Data Quality Issues in Current Nanopublications

Data Quality Issues in Current Nanopublications

Imran Asif
(0000-0002-1144-6265)
School of Computer Science and Mathematics
Heriot-Watt University
Edinburgh, UK
Email: ia48@hw.ac.uk

Jessica Chen-Burger
(0000-0002-7909-0541)
School of Computer Science and Mathematics
Heriot-Watt University
Edinburgh, UK
Email: Y.J.ChenBurger@hw.ac.uk

Alasdair J G Gray
(0000-0002-5711-4872)
School of Computer Science and Mathematics
Heriot-Watt University
Edinburgh, UK
Email: A.J.G.Gray@hw.ac.uk

***Abstract*—Nanopublications are a granular way of publishing scientific claims together with their associated provenance and publication information. More than 10 million nanopublications have been published by a handful of researchers covering a wide range of topics within the life sciences. We were motivated to replicate an existing analysis of these nanopublications, but then went deeper into the structure of the existing nanopublications. In this paper, we analyse the usage of nanopublications by investigating the distribution of triples in each part and discuss the data quality issues that were subsequently revealed. We argue that there is a need for the community to develop a set of guidelines for the modelling of nanopublications.**

1. Introduction

Scientific research relies on sharing ideas and results between researchers so that they can be independently tested and verified. Traditionally, this has been done in paper publications that are generally made available as PDFs or more recently as HTML pages on the Web. Much of the scientific work is reliant on data that is either made available in a public repository or published alongside the research paper. However, these are often large collections of data containing multiple claims, potentially from several authors using different collection methods. These datasets are published as a single unit, often with only rudimentary provenance and author information.

Nanopublications [1] provide a mechanism to publish individual claims together with fine-grained provenance specific to the claim, and publication metadata. To date, there have been over 10 million nanopublications published on the nanopublication network[1] [2], by a handful of researchers mostly focused on the life sciences. It has been argued that this approach provides improved data quality and attribution since the provenance of each claim can be individually verified, rather than the traditional coarse grained provenance and metadata associated with large datasets. The drawback of the nanopublication approach is that it significantly increases the size of the dataset. However, Kuhn *et al* [3] have shown that for versioned datasets this overhead is actually less than publishing each complete version of the claims in the dataset as done by traditional data publishing, with the advantage of the increased provenance of the data.

In this paper we look to repeat the analysis of Kuhn *et al* [3]. However, we found ourselves asking more questions about the collection of nanopublications and thus present our extended analysis of the nanopublication collection. We revealed issues about the current practice of publishing nanopublications from traditional datasets and the overall quality of the collection.

2. Background

A Nanopublication [1] is a granular-level, semantic, scientific publication of a claim together with its provenance and publication information. They are represented in RDF and consist of three sub-graphs. The Assertion graph contains the claim being published in the nanopublication. The Provenance graph contains the evidence to support the claim. The Publication graph contains the metadata about the nanopublication itself, i.e. who published it and when. These are connected together in the Head graph.

To understand the nanopublication, we take a simple example of a scientific claim that was originally used in [1]. The claim is *"Malaria is transmitted by mosquito"*. In this example, we have three things; two concepts (Malaria and Mosquito) and one relationship that is "Transmitted by". This statement can be represented in RDF as a triple with the Subject (Malaria), Predicate (Transmitted by), and Object (Mosquito). To store this claim in a nanopublication four named RDF graphs are used [4] as shown in Figure 1.

The structure of a nanopublication adds a large overhead to the publication of each claim when compared with just publishing the claim triple as is done in traditional data publishing. However, the benefit is that each claim is published with provenance and publication information pertinent to the claim. Kuhn *et al* [5] introduced a mechanism for indexing and reusing nanopublications which they showed eliminates

1. http://npmonitor.inn.ac/ accessed 21 June 2019

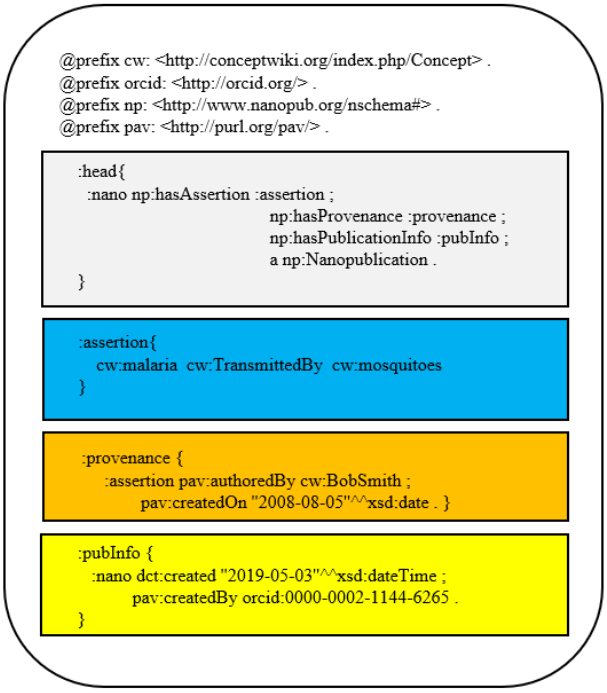

Figure 1. Example Nanopublication derived from [1]. The grey box depicts the head graph, the blue the assertion graph, the orange the provenance graph, and the yellow the publication information graph.

this overhead when compared to the traditional approach of republishing all triples in each version of a dataset.

Nanopublications can be published through a distributed peer-to-peer network called the nanopub network [2]. To date, there are over 10 million nanopublications that have been published on the nanopub network, mostly containing data from different life sciences datasets, including DisGeNET [6], neXtProt [7], and WikiPathways [8]. These nanopublications are additionally published using Trusty URIs [9] which provide a way for digitally signing the content of the publication and encoding this in the URI of the publication. Nanopublications that are published to the nanopub network using TrustyURIs are immutable, permanent, verifiable, and decentralized.

3. Data and Experiment Methodology

In this paper we were motivated to replicate some of the analysis presented in [3] and [5]. This involves reusing a subset of the data on the nanopublication network. We will now briefly describe the data used with a summary given in Table 1. Full details of the datasets and how they are generated can be found in [3], [5]. We will provide a fuller discussion of Table 1 in Section 4.

The datasets used in this paper are DisGeNET[2] version 4.0 [6], neXtProt[3] version 19001_20000 [7], WikiPathways[4] version 20170513 [8], OpenBEL large and small corpus[5] version 20131211 [10], and LIDDI[6] version 1.02 [11]. We note that DisGeNET is now at version 6.0 and WikiPathways is at version 20190510. However, our motivation was to replicate the work of Kuhn *et al*, and thus, we reuse the same versions of DisGeNET and WikiPathways. All the datasets used in this study come from the life sciences domain.

DisGeNET, neXtProt, and WikiPathways are all generated by a script that creates nanopublications based on the content of a traditional data store. This script is (typically) run with each data release, creating a new set of nanopublications for the dataset. The OpenBEL nanopublications were generated by Tobias Kuhn using the bel2nanopub[7] script. The LIDDI nanopublications were generated by Juan M. Banda.

The nanopublications were downloaded and stored into a triplestore. We are using two triplestores to save the data: Virtuoso [12] and Jena Fuseki [13]. Jena Fuseki provides good performance on smaller datasets, and supports multiple datasets within the same running instance. Within each Jena dataset we store one collection of nanopublications; with each nanopublication consisting of multiple named graphs. Due to the size of the DisGeNET 4.0 dataset, it was not possible to store this in Jena on our test machine. Therefore, we stored the DisGeNET dataset in a Virtuoso triplestore due to its abilities to efficiently store and query large datasets. We could not store all the datasets in a single Virtuoso instance, since we needed multiple data collections, each using named graphs within them, and Virtuoso's mechanism to support multiple datasets is to use named graphs.

Based on the previous work by Kuhn *et al*, it is our hypothesis that insights into the nanopublication collection can be gained by observing, analysing, and comparing the distributions of triples, the predicates used, and data being represented in the nanopublication collection. We wish to identify similarities as well as differences in each of these categories and derive conclusions based on them.

The code for our analysis was developed within a Jupyter Notebook [14] which is available from GitHub[8]. We note that to reuse our notebook you must first download and store the datasets in your own triplestore, and then change the URLs for the SPARQL endpoints within the notebook.

4. Results and Analysis

A summary of the nanopublications considered in our analysis is given in Table 1. Row 1 gives the total number of nanopublications in each of the datasets, and is plotted

2. http://rdf.disgenet.org/download/v4.0.0/ accessed 27 June 2019

3. https://sourceforge.net/projects/nextprot2rdf/files/data/nextprot/releases/2014-09/ accessed 27 June 2019

4. https://github.com/peta-pico/wikipathways-nanopubs/tree/master/output/combined accessed 27 June 2019

5. https://github.com/tkuhn/bel2nanopub/releases/ accessed 27 June 2019

6. https://github.com/jmbanda/LInked-Drug-Drug-Interactions accessed 27 June 2019

7. https://github.com/tkuhn/bel2nanopub accessed 27 June 2019

8. https://github.com/ImranAsif48/RO2019

TABLE 1. COMPLETE SUMMARY OF NANOPUBLICATION TRIPLES DISTRIBUTION IN EACH GRAPH OF DIFFERENT DATASETS

	Datasets				
	DisGeNET 4.0	***neXtProt 19001_20000***	***WikiPathways 20170513***	***OpenBEL 20131211***	***LIDDI V1.02***
Total Number of Nanopublications	1,414,902	220,916	26,934	74,173	98,085
Total Number of Triples	48,106,668	8,634,736	781,772	2,186,874	2,051,959
Average Triples per Nanopublication	48.0	39.1	29.0	29.5	20.9
Head Triples	9,904,314	883,664	107,736	296,692	392,340
Assertion Triples	7,074,510	899,013	354,139	845,272	678,414
Provenance Triples	12,734,118	3,653,161	127,289	822,391	686,950
Publication Info Triples	18,393,726	3,198,898	192,608	222,519	294,255
Assertion Min/Max	1/5	2/43	2/1,001	6/55	6/8
Provenance Min/Max	8/9	6/86	1/65	11/14	7/8
Publication Info Min/Max	11/13	12/42	3/39	3/3	3/3
Assertion Outliers	0	$\approx 54,457$	$\approx 10,998$	$\approx 32,311$	≈ 345
Provenance Outliers	0	$\approx 91,740$	$\approx 8,992$	$\approx 2,592$	≈ 355
Publication Info Outliers	0	$\approx 88,859$	$\approx 1,433$	0	0

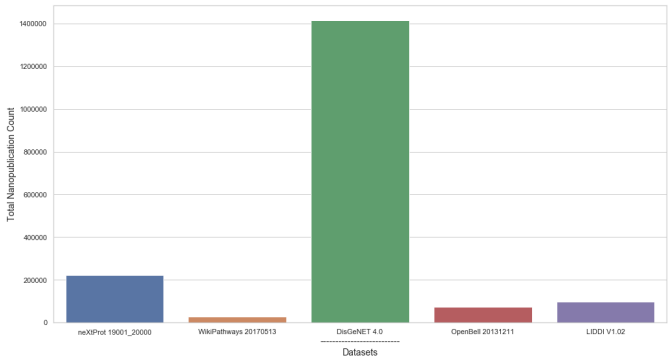

Figure 2. Total number of Nanopublications in each dataset

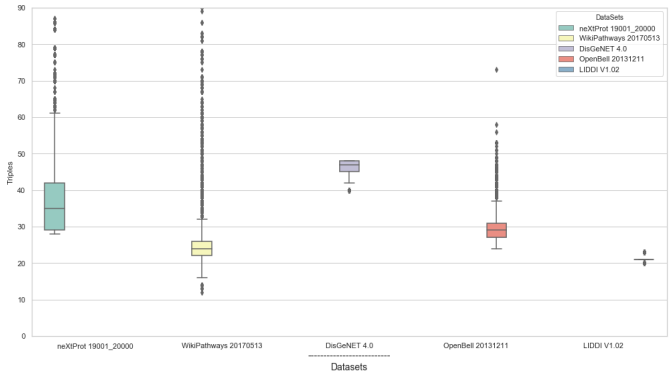

Figure 3. Frequency distribution of the number of triples per nanopublication in each dataset

in Figure 2. The plot shows that DisGeNET is published as significantly more nanopublications than the other datasets. This is expected due to the underlying size of each of the datasets. Row 2 presents the total number of triples used to represent the nanopublications in each dataset and Row 3 presents the average number of triples used per nanopublication. We can see from this data that there is a wide variance in the size of the representation of the nanopublications ranging between 20.9 and 48.0. Figure 3 plots the frequency distribution of the number of triples per nanopublication as a boxplot [15]. This highlights that there are a significant number of outliers (shown as dots) in the neXtProt, WikiPathways, and OpenBEL nanopublications, whereas DisGeNET and LIDDI are very consistent.

Rows 4 to 7 of Table 1 represent the total number of triples in each graph of the nanopublications. Rows 8 to 10 represent the minimum and maximum number of triples in the assertion, provenance, and publication information graphs respectively. The remaining rows of Table 1 represent the approximate number of outliers in each of the three named sub-graphs of a nanopublication.

4.1. Distribution Analysis

We first aim to replicate Figure 1 from [5] which presents a stacked bar chart of the count of triples in each part of a nanopublication, broken down by dataset. Figure 4 represents the average number of triples in each named graph of the nanopublication for each dataset, i.e. it is equivalent to the stacked bar chart from [5]. By unstacking the bar chart, it is easier to compare the different components of the nanopublications across the datasets. We can see that with the exception of DisGeNET, the head graphs contain on average the same number of triples (4 triples). DisGeNET contains seven triples on average in the head graph. The

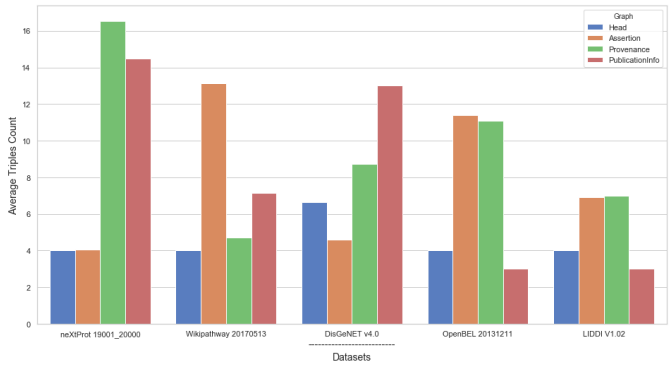

Figure 4. Average number of triples in each graph of the nanopublications by dataset.

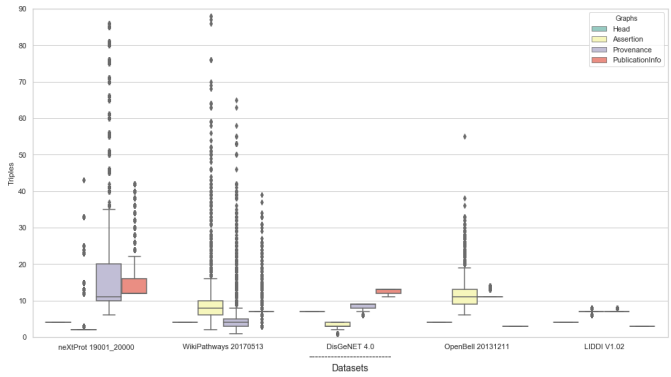

Figure 5. Frequency distribution of the number of triples in each graph of the nanopublications by dataset. Note that the y-axis is cut-off at a count of 90. The WikiPathways nanopublications include up to 1,001 triples in the assertion graph, as can be seen in the supplemental Jupyter notebook[8].

average number of triples in each of the other sub-graphs varies between the datasets with no discernible pattern.

The averages by graph are rather course and reveal little about the nature of the nanopublications. To investigate in more detail, we did a boxplot of the distribution of the count of triples in each of the graphs, see Figure 5. The boxplot shows the minimum value, lower quartile, median, upper quartile, and maximum value of each distribution. It also shows outliers (dotted points).

We first reanalyze the head graph of each dataset. From Figure 5 we can see that the head graph of each of the datasets has been uniformly represented, i.e. they have been represented using the same number of triples – this is shown as the first horizontal line in each of the dataset plots. We can see that each dataset has used four triples except DisGeNET which contains seven triples in the head graph. The use of four triples is expected as they declare the type of the data and link each of the sub-graphs in the nanopublication to the head graph, as per the nanopublication guidelines [16]. On further investigation of the DisGeNET nanopublications, we found that the three extra triples are used to assert the types of the sub-graphs.

Second, we analyze the assertion graph. We note that for all the nanopublications, the assertion graph can be considered to be small, the vast majority containing between 7 and 20 triples. The boxplot shows us that the assertion graph in neXtProt, DisGeNET, and LIDDI is more uniformly represented than the other two datasets, this is shown as a line for neXtProt and LIDDI and a small box for DisGeNET. The assertion graph in neXtProt has several outliers, shown by the dotted line coming from the top of the box, with the largest outlier containing 43 triples in the assertion graph. We looked at the content of this nanopublication http://np.inn.ac/RABK-HRA-95Nj1dNzH-5c9a2J92N2OrtOK8N6GuC7Qvmg and note that it contains information about ATPase activities and their number values. It appears to us that this nanopublication is providing a different type of information when compared to the core neXtProt nanopublications, e.g. http://np.inn.ac/RAB-Q5TQQdY0n4kF2LB4o-o49yr4Vbg6EFMdEFU5LckxI. We note that the generation of the neXtProt nanopublications is automatic, potentially with no check and balance when exporting all the records from the database as nanopublications.

The WikiPathways and OpenBEL assertion graphs have more variation than the other datasets. These datasets use 7 to 13 triples in the majority of the assertion graphs, but with a larger set of outliers, particularly in the case of WikiPathways where the largest is 1,001 triples. The largest of the WikiPathways outliers can be explained by the indexing approach used, see [5] for details. We believe that the other outliers are due to more variation in the content of the underlying databases. For example, WikiPathways contains details of biological pathways that can be of variable length; hence the number of triples needed to make an assertion is likely to be dependent on the length of the pathway. However, we have not investigated this in more detail.

Next we analyse the provenance graph. As we can see, the neXtProt provenance graph shows a large variation in the number of triples (shown by the large box). We believe that this large variation arises from the fact that neXtProt provides detailed evidence to support each claim, and the amount of evidence is not consistent from one claim to another. The WikiPathways provenance graph shows some variation and a large tail of outliers. On inspection of some nanopublications in the collection, we believe this is due to the majority of pathways linking to the scholarly articles where the pathway was published. The information provided consists of the pathway title, PubMed Identifiers for supporting articles, and other WikiPathways instance identifiers. The other datasets all have consistent provenance graphs, with only a handful of triples in each. We believe this is due to the underlying databases either not capturing, or not exposing, the detailed provenance for each claim. Thus, the provenance consists of linking back to the underlying database.

Finally, we analyse the publication information graph. The publication information graph contains the metadata information about the nanopublication itself, i.e. who created the nanopublication, who is the author of the knowledge content of the nanopublication, and when was the

nanopublication published. As we can see, WikiPathways, DisGeNET, OpenBEL, and LIDDI each have a consistent number of triples in the publication information graph, although with a significant number of outliers in the WikiPathways case. This is due to the use of `prov:Activity` to introduce the activity with additional information such as `prov:atLocation` and `prov:used`. neXtProt has some variation in the publication information graph. On inspection, this was found to be due to the neXtProt nanopublications containing more publication information using `prov:usedData`, `pav:authoredBy`, `pav:versionNumber`, and `prov:wasGeneratedBy`, as well as the creators' information, i.e. they contain information about the original authors of the knowledge content, not just who generated the nanopublication.

4.2. Authorship Analysis

Based on the above analysis, we decided to investigate the use of vocabulary terms in the publication information graph. We hypothesise that since there is little variation in the number of triples in the publication information graph that there are issues with the data quality. We use the definitions from [17] for the different roles.

- Author: the persons who generate the new knowledge or concept.
- Curator: the persons who assemble the knowledge that is published by the authors and then represent that knowledge in a meaningful way such as claim, hypothesis or research questions.
- Creator: the persons who stored this representation in some physical database.

Figure 6 depicts the distribution of the authors of the nanopublications in each dataset. To achieve this graph, we performed the SPARQL query with the predicate `pav:authoredBy`. Here `pav` is the Provenance, Authoring and Versioning (PAV) ontology [17]. We can see that two datasets, LIDDI and WikiPathways, have no authors using the `pav:authoredBy`, but the remaining have some authors. We will now look in more detail at each of the nanopublication collections.

In LIDDI, the publication information graph uses `prov:wasAttributedTo` to connect the nanopublication with the ORCID ID of Juan M. Banda. It does not claim authorship of the nanopublication or the knowledge content. The provenance graph includes details of how the nanopublication was generated rather than evidence in support of the claim. It also contains some errors such as `prov:Location` being used as a property.

We found that WikiPathways store the author information using the SemanticScience Interoperability Ontology (SIO) [18] term `sio:has-source`. This provides a link between the assertion and a PubMed ID and URL. This is following a Linked Data approach. However, it means that a further resource must be retrieved by the consumer in order to discover the authorship information.

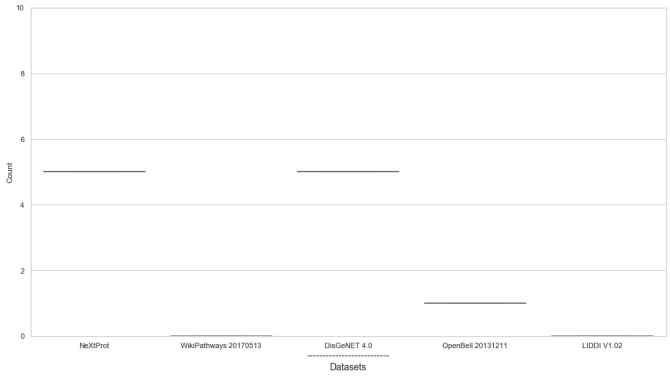

Figure 6. Frequency distribution of the `pav:authoredBy` property in each dataset

For the neXtProt dataset, we can see that each nanopublication claims to have five authors who generated the claim. These five authors are the same in all the nanopublications and correspond to people working on the CALIPHO project[9], i.e. the group who maintain the neXtProt database. This is inconsistent with the definition of authorship given for the `pav:authoredBy` property. It would be more correct to use the `pav:createdBy` property. Similarly for DisGeNET, there are five authors and they are the same for all the nanopublications. Again the usage of `pav:authoredBy` is incorrect.

For the OpenBEL small and large corpus, there is just one author. This is the Selventa project[10]. In this case the nanopublication does not provide details of who authored the content, but just the project in which it was done. Again, this is inappropriate usage of the `pav:authoredBy` property.

4.3. Summary

From the above analysis, we conclude that the majority of nanopublications considered in this study do not provide high quality information about the provenance of the claim nor the publication of the nanopublication. Nanopublications are supposed to provide granular publication of a claim together with evidence about the claim, and metadata about the nanopublication. The usage that we observe does not provide this. While we recognise the merit of the Linked Data approach followed by WikiPathways for providing authoring information, it increases the complexity for the consuming agent as it must recognise that it needs to retrieve another resource in order to discover the authorship information. Thus, from the triples contained in the published nanopublications we cannot see the complete picture in one nanopublication.

9. https://web.expasy.org/groups/calipho/
10. http://www.selventa.com/

5. Conclusions

Nanopublications are intended to be used to publish a claim together with its provenance and publication metadata. More than 10 million nanopublications have been published in the life sciences domain. In this study, we were initially motivated to repeat the analysis of Kuhn *et al* published in [5]. We were able to regenerate their figure showing the average number of triples used to represent each graph in a nanopublication, although we chose to display this as an unstacked bar chart. We were then motivated to look deeper into the distribution of the number of triples used in each graph. We found that this revealed interesting patterns that pointed to quality issues in the collection of nanopublications. In particular, the lack of variance in the number of triples used in the provenance and publication information graphs indicated that detailed provenance and metadata are not being provided.

Each of the nanopublication collections considered were generated using a script from some underlying database. The quality issues identified could be indicative of the limitations of these scripts, or due to the underlying data sources not containing sufficient data to generate high-quality nanopublications. This is supported by the neXtProt collection having the richest provenance and publication information graphs since the underlying data source captures this data. Our analysis also revealed that the nanopublications considered have not all used the authorship properties correctly. This may have been due to pragmatic approaches when developing the scripts, e.g. given the lack of data captured in the underlying source, or due to limited expertise available to them. In these nanopublications, the claimed authors actually seem to be the curators or creators of the nanopublication, but not the actual author of the claim. Finally, the WikiPathways nanopublications use a methodology that overcome the perceived large overhead of nanopublications. They publish nanopublications that contain indexes of collections of nanopublications, corresponding to different releases of the underlying dataset. We believe that there are issues in using nanopublications for both indexing a collection and publishing the content of the dataset, but this requires further investigation.

In this paper, we have pointed out some potential issues that may have occurred during the generation of nanopublications. Such issues can be caused by the content (or the lack of content) of databases that store the original data or the lack of expertise of the described domain that may have forced pragmatic approaches to be taken. Consequently, we believe that more detailed guidelines are required for the creation of high-quality nanopublications that encourage the supply of provenance data and accurately model the publication metadata.

References

[1] P. Groth, A. Gibson, and J. Velterop, "The anatomy of a nanopublication," *Information Services and Use*, vol. 30, no. 1-2, pp. 51–56, 2010. [Online]. Available: 10.3233/ISU-2010-0613https://content.iospress.com/articles/information-services-and-use/isu613

[2] T. Kuhn, C. Chichester, M. Krauthammer, N. Queralt-Rosinach, R. Verborgh, G. Giannakopoulos, A.-C. N. Ngomo, R. Viglianti, and M. Dumontier, "Decentralized provenance-aware publishing with nanopublications," *PeerJ Computer Science*, vol. 2, p. e78, 2016.

[3] T. Kuhn, A. Meroño-Peñuela, A. Malic, J. H. Poelen, A. H. Hurlbert, E. C. Ortiz, L. I. Furlong, N. Queralt-Rosinach, C. Chichester, J. M. Banda, E. Willighagen, F. Ehrhart, C. Evelo, T. B. Malas, and M. Dumontier, "Nanopublications: A growing resource of provenance-centric scientific linked data," in *14th International Conference on e-Science (e-Science)*. Amsterdam, Netherlands: IEEE, 2018. [Online]. Available: http://arxiv.org/abs/1809.06532

[4] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler, "Named graphs, provenance and trust," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 613–622.

[5] T. Kuhn, E. Willighagen, C. Evelo, N. Queralt-Rosinach, E. Centeno, and L. I. Furlong, "Reliable granular references to changing linked data," in *International Semantic Web Conference*. Springer, 2017, pp. 436–451.

[6] J. Piñero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong, "Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, 2015.

[7] L. Lane, G. Argoud-Puy, A. Britan, I. Cusin, P. D. Duek, O. Evalet, A. Gateau, P. Gaudet, A. Gleizes, A. Masselot *et al.*, "nextprot: a knowledge platform for human proteins," *Nucleic acids research*, vol. 40, no. D1, pp. D76–D83, 2011.

[8] A. R. Pico, T. Kelder, M. P. Van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo, "Wikipathways: pathway editing for the people," *PLoS biology*, vol. 6, no. 7, p. e184, 2008.

[9] T. Kuhn and M. Dumontier, "Making Digital Artifacts on the Web Verifiable and Reliable," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 9, pp. 2390–2400, 2015. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7079484https://doi.org/10.1109/TKDE.2015.2419657

[10] J. Fluck, S. Madan, S. Ansari, R. Karki, M. Rastegar-Mojarad, N. L. Catlett, W. Hayes, J. Szostak, J. Hoeng, M. Peitsch *et al.*, "Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (bel)," *Database*, vol. 2016, 2016.

[11] J. Schneider, P. Ciccarese, T. Clark, and R. D. Boyce, "Using the Micropublications ontology and the Open Annotation Data Model to represent evidence within a drug-drug interaction knowledge base," in *CEUR Workshop Proceedings*, vol. 1282, Riva de Garda, Italy, 2014, pp. 60–70. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01076282/file/lisc2014.pdf

[12] M. Saleem, Y. Khan, A. Hasnain, I. Ermilov, and A.-C. Ngonga Ngomo, "A fine-grained evaluation of sparql endpoint federation systems," *Semantic Web*, vol. 7, no. 5, pp. 493–518, 2016.

[13] "Apache Jena." [Online]. Available: https://jena.apache.org/

[14] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay *et al.*, "Jupyter notebooks-a publishing format for reproducible computational workflows." in *ELPUB*, 2016, pp. 87–90.

[15] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978. [Online]. Available: http://www.jstor.org/stable/2683468

[16] "Guidelines for Nanopublications," Concept Web Alliance, Working Draft, Sep. 2018. [Online]. Available: http://www.nanopub.org/2018/WD-guidelines-20180926/

[17] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A. J. Gray, C. Goble, and T. Clark, "Pav ontology: provenance, authoring and versioning," *Journal of biomedical semantics*, vol. 4, no. 1, p. 37, 2013.

[18] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath *et al.*, "The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery," *Journal of biomedical semantics*, vol. 5, no. 1, p. 14, 2014.