp), which can be adapted to pre- and post-publication peer review, if treated as stand-off annotation, using ontologies such as W3C Open Annotation [18]. Such argumentation models require characterization of the purpose of the citation – is it citing material as support, or is it challenging or discussing the cited material.

The articles by Angrosh et al. and Ciccarese et al. in this issue, both deal with this problem. Angrosh and colleagues developed a method for automatically classifying the type citations by characterizing their context. Ciccarese et al. used David Shotton’s Citation Typing Ontology (CiTO)[19] and the Peroni et al. FaBIO bibliographic ontology[20], to introduce a standardized method of inter-claim/inter-hypothesis relationship to the SWAN model [21,22] of hypotheses and claims. Methods such as these will be essential to develop robust ways of determining the real lineage of scientific claims, and the relationship of arguments, in an automated manner.

It has often been reported that scientific publication volumes are intractable. Hunter et al. for example, reported [23] that growth in Medline/PubMed articles has been double-exponential in the recent period. Improvements to queryability is of articles is one method to address this problem. The Marcondes et al. article in this issue exploits the “high-leverage” point at final article submission time, to require a fairly simple software-supported semantic enhancement of each article by its author, to this end. An important insight of this article is this point of leverage. The question in practice, as some publishers have discovered, is how to keep authors from “gaming the system” – as they have incentive to be published, not to be correctly annotated from an ontological perspective [24]. It may be that feedback from author citation metrics could help establish a better incentive for proper annotation term choice, assuming that authors can be kept in mind of the relationship between “findability” and citation counts.

As a final note, a very significant development since the articles in this Special Issue were accepted, has been the appearance of standardized interoperable methods for representing semantic annotation on linear text documents [25-27]. The W3C Open Annotation Model (OA), developed in a large (100+member, 50+ organizations) Community Group, is now on standards track. At this writing several colleagues are attending a workshop on using OA and various software tools for post-publication peer review. This model is intriguing and may be successful because it can integrate new semantic models into the current publication process as transparent semantic “overlays” on the existing linear document.

As noted in [25], the OA model is stand-off annotation and can be aggregated and mashed up independently of the site it was originally generated on. It is a first-class web object. This will allow new semantic models to be introduced to the scientific communications ecosystem, in a “backward-compatible” way.

Perhaps establishing backward compatibility for new methods is a good way to move forward to the next generation of scientific publishing.

References

1. Begley CG, Ellis LM (2012) Drug development: Raise standards for preclinical cancer research. Nature 483: 531-533.

2. Prinz F, Schlange T, Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 10: 712.

3. Obokata H, Kojima K, Westerman K, Yamato M, Okano T, et al. (2011) The potential of stem cells in adult tissues representative of the three germ layers. Tissue Eng Part A 17: 607-615.

4. Obokata H, Sasai Y, Niwa H, Kadota M, Andrabi M, et al. (2014) Bidirectional developmental potential in reprogrammed cells with acquired pluripotency. Nature 505: 676-680.

5. CODATA (2013) Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation. Data Science 12: 1-75.

6. Altman M, King G (2006) A Proposed Standard for the Scholarly Citation of Quantitative Data. DLib Magazine 13.

7. Force11 (2014) Joint Declaration of Data Citation Principles. http://force11.org/datacitation

8. Uhlir P (2012) For Attribution - Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop (2012) The National Academies Press. 220 p.

9. Helsby MA, Fenn JR and Chalmers AD (2013) Reporting research antibody use: how to increase experimental reproducibility [v2; ref status: indexed, **http://f1000r.es/1np]** *F1000Research* 2013, **2**:153 (doi: 10.12688/f1000research.2-153.v2)

10. Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, et al. (2013) On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ 1: e148. http://dx.doi.org/10.7717/peerj.148

11. Groth P, Gibson A, Velterop J (2010) The Anatomy of a Nano-publication. Information Services and Use 30: 51-56.

12. Schultes E, Chichester C, Burger K, Kotoulas S, Loizou A, et al. (2012) The Open PHACTS Nanopublication Guidelines V1.8. EU Innovative Medicines Initiative - Open PHACTS Project.

13. Greenberg SA (2009) How citation distortions create unfounded authority: analysis of a citation network. British Medical Journal 339: b2680.

14. Greenberg SA (2011) Understanding belief using citation networks. Journal of Evaluation in Clinical Practice 17: 389-393.

15. Ramos M, Melo J, Albuquerque U (2012) Citation behavior in popular scientific papers: what is behind obscure citations? The case of ethnobotany. Scientometrics 92: 711-719.

16. Simkin MV, Roychowdhury VP (2005) Stochastic modeling of citation slips. Scientometrics 62: 367-384

17. Clark T, Ciccarese P, Goble C (submitted) Micropublications: a Semantic Model for Claims, Evidence, Arguments and Annotations in Biomedical Communications. Journal of Biomedical Semantics. http://arxiv.org/abs/1305.3506

18. Sanderson R, Ciccarese P, Sompel HVd, Bradshaw S, Brickley D, et al. (2013) W3C Open Annotation Data Model, Community Draft, 08 February 2013. World Wide Web Consortium.

19. Shotton D (2010) CiTO, the Citation Typing Ontology. J Biomed Semantics 1 Suppl 1: S6. http://www.ncbi.nlm.nih.gov/pubmed/20626926

20. Shotton D, Peroni S, Ciccarese P, Clark T (2011) FaBiO, the FRBR-aligned Bibliographic Ontology, version 1.2.

21. Ciccarese P, Wu E, Wong G, Ocana M, Kinoshita J, et al. (2008) The SWAN biomedical discourse ontology. J Biomed Inform 41: 739-751.

22. Ciccarese P, Ocana M, Clark T (2009) Semantic Web Applications in Neuromedicine (SWAN) Ontology, W3C Interest Group Note 20 October 2009. World Wide Web Consortium.