Information Services & Use 30 (2010) 51–56
DOI 10.3233/ISU-2010-0613
IOS Press

The anatomy of a nanopublication

Paul Groth [a], Andrew Gibson [b] and Jan Velterop [c,*]
[a] *Free University Amsterdam, Amsterdam, The Netherlands*
[b] *University of Amsterdam, Amsterdam, The Netherlands*
[c] *Concept Web Alliance, NBIC, Nijmegen, The Netherlands*

Abstract. As the amount of scholarly communication increases, it is increasingly difficult for specific core scientific statements to be found, connected and curated. Additionally, the redundancy of these statements in multiple fora makes it difficult to determine attribution, quality and provenance. To tackle these challenges, the Concept Web Alliance has promoted the notion of nanopublications (core scientific statements with associated context). In this document, we present a model of nanopublications along with a Named Graph/RDF serialization of the model. Importantly, the serialization is defined completely using already existing community-developed technologies. Finally, we discuss the importance of aggregating nanopublications and the role that the Concept Wiki plays in facilitating it.

Keywords: Semantic Web, rich RDF-triples, disambiguation, publication

1. Introduction

Consider two distinct concepts, malaria and mosquitoes, and a relationship of 'is transmitted by' that together form a statement:

Malaria is transmitted by mosquitoes.

On its own, this statement – a scientific assertion – exists many times over in the published literature. The statement itself is what is common to all of the sources of the statement, but the statement can only be validated scientifically if you take into consideration its context. Traditionally, the context of a scientific statement is implicit in its immediate environment: the scientific publication. The details of the publication provide the different kinds of metadata that are required before it can be considered credible enough to be used in a new hypothesis.

However, this means that statements need to be taken into account in full view of their context, which increasingly becomes a practical impossibility, as the number of relevant published articles is overwhelming the typical research scientist. Not only that, the Semantic Web is providing the platform in which people can more easily generate statements, extract statements from existing literature and share them in a way that will allow computational agents to discover, aggregate and interpret these statements. The value and advantages of this are clear, but the lack of contextual validation of these statements presents a problem for their use in building hypotheses. Ideally, the concepts in a statement and the statement itself will have some unique identity that connects each instance of an statement across the Web of (formally as well as informally) published material.

*Corresponding author: Jan Velterop, Concept Web Alliance, NBIC, Nijmegen, The Netherlands. E-mail: velterop@conceptweballiance.org.

It can be expected that the number of systems that facilitate the creation of statements will increase further. These will come in the form of both processes designed to extract statements from existing material, and systems that facilitate *de novo* statement creation. Newer standards like RDFa also facilitate this and integrate with current html documents.

The challenge now becomes: what needs to be done to put the context that was formerly provided by a document back in to a statement. In this paper we explore the extra components that would need to be available to reinforce the value of a statement to the point where it could in itself be considered a publication. Due to its small size relative to a full paper, this is termed a 'nanopublication'. We separate out goals from implementations and consider the applicability of current standards to requirements.

This paper serves a dual role. One role is to define a model for nanopublications and illustrate how existing Semantic Web technologies could be used to implement it. The second and perhaps more important role is to act as an impetus for discussion between the Web community, the research community, science publishers and the Concept Web Alliance around the concept of nanopublications. The Concept Web Alliance (CWA) is a non-profit organization whose mission is "to enable an open collaborative environment to jointly address the challenges associated with high volume scholarly and professional data production, storage, interoperability and analyses for knowledge discovery".[1]

2. Core model

Our core model addresses some key requirements that stem from existing publication practices in peer-reviewed journals and the need to aggregate information from distributed sources. Similar to standard scientific publications, nanopublications need to be citable, attributable and reviewable. Furthermore, they need to be easily curated. Nanopublications must be easily aggregated and identified across the Web. Finally, they need to be extensible to cater for new forms of both metadata and description.

We begin with a core set of definitions:

- Concept – a concept is the smallest, unambiguous unit of thought. A concept is uniquely identifiable.
- Triple – is a tuple of three concepts (subject, predicate, object).
- Statement – a triple that is uniquely identifiable.
- Annotation – a triple such that the subject of the triple is a statement.
- Nanopublication – a set of annotations that refer to the same statement and contains a minimum set of (community) agreed-upon annotations.
- S-Evidence – all the nanopublications that refer the same statement.

Figure 1 depicts the relationship between these definitions.

Within this model, different communities may require different sets of annotations beyond those that are core to the definition. This allows for the expression of different types of nanopublications, for example, curated, observational and hypothetical nanopublications, as suggested by [4].

This proposed model could be instantiated into a number of different formats. However, there are some basic requirements that the model places on any format:

- The ability to uniquely identify a concept.
- The ability to uniquely identify a statement.

[1] From the CWA Declaration, available at: http://www.nbic.nl/about-nbic/affiliated-organisations/cwa/declaration/.

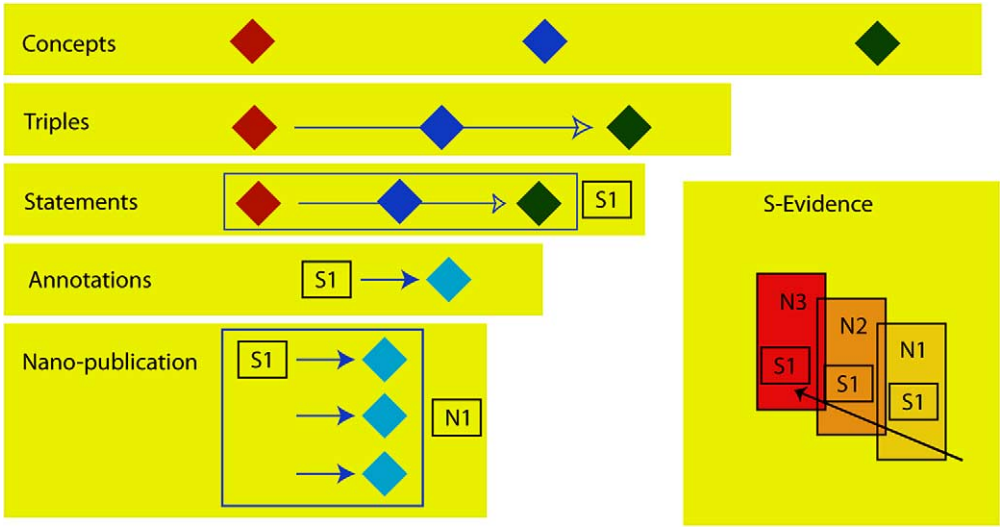

Fig. 1. The nanopublication model. (The colors are visible in the online version of the article.)

- The ability to refer to all uniquely identified entities.

We note that this could be satisfied by any number of formats. While using a common format is important, it is more important that the community come to an agreement on the vocabulary of annotations to be used in defining a nanopublication. We now discuss a possible realisation of this model using Semantic Web technologies.

3. A realization as Named Graphs

Named Graphs [1] is a simple extension to RDF adding the ability to assign a URI to a given RDF graph. Named Graphs are specifically designed with use cases similar to those posed by nanopublications in mind. In particular, Named Graphs were designed to support keeping track of provenance during aggregation and the definition of context for a particular graph. While Named Graphs are not yet a W3C standard they are widely supported by many implementations of the Semantic Web infrastructure (e.g., quad stores such as Virtuoso, 4store and NG4J).

The nanopublication model maps simply to Named Graphs.

- Each triple is an RDF-triple.
- Each statement is a separate Named Graph.
- Each annotation has as its subject the URI of a Named Graph.
- All annotations belonging to a nanopublication should be part of the same Named Graph.

Thus, the simplest nanopublication has two Named Graphs, one with a statement and another containing the annotations on that statement. While Named Graphs provide a convenient serialization for nanopublications, the key to enabling nanopublications to be aggregated is for their context to be well defined. We now discuss a possible set of annotations for the nanopublications.

4. Annotations

There has already been much work done on representing scientific discourse on the Web [3]. We propose to adopt wholesale wherever possible artifacts from that work. In particular, we believe that the SWAN series of ontologies [2] and its mapping to the SIOC [6] provide a comprehensive starting point. We extract a subset from these ontologies and extend where necessary with external ontologies.

From SWAN, we use the Scientific Discourse ontology and its requirements. Specifically, we define all core statements as a SWAN Research Statement. While SWAN enables one to describe complex associations between research statements to build a larger model of scientific discourse, we propose not to use the capabilities for nanopublications to decrease the overhead on aggregators. Instead, we use the provenance, annotation and versioning of the SWAN ontology.[2] Examples of the annotations provided are `importedFromSource` (identifies where the research statement was extracted from), `importedBy` (identifies what entity is responsible for importing an statement), `authoredBy` (identifies the author of a research statement). We refer readers to the ontology documentation for a complete list of annotations.

We note that SWAN extends FOAF [8], so people, organizations and software agents can be represented. Specifically, in order to understand a nanopublication, a system should understand the subclasses of FOAF Agents such as Person, Organization and Group.

5. Attribution, review, citation

Annotations provide a mechanism to describe information about a statement. For example, who authored the statement, when was the statement created, what software was used in creating the statement and so on. However, in a number of cases it is useful to be able to discuss a nanopublication as a whole, for example, to claim attribution on it, allow a reviewer to approve it, or to provide a way for people to vote for or cite a nanopublication. Here, we use attribution as an example.

While the provenance ontology from SWAN provides a reasonable set of information describing the annotations within a nanopublication, it does not yet provide a good mechanism for claiming the contents of a nanopublication.

To support this, we propose to use the Semantic Web Publishing ontology.[3] This ontology provides as `assertedBy` relationship, which relates a particular Named Graph to an entity (i.e., an authority). Thus, an entity can state that they asserted a nanopublication and thus claim it. Furthermore, this ontology provides the capability to express digital signatures on each of the graphs. This signature capability may be important in verifying claims.

There may be more than one nanopublication about the same statement (there often is). Through this 'asserted by' mechanism, it becomes easier to distinguish the origins of these different accounts of the same statement. Indeed, users (software or human agents) of a nanopublication may decide which accounts they trust and which they do not based on any number of heuristics. This notion of different views or accounts of the same statement is inspired by the Open Provenance Model [5].

We believe that attribution is an essential part to nanopublications; however, the community may decide that other metadata on nanopublications may be necessary, for example, reviews, or institutional association. Other uses may be to enable the construction of collections of nanopublications.

[2] Available at: http://swan.mindinformatics.org/spec/1.2/pav.html.

[3] Available at: http://www.w3.org/2004/03/trix/swp-1/.

```
@prefix swan: <http://swan.mindinformatics.org/ontologies/1.2/pav.owl> .
@prefix cw:   <http://conceptwiki.org/index.php/Concept>.
@prefix swp:  <http://www.w3.org/2004/03/trix/swp-1/>.
@prefix : <http://www.example.org/thisDocument#> .

:G1  = { cw:malaria cw:isTransmittedBy cw:mosquitoes }

:G2  = { :G1 swan:importedBy cw:TextExtractor,
         :G1 swan:createdOn "2009-09-03"^^xsd:date,
         :G1 swan:authoredBy cw:BobSmith }

:G3  = { :G2 ann:assertedBy cw:SomeOrganization }


:G9  = { :G1 ann:isApprovedBy cw:JohnSmith }
:G10 = { :G9 ann:isAssertedBy cw:ApprovalTrackingSystem }
```

Fig. 2. Example of nanopublication.

6. Example

To illustrate our model, Fig. 2 is a small example of nanopublication about the statement that malaria is transmitted by mosquitoes. Bob Smith authored the statement. It was imported by a text extractor and was created in September 2009. The nanopublication was asserted by *Some Organization*. The example uses TRIG syntax [9].

7. Aggregation and the Concept Wiki

The nanopublications and model should help facilitate the aggregation of fine-grained scientific information across the web. In the model we introduce the notion of S-Evidence, which is all the nanopublications that are about the same statement. A key role for aggregators will be to find, filter and combine all the evidence for a statement from a variety of nanopublications to ascertain the veracity of a statement. A benefit of separating statements from their various annotations is that it allows reasoning on only the statements themselves or on a condensed version of the annotations. A key to making S-Evidence practical is for publishers to use the same identifiers for statements and concepts.

However, in the model there is no requirement to use the same identifiers. Indeed, any Semantic Web resource can be used. Thus, to make aggregation easier, publishers should follow Linked Data principles by pointing to resources already available on the Web. To provide a repository of such resources, the CWA hosts the Concept Wiki. This wiki provides uniquely identifiable and unambiguous URLs for concepts. By referring to concepts on the Concept Wiki, publishers of nanopublications can facilitate their aggregation. Furthermore, the Concept Web Alliance will operate as an aggregator that takes nanopublications and makes their content available on the Concept Wiki. This aggregator will map from the resources used in a nanopublication to Concept Wiki concepts. Clearly, a nanopublication that already uses Concept Wiki concept identifiers will be better placed to be properly included in aggregated data. Thus, we introduce three types of nanopublications:

- Transformation Compatible – data that can be transformed to CWA format where a tool exists to perform the transformation.
- Format Compatible – these nanopublications use the CWA model and endorsed serialization nanopublications.

- Concept Wiki Compatible – these nanopublications are not only format compatible but also only use Concept Wiki identifiers.

Additionally, the Concept Wiki provides a place for users to easily create nanopublications. Finally, the Concept Wiki will follow the principles of Linked Data. It should also provide programmatic access to nanopublications following the format specified by the CWA (i.e., the successor to the one above).

We hope that our format would be suitable or even compatible with approaches such as aTags, a simple convention for representing annotated research statements with the SIOC vocabulary [7]. There are also tools that work with aTags that allow users to easily extract information from existing Web data. We would like to see such tools support nanopublications as well.

8. Conclusion

Here we have proposed an initial model and format for nanopublications. The format is based on existing community-produced ontologies and technologies. The role of the CWA-format working group is to specify a minimal common format for nanopublications that enables their aggregation and the correct preservation of the associated provenance. The CWA working group aims not to develop new specifications but instead to identify existing technology and formats that can be used for aggregating nanopublications.

Acknowledgements

We are grateful to the members of the CWA working group on nanopublications (http://www.myexperiment.org/groups/192) for their comments on this paper.

References

[1] J.J. Carroll, C. Bizer, P. Hayes and P. Stickler, Named graphs, provenance and trust, in: *International World Wide Web Conference*, Chiba, Japan, 2005, available at: http://www2005.org/cdrom/docs/p613.pdf.
[2] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg and T. Clark, The SWAN biomedical discourse ontology, *Journal of Biomedical Informatics* **41** (2008), 739–751.
[3] T. Groza, S. Handschuh, T. Clark, S.B. Shum and A.D. Waard, A short survey of discourse representation models, in: *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, Washington, DC, USA, 2009.
[4] B. Mons and J. Velterop, Nano-publication in the e-science era, in: *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, Washington, DC, USA, 2009.
[5] Open Provenance Model, available at: http://eprints.ecs.soton.ac.uk/18332/.
[6] A. Passant, P. Ciccarese, J.G. Breslin and T. Clark, SWAN/SIOC: aligning scientific discourse representation and social semantics, in: *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, Washington, DC, USA, 2009.
[7] M. Samwald and H. Stenzhorn, Simple, ontology-based representation of biomedical statements through fine-granular entity tagging and new web standards, in: *Bio-Ontologies 2009*, Stockholm, Sweden, 2009.
[8] The Friend of a Friend (FOAF) project, available at: http://www. foaf-project.org/.
[9] TriG, available at: http://www4.wiwiss.fu-berlin.de/bizer/TriG/.