Extending the Use of Nanopublications: Retrieval, Citation and Statement Verification

Erika Fabris[0000−0003−1991−3267]

Department of Information Engineering, University of Padua, Padua, Italy
erika.fabris@unipd.it

Abstract. In a scenario where data is as central as publications are, a novel model to represent scientific results has been proposed, the nanopublication model. This model has enormous potential for the representation of "atomic" scientific results allowing interoperability, data integration and exchange of scientific findings by using a structure that can easily be interpreted by machines. We present our contribution to overcome the lack of a citation standard for this model by defining a framework to automatically generate citations for nanopublications and to widen their use of by specifying methods and tools to ease their access, usability and comprehension also to non-expert users.

Keywords: nanopublication · data citation · statement verification

1 Introduction

The number of scholarly publications and available scientific data and results is growing continuously, as evidenced in [1] where the publishing rate is estimated of one new paper every 20 seconds. Thus, nowadays, researchers have to deal with an overwhelming amount of information and data to carry out a research project, to find relevant information for their research or to keep up to date.

Moreover, in recent years we have seen a shift in the nature of science due to the increasing use of data which has led to a change in the nature of scientific publications as well. This change has made it necessary to adapt the infrastructure for managing the growing amount of scientific data [2], led to the definition and adoption of open access policies for the access to scholarly data, to new concepts of data scholarship [3] and sanctioned the transition to data-intensive research where data are as essential as scientific publications [4].

In this scenario, a new model of publication, the nanopublication model, has been proposed to overcome the increasing difficulties introduced from the growing amount of scientific results and data in founding, connecting and curating scientific statements and in determining their provenance [5]. The model is based on the idea that the data and atomic statements are themselves a publication and it aims to represent scientific statements in a machine-readable format together with attribution and provenance metadata and to make them accessible, uniquely identifiable, citable and attributable and to promote interoperability among scientific results, data integration and sharing of scientific results.

However, since the concept of nanopublication is relatively novel, some facets of the topic have been already consolidated – e.g. their structure and storing –, on the other hand, there are diverse unresolved open challenges related to nanopublications – e.g. how to properly cite a nanopublication or a set of nanopublications referenced within a scientific publication? How can nanopublications be used by users who do not know their structure to access some pieces of specific information? How to display information within a nanopublication so that it is understood by human users? Is it possible to verify statements by using data from the existing nanopublications? How to allow non-expert users to retrieve statements from the published nanopublications relative or similar to a given statement?

Our work tackles these open challenges by defining a framework which outlines methods and tools to ease the access, the usability and the comprehension of nanopublication also to non-expert users.

2 Background

Nanopublication Nanopublication is a novel publishing model to represent minimal scientific assertions or statements together with its attribution and provenance information [6]. A nanopublication is essentially a single assertion in the form of an atomic statement (subject, predicate and object) associated with its provenance, which contains how its origin and generation process, and with publication information metadata about the creation and publication. The nanopublication schema is formally structured by making use of Semantic Web technologies as three W3Cs Resource Description Framework (RDF) graphs, which are the assertion, the provenance graph and publication information graph, containing information serialized as RDF triples. Each RDF triple can be represented as a node-arc-node link of an RDF graph and consists of a subject, a predicate and an object. Each element is represented by means of an Internationalized Resource Identifier (IRI) or an ontology term (specified using Web Ontology Language (OWL) semantics).

The nanopublication model was introduced to be used in scholarly communication, based on the idea that scientific results can be split into atomic sentences. The model aims to overcome the difficulties due to the growing amount of scientific results, to promote data interoperability, data integration, the exchange of scientific results, to provide a machine-readable format of representing scientific results, and to enable the distribution of scientific statements as independent publications even without the related research article.

Today more than 10 M nanopublications are freely available on a server network and other 200 M are available as independent private datasets[1]. The majority of them was generated by performing automatic extraction of atomic assertion

[1] http://nanopub.org/wordpress/?page_id=749

from scientific publications, the others are manually created. Nanopublications come from several domains: so far mostly from Life Sciences domain such as pharmacology, genomics and proteinomics, and minor datasets coming from humanities domain such as philosophy, archaeology and musicology.

Several management tools are provided to perform different tasks such as the validation of nanopublication structure, the access to subsets of nanopublications by performing SPARQL queries or the access to specific nanopublication from its identifier and the publication of new nanopublications. It is worth noting that all the provided tools are addressed to expert users and in order to make use of those tools, a full knowledge of the nanopublication anatomy, RDF structures and OWL semantics is required. Yet not every related aspect has been already consolidated: neither a standard and structured way for the citation nor easy access and inspection tools for non-expert users have been provided so far.

Data Citation Citations are one of the main 'driving force' for the scientific progress, to promote the diffusion of knowledge, to ensure the transparency and reproducibility of the scientific findings, to verify research conclusions and support the reuse of the scientific results, to assure proper attribution and credit.

Since we were witnessing a radical change in the nature of science towards the fourth paradigm of science, which made data to be considered crucial for scientific progress, data citation is gaining importance [8, 9]. Two are the main aspects of data citation that have been highlighted and studied: the definition of data citation principles and the development of solutions for computational problems related to the generation of citation.
In particular, two international initiatives promote the definition of standard principles for data citation: CODATA (Committee on Data of the International Council for Science) [13] and FORCE 11 (The Future of Research Communications and e-Scholarship) [14]. Until now, several solutions to automatically generate data citation snippets have been proposed mostly dealing with databases, considering provided queries and relying on views to generate a citation [10, 11, 12]. Unfortunately, none of these solutions can be applied to nanopublications. Moreover, even if the property of being citable is a prerogative of the nanopublications, nowadays the only way to cite them is by means of their identifiers or referring to the whole dataset. It is worth noting that, given that nanopublication are publications in their own right, citing a specific set of nanopublications is important to give credit to the people who contributed to their creation and to allow the reproducibility of works that rely on single or a finite set of nanopublications.

3 Objectives

Our work provides the solutions to the lack of a citation standard for the nanopublication model and some tools to extend the use of nanopublications.

Firstly we design a citation framework to automatically create human- and machine-readable reference snippets of a single nanopublication and, afterwards,

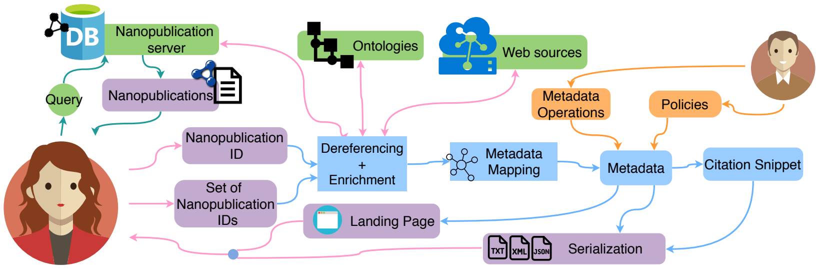

Fig. 1: Nanopublications Citation System framework workflow.

we extend the framework to handle a more common demand: the citation for a set of nanopublications. Then we develop the above frameworks and provide a citation system to be used to obtain nanopublications citation to be included in the reference list, to obtain machine-readable citation metadata and to get a visual picture of the content of nanopublications. Moreover, to widen the usability of nanopublications, our work presents the structure of a knowledge base with interlinked existing nanopublications as its base components and methods for access and retrieval purposes. In addition, our work provides a tool to verify if a given statement contradicts existing assertions within the nanopublication databases by making use of the just mentioned knowledge base.

4 Methods

Nanopublication Citation

We propose a framework, the *Nanocitation* framework, to automatically generate citation snippet by considering as the only input from the user the identifier of the nanopublication. Figure 1 shows the main components of the framework. The key idea is that every IRI defining an element of the RDF triple in the nanopublication can be dereferenced and other related details and information can be extracted by requesting them to external sources. The process starts from the user citation request, the raw nanopublication undergoes a `dereferencing and enrichment` process by which the information contained in the form of IRI identifiers is transformed into human-readable form and additional relative details are sought. Afterwards, through the `metadata mapping` process, the data within the output of the latter process (the `enriched nanopublication`) are integrated to form a structured human-readable entity, so-called `metadata`, that embodies more information than the original nanopublication. The semantics and constraints of the metadata are specified in an *ad-hoc* formal representation.

In order to generate the reference snippet, some relevant metadata fields need to be selected and optionally modified. The operations to be undertaken at this point for the creation of the reference follow the policies formally defined by the administrator database through relational algebra constructs which include three possible types of operations: selection and sorting (which field of metadata to integrate in the reference snippet and in what order), single-field operations (optional operations to be performed for each field) and presentation (reference rendering, e.g. comma/semicolon between elements).

Note that, as long as reasonable policies are used, generated references can be concise enough to be included in the reference list of a publication, furthermore, it allows the reader to identify the nanopublication and understand its content.

At the end of the procedure, to make up for the lack of a human-readable visualization of nanopublications and understandable by both expert and non-expert user, a way to access to more complete and specific information about the relative nanopublications is provided to the user: a web page which we call 'landing page' which provides a full picture of the nanopublication content in both a human- and machine-readable form. The purpose of this page is to support the user in the in-depth exploration of the nanopublication content and to obtain a metadata citation serialization. It is worth noting that given a nanopublication or a specific set, the system always returns the same landing page. The process to obtain a citation reference of a set of nanopublication differs by the need to define operations that allow to aggregate multiple citation metadata to obtain single metadata which identifies and describes the set at a certain level of completeness. These operations, as for the policies, need to be defined by the system administrator in a formal representation.

We implemented the *Nanocitation* framework [7] and publish it as a web application (`http://nanocitation.dei.unipd.it`) with a simple user interface and a RESTful API enables programmatic requests of citation snippets and serializations.

Knowledge base of Nanopublications and Statement Verification
We plan to create a knowledge graph considering the nanopublications as the base elements and interlinking elements by considering both relationships between information embedded in the RDF triples and both relationships between elements obtained from the procedure of deference and enrichment of the data within nanopublication. These relationships are formally created and maintained by an automatic linker to be executed just after some changes in the knowledge base such as the introduction of a new element.

Together with the knowledge base, as results of crawling and reasoning within the data in the knowledge base, we will provide tools to get: (i) all nanopublications containing relevant information about a given input; (ii) all scientific statements published that are similar to a statement inserted by the user or that contain similar or relevant elements. Moreover, we plan to design a system which can be used by researchers to automatically or semi-automatically verify if an input statement contradicts existing assertions in the knowledge base.

5 Final Remarks

With this ongoing work, we will provide solutions to some of the open challenges related to the novel model to publish and represent atomic scientific statements.

We contribute to the research in the field of data citation by providing *Nanocitation*, a general framework to obtain nanopublications citation, by outlining a formal procedure and by providing a web app where to automatically get citation snippets and a human-readable visual representation of the information contained in a nanopublication. Our implemented citation system embeds policies and operations that aim to create citation snippets which are short enough to be included in a paper reference list, but, on the other hand, satisfy the data citation requirements (identification and access of the source, persistence, citation completeness and interoperability). Besides, our platform will be useful for researchers for the retrieval and verification of scientific results. Our citation platform, knowledge base platform and methods will make up for the lack of a human-readable visualization and lack of a system of citation, access and search, thus we think that our work outcome will greatly extend the usability of nanopublications and will promote their use.

References

1. Larsen, P. O., von Ins, M.: The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics vol. 84(3), pp. 575-603 (2010)
2. Wood, J., Andersson, T., Bachem, A., Best, C., Genova, F., Lopez, D. R., ... Hudson, R. L.:Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. (2010)
3. Borgman, C.L.: Big Data, Little Data, No Data. MIT Press (2015)
4. Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, USA (2009)
5. Mons, B., van Haagen, H., Chichester, C., Hoen, P.B., *et al.*: The value of data. Nature Genetics **43**(4), 281–283 (4 2011)
6. Groth, P., Gibson, A., Velterop, J.: The Anatomy of a Nanopublication. Inf. Serv. Use **30**(1-2), 51–56 (2010)
7. Fabris, E. Kuhn, T. and Silvello, G.: A Framework for Citing Nanopublications. In: Proc. of the 23rd International Conference on Theory and Practice of Digital Libraries (TPDL 2019), (2019)
8. Silvello, G.: Theory and Practice of Data Citation. Journal of the Association for Information Science and Technology (JASIST) **69**(1), 6–20 (2018)
9. Yinjun Wu, Y., Alawini, A., Davidson, S.B., Silvello, G.,.: Data Citation: Giving Credit Where Credit is Due. SIGMOD '18. 99–114 (2018)
10. Buneman, P., Davidson, S.B., Frew, J.: Why data citation is a computational problem. Communications of the ACM (CACM) **59**(9), 50–57 (2016)
11. Davidson, S.B., Buneman, P., Deutch, D., Milo, T., Silvello, G.: Data Citation: A Computational Challenge. In: Proc. of the 36th ACM Symposium on Principles of Database Systems, PODS 2017. pp. 1–4. ACM Press (2017)
12. Silvello, G.: Learning to Cite Framework: How to Automatically Construct Citations for Hierarchical Data. Journal of the Association for Information Science and Technology (JASIST) **68**(6), 1505–1524 (2017)
13. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data, vol. 12. CODATA-ICSTI (September 2013)
14. FORCE-11: Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. FORCE11, San Diego, CA, USA (2014)