A Framework for Citing Nanopublications

Fabris, Erika; Kuhn, Tobias; Silvello, Gianmaria

published in
Digital Libraries for Open Knowledge
2019

DOI (link to publisher)
10.1007/978-3-030-30760-8_6

document version
Publisher's PDF, also known as Version of record

document license
Article 25fa Dutch Copyright Act

citation for published version (APA)
Fabris, E., Kuhn, T., & Silvello, G. (2019). A Framework for Citing Nanopublications. In A. Doucet, A. Isaac, K. Golub, T. Aalberg, & A. Jatowt (Eds.), *Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings* (pp. 70-83). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 11799 LNCS). Springer Verlag. https://doi.org/10.1007/978-3-030-30760-8_6

A Framework for Citing 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. In this paper we discuss the role of the Nanopublication (nanopub) model for scholarly publications with particular focus on the citation of nanopubs.

To this end, we contribute to the state-of-the-art in data citation by proposing: the *nanocitation* framework that defines the main steps to create a text snippet and a machine-readable citation given a single nanopub; an *ad-hoc* metadata schema for encoding nanopub citations; and, an open-source and publicly available citation system.

Keywords: Nanopublication · Data citation · DisGeNET

1 Introduction

Nowadays, the role of data in scientific research is central and the transition to the fourth paradigm of science marks a change even in academic publications. New concepts of data scholarship [5] and data-intensive research [16] have now become extremely important to the world of science. Experimental and observational data along with scientific models pertain mainly to a digital domain. Scholarly publications and credit attribution processes are increasingly considering this trend. Data can thus be regarded as first-class subjects of the system of sciences and, in many cases, are becoming scholarly publications in their own right [22]. Moreover, data publication is regarded as a prerequisite for data sharing and reuse, often through data journals that mirror the scientific publication model [7].

In this context, the connection between scientific claims and the underlying data that lead to their definition or discovery should be explicit and evident. As generally known, "we must respect the connection between the articles and the [underlying] data and value both appropriately" and that "*de novo* claims and the supporting data should be exchanged in machine-readable [and] unambiguous format" [22].

To this end, the Nanopublication (nanopub) model has been proposed as a means to represent scientific statements and results and make them uniquely

A. Doucet et al. (Eds.): TPDL 2019, LNCS 11799, pp. 70–83, 2019.
https://doi.org/10.1007/978-3-030-30760-8_6

identifiable, accessible, attributable, citable and reusable [14]. A nanopub is normally represented as three Resource Description Framework (RDF) graphs: the main one containing the scientific assertion; the second, the assertion source—e.g., where the assertion comes from and how it has been defined; the third one, called publication information, contains metadata of the nanopub indicating who curated the assertion and when the nanopub was created. A nanopub is a self-contained entity associated with a persistent identifier that can be used to explicitly connect articles with underlying data.

The use of nanopubs is becoming more popular, especially in the domain of life sciences where a growing dataset of more than 10 million nanopubs comes from heterogeneous sources such as DrugBank[1], DisGeNET[2], Global biotic interactions[3] and others [20]. Nanopubs are also increasingly prominent within the domain of digital humanities where they are used for philosophical[4], archaeological [12] and music notation[5] purposes. The use of nanopub is also found in the context of Europeana as a format for data on humanities[6].

The potential of nanopubs in scholarly publications is expanding since they enable article-data connections and credit attribution to data creators and curators. Their role is however limited to the identification, representation, and access to scientific assertions, but there is no generally-agreed method to cite them at a specific and general level of aggregation. Nanopubs are generally cited at the dataset level and not at the single statement level, which is the very purpose on the nanopub model [20]. This limits the chance to attribute credit to data creators and curators and estimate the impact of data themselves.

The aim of our work is to define a model for single nanopubs citations and provide a system to automatically create references to cite nanopubs. A data citation system has to meet some basic requirements [1,10,25]: (i) identification (with variable granularity) of the cited data; (ii) access to the cited data in the form they were cited; (iii) create human-readable citation text snippets to be used as reference in the articles; (iv) develop machine-readable citations/references to enable interoperability. The nanopub model and its supporting services guarantee data identification and access, we therefore focus on the definition and creation of human- and machine-readable citations at the single nanopub level.

The citation text snippet has to address the trade-off between completeness of the information it reports and its length. A dataset usually has hundreds of creators, curators, and contributors who cannot be exhaustively reported in a scientific paper due to space limitations. On the other hand, every contributor counts and should get credit. For this reason, we define the citation text snippet based on custom citation policies, established by data curators or database

[1] https://www.drugbank.ca/.
[2] http://www.disgenet.org.
[3] https://www.globalbioticinteractions.org/.
[4] http://emto-nanopub.referata.com/wiki/EMTO_Nanopub.
[5] https://mith.umd.edu/research/enhancing-music-notation-addressability/.
[6] https://dm2e.eu/open-humanities-awards-early-modern-european-peace-treaties-online-final-update/.

administrators, aimed at finding the best trade-off between completeness and length of the citation text snippet. Moreover, the citation system we propose, together with the citation text snippet, returns a web landing page containing all the information regarding the cited nanopub and the citation itself.

Contributions of this work include:

1. a general framework (i.e. the *nanocitation* framework) that defines the main steps to create a citation text snippet and a machine-readable citation given a single nanopub;
2. an *ad-hoc* metadata scheme for encoding nanopub citations;
3. an open-source nanopub citation system and a publicly available WebApp that given a nanopub identifier, returns a text snippet and a web landing page.

Outline. The other part of the paper is organized as follows: Sect. 2 discusses related work about data citation and nanopubs. Section 3 introduces a use case based on DisGeNET gene-disease association data and a running example used throughout the paper. Section 4 presents the nanocitation framework and the metadata scheme. Section 5 describes how the citation system is realized, and finally, Sect. 6 draws some final remarks and outlines for future work.

2 Background

Linked Open Data and RDF. The Linked Open Data (LOD) paradigm [15] refers to a set of best practices for the publication of data on the Web, based on a standardized data model, RDF [17]. RDF is designed to represent information in a minimally constraining way. It is based on the following building blocks: graph data model, IRI-based vocabulary, data types, literals, and several serialization syntaxes. The basic structural construct of RDF is a triple: subject, predicate, and object. It can be represented in a graph where subjects and objects are the nodes and the predicates the arcs. An RDF dataset is a collection of RDF graphs composed of a default RDF graph and a set of named graphs [8]. The latter are pairs consisting of an IRI (i.e. the name of the graph) and an RDF graph.

Nanopublication. Nanopubs are "the smallest unit of publishable information: an assertion about anything that can be uniquely identified and attributed to its author"[7]. The nanopub model was introduced to overcome the increasing difficulties to retrieve, exchange and connect scientific results with the underlying data as a consequence of an ever-growing amount of scientific papers and datasets [14]. A nanopub is a publication itself that carries all the information to be understood and re-used by humans and machines.

The key idea behind the nanopub model is that a scientific result can be divided into individual atomic statements—i.e., assertions—represented as RDF triples. A nanopub comprehends all the information related to a single assertion

[7] http://nanopub.org/.

which is structured into three main named graphs: (i) the *assertion* graph containing the information related to the main scientific statement of the nanopub; (ii) the *provenance* graph containing information on the origin and creation process of the assertion; and (iii) the *publication info* graph including the metadata about the creation and publication of the nanopub. The above three graphs are interconnected by means of an additional graph: the *head* graph.

The main goals of the nanopub model are to promote interoperability among scientific results, data integration and trustworthiness, to ease the access to a scientific statement and to enable the citation of atomic statements allowing for fine-grained citation metrics on the level of individual claims [11,14,19].

The nanopub model has been used to represent data and content of sources from different fields, so far mainly in biomedicine and bioinformatics, e.g. WikiPathways[8], DisGeNET, and neXtProt[9]. In general, there are currently more than 10M nanopubs publicly accessible [20] at http://npmonitor.inn.ac/ and mirrored on 13 different server instances.

Data Citation. Citations are a cornerstone for the diffusion of knowledge in science and a pivotal tool to assess the quality of research as well as to direct investments in science. In this transition phase towards the fourth paradigm of science, data are as vital to scientific progress as traditional publications. For this reason, data are considered as first-class subjects of the system of sciences and data citation is gaining ground in all scientific fields [27].

Until now, data citation has been studied from two main perspectives: defining core principles and developing computational solutions. Two major international initiatives have focused on defining the core principles for data citation: CODATA, which published a report on data citation principles [1]; and FORCE 11, which published a list of principles summarising the ideas of a number of working groups [10]. The principles not only highlight the idea that data is a research object that should be citable, giving credit to data curators, but they also identify the criteria that a citation should follow: (i) the identification of and access to the cited data; (ii) the persistence of data identifiers as well as related metadata; (iii) the completeness of the reference, thus data citation that contains all the necessary information to interpret and understand the data; and, (iv) the interoperability of citations, thus interpretable both by humans and machines.

Data citation has been recently defined also as a computational problem [6] and some solutions to automatically generate data citations have been proposed in the literature [3,9,26,29]. Most of the solutions proposed focus on relational and graph databases where a citation for a given query must be provided, exploiting database views to build a viable reference. In [25] a system to store user queries is defined and these queries are treated as proxies of cited data, but it does not provide a solution to produce human- and machine-readable references. In general, none of the existing solutions can be easily applied to nanopubs. Whenever a scientific assertion is central, information from named graphs must automatically be extracted and no user query or views can be leveraged on.

[8] https://www.wikipathways.org.
[9] https://www.nextprot.org.

Although one of the main goals of nanopubs is to enable the citation of atomic scientific statements, no solution to automatically build human- and machine-readable nanopub citations has been found so far.

3 Use Case: DisGeNET

We consider the biomedical platform DisGeNET (ver. 4.0.0) as a use case and we employ it as a running example throughout the paper.

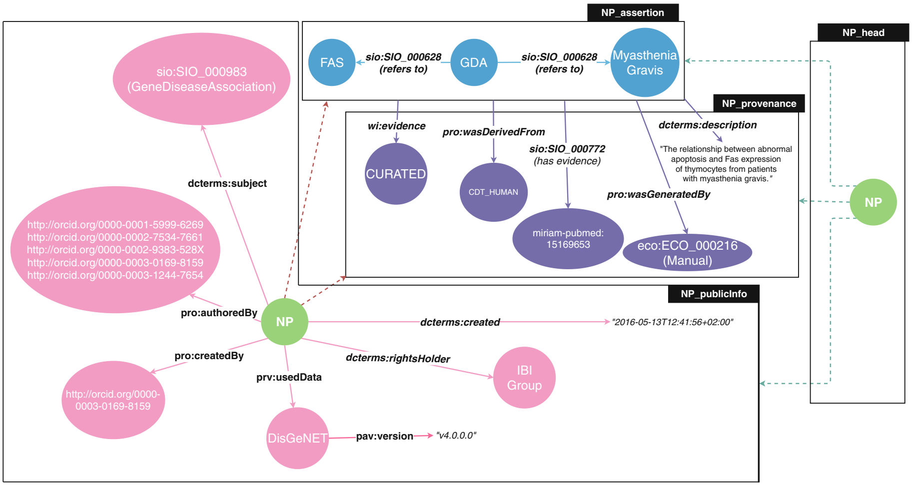

Fig. 1. Running example: A simplified graphical representation of DisGeNET nanopub.

DisGeNET is an open discovery platform containing comprehensive information on gene-disease associations (GDA) and variant-disease associations (VDAs) gathered through extraction and text-mining procedures from expert curated databases, catalogues and the scientific literature [23]. DisGeNET is widely used in several domains because it provides: free access to one of the most comprehensive collections of GDAs and VDAs (i.e. 430K GDAs, linking 17K genes to 15K diseases, and 72K VDAs), access to heterogeneous resources and external datasets, a web interface with search functionalities, an automatic generator of customized scripts, suitable tools to visualize and explore GDAs networks, and machine-readable version of the data as RDF. Moreover, DisGeNET is one of the foremost sources providing access to its data as nanopubs [24].

Figure 1 shows a simplified graphic representation of a typical DisGeNET nanopub used as a running example in the paper. It consists of three named graphs linked together by a fourth named graph called "head graph". The publication info graph provides all the information related to the publication of the DisGeNET nanopubs such as creation date, owner rights, names of creators and contributors. Worth noting that all the DisGeNET nanopubs are authored by five contributors and one creator. They target one general assertion topic—i.e., gene-disease association. The provenance graph contains all the information about the origin of the assertion and its generation process—i.e., the identifier of the scientific paper containing the evidence of the assertion, classification derivation and the way the assertion was created, either automatically or manually. The assertion graph maps the details about the gene-disease association. In Fig. 1 the assertion consists on the relationship (i.e., `GDA`) between the gene `Myasthenia Gravis` and the disease `FAS`. Currently, in DisGeNET all nanopubs have the same structure and number of triples.

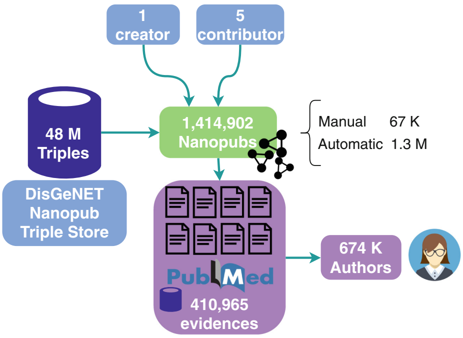

Fig. 2. DisGeNET nanopubs v. 4.0.0 statistics.

In Fig. 2, some basic statistics about DisGeNET and its nanopubs are provided. There are 1,414,909 nanopubs for a total of about 48M of triples. The majority of nanopubs are automatically created – 1.3M – and nanopubs refer to 410K PubMed evidences (citations and abstracts for biomedical literature) written by a total of 674K authors.

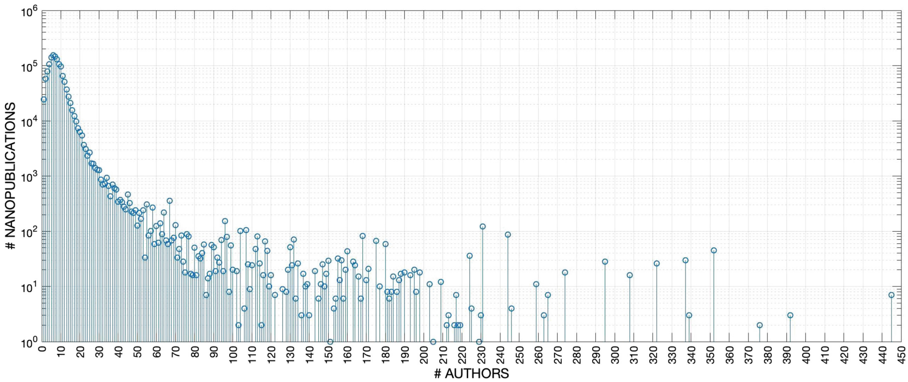

Fig. 3. Number of nanopubs grouped by the same number of evidence authors (y-axis is log-based) – e.g., ∼150K nanopubs are based on an evidence with 6 authors.

Two key facts are therefore worth highlighting: (i) many nanopubs are based on the same evidence and a nanopub can be based on several evidences; and, (ii) many nanopubs refer to evidences written by a large number of authors.

Figure 3 reports the number of nanopubs that share the same number of evidence authors. As can be seen, there are thousands of nanopubs with dozens of evidence authors and that in general, almost all nanopubs have more than ten evidence authors. For this reason, all the evidence authors in the text snippet of a nanopub cannot be reported properly.

4 The Nanocitation Framework

We propose *nanocitation*, a framework to cite nanopubs composed of four main components that takes the URI of a single nanopub and a set of citation policies as inputs and returns a reference (i.e. text snippet) and a web landing page as outputs.

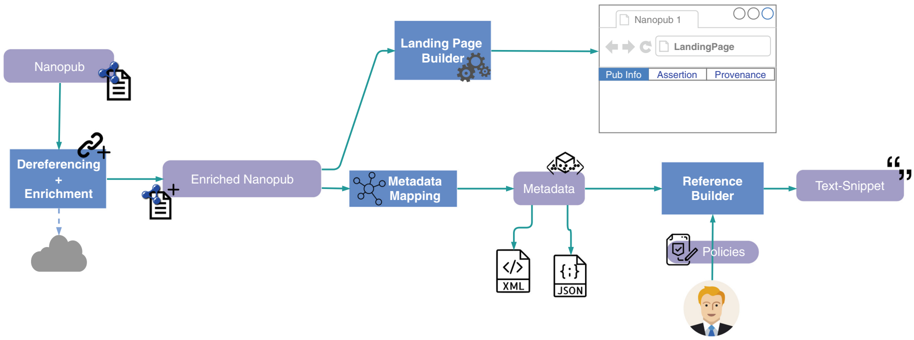

Fig. 4. Framework schema for nanopub citation.

Figure 4 shows the main components of the nanocitation framework. The citation process is created starting from a raw nanopub – i.e., the object of the citation. The `Dereferencing and Enrichment` component dereferences all the URIs composing the nanopub triples and searches for relevant information on the Web to enrich the nanopub. Afterwards, the enriched nanopub is mapped into metadata that encode all the necessary information to create the text snippet. The metadata can also be serialized as XML or JSON files. The text snippet is created by selecting and transforming some fields of metadata, on the basis of citation procedures defined by the database administrator or the nanopubs curator. The output of the `reference builder` component is the final text snippet. Moreover, the enriched nanopub is the input of the `landing page builder` component employed to create the landing page.

Enriched nanopub. An enriched nanopub is a human-readable version of the nanopub where all the URIs have been dereferenced by accessing external sources such as: ORCID[10] to get the nanopub creators and curators names, PubMed to get evidence of the assertion; relevant ontologies such as ProteinAtlas[11] and NCBI[12] to get details on assertion terms on genes, diseases and proteins. The enriched nanopub contains more information than the original one: it offers the list of evidence-authors extracted from PubMed, detailed and related information on genes and diseases found in the assertion.

Metadata. Human-readable information composing the enriched nanopub are structured as metadata. For data citation several metadata formats are proposed by the literature [4,13,28]. DataCite [2], the most recent and widely recognized metadata format proposal for citing data, provides a common set of fifteen fields. Nevertheless, the DataCite format cannot be used to cite a nanopub since several data within the enriched nanopub do not find any correspondent field or lead to field overload as reported in Table 1.

Hence, a custom metadata scheme was defined as a Dublin Core Application Profile. Table 1 illustrates the core terms of this scheme and a concrete metadata instance based on the running example reported above. All literal terms are formatted strings based on standards that depend on the entity they represent – i.e. W3CDTF standard for date types, RFC3986 syntax for URIs or identifiers. Other terms refer to non-literal customized types. These are: `creator` and `contributor`, defined as `person` type (composed of `givenName`, `familyName` and `personID` string-fields) or `organisation` type (composed of a string field `orgName`), `evidenceAuthor`, defined as `person` type and `content`, defined as set of pairs `subject–assertion(s)` (there may exist multiple assertions associated to a single subject).

The metadata description set and the XML/JSON scheme are publicly available from the homepage of the *nanocitation* WebApp.

Policies. To create the text snippet some relevant metadata fields need to be selected and transformed. This process is controlled by the database administrator or the nanopub creator using a set of predefined data citation policies. Data citation policies are constituted by a set of operations on the metadata fields.

Let us consider the instance in Table 1, a viable citation policy is:

Selection and ordering: `creator`, `contributor`, `creationDate`, `rightsHolder`, `content`, `platform`, `version`, `evidenceAuthor`, `landingpageUrl`.

Operations:
- `creator`, `contributor` and `evidenceAuthor`: select `givenName` and `familyName` of the first element and, in the case of more than one creator (curator), affix "et al." followed by brackets containing the total number of creators (curators).
- `creationDate`, `rightsHolder`, `platform`, `version` and `landingpageUrl`: use the content field as it is.

[10] https://orcid.org/.
[11] https://www.proteinatlas.org/.
[12] https://www.ncbi.nlm.nih.gov/.

Table 1. Nanopublication metadata schema and a description of its use based on the running example.

Term	Min-Max	Lit.	Instance	DataCite
NanopubIdentifier	1-1	l	http://rdf.disgenet.org/resource/nanopub/NP14146.RALnE6gfq0eJ45bgXkmCLy3rV9GT2VSRUMfRYS9qR8KQk	Identifier
Creator	1-∞	non-l	[(Núria, Queralt Rosinach, http://orcid.org/0000-0003-0169-8159)]	Creator
Contributor	0-∞	non-l	[(Álex, Bravo Serrano, http://orcid.org/0000-0001-5999-6269), (Ferran, Sanz, http://orcid.org/0000-0002-7534-7661), (Laura I., . . .]	Contributor
CreationDate	0-1	l	2016-05-13	Date
RightsHolder	0-1	l	IBIGroup	Contributor
Content	1-1	non-l	[(gene-disease association (gene-disease biomarker association), FAS Fas cell surface death receptor - Myasthenia Gravis)]	–
Platform	0-1	l	Disgenet	–
Version	0-1	l	v4.0.0.0	Version
EvidenceReference	0-∞	l	[http://identifiers.org/pubmed/15169653]	RelatedIdentifier
EvidenceAuthor	0-∞	non-l	[(Y, Du, -), (QY, Zhang, -), (LR, Ruan, -), (CC, Liang, -), (W, He, -)]	Contributor
LandingpageUrl	1-1	l	nanocitation.dei.unipd.it/landingpage/RALnE6gfq0eJ45bgXkmCLy3rV9GT2VSRUMfRYS9qR8KQk	RelatedIdentifier

– `content`: select the `subject` followed by a colon and the `assertions` separated by comma.

Presentation: Separate text-snippet elements with a comma.

The *selection and ordering* component defines the list of fields that are selected from the metadata and the order in which they appear in the text snippet. The *operations* component defines the operations to be performed on the single selected fields, and the *presentation* component defines how the field's content is rendered in the text snippet. For sake of clarity, we have not presented the specific syntax that define the operations over the fields, but have focused on the idea of citation policy at a higher abstraction level. By applying this policy to the metadata reported in Table 1 the following text-snippet is obtained:

> Queralt Rosinach Núria, Piñero Janet et al. (5), 2016-05-13, IBIGroup, gene-disease association (gene-disease biomarker association): FAS Fas cell surface death receptor - Myasthenia Gravis, disgenet, v4.0.0, Du Y et al. (6). nanocitation.dei.unipd.it/landingpage/RALnE6gfq0eJ45bgXkmCLy3rV9GT2VSRUMfRYS9qR8KQk

Worth noting that the text-snippet contains all the information that allows the reader to easily identify the authors and contributors who provided the data, the date of creation, the publish information of the source, the general content and the authors' evidence of the content, alongside the link to the landing page where more information is provided. Text-snippets may vary in length, depending on the policies applied: the above example is concise enough to be included in the reference list of a paper.

5 The Nanocitation System

We implemented the nanocitation framework as a WebApp accessible at the URL: nanocitation.dei.unipd.it

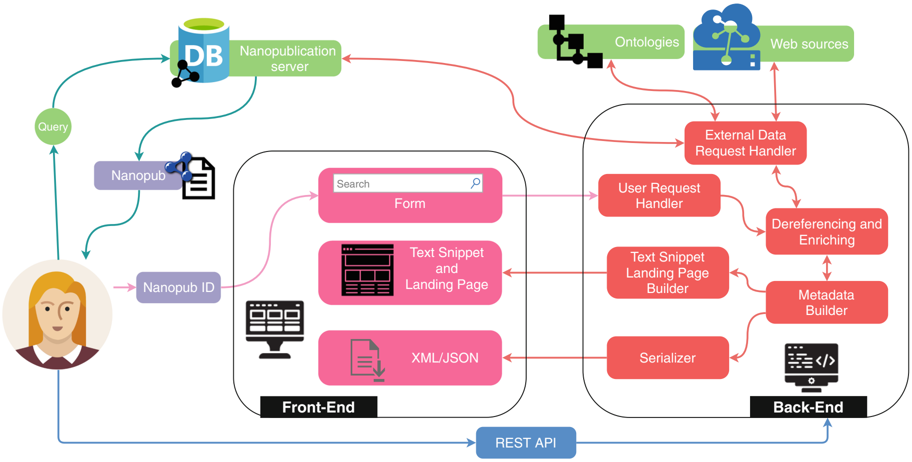

Fig. 5. Nanopublication citation system architecture.

Figure 5 illustrates the overall structure of the nanocitation system composed of a front-end and a back-end component.

The front-end provides a user interface which takes the identifier of the nanopub to be cited as input, returning the landing page of the citation. The landing page is structured as shown in Fig. 6. It provides the citation text-snippet, the link to the nanopub and three main collapsible elements where all the information contained in the nanopub (named graphs) and the additional information added in the enrichment phase are reported. From the landing page, the user can access external sources containing a more detailed definition or further data related to the nanopub. For instance, the authors are connected to their ORCID pages, and the assertion elements are connected to their ontological definitions. The landing page offers the possibility to download the XML and JSON serialization of the citation metadata.

The back-end comprises six main components. The *User Request Handler* receives the identifier requested by the user and activates the citation mechanism by calling the *Dereferencing and Enrichment* component. This component

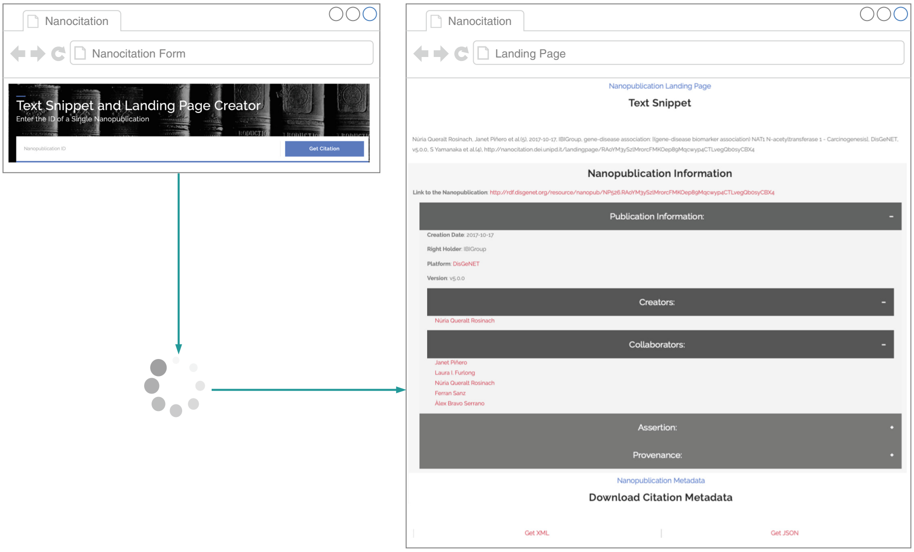

Fig. 6. Two screen-shots showing the primary user interfaces of the nanocitation system.

gathers data from external web sources – i.e., platforms such as Linked Life Data[13], National Center for Biotechnology Information (NCBI)[14], ontologies and vocabulary resources including SemanticScience Integrated Ontology (SIO) and NCI Thesaurus. The whole process is handled by an *External Data Request Handler*. Once the nanopub elements have been dereferenced and enriched, the *Metadata Builder* component creates the metadata object as defined above. The *Text Snippet/Landing Page Builder* component creates the text snippet and the landing page by applying the policies specified by the database administrator or the data curator. As a result, the landing page provides additional data and references on the aforementioned nanopub in a readable format for both users who do not know the structure of nanopub as well as for informed users.

Moreover, a RESTful API enables programmatic requests of the citation text snippet, the citation landing page and the XML or JSON serialization of the citation metadata.

The back-end component is implemented in Java 8 and relies on the `nanopub-java` external library [18]. The WebApp runs on an Apache Tomcat ver. 9 web server based on the Spring MVC framework ver. 5[15] and on Thymeleaf ver. 3[16]. The code is available as open-source at: https://github.com/erikafab/nanocitation/.

13 http://linkedlifedata.com.
14 https://www.ncbi.nlm.nih.gov.
15 https://spring.io/.
16 https://www.thymeleaf.org/.

We tested the system by randomly sampling 1000 DisGeNET nanopubs, recording the time needed to generate the text snippet, the number of empty fields in the text snippets and in the citation metadata. On average, text snippets are 335-characters long; this length is comparable to a conference bibliographic citation. The average generation time is 10 s, due mainly to the delay time to get data from external resources. The percentage of citation metadata with missing fields is 2% of the total, where the only missing field is `evidenceAuthor` due to nanopubs without explicit reference to an evidence paper.

6 Conclusion and Future Work

In this paper we discuss the nanopub model in the context of scholarly publications highlighting the importance of data citation for credit attribution at the data level. As contribution to the research in the field of data citation, we propose the *nanocitation* framework for citing single nanopubs and an open-source system that enables the automatic creation of citation text snippets and landing pages. Until now, the nanopublications were referenced at the datasets level or just by the means of their identifiers; with our system researchers can create references to the specific nanopubs used or mentioned in their works. In addition, our work offers a system for in-depth exploration of their content (landing page) that is missing in the current state of the art in nanopubs.

The proposed nanocitation framework meets data citation requirements discussed in the Background section, namely: (i) identification and access of the cited data; (ii) persistence of data ids and related metadata; (iii) completeness of the citation; and, (iv) interoperability. In fact, a nanopub is uniquely identified by its URL and the citation landing page is uniquely connected to it because its URL includes the identifier of the nanopub. This guarantees that a landing page is always associated with one and only one nanopub. Moreover, given a nanopub identifier, the nanocitation system always returns the same landing page and citation metadata (the text snippet depends on the specific citation policy). Currently, the page and the metadata are generated dynamically, so they rely on the availability of the nanopub and external information. Nevertheless, without weakening the general framework and avoiding any significant changes, the landing page and the citation metadata can be dumped and stored in a database to guarantee their availability even beyond the nanopub lifespan.

Nanopubs persistence is guaranteed by the nanopub specification and the storage system put in place to deal with fixity of nanopubs, as described in [19]. When a nanopub is updated, a new nanopub is minted or a versioning system based on the creation date is activated [21]. In the first case, the citation produced by the nanocitation system always refers to the initially cited nanopub; whereas, in the second case, the creation date of the nanopub stored in the citation metadata refers to the initially cited nanopub. In the latter case, we cannot rely on a dynamic citation system, and all citation metadata at creation time must be dumped and stored.

The completeness of the citation text snippet depends on the citation policy. Nevertheless, the citation metadata contains all the available information about the nanopub and related resources. Consequently, with adequate citation policies (i.e., selection and operations over the metadata fields) the text snippet can be correctly interpreted by the users.

What guarantees interoperability is the nanocitation system which returns a human-readable text snippet and a landing page as well as a machine-readable serialized version of the citation metadata. The metadata can be automatically processed by bibliometric systems to count citations or to calculate other relevant metrics.

As future work we intend to extend the citation framework and the citation system of nanopub aggregation with variable granularity and to provide a language to define citation policies.

Acknowledgments. 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.

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. Alawini, A., Davidson, S.B., Silvello, G., Tannen, V., Wu, Y.: Data citation: a new provenance challenge. IEEE Data Eng. Bull. **41**(1), 27–38 (2018)
4. Altman, M., King, G.: A proposed standard for the scholarly citation of quantitative data. D-Lib Mag. **13**(3/4) (2007)
5. Borgman, C.L.: Big Data, Little Data, No Data. MIT Press, Cambridge (2015)
6. Buneman, P., Davidson, S.B., Frew, J.: Why data citation is a computational problem. Commun. ACM (CACM) **59**(9), 50–57 (2016)
7. Candela, L., Castelli, D., Manghi, P., Tani, A.: Data journals: a survey. J. Assoc. Inf. Sci. Technol. **66**(9), 1747–1762 (2015)
8. Carroll, J., Bizer, C., Hayes, P., Stickler, P.: Semantic web publishing using named graphs. In: Proceedings of the ISWC 2004 Workshop on Trust, Security, and Reputation on the Semantic Web. CEUR Workshop Proceedings, vol. 127. CEUR-WS.org (2004)
9. Davidson, S.B., Buneman, P., Deutch, D., Milo, T., Silvello, G.: Data citation: a computational challenge. In: Proceedings of the 36th ACM Symposium on Principles of Database Systems, PODS 2017, pp. 1–4. ACM Press (2017)
10. FORCE-11: Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. FORCE11, San Diego, CA, USA (2014)
11. Gibson, A.P., van Dam, J.C.J., Schultes, E., Roos, M., Mons, B.: Towards computational evaluation of evidence for scientific assertions with nanopublications. In: Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences (2012)

12. Golden, P., Shaw, R.: Period assertion as nanopublication: the PeriodO period gazetteer. In: Proceedings of the 24th International Conference on World Wide Web, WWW 2015 Companion, pp. 1013–1018. ACM Press (2015)
13. Green, T.: We need publishing standards for datasets and data tables. Technical report. OECD Publishing (2010)
14. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Inf. Serv. Use **30**(1–2), 51–56 (2010)
15. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers, San Rafael (2011)
16. Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond (2009)
17. Klyne, G., Carroll, J.J.: Resource description framework (RDF): concepts and abstract syntax. Technical report W3C (2004)
18. Kuhn, T.: nanopub-java: a Java library for nanopublications. In: Proceedings of the 5th Workshop on Linked Science 2015 co-located with (ISWC 2015). CEUR Workshop Proceedings, vol. 1572, pp. 19–25. CEUR-WS.org (2015)
19. Kuhn, T., et al.: Decentralized provenance-aware publishing with nanopublications. PeerJ Comput. Sci. **2**, e78 (2016)
20. Kuhn, T., et al.: Nanopublications: a growing resource of provenance-centric scientific linked data. In: 14th IEEE International Conference on e-Science, pp. 83–92. IEEE Computer Society (2018)
21. Kuhn, T., Willighagen, E., Evelo, C., Queralt-Rosinach, N., Centeno, E., Furlong, L.I.: Reliable granular references to changing linked data. In: d'Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10587, pp. 436–451. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68288-4_26
22. Mons, B., et al.: The value of data. Nat. Genetics **43**(4), 281–283 (2011)
23. 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)
24. Queralt-Rosinach, N., Piñero, J., Bravo, À., Sanz, F., Furlong, L.: DisGeNET-RDF: harnessing the innovative power of the semantic web to explore the genetic basis of diseases. Bioinformatics **32**(14), 2236–2238 (2016)
25. Rauber, A., Ari, A., van Uytvanck, D., Pröll, S.: Identification of reproducible subsets for data citation, sharing and re-use. Bull. IEEE Techn. Comm. Digit. Libr. Spec. Issue Data Cit. **12**(1), 6–15 (2016)
26. Silvello, G.: Learning to cite framework: how to automatically construct citations for hierarchical data. J. Assoc. Inf. Sci. Technol. (JASIST) **68**(6), 1505–1524 (2017)
27. Silvello, G.: Theory and practice of data citation. J. Assoc. Inf. Sci. Technol. (JASIST) **69**(1), 6–20 (2018)
28. Starr, J., Gastl, A.: isCitedBy: a metadata scheme for DataCite. D-Lib Mag. **17**(1/2) (2011)
29. 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 (2018)