Using Nanopublications to Incentivize the Semantic Exposure of Life Science Information

Mark Thompson and Erik A. Schultes

Leiden University Medical Center

Abstract. The growing rate of data production in the life sciences creates an urgent need for semantic integration of information. Although the development of tools and infrastructure will make semantic data exposure easier with time, presently the effort associated with creating linked data remains largely unrecognized by peer-review processes, publishers, and promotion committees. Here, we describe a novel data publishing framework called nanopublications that provides incentives for researchers to expose their data in semantic form. A nanopublication is the smallest unit of publishable information and is composed of an assertion (a semantic triple subject-predicate-object combination) and provenance metadata such as personal and institutional attribution (which also uses triples). As RDF named graphs, nanopublications are fully interoperable and machine readable, and need not be tethered to centralized databases, research articles or other schema for their retrieval and use. Hence, individual nanopublications can be cited and their impact tracked, creating powerful incentives for compliance with open standards and driving data interoperability.

1 Background

Large, harmonized datasets, especially from heterogeneous sources, promise to accelerate discovery in the life sciences, and offer new approaches to managing intrinsically complex biomedical systems [1]. However, the channels for data consumption have not scaled with data production leading to the loss of valuable data from scientific discourse [1, 2]. Given the magnitude and diversity of data production in the life sciences, the identification of trends and the inference of novel and relevant associations demands automated approaches to analysis and reasoning. In turn, this requires the automatic and universal interoperability of data [3]. Semantic technologies have emerged that effectively address these issues, but the legacies of scholarly communication continue to preempt efforts of data integration [4]. The narrative research article has been, for over a century, the accepted unit of attribution and scientific productivity. This made sense when typical datasets were small enough to be included in the research article itself (as tables or figures). However, as data production becomes increasingly automated, large-scale datasets must necessarily be hosted independently of the research article [5, 6]. In response, a dynamic ecosystem of technological solutions to large-scale data deposition, archival, persistence, licensing, access and

attribution has emerged[1]. Yet, no consensus around data representation, protocols for data linking or citation has crystallized among the research community or publishers. Thus, the lack of data interoperability continues to persist as a sociological, rather than as a technological problem.

2 What is a nanopublication?

Since 2009, the Dutch BioSemantics Group [7] has been developing a data format standard that scales with the demands of Big Data. This standard, called nanopublications [5, 8, 9], attaches to individual datum provenance metadata such that data no longer need to be tethered to centralized databases, research articles or other schema for their retrieval and use. Furthermore, exploiting off-the-shelf semantic technology, nanopublications are fully interoperable and machine readable. Hence, individual nanopublications can be cited and their impact tracked, creating incentives for individuals and institutions to exchange appropriate data.

Nanopublication packages individual datum as citable, stand-alone publications using semantic representations. Nanopublication is a schema on top of existing semantic technology using controlled vocabularies and ontologies. A nanopublication has two parts: the assertion (datum) and provenance (metadata). The assertion and provenance are RDF named graphs composed of semantic triples (subject-predicate-object combinations) [10]. The assertion describes a minimal unit of actionable scientific information such as a controlled observation (from the field or the laboratory) or a simple hypothesis (that can later be tested). The provenance describes how the assertion came to be, and includes both supporting information (e.g., context, parameter settings, a description of methods) and attribution information including fine-grained acknowledgment of institutions supporting the work, funding sources and other information like date and time stamps and certification [11–13]. A nanopublication represents the smallest unit of actionable information and combines both the technical solution for interoperability (semantic web representations) with the incentives (attribution) as a single publishable unit.

3 How to use nanopublications?

Creating a nanopublication requires a one-time effort to model the scientific assertion and provenance as RDF named graphs. After submission to an open, decentralized nanopublication store (essentially a triple store), nanopublications will be available as both human-readable and machine-readable information and will be fully interoperable under semantic queries and to automated inferencing engines. Nanopublications can be used to expose any data type whatsoever, including quantitative and qualitative data, experimental data as well as hypotheses, novel or legacy data and even negative results that usually go unpublished.

[1] Some examples are available at www.datadryad.org, www.foaf-project.org, www.thedatahub.org, www.datamarket.com, www.thedata.org and www.gigasciencejournal.com

The nanopublication framework can be used to expose data streams from curated databases as well as from instrumentation (sensors) and communication sources (internet transactions, email, video, click streams, or other digital sources available today and in the future). As a data publishing framework, nanopublications are meant to augment (not replace) traditional narrative research articles, although nanopublications can be used to expose individual assertions from narrative text.

By linking assertions and provenance using semantic representations, not only do data become interoperable, but their value can be independently estimated. Nanopublications provide a common currency for the exchange of data and thus allow crowd sourced or market-driven assignment of value to individual datum [4, 5, 14–17]. This is in contrast to traditional peer-review which has not scaled with the demands of data production and increasingly shows signs of bias and failure [18–22]. Based on this estimated value, nanopublications can be filtered and prioritized for the purposes of search and inclusion in automated inferencing algorithms. Large networks of custom nanopublication mash-ups from diverse sources can be constructed and searched for novel (implied) associations that would otherwise escape the human reasoning. Indeed, newly discovered associations can themselves be represented and shared as nanopublications. In turn, the value of individual datum can be translated into citation metrics, measures of scientific impact and other professional and economic indicators incentivizing interoperability and sharing [14].

References

1. Wild, D.J.: Mining large heterogeneous data sets in drug discovery. Expert Opinion on Drug Discovery **4**(10) (2009) 995–1004
2. `http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html`
3. Sansone, S.A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.A., Copeland, J., Das, S., De Daruvar, A., De Matos, P., Dix, I., Edmunds, S., Evelo, C.T., Forster, M.J., Gaudet, P., Gilbert, J., Goble, C., Griffin, J.L., Jacob, D., Kleinjans, J., Harland, L., Haug, K., Hermjakob, H., Sui, S.J.H., Laederach, A., Liang, S., Marshall, S., McGrath, A., Merrill, E., Reilly, D., Roux, M., Shamu, C.E., Shang, C.A., Steinbeck, C., Trefethen, A., Williams-Jones, B., Wolstencroft, K., Xenarios, I., Hide, W.: Toward interoperable bioscience data. Nature Genetics **44**(2) (2012) 121–126
4. Marx, V.: My data are your data. Nat Biotech **30**(6) (June 2012) 509–511
5. Mons, B., van Haagen, H., Chichester, C., Hoen, P.B.t., den Dunnen, J.T., van Ommen, G., van Mulligen, E., Singh, B., Hooft, R., Roos, M., Hammond, J., Kiesel, B., Giardine, B., Velterop, J., Groth, P., Schultes, E.: The value of data. Nat Genet **43**(4) (April 2011) 281–283
6. World Economic Forum: Big Data , Big Impact : New Possibilities for International Development. Agenda (2012) 0–9
7. `http://www.biosemantics.org`
8. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use **30**(1) (2010) 51–56

9. http://www.nanopub.org
10. Klyne, G., Carroll, J.J.: Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report
11. Taylor, C.F., Field, D., Sansone, S.A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C.A., Binz, P.A., Bogue, M., Booth, T., Brazma, A., Brinkman, R.R., Michael Clark, A., Deutsch, E.W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., Grimes, G., Hancock, J.M., Hardy, N.W., Hermjakob, H., Julian, R.K., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., Kuiper, M., Novere, N.L., Leebens-Mack, J., Lewis, S.E., Lord, P., Mallon, A.M., Marthandan, N., Masuya, H., McNally, R., Mehrle, A., Morrison, N., Orchard, S., Quackenbush, J., Reecy, J.M., Robertson, D.G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R.H., Schober, D., Smith, B., Snape, J., Stoeckert, C.J., Tipton, K., Sterk, P., Untergasser, A., Vandesompele, J., Wiemann, S.: Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotech **26**(8) (August 2008) 889–896
12. http://isa-tools.org
13. http://www.w3.org/2004/02/skos
14. Bartolini, C., Vukovic, M.: Crowdsourcing human mutations. (2011)
15. Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., Weeber, M., den Dunnen, J., van Ommen, G.J., Musen, M., Cockerill, M., Hermjakob, H., Mons, A., Packer, A., Pacheco, R., Lewis, S., Berkeley, A., Melton, W., Barris, N., Wales, J., Meijssen, G., Moeller, E., Roes, P.J., Borner, K., Bairoch, A.: Calling on a million minds for community annotation in WikiProteins. Genome biology **9**(5) (January 2008) R89
16. Hoffmann, R.: A wiki for the life sciences where authorship matters. Nat Genet **40**(9) (September 2008) 1047–1051
17. Oprea, T.I., Bologa, C.G., Boyer, S., Curpan, R.F., Glen, R.C., Hopkins, A.L., Lipinski, C.A., Marshall, G.R., Martin, Y.C., Ostopovici-Halip, L., Rishton, G., Ursu, O., Vaz, R.J., Waller, C., Waldmann, H., Sklar, L.A.: A crowdsourcing evaluation of the NIH chemical probes. Nature Chemical Biology **5**(7) (2009) 441–447
18. Miller, A., Barwell, G.: Science and Technology Peer review in scientific publications Eighth Report of Session 201012. (July) (2011)
19. Mullard, A.: Reliability of 'new drug target' claims called into question. Nat Rev Drug Discov **10**(9) (September 2011) 643–644
20. Prinz, F., Schlange, T., Asadullah, K.: Believe it or not: how much can we rely on published data on potential drug targets? Nature reviews. Drug discovery **10**(9) (September 2011) 712
21. Booth, B.: Academic Bias & Biotech Failures. http://lifescivc.com/2011/03/academic-bias-biotech-failures/ (2011)
22. Fanelli, D.: How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. PLoS ONE **4**(5) (2009) e5738