

# Towards Computational Evaluation of Evidence for Scientific Assertions with Nanopublications and Cardinal Assertions

Andrew Gibson<sup>1</sup>, Jesse C.J. van Dam<sup>1,2</sup>, Erik A. Schultes<sup>1</sup>, Marco Roos<sup>1</sup>, and Barend Mons<sup>1,3</sup>

<sup>1</sup> Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands

<sup>2</sup> Wageningen University, Wageningen, the Netherlands

<sup>3</sup> Netherlands Bioinformatics Center, Nijmegen, the Netherlands

**Abstract.** On the Web, it is possible for anyone to publish linked open data as RDF. Whilst this has huge potential to benefit data integration efforts, it highlights challenges of assessing data quality and trust. Nanopublication is an approach to data and knowledge publication in which assertions are individually encoded in RDF along with details about provenance, context and attribution. Collectively these details form a body of evidence for (or against) an assertion, which can be used as quality and trust criteria during data integration. In this position paper, we highlight the features of the Nanopublication specification that can be used as quality and trust criteria for life science data. We introduce the concept of cardinal assertions; assertions that are derived from the aggregation of multiple nanopublications to give an evidence value. We also identify a role for cardinal assertions in the evolution of evidence over time, supporting the re-evaluation of data and hypotheses.

## 1 Introduction

As the corpus of life science knowledge grows, along with the increasing amount of structured and unstructured life science data available on the Web, one of the challenges faced by life science researchers is the evaluation of evidence for biological assertions [1, 2]. Even simple prerequisite tasks of compiling sets of, for example, functional annotations of genes, protein-protein interactions or drug-target associations remain technically challenging, with relevant data often being distributed over different databases. Some of the overhead of life science data integration is being reduced by the increasing use and coverage of bio-ontologies that provide common terms, semantic types and properties for data annotation, classification and linking [3]. Semantic Web standards and Linked Data approaches can reduce the technical overhead of data publishing and integration, though there are still many unresolved issues [4]. Many life science datasets have been converted to an RDF representation, with several warehouses of linked life science datasets available [5, 6]. Efforts aimed at RDF data warehousing of life

science data are showing signs of becoming more targeted to specific research questions, as shown by LODD [7] and SLAP [8].

Data integration is not only technically challenging, but the result is also subjective, as data sources are curated to different standards and integrators have different motivations. In a scientific context, where quality and trust are paramount [9], it is important to be able to discover what has been integrated, and why. Annotations of assertions, and the methods used to derive them, are important factors in deciding their scientific quality. In many databases, bibliographic references are associated with assertions, which gives a researcher a broad indication of where they can find out more about their provenance. The importance of increasing the resolution of evidence for an assertion is apparent in the use of evidence codes by curated databases, such as those used for Gene Ontology annotations [10] or by the BioGRID Interaction Database [11].

In this position paper we consider how evidence is currently used to support scientific assertions in the life sciences, and the potential impact that the structured provenance for assertions provided by the Nanopublication specification might have on the way research data is published, used and integrated.

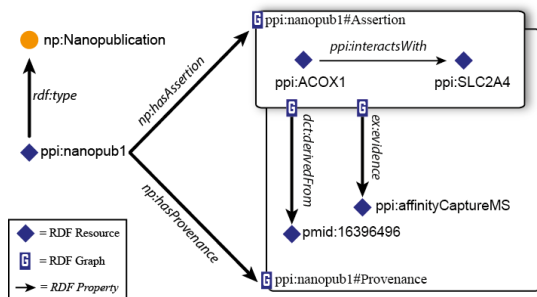
## 2 Nanopublications and Life Science Data

A lack of fundamental information about biological assertions, such as their source and the date that they were last updated, can make it difficult to assess their quality [12]. Nanopublications<sup>4</sup> have been proposed as a way to encode and publish individual assertions using Semantic Web and Linked Data principles [13, 14, 15]. Authorship and timestamp metadata are mandatory components of Nanopublications, ensuring a basic level of trust for any Nanopublication. Each Nanopublication contains exactly one assertion, which may be encoded as one or more RDF triples in a named graph; the assertion graph. Further provenance of the assertion is specified as annotations of the assertion graph in a second named graph; the provenance graph. Figure 1 shows components of a Nanopublication that might be published by a database containing protein-protein interactions<sup>5</sup>. In this case a reference to a publication and a laboratory technique, elements that are commonly found as supporting evidence, are encoded as provenance. Other types of assertion such as functional annotations of genes or drug-target associations can be represented just as easily in the Nanopublication framework. With the structure of Nanopublications, it is possible to list the evidence for an assertion by querying for the provenance of the assertion across known Nanopublications. Simply listing all of the associated evidence is a lightweight approach, but it can be used to make an assessment of the level of evidence available. Like with evidence codes, researchers and application developers will be able to make broad but valuable distinctions between assertions that have no evidence, are

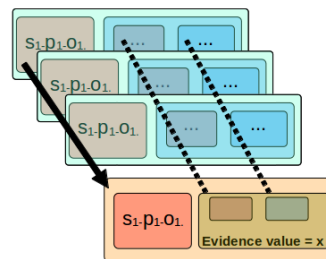
---

<sup>4</sup> Fine details of Nanopublication structure and content are yet to go through a full standardization process, so we only highlight broadly agreed principles here.

<sup>5</sup> The example is adapted from an entry in the BioGRID database



**Fig. 1.** A partial Nanopublication that might be published in the namespace of a protein-protein interaction database (ppi). The assertion graph (top right) contains the assertion that the proteins ACOX1 and SLC2A4 interact. The provenance graph contains statements that the assertion is derived from a paper, and that the type of evidence is affinity capture mass spectrometry. Attribution statements are not shown. It is assumed that extra protein information, such as species, is specified elsewhere in the database and is accessible for querying.



**Fig. 2.** Structure of an aggregate nanopublication containing a cardinal assertion. Black arrow indicates the creation of summary evidence by collapsing information from aggregated nanopublications. The dotted lines indicate the possibility to access the supporting information of the original nanopublications.

predictions, author statements, and assertions that have been derived experimentally. In the context of RDF warehouses, the consistent representation of evidence in provenance graphs enables researchers to include queries that reveal the level of support for interesting connections.

### 3 Cardinal Assertions

So far we have considered how Nanopublications can provide a means to collect and integrate assertions published on the Web. Beyond integration, evidence is also often aggregated computationally such that one value of overall confidence is produced for an assertion. An example of evidence aggregation is provided by StringDB [16], which contains data about associations between proteins. For each type of evidence, a probabilistic score indicates how likely a functional association is considered to be. In terms of Nanopublications, these associations could be published as assertions that are supported by the type of evidence and confidence score as part of the provenance graph. In addition, a combined score is calculated that takes all of the evidence into account. This kind of combination and evaluation of evidence from different sources is not unique to StringDB. Another example of aggregation comes from neXtProt [17], where evidence is assigned manually by a Gold, Silver or Bronze rating. These practices have prompted us to consider how such aggregated evidence fits into the Nanopublication framework.

Here we define a cardinal assertion. Cardinal assertions are the result of aggregating the evidence associated with identical assertions to produce a new measure that represents a consensus of that evidence (Figure 2). A cardinal assertion comes with two specific criteria: 1) it links to the Nanopublications containing the source assertions 2) the method of generating the overall confidence score should be clearly defined and linked to. The links should be encoded in a provenance graph and published in the Nanopublication format. These criteria ensure that the component assertions and their provenance can be identified at a later date, and that the aggregation method can be evaluated and repeated. As cardinal assertions use the Nanopublication approach, they also inherit the ability to be published as citable entities for which the aggregator can get direct attribution when they are used or referenced. A set of cardinal assertions represents the judgement of a data aggregator as to what available evidence should contribute to the state of the art of knowledge and by how much. For instance, StringDB could aggregate evidence from Nanopublications about interactions and expose its combined evidence scores as Cardinal assertions in RDF. Through the creation of a set of cardinal assertions, a data aggregator removes the need to repeatedly evaluate all of the Nanopublications in a set.

## 4 Future Perspective I: Curation of Cardinal Assertions

With open access to Nanopublications, researchers have the opportunity to collect and aggregate them into Cardinal assertions to summarize evidence. An authoritative collection of cardinal assertions could represent a valuable, structured interpretation of the state-of-the-art of knowledge in a particular domain of the life sciences. Such data sources could form the foundation of a new level of trust and quality of data used to form new hypotheses and analyze data that is more in line with that expected from the scientific method. Here we consider some of the aspects of managing Nanopublications and cardinal assertions that an authoritative data provider might need to consider.

A data aggregator will be able to curate nanopublications as trustworthy and reject others based on their provenance, for example where provenance is insufficiently described, or indicates that an assertion is derived from a methodology that has later been identified as unreliable. Different aggregators may have different parameters for what constitutes trustworthy evidence, for example, whether text-mining predictions are included. The ability to make these decisions is provided by the content of the Nanopublication provenance graph.

If a data aggregator chooses not to include a particular (type of) Nanopublication in their dataset, they also have the opportunity to publish this fact as further data so that 1) others can see that a particular Nanopublication has been rejected or not considered by them and 2) the reason for its absence is clear. Similarly, accepting a particular Nanopublication during the curation process represents an endorsement of its quality. For consumers of the data it is important to be able to establish which nanopublications were or were not accepted by the data provider, and why, so this should be a queryable part of the

structured data. From this management perspective, we are working on a system that can keep track of which nanopublications have been accepted and which have not. We intend to integrate this functionality with an RDF triplestore to transparently add metadata about this process.

## 5 Future Perspective II: Cardinal Assertions and the Evolution of Knowledge

Life science knowledge is constantly evolving as new experiments are performed, new data are produced and our ability to more accurately describe existing knowledge improves. This evolution is increasingly characterized by evidence derived from high-throughput experimentation and bioinformatics methods [18]. For example, it is reported that some ninety-eight percent of all Gene Ontology annotations are uncurated and inferred through in silico processes [2]. Interestingly the same study reports that the quality of these is higher than generally perceived, though this varies with the methods used and the type of assertion. This insight may change the evidence value of automated annotations and could cause aggregators to re-evaluate their trust of these assertions, which may previously have been rejected because of their evidence code [19].

A second example of knowledge evolution comes from the interpretation of high-dimensional datasets. For example, in 1998 Spellman *et al.* claimed to have elucidated the set of cell-cycle regulated genes of the yeast *Saccharomyces cerevisiae* in a landmark gene expression paper [20]. Alternative transcriptomic experiments have subsequently implicated different, but overlapping sets of genes. In addition, different data analysis methods produce different sets of genes from the same experimental data [21] leading to uncertainty about the meaning of the data [22].

If assertions like those described above were published as Nanopublications over time, a data aggregator can use the provenance information to recalculate the evidence values for their assertions and then publish them as Cardinal assertions. As evidence values change over time, it will influence the interpretation and use of those of cardinal assertions, which can be very significant if they were used in a computational analysis pipeline e.g. [23].

Nanopublications themselves have been proposed as immutable things; once they are published their content should not change. This is also true for cardinal assertions, in that as new evidence is taken into consideration, new editions of cardinal assertions should be generated. This will create a chain of assertions that represent the state of the overall evidence for a cardinal assertion over time. To be able to compare between editions through querying it is important that each can be accessed or reconstructed, and that the nanopublications that affected the evidence value are clear.

## 6 Discussion

As life science datasets are often curated to different standards, and descriptions of quality are sparse, integrators and users of integrated RDF datasets would have trouble measuring the quality of assertions in an aggregated dataset. Subsequently it would be difficult to query or inference over them with any degree of trust for hypothesis generation or data analysis. Poorly supported, erroneous or obsolete assertions would be difficult to identify. The Nanopublication specification was in part designed with this issue in mind, enabling trust through clear provenance statements for assertions. By introducing Cardinal assertions we hope to extend the functionality of Nanopublications for data providers and consumers who want to generate and use combined evidence. We acknowledge that Nanopublications are only a piece of a solution as the evaluation of scientific assertions can require deep insights into data provenance that may not be encoded with recognized standards by producers and providers of data. However, in this position paper we have shown potential applications of Nanopublications and Cardinal assertions to reinforce quality and trust aspects of shared life science data.

## References

- [1] Bell, M.J., Gillespie, C.S., Swan, D., Lord, P.: An approach to describing and analysing bulk biological annotation quality: a case study using UniProtKB. *Bioinformatics* **28**(18) (September 2012) i562–i568 PMID: 22962482 PMCID: PMC3436799.
- [2] kunca, N., Altenhoff, A., Dessimoz, C.: Quality of computationally inferred gene ontology annotations. *PLoS Computational Biology* **8**(5) (May 2012) PMID: 22693439 PMCID: PMC3364937.
- [3] Blake, J.A., Bult, C.J.: Beyond the data deluge: data integration and bio-ontologies. *Journal of biomedical informatics* **39**(3) (June 2006) 314–320 PMID: 16564748.
- [4] Goble, C., Stevens, R.: State of the nation in data integration for bioinformatics. *Journal of biomedical informatics* **41**(5) (October 2008) 687–693 PMID: 18358788.
- [5] Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics* **41**(5) (October 2008) 706–716 PMID: 18472304.
- [6] Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., Wild, D.J.: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC bioinformatics* **11** (2010) 255 PMID: 20478034.
- [7] Samwald, M., Jentzsch, A., Bouton, C., Kallese, C.S., Willighagen, E., Hajagos, J., Marshall, M.S., Prud’hommeaux, E., Hassenzadeh, O., Pichler, E., Stephens, S.: Linked open drug data for pharmaceutical research and development. *Journal of cheminformatics* **3**(1) (2011) 19 PMID: 21575203.
- [8] Chen, B., Ding, Y., Wild, D.J.: Assessing drug target association using semantic linked data. *PLoS Computational Biology* **8**(7) (July 2012) PMID: 22859915 PMCID: PMC3390390.

- [9] Gamble, M., Goble, C.: Quality, trust, and utility of scientific data on the web: Towards a joint model. In: Proceedings of the ACM WebSci'11. Volume Proceedings of the ACM WebSci'11., Koblenz, Germany (June 2011) 1–8
- [10] du Plessis, L., kunca, N., Dessimoz, C.: The what, where, how and why of gene ontology primer for bioinformaticians. Briefings in Bioinformatics **12**(6) (November 2011) 723–735 PMID: 21330331 PMID: PMC3220872.
- [11] Stark, C., Breitkreutz, B.J., Chatr-aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J.M., Winter, A., Dolinski, K., Tyers, M.: The BioGRID interaction database: 2011 update. Nucleic Acids Research **39**(Database issue) (January 2011) D698–D704 PMID: 21071413 PMID: PMC3013707.
- [12] Buza, T.J., McCarthy, F.M., Wang, N., Bridges, S.M., Burgess, S.C.: Gene ontology annotation quality analysis in model eukaryotes. Nucleic acids research **36**(2) (February 2008) e12 PMID: 18187504.
- [13] Mons, B., Velterop, J.: Nano-publication in the e-science era. In: Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse, Washington DC, USA (October 2009)
- [14] Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use **30**(1) (January 2010) 51–56
- [15] Mons, B., Haagen, H.v., Chichester, C., Hoen, P.B.t., Dunnen, J.T.d., Ommen, G.v., Mulligen, E.v., Singh, B., Hooft, R., Roos, M., Hammond, J., Kiesel, B., Giardine, B., Velterop, J., Groth, P., Schultes, E.: The value of data. Nature Genetics **43**(4) (2011) 281–283
- [16] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguéz, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L.J., von Mering, C.: The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research **39**(Database issue) (January 2011) D561–568 PMID: 21045058.
- [17] Lane, L., Argoud-Puy, G., Britan, A., Cusin, I., Duek, P.D., Evalet, O., Gateau, A., Gaudet, P., Gleizes, A., Masselot, A., Zwahlen, C., Bairoch, A.: neXtProt: a knowledge platform for human proteins. Nucleic Acids Research **40**(D1) (December 2011) D76–D83
- [18] Kell, D.B., Oliver, S.G.: Here is the evidence, now what is the hypothesis? the complementary roles of inductive and hypothesis-driven science in the post-genomic era. BioEssays: news and reviews in molecular, cellular and developmental biology **26**(1) (January 2004) 99–105 PMID: 14696046.
- [19] Rhee, S.Y., Wood, V., Dolinski, K., Draghici, S.: Use and misuse of the gene ontology annotations. Nature Reviews Genetics **9**(7) (July 2008) 509–515
- [20] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. Molecular biology of the cell **9**(12) (December 1998) 3273–3297 PMID: 9843569.
- [21] Lichtenberg, U.d., Jensen, L.J., Fausbll, A., Jensen, T.S., Bork, P., Brunak, S.: Comparison of computational methods for the identification of cell cycle-regulated genes. Bioinformatics **21**(7) (April 2005) 1164–1171
- [22] Futschik, M.E., Herzel, H.: Are we overestimating the number of cell-cycling genes? the impact of background models on time-series analysis. Bioinformatics **24**(8) (April 2008) 1063–1069
- [23] van den Berg, B.H.J., McCarthy, F.M., Lamont, S.J., Burgess, S.C.: Re-annotation is an essential step in systems biology modeling of functional genomics data. PloS one **5**(5) (2010) e10642 PMID: 20498845.