Nanocitation: Complete and Interoperable Citations of Nanopublications

Erika Fabris[1](✉), Tobias Kuhn[2], and Gianmaria Silvello[1]

[1] Department of Information Engineering, University of Padua, Padua, Italy
{erika.fabris,gianmaria.silvello}@unipd.it
[2] Department of Computer Science, VU University Amsterdam,
Amsterdam, The Netherlands
t.kuhn@vu.nl

Abstract. Nanopublication is a data publishing model which has a great potential for the representation of scientific results allowing interoperability, data integration and exchange of scientific findings. But this model suffer of the lack of an appropriate standard methodology to produce complete and interoperable citations providing both data identification and access. In this paper we introduce *nanocitation*, a framework to automatically get human-readable text-snippet snippet and machine-readable citations of nanopublications.

Keywords: Nanopublication · Data citation · DisGeNET

1 Introduction

We have recently witnessed a transition to the data-intensive scientific research paradigm – i.e. the fourth paradigm [8] of science – which led to a change in the nature of scientific discovery and publication. This paradigm shift has made it necessary to adapt the infrastructure for managing the growing amount of scientific data, 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 [8].

One of the grand challenges of data-intensive science is to ease knowledge discovery, evaluation, propagation and reuse. To this end, international initiatives together with academia, industry and publishers designed the so-called

The full paper was presented at TPDL 2019 [5].

The work was partially funded by the "Computational Data Citation" (CDC) STARS-StG project of the University of Padua. The work was also partially funded by the EXAMODE (contract n. 825292) part of the H2020-ICT-2018-2 call of the European Commission.

© Springer Nature Switzerland AG 2020

M. Ceci et al. (Eds.): IRCDL 2020, CCIS 1177, pp. 182–187, 2020.
https://doi.org/10.1007/978-3-030-39905-4_18

FAIR (Findability, Accessibility, Interoperability and Reusability) principles to be followed when producing and storing data.

In this scenario, the nanopublication model has been proposed as a means to represent and publish individual scientific claims or statements together with their provenance specification and publication information. This model allows individual scientific results to be uniquely identifiable, accessible, attributable, citable and reusable [7,11].

Alongside the improvement of scientific data management and infrastructure, another problem has gained importance: data citation [15]. Citations are one of the main "driving force" for scientific progress and, since data has gained the same scholarly status of traditional publications, even data citation has become a "driving force" for scientific progress as well. Data citation is central to enable: (i) credit attribution to database creators and curators not only to papers authors; (ii) connection between scientific papers and used data; (iii) identifiability, reachability and accessibility of data; (iv) knowledge sharing and propagation; (v) evaluation of the impact of the data; (vi) reproducibility of the experiments. To date, two are the facets of data citation which have been studied: the definition of data citation principles and the development of solutions for computational problems related to the automatic or semi-automatic generation of citations. Two main international initiatives (CODATA [1] and FORCE 11 [6]) have defined core principles and criteria for data citation which are: (i) through the citation data should be identifiable on variable granularity; (ii) both citation metadata and cited data should be persistent and accessible; (iii) every citation should come with a citation reference (text-snippet) which describes the cited data and should be complete enough to attribute credit and to interpret the data content; (iv) citations should be flexible and both machine- and human-readable. Moreover, data citation and the design of systems for the automatic creation of data references are considered a computational problem [4]. Some solutions to automatically generate data citations have been proposed in the literature, mostly to create citations of subsets of relational and graph databases [14,16].

Nevertheless, up to now there is no automatic solution to create complete, consistent, interoperable textual citations to single nanopublications. Thus, we introduce a framework to automatically create a citation text-snippet and a machine-readable citation given a single nanopublication and a landing page where all the information within the nanopublication is shown to the user.

The rest of the paper is organized as follows: Sect. 2 introduces the concept of nanopublication, Sect. 3 presents the *nanocitation* framework to obtain citation of a single nanopublication and Sect. 4 draws some conclusions and presents an outlook to future work.

2 Background

A nanopublication is a granular-level publication containing an individual scientific claim (an atomic statement in the form of subject-predicate-object, e.g. *malaria is transmitted by mosquitoes* together with its provenance (its origin and generation process) and publication information.

The nanopublication model makes use of Linked Data W3Cs Resource Description Framework(RDF) specification representing its contents (scientific statement, provenance and publication information) within three graphs in the form of RDF triples (two entities/resources with a certain relation) where each element or ontology term is represented through an Internationalized Resource Identifier (IRI). The potential of the nanopublication model is that each scientific claim can be individually represented, linked to its evidence and can be independently addressed allowing fine-grained citation metrics on the level of individual claims and enable article-data connections [7].

Today, over 10M nanopublications are freely accessible and hosted on a nanopublication network[1] [9] and other 200M are available as independent private datasets. Published nanopublications represent data and scientific claims extracted from datasets from several domains, mostly from Life Science domain dataset including DisGeNET [13], neXtProt [10] and WikiPathways [12], but also smaller nanopublication datasets from digital humanities domain (philosophy, archaeology and music) have published.

Nowadays all the solutions to cite nanopublications (through their identifiers or citing the data papers or the whole dataset where they are stored) guarantee only up to two data citation requirements, i.e. the accessibility of data and persistence of data. But these solutions do not guarantee completeness and interoperability requirements, they do not provide necessary information or provide partial information to attribute the credit to all contributors nor provide content information as well as they raise up issues about the loss of specificity. We tackle these problems by proposing a citation framework which meets all the citation criteria.

3 The Nanocitation Framework

We design *nanocitation*, a framework to automatically obtain citations of nanopublications as illustrated in Fig. 1. This framework concerns the creation of citations for single nanopublications (see violet and blue components in Fig. 1).

The nanocitation framework is composed of four main components, as input it receives the identifier of the nanopublication to be cited (i.e. the URI) and a set of citation policies and produces three outputs:

1. a text-snippet citation to be included in reference lists;
2. the nanopublication citation metadata in machine-readable formats (i.e. XML and JSON);
3. a landing page where the user can explore the content of the nanopublication in a human-readable form.

Dereferencing and Enrichment – The first module of the framework aims at dereferencing all the identifiers of the resources and entities within the nanopublication RDF triples of the nanopublication and aims at searching for all the

[1] http://npmonitor.inn.ac/ accessed on 09/25/2019.

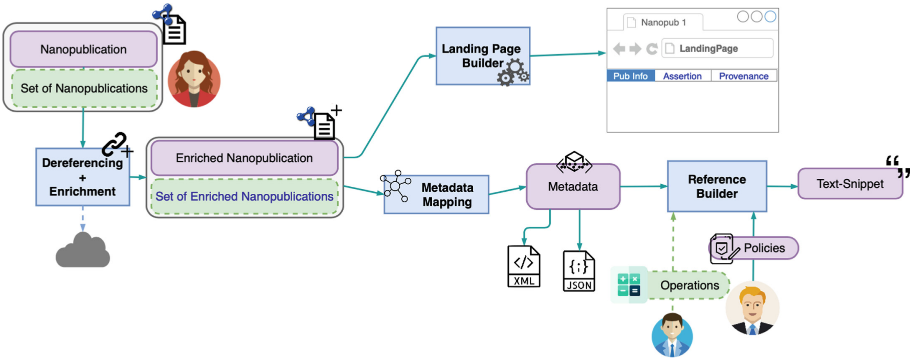

Fig. 1. Framework schema for nanopublication citation. Violet and blue nodes represents the components involved in citation creation, whereas green and blue nodes concern citation creation for a set of nanopublications. (Color figure online)

relevant information related to them from external sources. At the end of the process, the module produces a human-readable and enriched version of the nanopublication, i.e. the *enriched nanopublication*.

Metadata Mapping – Human-readable information within the enriched nanopublication are structured as metadata. For data citation, several metadata formats are proposed by the literature. The most recent and widely recognized metadata format for citing data, DataCite [2] needs to be extended in order to represent the nanopublication citation metadata since several data within the enriched nanopublication do not find any correspondent metadata fields and some metadata fields would need to be overloaded. Thus, we define the semantics and constraint of the metadata by defining an *ad-hoc* metadata schema as a Dublin Core Application Profile. Once the metadata has been created, this can be used as a machine-readable serialization of the content of the nanopublication.

Reference Builder – The reference builder module performs the creation of the citation text-snippet according to some citation policies, which has to be defined by the database administrator in the form of selection, ordering of the fields of the metadata and fields operations (e.g. concatenation).

Landing Page Builder – Moreover, the enriched nanopublication is the input of the landing page builder component which is employed to create the landing page. The landing page is provided to the user to better and fully explore the content of the nanopublication. Through the landing page the user can browse all the content information and, throughout provided links, get to the original sources used to build the nanopublication. The landing page provides a user-friendly interface and human-readable visualization of the complete and specific information about the nanopublication together with the possibility to get the machine-readable form of the landing page content and citation serialization.

We implemented the *nanocitation* framework as a web application freely accessible at nanocitation.dei.unipd.it and we provide also a RESTful API which enables programmatic requests of text-snippet citations, landing pages and citation metadata serializations in JSON and XML format.

4 Discussion

The requirements which a data citation system has to meet: (i) identification of the cited data; (ii) persistence and accessibility of both cited data and citation metadata; (iii) complete and understandable citation text-snippet; (iv) interoperability of the citation in both human- and machine-readable format. Unlike existing solutions to cite nanopublications, our framework satisfies all the above requirements. The identification of the nanopublication is guaranteed by its unique identifier which is reported in the landing page, in the citation metadata and text-snippet. The data is persistent due to the nature of the nanopublication and by the nanopublication network and specification [9], besides, the citation is persistent due to the one-to-one correlation between a given nanopublication and its citation metadata and landing page (a nanopublication is always associated to the same metadata). Moreover, the citation completeness is guaranteed by the presence of all the information relative to the nanopublication in the citation metadata as well as in the landing page, which provides all the information needed to attribute the credit to whom was committed to the creation of the nanopublication and its scientific claim. Furthermore, the human- and machine-readable citations provided as outputs of the framework ensure the interoperability requirements.

To date, our implementation of the framework allows a user to cite single nanopublications, but we are committed to extending the framework to handle the creation of citations of sets/aggregations of nanopublications. Thus, we have to face several problems which are the following: (a) ensure the presence of information of the overall content of the set of nanopublications in the citation even on a situation of heterogeneous content of the single nanopublications; (b) guarantee the completeness of the citation and the creation of a text-snippet concise enough to be integrated in a reference list, which is threatened by the volume of information contained in the set; (c) ensure the presence of the identifiers of all the nanopublications of the set in the citation. To solve all these problems we plan to extend the framework by considering some improvements as shown in Fig. 1 (green and blue nodes). Each nanopublication in the given set undergoes a dereferencing and enriching process and then is mapped into a separate metadata schema. Afterwards, the set of separate metadata are aggregated to form single metadata containing citation information common to all the nanopublications in the set. This cannot be done by performing a mere concatenation of the metadata, but by performing some operations at field-level in the set of metadata. As for the citation policies, the operations to be executed are defined by the system administrator. Then, by applying citation policies to the metadata of the overall set the text-snippet is created. Additionally, the landing page provides complete and understandable information about the content of the nanopublications in the set, alongside the link to the access to landing pages and citations of the single nanopublications.

References

1. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data, vol. 12. CODATA-ICSTI Task Group on Data Citation Standards and Practices, September 2013
2. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data, Version 4.0. Technical report, DataCite Metadata Working Group (2016)
3. Borgman, C.L.: Big Data, Little Data, No Data. MIT Press, Cambridge (2015)
4. Buneman, P., Davidson, S.B., Frew, J.: Why data citation is a computational problem. Commun. ACM (CACM) **59**(9), 50–57 (2016)
5. Fabris, E., Kuhn, T., Silvello, G.: A framework for citing nanopublications. In: Doucet, A., Isaac, A., Golub, K., Aalberg, T., Jatowt, A. (eds.) TPDL 2019. LNCS, vol. 11799, pp. 70–83. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30760-8_6
6. FORCE-11: Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. FORCE11, San Diego, CA, USA (2014)
7. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Inf. Serv. Use **30**(1–2), 51–56 (2010)
8. Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond (2009)
9. Kuhn, T., et al.: Decentralized provenance-aware publishing with nanopublications. PeerJ Comput. Sci. **2**, e78 (2016)
10. Lane, L., et al.: Nextprot: a knowledge platform for human proteins. Nucleic Acids Res. **40**(Database-Issue), 76–83 (2012)
11. Mons, B., et al.: The value of data. Nat. Genet. **43**(4), 281–283 (2011)
12. Pico, A.R., et al.: WikiPathways: pathway editing for the people. PLoS Biol. **22**, e184 (2008)
13. Piñero, J., et al.: DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. **45**(D1), D833–D839 (2017)
14. Silvello, G.: Learning to cite framework: how to automatically construct citations for hierarchical data. J. Am. Soc. Inf. Sci. Technol. (JASIST) **68**(6), 1505–1524 (2017)
15. Silvello, G.: Theory and practice of data citation. J. Am. Soc. Inf. Sci. Technol. (JASIST) **69**(1), 6–20 (2018)
16. Wu, Y., Alawini, A., Davidson, S.B., Silvello, G.: Data citation: giving credit where credit is due. In: Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, pp. 99–114. ACM Press, New York (2018)