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.

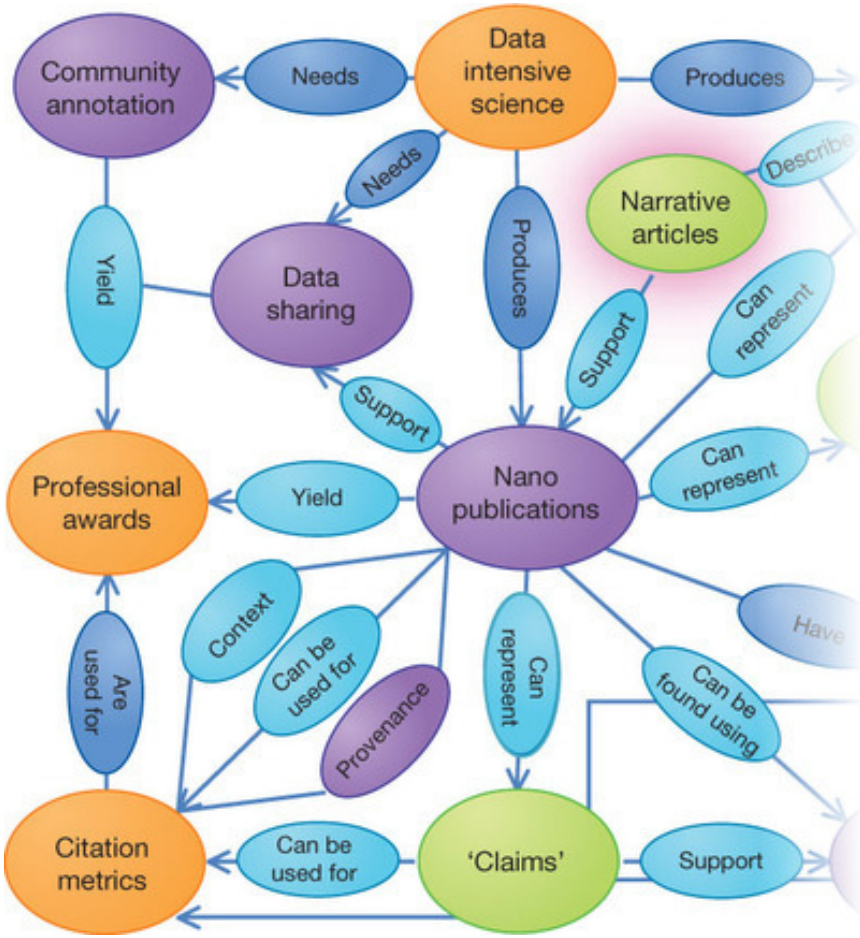
Vision: Changing Scholarly Communication

Now

Narrative articles at the center

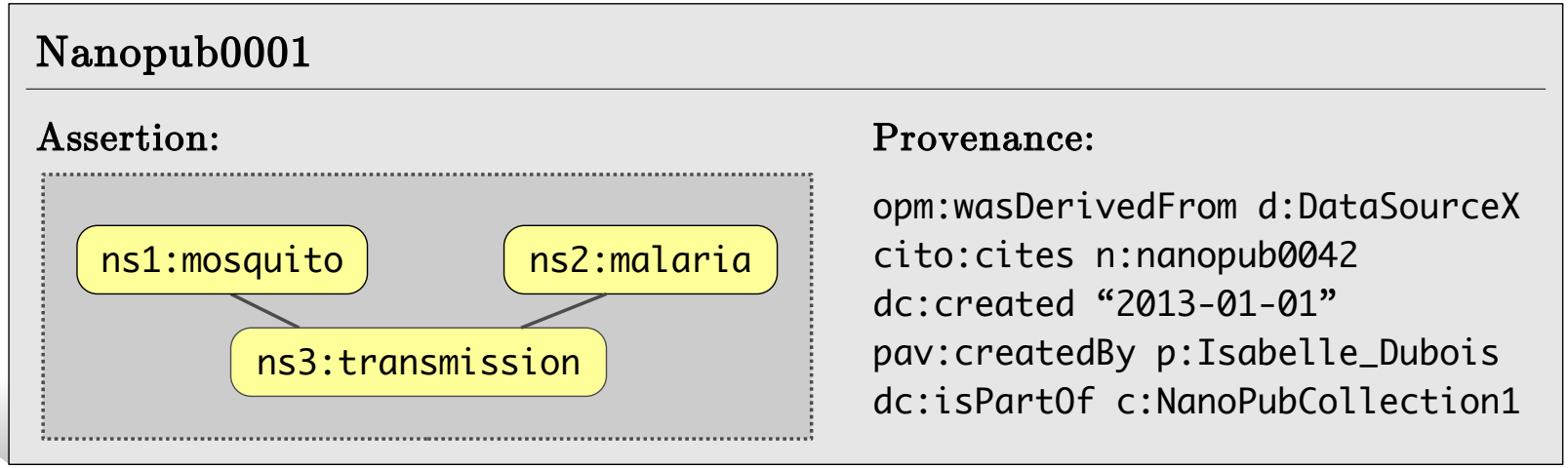
Future

Nanopublications at the center



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

Structure of Nanopublications



Assertion:

- Formalized scientific claim (or hypothesis)

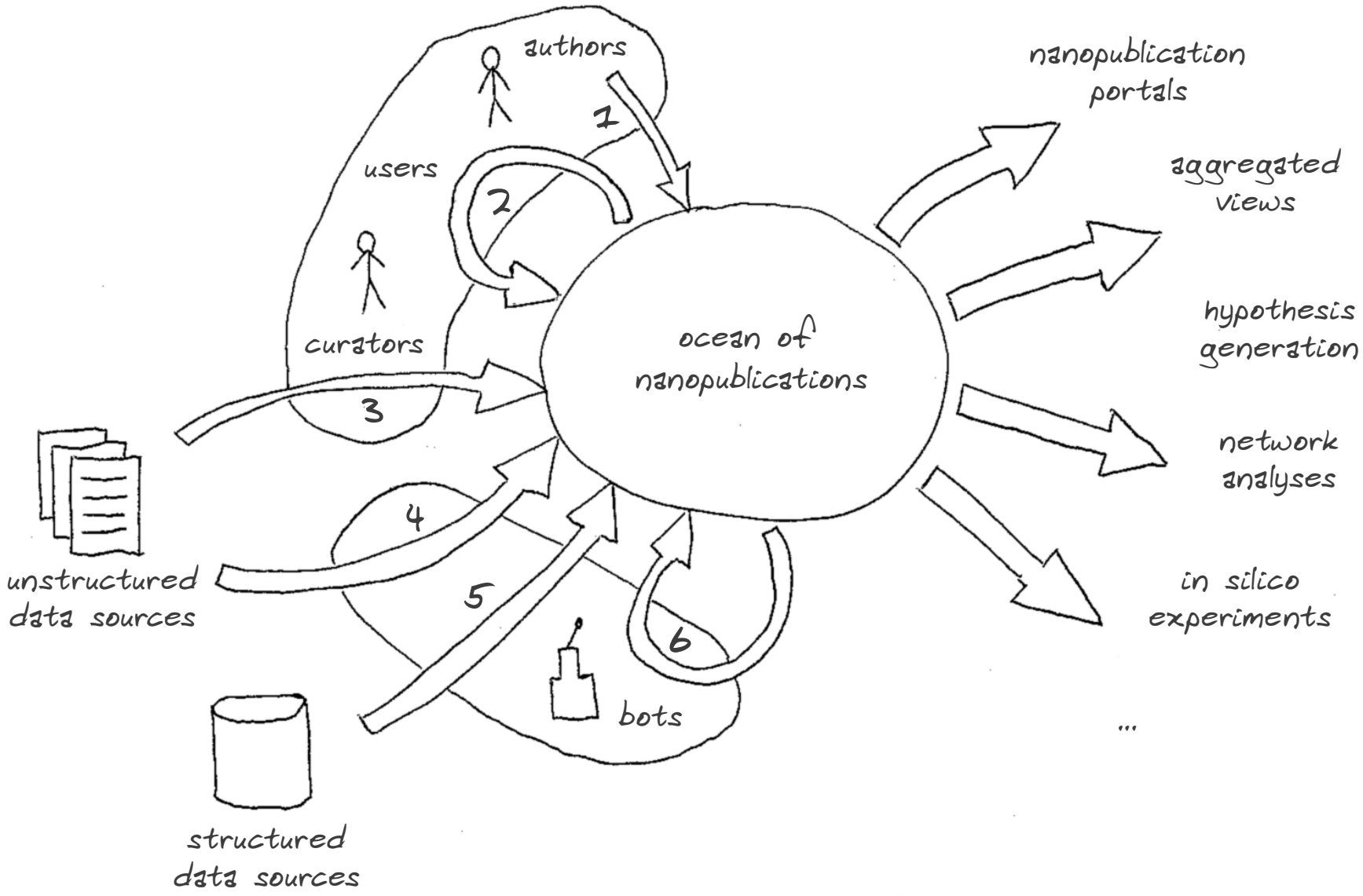
Provenance:

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

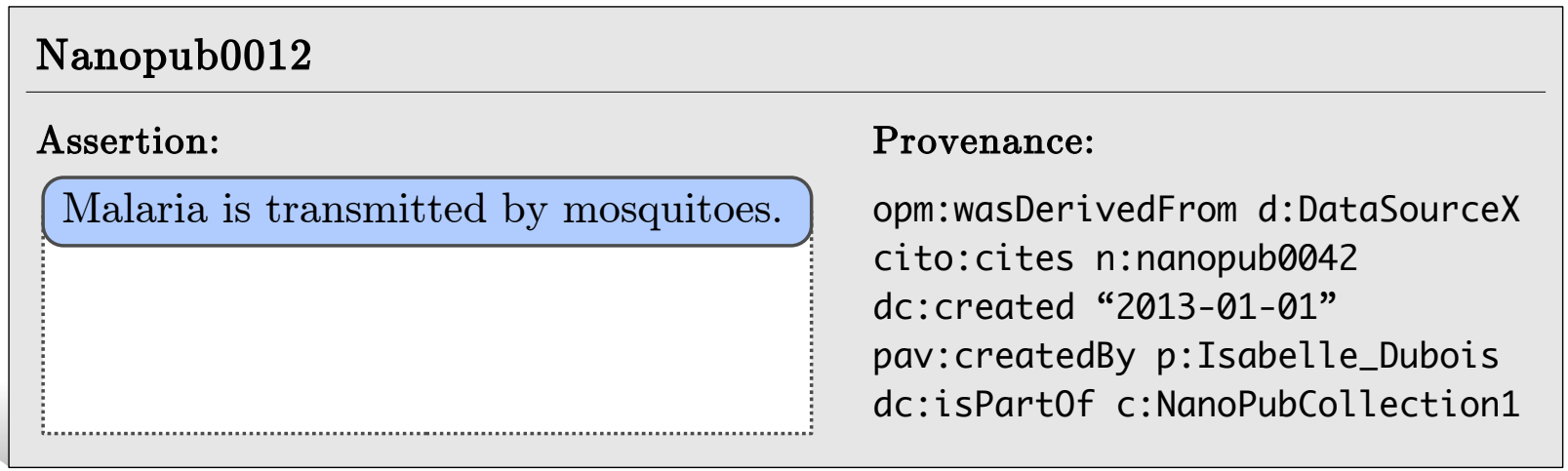
nanobrowser: Classical Nanopublication

http://nanobrowser.inn.ac

Ocean of Nanopublications



Proposed Extension 1: Informal Assertions

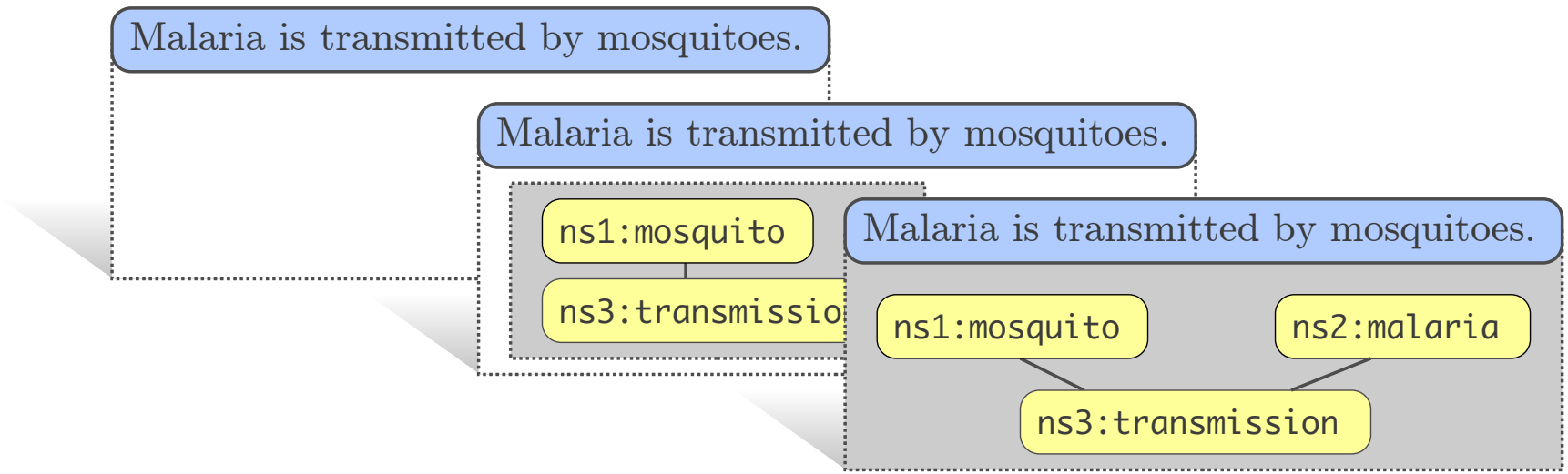


Assertion:

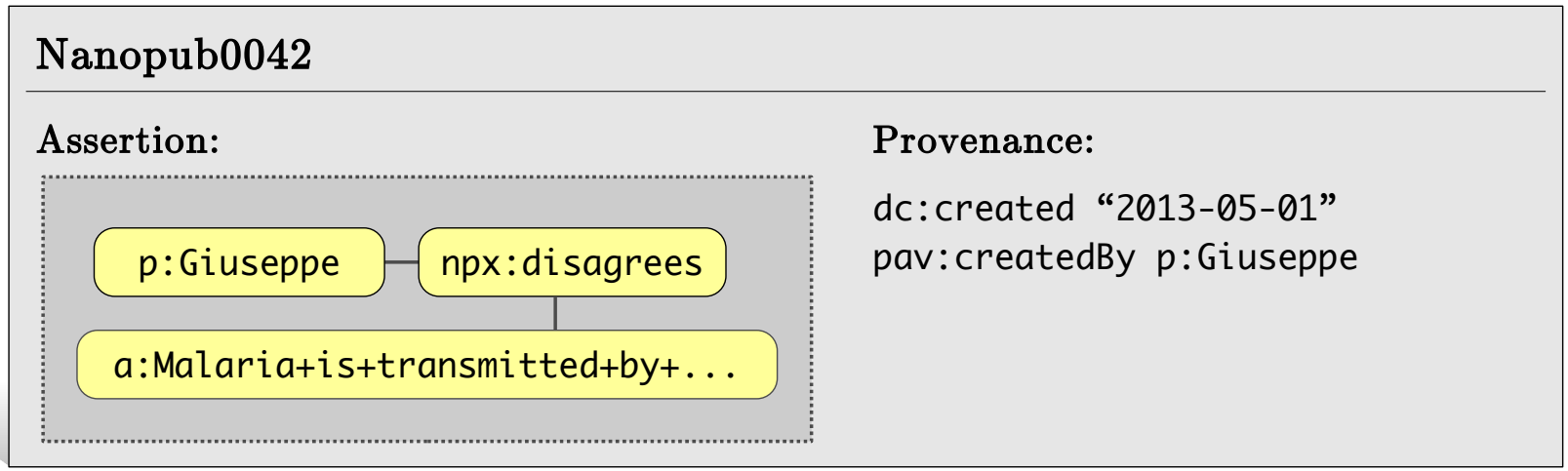
- Informal English sentence
- Sentences are independent entities and represented by URIs:
  `http://purl.org/aida/Malaria+is+transmitted+by+mosquitoes.`

Levels of Formalization

- Informal (only an AIDA sentence)
- Underspecified (formal representation for part of the sentence)
- Fully formal (formal representation for the complete sentence)



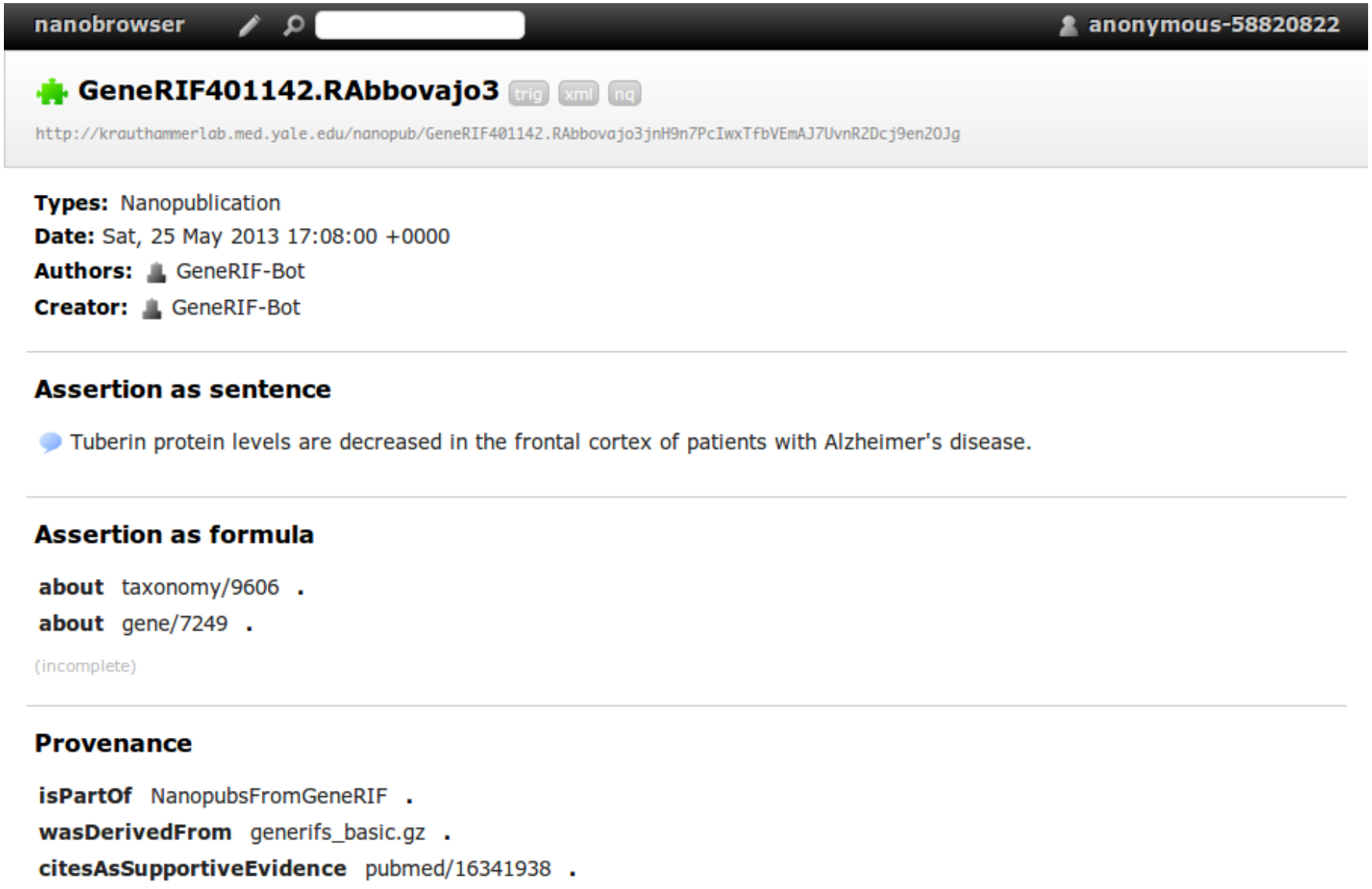
Proposed Extension 2: Non-Scientific Assertions



Assertion:

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

nanobrowser: Nanopublication with Sentence



Approach

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

AIDA Sentences

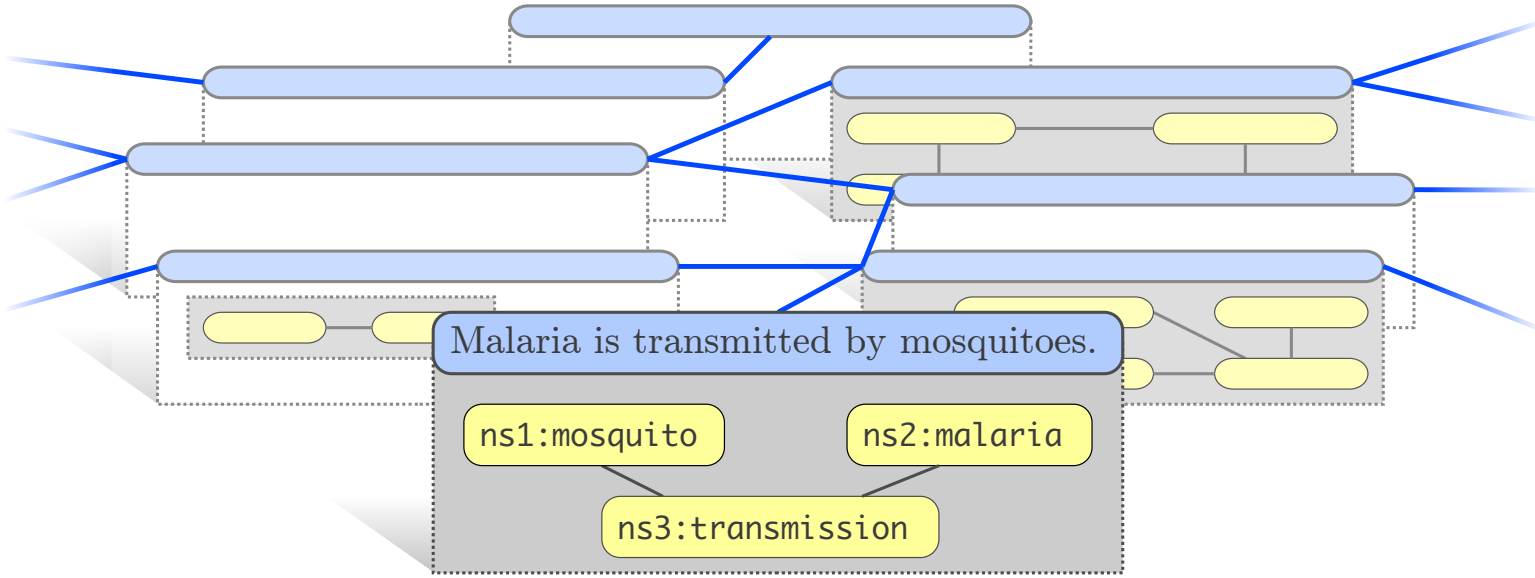
To fit into the nanopublication concept, the sentences of our approach should be **AIDA**:

- **A**tomic: a sentence describing one thought that cannot be further broken down in a practical way
- **I**ndependent: a sentence that can stand on its own, without external references like "this effect" or "we"
- **D**eclarative: a complete sentence ending with a full stop that could in theory be either true or false
- **A**bsolute: 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

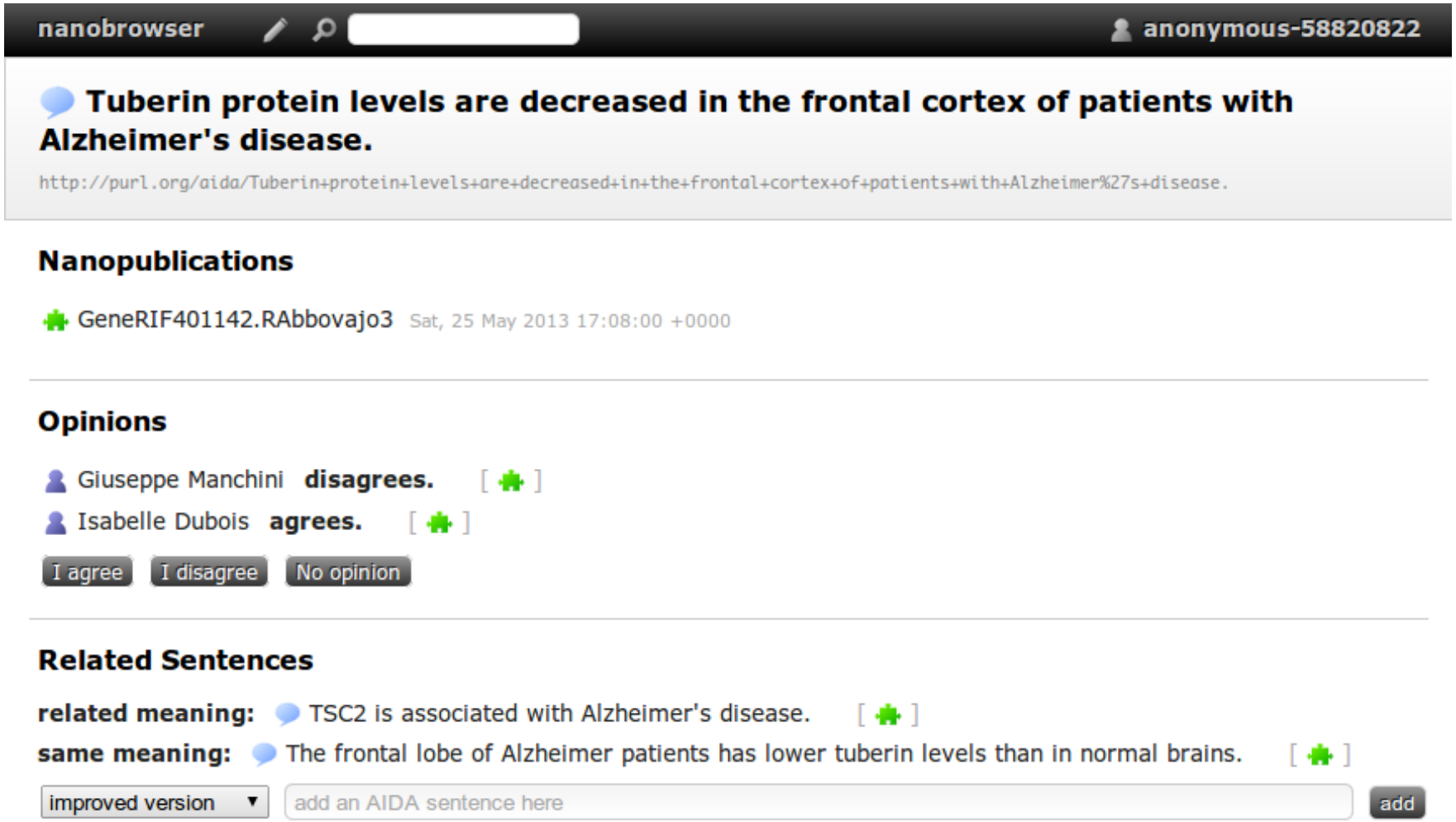


Possible relations:

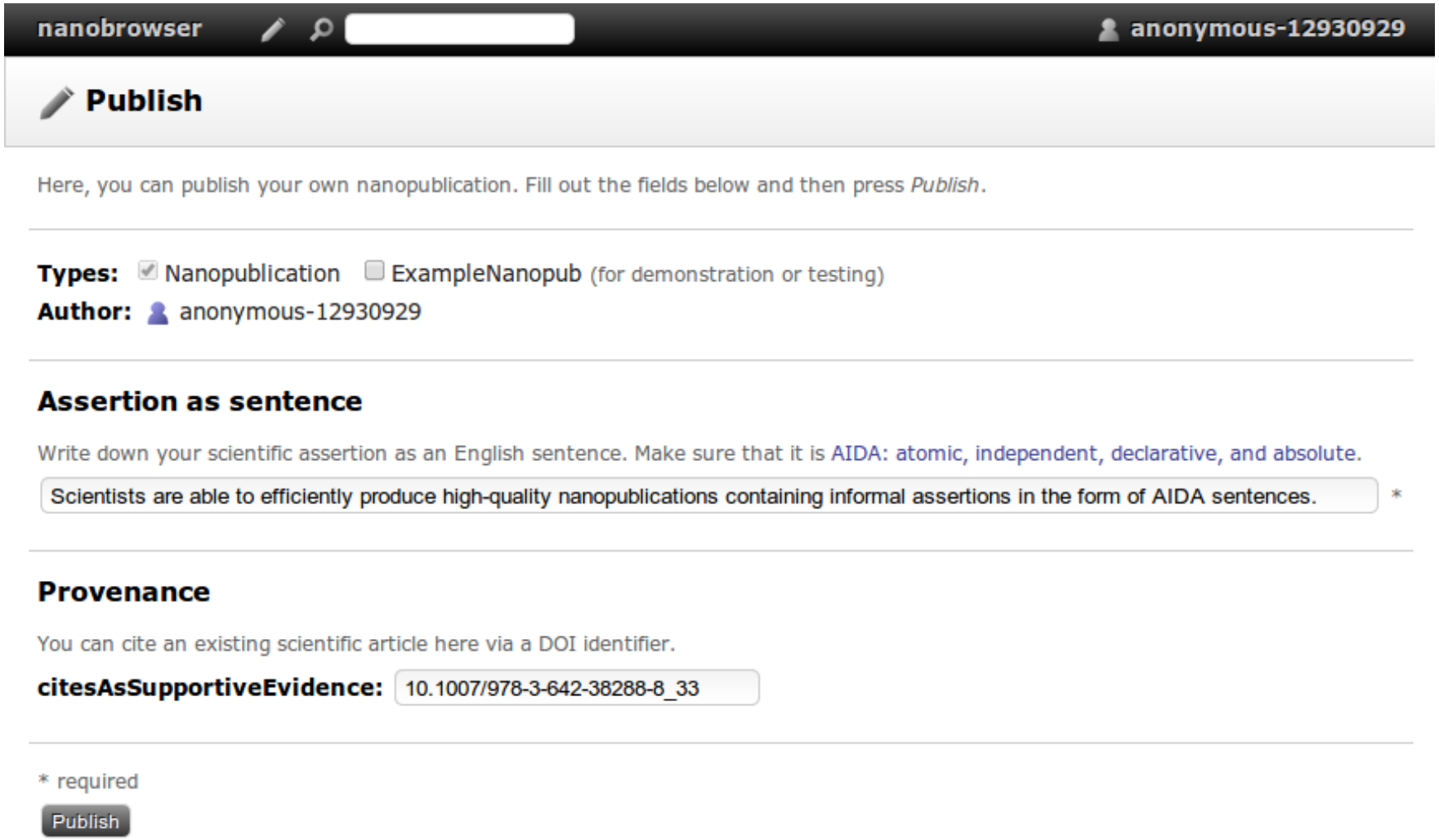
- [CLAIM] is equivalent to / contradicts / is similar to [CLAIM]
- [PERSON] agrees with / disagrees with / challenges [CLAIM]
- [STUDY] provides (counter-)evidence for [CLAIM]

These relations can be published as nanopublications too!

nanobrowser: Opinions and Sentence Relations

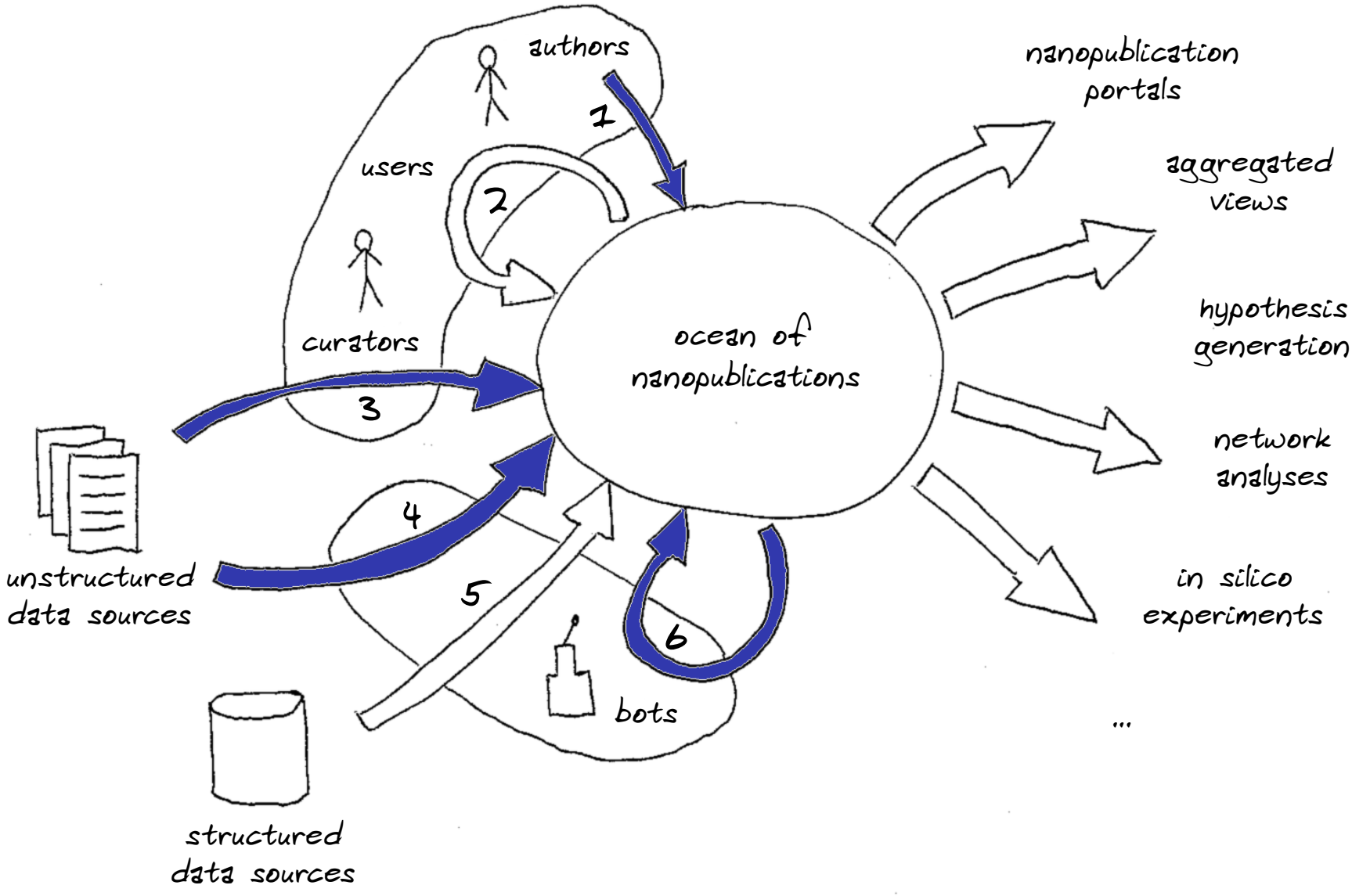


Publishing AIDA-Nanopubs



http://nanobrowser.inn.ac/publish

Evaluation


authors
users
curators
bots
1
2
3
4
5
6
ocean of nanopublications
unstructured data sources
structured data sources
nanopublication portals
aggregated views
hypothesis generation
network analyses
in silico experiments
...

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

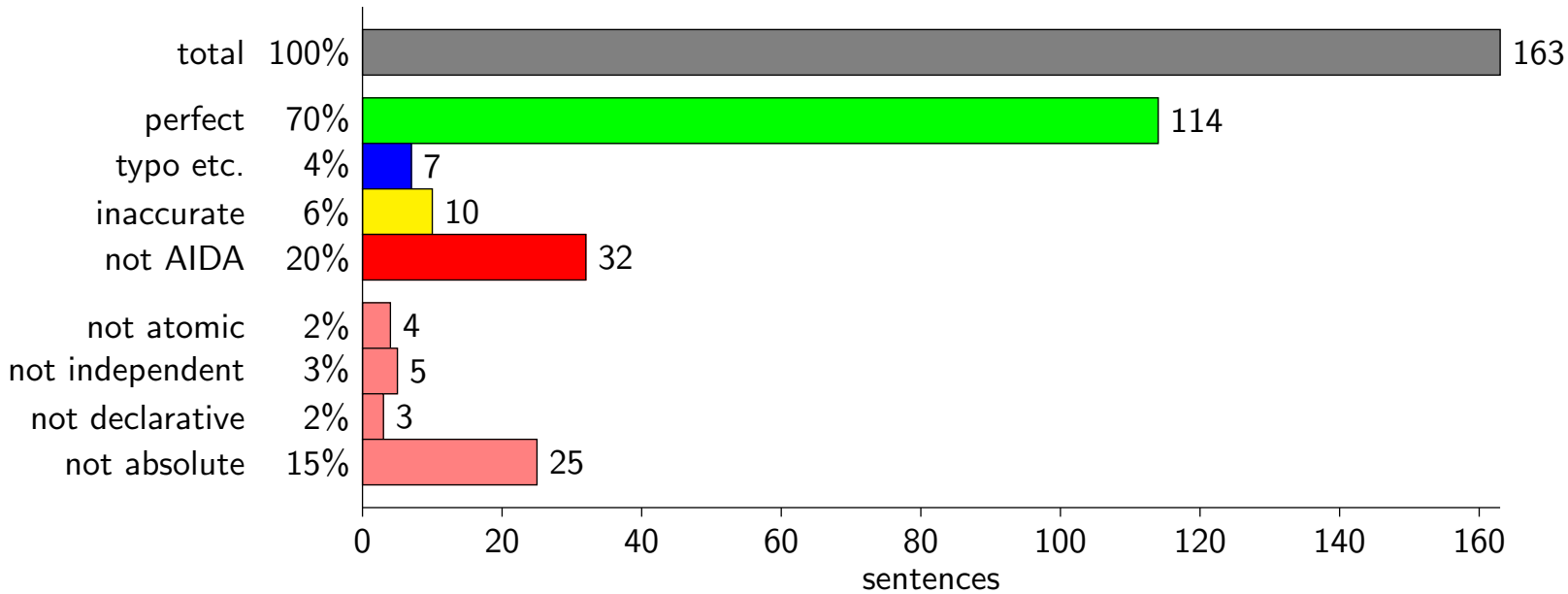
Original text: The results of this study showed that the hepatic reticuloendothelial function is impaired in cirrhotic patients, but the degree of impairment does not differ between patients with and without previous history of SBP.

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:



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

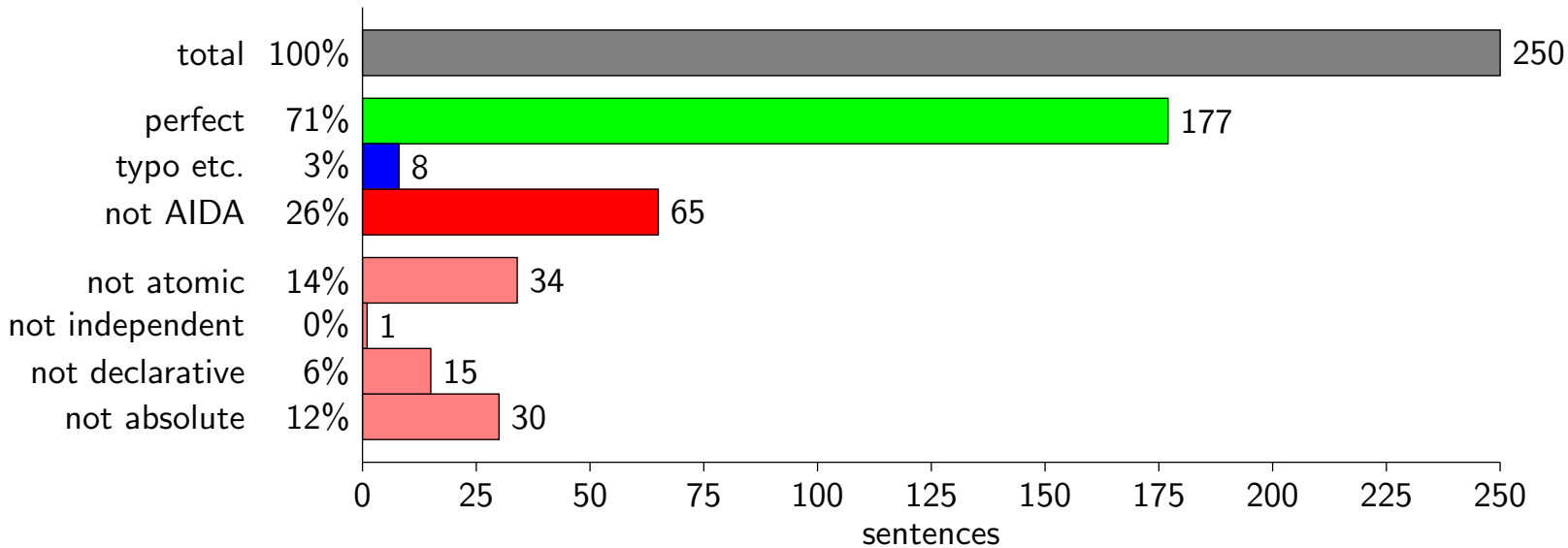
Example

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:



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?