The Anatomy of a Nano-publication

Paul Groth
VU University Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam,
The Netherlands
pgroth@few.vu.nl

Andrew Gibson
University of Amsterdam
Nieuwe Achtergracht 166, C-712
1018 WV Amsterdam
The Netherlands
a.p.gibson@uva.nl

Johannes Velterop
Concept Web Alliance / NBIC
9 Benfleet Close
Cobham, Surrey, KT11 2NR
United Kingdom
jan.velterop@nbic.nl

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 nano-publications and the role that the Concept Wiki plays in facilitating it.

Categories and Subject Descriptors

General Terms

Management, Documentation, Languages

Keywords

Keywords are your own designated keywords.

1. INTRODUCTION

Consider two distinct concepts, malaria and mosquitos, and a relationship of ‘is transmitted by’ that together form a statement:

malaria is transmitted by mosquitos

On its own, this statement exists many times over in 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 kind of metadata that are required before it can be considered credible enough to be used in a new hypothesis.

However, 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 advantages of this are clear, and ideally, the concepts in a statement and the statement itself will have some unique identity that connects each instance of a statement across the web of (formally as well as informally) published material.

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

The challenge now becomes; what needs to be done to put the context back in to a statement that was formerly provided by a document. 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. This is termed a nano-publication. 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 nano-publications 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 Health Care and Life Science community and the Concept Web Alliance around the concept of nano-publications. 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 and the need to aggregate information from distributed sources. Similar to standard scientific publications, nano-publications need to be citable, attributable, and reviewable. Furthermore, they need to be easily curated. Nano-publications 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.

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

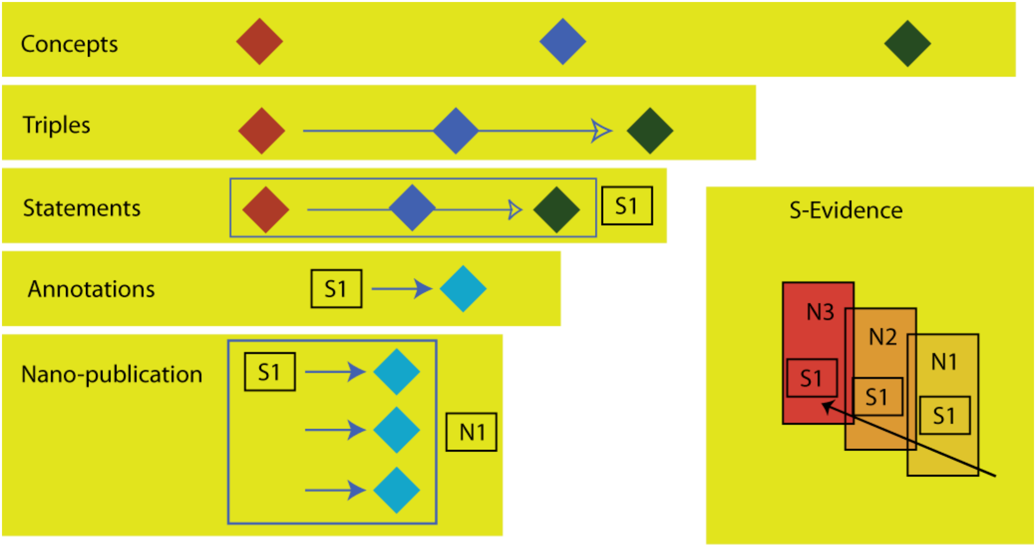

Figure 0: The Nano-publication Model

- 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 nano-publications, for example, curated, observational, and hypothetical nano-publications, as suggested by [1].

This proposed model can 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.
- The ability to refer to all uniquely identified concepts and statements.

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 nano-publication. We now discuss a possible realization of this model using Semantic Web technologies.

3. A REALIZATION AS NAMED GRAPHS

Named Graphs[2] 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 nano-publications 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 Virtuso, 4store, and NG4J).

The nano-publication 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 nano-publication should be part of the same Named Graph.

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

4. ANNOTATIONS

There has already been much work 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 [4] and its mapping to the SIOC [5] 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 nano-publications to decrease the overhead on aggregators. Instead, we use the provenance, annotation and versioning 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 a statement), `authoredBy` (identifies the author of a research statement). We refer readers to the ontology documentation for a complete list of annotations.

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

```
@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 }
```

Figure 2: Example Nano-publication

We note that SWAN extends FOAF [6], so people, organizations, and software agents can be represented. Specifically, to understand a nano-publication a system should understand the subclasses of FOAF Agent 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 useful to be able to discuss a nano-publication 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 nano-publication. Here, we use attribution as an example.

While the provenance ontology from SWAN provides a reasonable set of information describing the annotations within a nano-publication. It does not yet provide a good mechanism to for claiming the contents of a nano-publication.

To support this, we propose to use the Semantic Web Publishing ontology[3]. This ontology provides as `assertedBy` relationship, which relates a particular NamedGraph to an entity (i.e. an authority). Thus, an entity can state that they asserted a nano-publication and thus claim. 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 nano-publication about the same statement. 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 nano-publication may decide which accounts they trust and which they don't based on any number of heuristics. This notion of different views or accounts of the same statement is inspired by the Open Provenance Model [7].

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

6. EXAMPLE

To illustrate our model, Figure 2 provides is a small example nano-publication 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 nano-publication was asserted by Some Organization. The example uses TRIG syntax [8].

7. AGGREGATION AND THE CONCEPT WIKI

The nano-publications 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 nano-publications 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 nano-publications 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 nano-publications can facilitate their aggregation. Furthermore, the Concept Web Alliance will operate an aggregator that takes nano-publications and makes their content available on the Concept Wiki. This aggregator will map from the resources used in a nano-publication to Concept Wiki concepts. We are currently investigating approaches to implement this mapping. However, a nano-publication that already uses Concept Wiki concepts will be better placed to be aggregated. Thus, we introduce three types of nano-publications:

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

- Transformation Compatible – data that can be transformed to CWA format where a tool exists to perform the transformation.
- Format Compatible – these nano-publications use the CWA model and endorsed serialization nano-publications.
- Concept Wiki Compatible – these nano-publications are not only format compatible but also only use Concept Wiki URLs.

Additionally, the Concept Wiki provides a place for users to easily create nano-publications. Finally, the Concept Wiki will follow the principles of Linked Data. Additionally, it should provide programmatic access to nano-publications 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 [9]. There are also tools that work with aTags that which allow users to easily extract information from existing Web data. We would like to see such tools support nano-publications as well.

8. CONCLUSION

Here, we have proposed an initial nano-publication model, a format instantiation, and how the Concept Wiki can be used to facilate aggregation. 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 nano-publications 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 nano-publications.

Finally, the role of 'traditional' publications has always been a combination of record keeping [10] and knowledge transfer. The sheer volume of science articles published every day makes the efficacy of the latter part of this traditional role all but disappear. Nano-publications, with their attributes of publications, make a separation of these two article functions possible, and yet maintain their natural connectedness. With the 'macro-publication' still being the version of record, the nano-publications derived from them can be efficient vehicles for knowledge dissemination and large-scale aggregation (including with nano-publications from sources other than peer-reviewed published articles), due to their machine-readable characteristics.

9. ACKNOWLEDGMENTS

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

10. REFERENCES

[1] B. Mons and J. Velterop, "Nano-Publication in the e-Science Era," *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, 2009.

[2] J.J. Carroll, C. Bizer, P. Hayes, and P. Stickler, "Named graphs, provenance and trust," *International World Wide Web Conference*, 2005.

[3] T. Groza, S. Handschuh, T. Clark, S.B. Shum, and A.D. Waard, "A Short Survey of Discourse Representation Models," *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, 2009

[4] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, and T. Clark, "The SWAN biomedical discourse ontology.," *Journal of biomedical informatics*, vol. 41, 2008, pp. 739-51.

[5] A. Passant, P. Ciccarese, J.G. Breslin, and T. Clark, "SWAN/SIOC: Aligning Scientific Discourse Representation and Social Semantics," *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, 2009.

[6] The Friend of a Friend (FOAF) project. Available from: http://www. foaf-project.org/.

[7] Open Provenance Model. http://eprints.ecs.soton.ac.uk/18332/

[8] TriG. http://www4.wiwiss.fu-berlin.de/bizer/TriG/

[9] M. Samwald and H. Stenzhorn, "Simple, ontology-based representation of biomedical statements through fine-granular entity tagging and new web standards," *Bio-Ontologies 2009,* 2009.

[10] "Keeping the Minutes of Science" – in: Proceedings of Electronic Libraries and Visual Information Research (ELVIRA) Conference, Aslib, London, No. 2-14 May 1995.