Advances in Nanopublication and Its Applications: Postprint

Authors: Lihui Niu, Ou Shiyan

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] With the massive growth of academic journal literature, researchers need to spend considerable time searching for, acquiring, and interpreting required information from literature under the traditional scientific literature publishing model. To facilitate the dissemination and exchange of scientific information, fine-grained semantic publishing oriented towards scientific literature content has become a new trend. This paper introduces "nanopublication," a representative publishing model in semantic publishing, and analyzes the possibility and application characteristics of nanopublication across different disciplinary fields. [Method/Process] First, the nanopublication model is introduced. Then, through literature review, the current application status of nanopublication is reviewed. Finally, examples are used to illustrate the application characteristics of nanopublication in different disciplinary fields. [Results/Conclusion] Research results indicate that: nanopublication is currently mainly applied in the biomedical field, with limited applications in computer science and humanities, and almost no applications in other fields; nanopublication can be extended for application in other disciplinary fields, but it is necessary to construct nanopublications that conform to the characteristics of the disciplinary field based on its features.

Full Text

Preamble

ChinaXiv Partner Journal
Vol. 62, No. 7, April 2018

Research Advances in Nanopublication and Its Applications

Niu Lihui, Ou Shiyan
School of Information Management, Nanjing University, Nanjing 210023

Abstract

[**Purpose/Significance**] With the exponential growth of academic journal literature, researchers must spend considerable time searching for, retrieving, and interpreting needed information under traditional scientific publishing models. To facilitate the dissemination and exchange of scientific information, fine-grained semantic publishing oriented toward scholarly content has emerged as a new trend. This paper introduces "nanopublication," a representative publishing model within semantic publishing, and analyzes its potential applications and application characteristics across different disciplinary fields.

[**Method/Process**] We first introduce the nanopublication model, then review the current state of nanopublication applications through literature survey, and finally illustrate application characteristics in different disciplines with concrete examples.

[**Result/Conclusion**] Our findings indicate that: (1) Nanopublication is currently applied primarily in the biomedical domain, with limited use in computer science and humanities, and virtually no application in other fields; (2) Nanopublication can be extended to other disciplines, but requires constructing nanopublications that align with the specific characteristics of each domain.

Keywords: nanopublication, semantic publishing, knowledge representation

---

In the era of big data, users have access to increasingly vast information resources, and the development of digital publishing technologies has led to an explosion of scientific literature information. Under current digital publishing models, scientific documents are predominantly published online in unstructured formats similar to traditional print publications (such as HTML and PDF), causing numerous scientific discoveries and conclusions to be buried within massive scientific texts. Researchers must invest substantial time locating literature, reading, and identifying key information, making it difficult to rapidly obtain needed scientific information and its interconnections. This hinders scientific knowledge exchange, sharing, and reuse.

In 2009, Professor D. Shotton of Oxford University formally introduced the concept of "Semantic Publishing," which has since garnered widespread attention in academic and publishing circles. Shotton defined semantic publishing as a form of semantically enhanced journal publishing that employs Semantic Web technologies to annotate and interlink information in journal articles, enriching publication content, enhancing article semantics, and promoting knowledge dissemination and academic exchange. Building upon this concept, the non-profit academic organization Concept Web Alliance proposed a new scientific information publishing model in 2009—**nanopublication**—which aims to establish a structured and semantic fine-grained knowledge representation and dissemination method. In the nanopublication model, scientific facts or conclusions buried in unstructured scientific texts or massive datasets are extracted, structurally and semantically represented, and formed into independent nanopublications with context and supporting information. This approach enhances machine readability of scientific information while promoting knowledge dissemination and exchange.

2. Structure and Technical Implementation of Nanopublications

To address the challenges of discovering, connecting, and curating scientific knowledge in the big data context, the Concept Web Alliance first proposed the concept of nanopublications in May 2009, defining them as the smallest units of publishable information—unique, attributable assertions about anything. Building on this concept, P. Groth et al. from VU University Amsterdam proposed a nanopublication model in 2010 and developed initial technical specifications for generating nanopublications. In March 2011, the OpenPHACTS project (Open Pharmacological Concepts Triple Store), funded by the European Innovative Medicines Initiative, began implementing nanopublication as a unified representation format for pharmaceutical data to facilitate data integration and interoperability, providing semantic infrastructure for drug discovery research and marking the first practical application of nanopublications. In June 2012, the OpenPHACTS project released the Nanopublication Guidelines, defining the constituent elements and representation methods of nanopublications, with the latest version published in 2015 and the creation of the nanopub.org website, which has strongly promoted nanopublication development.

2.1 Structure of Nanopublications

According to the latest Nanopublication Guidelines released by the Concept Web Alliance in 2015, the composition structure of a nanopublication is shown in Figure 1 [Figure 1: see original paper]. A nanopublication contains three fundamental components: (1) **Assertion**: The smallest unambiguous information unit, representing relationships between concepts in subject-predicate-object triple form. As the core of a nanopublication, it typically expresses scientific claims or conclusions. (2) **Provenance**: Indicates the assertion's source, including author, institution, time, and location. (3) **Publication Information**: Metadata about the nanopublication itself, including creator, creation time, copyright, and version information.

Additionally, nanopublications include two auxiliary elements: **Head** and **Nanopublication ID**. The Head establishes relationships between the nanopublication and its assertion, provenance, and publication information. The Nanopublication ID provides a unique identifier, typically a URI, ensuring uniqueness. The Integrity Key, part of the Nanopublication ID, uses cryptographic hash algorithms to add a hash value to the nanopublication content, ensuring verifiability and permanence. Scholars at ETH Zurich, such as T. Kuhn, recommend using Trusted URIs for this purpose.

The Concept Web Alliance uses RDF Named Graphs to formally represent nanopublications. RDF Named Graphs are a simple extension of the RDF triple model—a collection of RDF triples identified by a URI that can be referenced directly as subject or object in other RDF triples. Currently, there are three main serialization formats for named graphs: RDF/XML, TriX, and TriG. RDF/XML and TriX are XML-based, while TriG is a variant of RDF/Turtle format.

A nanopublication's formal representation consists of four RDF named graphs: the head graph, assertion graph, provenance graph, and publication information graph. The head graph contains four RDF triples defining a nanopublication instance and describing its relationships with the other three graphs. The assertion graph describes only one natural language claim, requiring domain ontologies and controlled vocabularies to normalize natural language terms, and can be formally expressed as one or more RDF triples. The provenance graph describes background information about the assertion (i.e., assertion metadata), consisting of one or more RDF triples with the assertion graph as subject, typically using properties from the Provenance Ontology (e.g., `prov:generatedAtTime` for generation time, `prov:wasDerivedFrom` for source, `prov:wasAttributedTo` for attribution). The publication information graph describes nanopublication metadata, consisting of one or more RDF triples with the nanopublication URI as subject, using properties from provenance ontologies or metadata standards like Dublin Core.

To illustrate the structure clearly, consider a biomedical scientific claim: "Trastuzumab is indicated for breast cancer." Represented as a nanopublication (Figure 3 [Figure 3: see original paper]), the assertion describes the relationship between entities `ex:trastuzumab` and `ex:breast_{cancer}` via `ex:is_{{{indicated}}_{{for}}}`. Provenance states the assertion was generated on February 3, 2012, at 14:38, derived from an experimental conclusion, and attributed to a scientist. Publication information indicates the nanopublication was created on October 26, 2012, at 12:45 by P. Groth.

2.2 Technical Implementation of Nanopublications

Nanopublication implementation involves four main steps:

(1) Scientific Assertion Extraction: The primary task is identifying and extracting important scientific claims or conclusions from unstructured scientific literature to serve as nanopublication assertions. Since scientific literature typically follows fixed structural patterns, with claims usually appearing in specific sections (e.g., conclusions), parsing document structure helps rapidly locate key assertions. Representative models for scientific document structure include CISP (Core Information about Scientific Papers) developed by Aberystwyth University, which divides scientific research into eight categories: research purpose, motivation, object, method, experimental process, observation, results, and conclusion. CISP developers also created the SAPIENT tool (Semantic Annotation of Papers: Interface & Enrichment Tool) to assist manual identification of natural sentences representing these eight key categories.

(2) Semantic Representation of Assertions: Converting extracted natural language claims into RDF triples. First, natural language processing transforms statements into logical subject-predicate-object triples, where subjects/objects correspond to nouns/noun phrases, and predicates correspond to relational words (verbs, prepositional phrases). Second, natural language terms must be converted to controlled vocabulary with URIs for unique identification. Since the nanopublication model doesn't provide its own vocabularies, reusing existing open vocabularies and ontologies is recommended to facilitate integration. OpenPHACTS recommends common vocabularies for life sciences (Table 1). For unique identification, academic institutions have established URI services like Identifiers.org by the European Bioinformatics Institute, which provides URIs for life science entities in the form `http://identifiers.org/[namespace]/[entity]`. Finally, RDF triples are represented as RDF named graphs to form the assertion graph.

(3) Provenance Construction: Adding metadata to assertions from scientific literature, including author, institution, time, and location. The Provenance Ontology is primarily used to describe this metadata, which forms the provenance graph.

(4) Publication Information Construction: Adding metadata to the formalized nanopublication, including creator, creation time, copyright, and version information, using DCMI metadata and provenance ontology properties. This forms the publication information graph.

3. Current Application Status of Nanopublications

We collected 66 relevant papers on nanopublications using Google Scholar, of which 22 discuss the concept/model itself and 44 describe practical applications. Analysis of these 44 papers reveals that nanopublications are primarily applied in biomedicine, computer science, and humanities, as shown in Figure 4 [Figure 4: see original paper].

As Figure 4 shows, biomedical applications are relatively mature, accounting for approximately 80% of current usage. Data sources for nanopublications fall into two categories: (1) Researchers publishing their own findings to accelerate dissemination; (2) Converting existing relational database data to nanopublications to discover new relationships and improve credibility.

Representative cases include: In 2015, E. Mina et al. from Leiden University published Huntington's disease experimental results and processes as Linked Data, overcoming traditional journals' inability to present experimental data and enhancing data interconnectivity. They first converted natural language results into RDF triples as assertions, then used a workflow-centric research object model (WROM) to annotate experimental processes and data as provenance.

The EU Innovative Medicines Initiative's OpenPHACTS project (2005) applied nanopublications to drug discovery infrastructure, representing bioactivity data and conclusions to build a drug data integration platform, improving credibility and interoperability. In 2017, J.P. McCusker et al. from Rensselaer Polytechnic Institute used nanopublications to unify semantic representation and centralized storage of drug-target, protein-protein, and disease-gene relationships from diverse databases, solving data fragmentation for drug repurposing. Such applications rely on RDF conversion tools like McCusker's 2013 Prizms tool (converting CSV/XML/JSON to RDF) and G.P. Patrinos's 2012 tool for the LOVD database (converting human genetic variation data).

Beyond biomedicine, researchers have explored other domains. In 2014, A. Lipani et al. from Vienna University of Technology semantically annotated information retrieval papers using an IR ontology, publishing IR experimental data as nanopublications to enhance reproducibility. In 2015, P. Golden et al. from UNC Chapel Hill formalized historical period descriptions from archaeology, history, and art history as nanopublications and published them as Linked Data. Also in 2015, R. Viglianti from the University of Maryland used nanopublications to publish traditional musical scores as Linked Data, improving addressability.

Since 2013, Chinese scholars have begun studying nanopublications, though still in early stages. Wu Sizhu et al. from the National Science Library, Chinese Academy of Sciences (2013) first introduced nanopublication in China, discussing its concept, source files, semantic patterns, structural description, application value, and role in knowledge discovery. Su Yunmei and Wu Jianguang from Shanxi Medical University Library (2015) explored nanopublication's relationship with knowledge discovery, analyzing semantic association patterns using Seckel syndrome as a case study, but without practical implementation. Lü Yuanzhi from Shanghai Normal University (2015) constructed a semantic description framework for digital archival resources based on nanopublication, but this coarse-grained external structure description (titles, classification numbers, keywords) deviates significantly from nanopublication's fine-grained knowledge representation purpose and cannot be considered a true application.

4. Application Characteristics of Nanopublication in Different Disciplines

With rapid growth of scientific conclusions and data across disciplines and urgent needs for automatic semantic integration, linking, and reasoning, expanding nanopublication applications is essential. However, T. Kuhn et al. from ETH Zurich argue that representing all scientific data as nanopublications seems impractical and may limit application scope. Accordingly, P. Golden et al. from UNC Chapel Hill propose that nanopublication assertions should be constructed based on researchers' actual needs—useful for target users without excessive strict standardization that could create complexity and hinder use.

The core of nanopublication is the assertion, typically derived from claims in scientific literature. Concept Web Alliance member B. Mons categorizes scientific claims into three types: curated statements (peer-reviewed facts stored in professional databases like OMIM and UniProt, forming domain ontology foundations), observational statements (experimental or statistical conclusions without fixed databases), and hypothetical statements (derived from text mining or inference, serving as sources for knowledge discovery). Different disciplines describe and utilize knowledge differently, resulting in varied extracted assertion content. Below, we illustrate distinct characteristics across biomedicine, history, and sociology.

4.1 Characteristics and Formalization of Biomedical Assertions

Biomedical literature primarily contains scientific conclusions derived from experiments, often expressed as correlations, typically located in conclusions and mentioned in abstracts. For example, in a 2010 paper by P. Liu et al. from the University of Washington, the authors used cross-sectional studies to explore relationships between methylenetetrahydrofolate reductase gene polymorphisms and menarche/natural menopause timing in white women. The abstract contains the claim: "The results of our study suggest that the MTHFR gene may influence the onset of menarche and natural menopause."

This claim is first simplified to a logical representation: "MTHFR → influences → the onset of menarche and natural menopause." Using the National Cancer Institute Thesaurus (NCIT), we normalize "MTHFR" to `ncit:MethylenetetrahydrofolateReductase`, and use MEDDRA terms `meddra:Menarche` and `meddra:NaturalMenopause` for the physiological phenomena. The DbPedia ontology property `dbpedia:influence` expresses the relationship. This yields two RDF triples as the nanopublication assertion, shown in Figure 5 [Figure 5: see original paper] (Turtle format).

4.2 Characteristics and Formalization of Historical Assertions

Unlike natural sciences, humanities claims are often authors' empirical opinions or hypotheses that cannot be experimentally verified. For humanities literature, we primarily extract theoretical arguments or conclusions, which typically appear scattered throughout texts as descriptions of characteristics or attributes. For instance, in a 2010 paper by P. Beaujard from the French National Center for Scientific Research, the author states: "In China, the political fragmentation characterizing the period known as Spring-and-Autumn (771BC-481BC) went along with the development of the craft industry and exchange."

From this we extract the temporal span claim: China's historical period "Spring-and-Autumn" spanned from "771BC" to "481BC". Here, the temporal entity "Spring-and-Autumn" is represented as a `dbpedia:SpringAndAutumnPeriod` instance, the country "China" as `dbpedia:China`, using DCMI metadata property `dcterms:spatial` for spatiotemporal relationships and time ontology properties `time:intervalStartedBy` and `time:intervalFinishedBy` for temporal boundaries, with SKOS property `skos:prefLabel` for the Gregorian dates "771BC" and "481BC". This yields five RDF triples as the assertion, shown in Figure 6 [Figure 6: see original paper].

4.3 Characteristics and Formalization of Social Science Assertions

Social science literature includes two types: theory-based (logical argumentation) and data-based (experimental). Theory-based papers have diverse structures without fixed patterns, with conclusions typically being commentary or recommendations whose complex logic is difficult to reduce to triples. Moreover, nanopublication aims to publish fine-grained scientific knowledge as an objective knowledge graph, while commentary/recommendations are often too subjective. Conversely, data-based experimental papers have fixed structures, objective conclusions, and more easily tripleized results. Therefore, social science nanopublications primarily derive from experimental conclusions, usually located in conclusions and mentioned in abstracts.

For example, in a 2016 paper by M. Lawrence from Middlebury College, the author used causal inference to explore relationships between parental and children's higher education attainment. The abstract claims: "Results show that college increases male graduates' probability of having a child who completes college."

This simplifies to two logical representations: (1) College → increases → probability; (2) The probability → refers to → male college graduates' child completes college. Using the Psychology Ontology (APAONTO), DbPedia, Sequence Ontology (SO), and Medical Subject Headings (MeSH), we normalize entities. Since "male college graduates" lacks a direct match, we extend APAONTO by adding subclasses `MaleCollegeGraduates` and `FemaleCollegeGraduates` under `CollegeGraduates`. We represent "college" as `apaonto:College`, "male college graduates" as `apaonto:MaleCollegeGraduates`, and use MeSH terms `mesh:Probability` and `mesh:Child`. The resulting RDF triples are shown in Figure 7 [Figure 7: see original paper] (Turtle format).

Cross-Disciplinary Characteristics

Nanopublication applications across these three disciplines reveal several characteristics:

1. **Different assertion types**: Experimental fields produce curated and observational statements (in abstracts/conclusions), while non-experimental fields yield author viewpoints or consensus (location-variable, requiring cue-word extraction).
2. **Different precision levels**: Natural sciences describe objective facts with unique, reproducible conclusions, yielding consistent nanopublications with minimal conflict. Humanities claims are culturally influenced,

requiring tolerance for interpretation variance—overly precise constraints hinder application.

3. **Different use cases**: Natural sciences, with rapid knowledge updates and high output, benefit from nanopublications' minimal-unit rapid publishing for result validation and sharing. Humanities and social sciences can use nanopublications as a unified format for cultural heritage, financial, health data, etc., to achieve cross-domain knowledge linking and make tacit knowledge explicit, with provenance information ensuring data quality and credibility.

Thus, representing disciplinary claims as nanopublications requires capturing domain-specific features to construct appropriate nanopublications.

5. Conclusions and Future Work

Nanopublication enables the shift from full-text journal publishing to fine-grained knowledge unit publishing, independently releasing scientific facts and conclusions in structured form. Provenance information indicates claim origins to increase trust, while URIs provide unique identifiers for network retrieval and citation, potentially reducing infrastructure-based disparities among researchers worldwide.

However, limitations exist: (1) Publishing conclusions as nanopublications requires manual construction, yet researchers prefer natural language expression, and the tedious construction process outweighs benefits for most scientists; (2) Nanopublications cannot resolve assertion conflicts—they faithfully represent literature conclusions but cannot adjudicate correctness when authors disagree; (3) Nanopublications formalize conclusion statements but cannot represent the argumentative process through citations, experimental data, or methods, preventing experimental replication based solely on nanopublications.

Future work will explore nanopublication applications in social sciences to expand its scope. Given these limitations, we plan to develop a new semantic publishing model combining nanopublication with other models to semantically represent not only scientific conclusions but also their original sources and argumentative processes, thereby enhancing credibility and effectiveness.

References

[1] SHOTTON D. Semantic publishing: the coming revolution in scientific journal publishing [J]. Learned publishing, 2009, 22(2): 85-94.
[2] Concept Web Alliance [EB/OL]. [2017-07-16]. https://www.nbic.nl/about-nbic/affiliated-organisations/cwa/introduction/index.html.
[3] Nanopub.org [EB/OL]. [2017-07-16]. http://nanopub.org/wordpress/.
[4] GROTH P, GIBSON A, VELTEROP J. The anatomy of a nanopublication [J]. Information services & use, 2010, 30(1/2): 51-56.
[5] OpenPHACTS [EB/OL]. [2017-07-16]. http://www.openphacts.org/index.php.

[6] The open PHACTS nanopublication guidelines [EB/OL]. [2017-07-16]. http://www.nanopub.org/guidelines/1.8/.
[7] Nanopublication guidelines [EB/OL]. [2017-07-16]. http://nanopub.org/guidelines/working_{draft}/.
[8] KUHN T. Nanopub-java: a java library for nanopublications [EB/OL]. [2017-07-16]. http://dare.ubvu.vu.nl/handle/1871/53984.
[9] BIZER C, CARROLL J J, HAYES P, et al. Named graphs, provenance and trust [EB/OL]. [2017-07-16]. http://lists.w3.org/Archives/Public/www-archive/2004Mar/att-0118/ng.pdf.
[10] PROV-O: the PROV ontology [EB/OL]. [2017-07-16]. http://www.w3.org/TR/prov-o/.
[11] SOLDATOVA L, LIAKATA M. An ontology methodology and CISP-the proposed core information about scientific papers [EB/OL]. [2017-07-16]. http://www.aber.ac.uk/en/media/departmental/impacs/computerscience/pdfs/ReportCISPshort.pdf.
[12] LIAKATA M, CLAIRE Q, SOLDATOVA L N. Semantic annotation of papers: interface & enrichment tool (sapient) [EB/OL]. [2017-07-16]. http://www.aber.ac.uk/en/media/departmental/impacs/computerscience/pdfs/sapientBio2009Final.pdf.
[13] Identifiers.org [EB/OL]. [2017-07-16]. http://identifiers.org/.
[14] MINA E, THOMPSON M, KALIYAPERUMAL R, et al. Nanopublications for exposing experimental data in the life-sciences: a huntington's disease case study [J]. Journal of biomedical semantics, 2015, 6(1): 1-12.
[15] HARLAND L. Open PHACTS: a semantic knowledge infrastructure for public and commercial drug discovery research [EB/OL]. [2017-07-16]. http://www.openphacts.org/documents/publications/121008_{{{EKAW}}*{{Lee}}}%20Harland_A%20sema*
[16] MCCUSKER J P, DUMONTIER M, YAN R, et al. Finding melanoma drugs through a probabilistic knowledge graph [EB/OL]. [2017-07-16]. https://peerj.com/articles/cs-106.pdf.
[17] MCCUSKER J P, LEBO T, KRAUTHAMMER M, et al. Next generation cancer data discovery, access, and integration using prizms and nanopublications [EB/OL]. [2017-07-16]. http://pdfs.semanticscholar.org/92b0/b95eedac4cd3cbbb1e7ff73900f94ceda21e.pdf.
[18] PATRINOS G P, COOPER D N, MULLIGENE V, et al. Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain [J]. Human mutation, 2012, 33(11): 1-10.
[19] LIPANI A, PIROI F, ANDERSSON L, et al. Extracting nanopublications from IR papers [EB/OL]. [2017-07-16]. https://www.researchgate.net/publication/267402028{{{Extracting}}}{
[20] GOLDEN P, SHAW R. Nanopublication beyond the sciences [EB/OL]. [2017-07-16]. https://dx.doi.org/10.7287/peerj.preprints.1284v1.
[21] Enhancing music notation addressability [EB/OL]. [2017-07-16]. http://mith.umd.edu/research/enhancing-music-notation-addressability/.
[22] WU Sizhu, LI Feng, ZHANG Zhixiong. Research on semantic representation and publishing mode of knowledge resources: a case study of Nanopublication [J]. Journal of library science in China, 2013(4): 102-109.
[23] SU Yunmei, WU Jianguang. Nanopublication semantic pattern and its relationship with knowledge discovery [J]. Chinese journal of medical library and information science, 2015(12): 15-18.
[24] WU Jianguang, SU Yunmei. Research on Nanopublication semantic

pattern based on knowledge discovery [J]. Journal of Shanxi medical university, 2015(8): 833-836.
[25] LÜ Yuanzhi. Research on semantic description of digital archival resources based on Nanopublication framework [J]. Archives science bulletin, 2015(3): 57-62.
[26] KUHN T, BARBARO P E, NAGY M L, et al. Broadening the scope of nanopublications [J]. Lecture notes in computer science, 2013, 7882: 487-501.
[27] MONS B, VELTEROP J. Nano-publication in the e-science era [EB/OL]. [2017-07-16]. https://www.w3.org/wiki/images/4/4a/HCLS

$ISWC$2009

Workshop$Mons.pdf.
[28] LIU P, LU Y, RECKER R R, et al. Association analyses suggest multiple interaction effects of the methylenetetrahydrofolate reductase polymorphisms on timing of menarche and natural menopause in white women [J]. Menopause, 2010, 17(1): 185-190.
[29] PHILIPPE B. From three possible Iron-age world-systems to a single Afro-Eurasian world-system [J]. Journal of world history, 2010, 21(1): 1-43.
[30] LAWRENCE M, RICHARD B. And their children after them? The effect of college on educational reproduction [J]. American journal of sociology, 2016, 122(2): 532-572.

Author Contributions

Niu Lihui: Data collection, manuscript writing and revision.
Ou Shiyan: Research direction and conceptualization, manuscript revision and review.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.