Broadening the Scope of Nanopublications

Tobias Kuhn$^{1,2}$, Paolo Emilio Barbano$^3$, Mate Levente Nagy$^4$, Michael Krauthammer$^{4,1}$

$^1$Department of Pathology, Yale University
$^2$Chair of Sociology, in particular of Modeling and Simulation, ETH Zurich
$^3$Department of Mathematics, Yale University
$^4$Program for Computational Biology and Bioinformatics, Yale University

ESWC 2013, Montpellier (France)
29 May 2013
Motivation

The key problem of the current system of scholarly communication is that it is centered around narrative articles:

- They are good for individual consumption by human beings but very bad for aggregation or automatic processing
- They are very ineffective for sharing scientific information, especially in data-intensive sciences
- There are no rewards for providing, sharing, and maintaining datasets, software, and other digital artifacts

The current system is slow and inefficient.

Nanopublications have been proposed to solve these problems: they are minimal portions of scientific contributions in RDF.
Vision: Changing Scholarly Communication

Now
Narrative articles at the center

Future
Nanopublications at the center

Images from Mons et al. The value of data. *Nature genetics*, 43(4):281–283, 2011
Structure of Nanopublications

Nanopub0001

**Assertion:**
- Formalized scientific claim (or hypothesis)

**Provenance:**
- Reference to article, experimental methods, etc.
- Who published it when and how

Tobias Kuhn, Yale / ETH Zurich

Broadening the Scope of Nanopublications
nanobrowser: Classical Nanopublication

Types: Nanopublication
Date: Fri, 08 Jun 2012 14:15:31 +0000
Authors: Herman van Haagen Erik Schultes
Creator: Zuotian Tatum

Assertion as formula

association_000001 type statistical-association .
association_000001 has-measurement-value association_000001_p_value .
association_000001 refers-to geneid:6448 .
association_000001 refers-to omim:252900 .
association_000001 comment "This is a statistical association between a gene (Entrez gene id 6448) and a disease (OMIM 252900). It is generated by Concept Profile Matching method and has p-value of 0.0001." .
association_000001_p_value type probability_value .
association_000001_p_value has-value "6.56e-05" .

Provenance

isPartOf gene_disease_associations .
assertion wasDerivedFrom gene_disease_concept_profiles_1980_2010 .
assertion wasGeneratedBy gene_disease_concept_profiles_matching_1980_2010 .

http://nanobrowser.inn.ac
Ocean of Nanopublications

- Authors
- Users
- Curators

Structured data sources
Unstructured data sources

Ocean of nanopublications

Nanopublication portals
Aggregated views
Hypothesis generation
Network analyses
In silico experiments

...bots

1 2 3 4 5 6
Proposed Extension 1: Informal Assertions

Nanopub0012

Assertion:

Malaria is transmitted by mosquitoes.

Provenance:

opm:wasDerivedFrom d:DataSourceX
cito:cites n:nanopub0042
dc:created "2013-01-01"
pav:createdBy p:Isabelle_Dubois
dc:isPartOf c:NanoPubCollection1

Assertion:

- Informal English sentence
- Sentences are independent entities and represented by URIs:
  
  http://purl.org/aida/Malaria+is+transmitted+by+mosquitoes.
Levels of Formalization

- **Informal** (only an AIDA sentence)
- **Underspecified** (formal representation for part of the sentence)
- **Fully formal** (formal representation for the complete sentence)
Proposed Extension 2: Non-Scientific Assertions

Assertion:
- Meta-statement or other non-scientific assertion
- Can include opinions, social relations, introduction of new entities, ...

Nanopub0042

Assertion:
dc:created "2013-05-01"
pav:createdBy p:Giuseppe
Provenance:
p:Giuseppe npx:disagrees
a:Malaria+is+transmitted+by+...
nanobrowser: Nanopublication with Sentence

**Types:** Nanopublication

**Date:** Sat, 25 May 2013 17:08:00 +0000

**Authors:** GeneRIF-Bot

**Creator:** GeneRIF-Bot

---

**Assertion as sentence**

🔍 Tuberin protein levels are decreased in the frontal cortex of patients with Alzheimer's disease.

---

**Assertion as formula**

about taxonomy/9606.

about gene/7249.

(incomplete)

---

**Provenance**

isPartOf NanopubsFromGeneRIF.

wasDerivedFrom generifs_basic.gz.

citesAsSupportiveEvidence pubmed/16341938.
Related existing approaches: SWAN, EXPO, GeneRIF, BEL

Our approach differs from these approaches in the following respects:

- **Very broad application area** (science as a whole and beyond)
- **Sentences exist independently from authors** (no ownership of sentences)
- **Use of a controlled natural language**
- **Continuum** from informal over underspecified to fully formal statements
To fit into the nanopublication concept, the sentences of our approach should be **AIDA:**

- **Atomic:** a sentence describing one thought that cannot be further broken down in a practical way
- **Independent:** a sentence that can stand on its own, without external references like “this effect” or “we”
- **Declarative:** a complete sentence ending with a full stop that could in theory be either true or false
- **Absolute:** a sentence describing the core of a claim ignoring the uncertainty about its truth and how it was discovered (no “probably” or “evaluation showed”); typically in present tense

**Example**

The majority of patients with idiopathic REM sleep behavior disorder who develop a neurodegenerative disease develop Parkinson disease and Lewy body dementia.
Linking Scientific Claims

Possible relations:

- [CLAIM] is equivalent to / contradicts / is similar to [CLAIM]
- [PERSON] agrees with / disagrees with / challenges [CLAIM]
- [STUDY] provides (counter-)evidence for [CLAIM]

These relations can be published as nanopublications too!
nanobrowser: Opinions and Sentence Relations

Tuberin protein levels are decreased in the frontal cortex of patients with Alzheimer's disease.
http://purl.org/aida/Tuberin+protein+levels+are+decreased+in+the+frontal+cortex+of+patients+with+Alzheimer%27s+disease.

Nanopublications

GeneRIF401142.RAbbovajo3 Sat, 25 May 2013 17:08:00 +0000

Opinions

- Giuseppe Manchini disagrees.
- Isabelle Dubois agrees.

Related Sentences

- TSC2 is associated with Alzheimer's disease.
- The frontal lobe of Alzheimer patients has lower tuberin levels than in normal brains.

Tobias Kuhn, Yale / ETH Zurich
Publishing AIDA-Nanopubs

Here, you can publish your own nanopublication. Fill out the fields below and then press Publish.

Types: Nanopublication ExampleNanopub (for demonstration or testing)
Author: anonymous-12930929

Assertion as sentence
Write down your scientific assertion as an English sentence. Make sure that it is AIDA: atomic, independent, declarative, and absolute.

Scientists are able to efficiently produce high-quality nanopublications containing informal assertions in the form of AIDA sentences.

Provenance
You can cite an existing scientific article here via a DOI identifier.
citesAsSupportiveEvidence: 10.1007/978-3-642-38288-8_33

http://nanobrowser.inn.ac/publish
Evaluation

...
Evaluation of AIDA-Nanopublications

So far, the following aspects have been evaluated:

- How well can authors or curators express scientific findings as AIDA sentences? (channels 1 and 3)
- How well can AIDA sentences be automatically extracted from existing text sources? (channel 4)
- How well can similar AIDA sentences be automatically clustered? (channel 6)
User Study Design

Questionnaire-style online study:

- **16 participants**: Scientists with strong background in biology and/or medicine
- **Five short texts** (one or two sentences) from conclusion sections of structured abstracts of PubMed articles
- **Task**: Rewrite each short text as one or more AIDA sentences
- **Brief explanation** of the task and the AIDA concept

Example

**Original text**: The results of this study showed that the hepatic reticuloendothelial function is impaired in cirrhotic patients, but the degree of impairment does not differ between patients with and without previous history of SBP.

**AIDA 1**: The hepatic reticuloendothelial function is impaired in cirrhotic patients.

**AIDA 2**: The degree of hepatic reticuloendothelial function impairment does not differ between cirrhotic patients with and without previous history of SBP.
User Study Results

Quality of the sentences created within the user study:

- Total: 163 sentences (100%)
- Perfect: 114 sentences (70%)
- Typo etc.: 7 sentences (4%)
- Inaccurate: 10 sentences (6%)
- Not AIDA: 32 sentences (20%)
- Not atomic: 4 sentences (2%)
- Not independent: 5 sentences (3%)
- Not declarative: 3 sentences (2%)
- Not absolute: 25 sentences (15%)

Tobias Kuhn, Yale / ETH Zurich

Broadening the Scope of Nanopublications
Automatic Extraction

Evaluation of automatic extraction of AIDA-nanopubs:

- We used the GeneRIF dataset, which contains sentences describing the functions of genes and proteins
- Simple regular expressions to filter out sentences that are not AIDA-compliant
- Simple transformations on the sentences, such as dropping certain initial phrases

Example

Original GeneRIF sentence: We have established that LadC plays an important role in L. pneumophila infection.

Extracted AIDA sentence: LadC plays an important role in L. pneumophila infection.
Automatic Extraction Results

Quality of the sentences extracted from the GeneRIF dataset:

- Total: 250 sentences (100%)
- Perfect: 177 sentences (71%)
- Typo etc.: 8 sentences (3%)
- Not AIDA: 65 sentences (26%)
- Not atomic: 34 sentences (14%)
- Not independent: 15 sentences (6%)
- Not declarative: 15 sentences (6%)
- Not absolute: 30 sentences (12%)

Tobias Kuhn, Yale / ETH Zurich

Broadening the Scope of Nanopublications
Automatic Clustering

Evaluation of automatic clustering:

- **Sentences extracted from GeneRIF:** 119,088 unique sentences
- **Sentences from user study:** 94 unique sentences from five tasks
- **Results:** Sentences from a user study task were clustered almost exclusively (99.2%) with other sentences from the same task

**Example**

Hepatic reticuloendothelial function is impaired to the same degree in cirrhotic patients with or without a previous history of SBP.

History of spontaneous bacterial peritonitis does not affect impairment of hepatic reticuloendothelial function in cirrhotic patients.
Conclusions

As AIDA sentences:

- Nanopublications containing informal assertions in the form of AIDA sentences are a practical approach to advance scientific communication.
- Scientists are able to efficiently produce high-quality nanopublications containing informal assertions in the form of AIDA sentences.
- It is feasible to extract high-quality nanopublications containing informal assertions in the form of AIDA sentences from existing text resources.
Thank you for your Attention!

Questions?