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 raised by this analysis. From this analysis we argue that there is a need for the community to develop a set of community 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. Based on our analysis we raise questions about the current practice of publishing nanopublications from traditional datasets and the overall quality of the data.

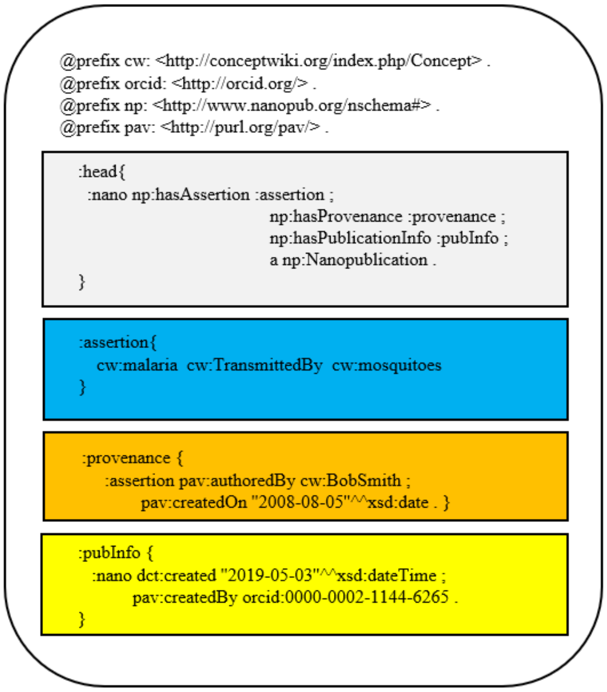
2. Background

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". So, this statement now easily represents in the RDF triple as Subject (Malaria), Predicate (Transmitted by), and Object (Mosquito). To store this claim in a nanopublication, we need four named RDF graphs [4] those are Head, Assertion, Provenance, and Publication Information. The full nanopublication is shown in Figure 1.

The structure of a nanopublication adds a large overhead to the publication of each claim when compared with 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

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



```
@prefix cw: <http://conceptwiki.org/index.php/Concept> .
@prefix orcid: <http://orcid.org/> .
@prefix np: <http://www.nanopub.org/nschema#> .
@prefix pav: <http://purl.org/pav/> .

:head{
  :nano np:hasAssertion :assertion ;
                        np:hasProvenance :provenance ;
                        np:hasPublicationInfo :pubInfo ;
                        a np:Nanopublication .
}

:assertion{
   cw:malaria  cw:TransmittedBy  cw:mosquitoes
}

:provenance {
     :assertion pav:authoredBy cw:BobSmith ;
            pav:createdOn "2008-08-05"^^xsd:date . }

:pubInfo {
   :nano dct:created "2019-05-03"^^xsd:dateTime ;
          pav:createdBy orcid:0000-0002-1144-6265 .
}
```

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.

a mechanism for indexing and reusing nanopublications. They have argued that their approach to reusing nanopublications across multiple dataset version indexes eliminates this overhead when compared to the traditional approach of publishing 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 bel2nanopub[7] script. The LIDDI nanopublications were generated by Juan M. Banda.

The datasets 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 the multiple named graphs of the nanopublications. However, due to the size of the DisGeNet 4.0 dataset, it was not possible to store this on Jena on our test machine. Therefore, we stored the DisGeNET dataset on the Virtuoso server due to its abilities to efficiently store and query large datasets. We note that with Virtuoso it is difficult to store multiple datasets in different named graph.

Following approaches taken in the previous papers, it is our hypothesis that it is possible to gain insights into the data quality of nanopublications purely based on observation and without having to have expertise of the described domain. We therefore plan to observe, analyse and compare the distributions of triples in nanopublications, the predicates used, and data being represented in the above different datasets. 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

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. The raw count of the number of nanopublications by dataset is given in row 1 of Table 1 and plotted in Figure 2. The plot

[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

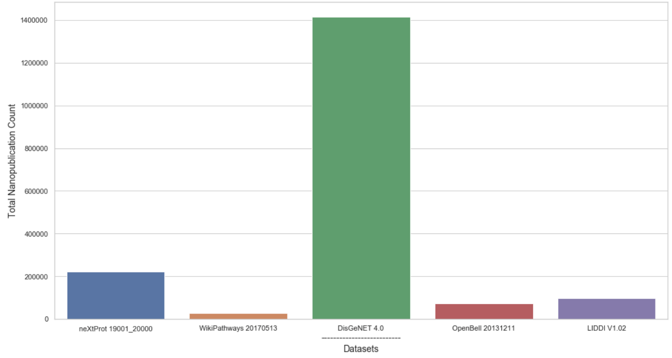


Figure 2. Total number of Nanopublications in each dataset

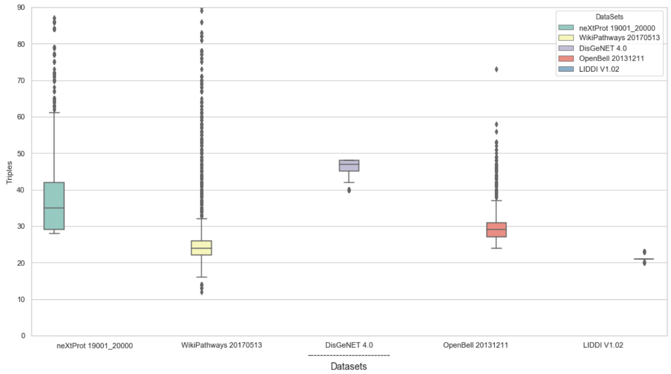


Figure 3. Frequency distribution of the number of triples per nanopublication in each dataset

shows us that DisGeNET is published as significantly more nanopublications than the other datasets, but this is expected due to the underlying size of each of the datasets. Row 3 of Table 1 presents the average number of triples used per nanopublication in each of the datasets. 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, on average shown in Figure 3. From row 4 to 7 in Table 1 represent the total number of triples in each graph of different datasets and this distribution is shown in Figure 5. Row 8 to 9 represent the minimum and maximum triples count of assertion, provenance and publication information graph. The remaining rows of the Table 1 represent the approximately outliers of the three main graphs. As we can see that DisGeNET has no outliers in each graph that gives the consistency of the dataset. OpenBEL and LIDDI have no outliers in publication Info graph that shows the consistency in the one graph. The distribution of the outliers in each graph of different datasets are shown in Figure 5.

We will now investigate the breakdown of triples within the different parts of the nanopublication as was done by Kuhn et al. Figure 4 represents the average number of triples in each part of the nanopublication for each dataset, i.e. 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). These triples link each of the sub-graphs in the nanopublication to the head graph. DisGeNET contains on average three more triples in the head graph to define the sub-graphs such as assertion, provenance and publication info separately.

The average number of triples in each of the other sub-graphs varies between the datasets with no discernable pattern. To investigate this in more detail, we generated a

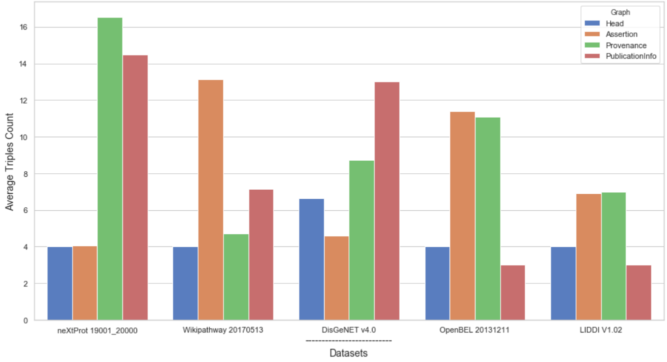


Figure 4. Average number of triples in each graph of the nanopublications by dataset.

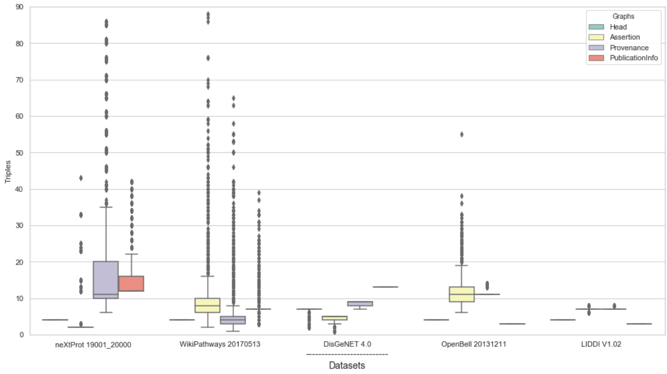


Figure 5. Frequency distribution of the number of triples in each graph of the nanopublications by dataset.

boxplot [15] to show the distribution of the count of triples in each of the graphs Figure 5. The boxplot represents five number summary that include the minimum value, lower quartile, median, upper quartile, and maximum value of the dataset. From this figure, we first reanalyze the head graph of each dataset and as one can see that the head graph of each of the nanopublications 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. As one can see, each dataset has used four triples except DisGeNET which contains seven triples in the head graph because head graph of DisGetNet nanopublication contains the extra three triples about to define the assertion, provenance and publication graphs[9].

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 only outliers, shown by the dotted line coming from the box, with the top one 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 thier 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 other datasets. These datasets store 7 to 13 triples in the assertion graph, but with a larger set of outliers particularly in Wikipathways. We believe that this is due to more variation in the content of the underlying databases.

Next we analyse the provenance graph. As we can see, the neXtProt provenance graph shows a large variation in the number of triples that support the assertion graph. The variation in this graph arises from the need to show the evidence to support the claim and link to the data from which it is derived. The WikiPathways provenance graph shows some variation and a large tail of outliers because in this graph contains the information about the WikiPathways instance title, PubMed Ids and other WikiPathways instance identifiers. The other datasets all have consistent provenance graphs, with only a handful of triples in each. This is because in these datasets, the exported nanopublications are not giving detailed provenance information, possibly because it is not captured in 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 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. This is because the neXtProt nanopublications contain more publication information using `prov:usedData`, `pav:authoredBy`, `pav:versionNumber`, and `prov:wasGeneratedBy`, as well as the creators' information.

Based on the above analysis, we decided to investigate further into the publication graph. We hypothesise that since there is little variation in the number of triples in the publication graph that there are issues with the data quality. However, first we must understand about some terms commonly used in the publication graph such as creator, author, and curator. We use the definition from [16].

Author: the persons who generate the new knowledge or concept.

[9] http://www.nanopub.org/nschema

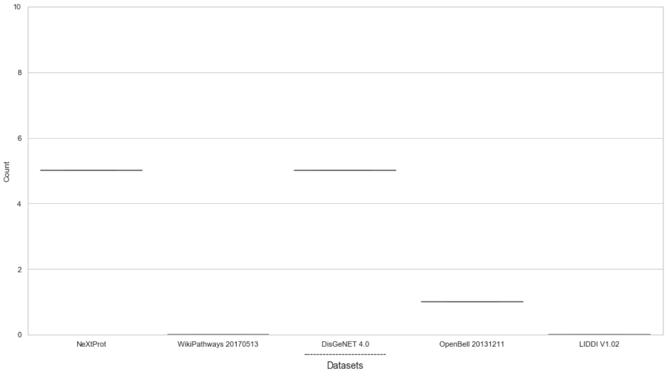


Figure 6. Frequency distribution of the `pav:authoredBy` property in each dataset

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 [16]. We can see that two datasets, LIDDI and Wikipathways, have no authors using the `pav:authoredBy`, but the remaining have some authors. We looked in more detail at the structure of these nanopublications.

For LIDDI we discovered that it is fully generated by the software that was made by Juan M. Banda. In this case we can say that Jaun M. Banda is the creator or curator but not the author. Some of them using the URL of the software that are available on the GitHub that does not represent the meaning and authoring of the claim. In some nonpublication, author information represents in the form of PubMed id that represent the research article link on PubMed. So, this is also not the meaningful way to store the author information in the nanopublication because we cannot fetch the number of authors from the PubMed Id.

Similarly, we investigated how Wikipathways represent the author information. We found that WikiPathways store the author information using the SemanticScience Interoperability Ontology [17] using the `sio:has-source` property to provide a link between the assertion and a PubMed id and URL. This means that a further resource must be retrieved and then analysed.

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[10], 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.

For the OpenBEL small and large corpus, there is just one author. This is the Selventa project[11]. In this caes it does not provide the author name who generated the claim so it means that the nanopublication are automatically generated by the code script that is bel2nanopub[12]. Again, this seems to be inappropriate usage of the `pav:authoredBy` property.

Similarly for DisGeNET, there are five authors and they are the same for all the nanopublications. To give the evidence about the claim they use the PubMed id to link to a paper where the knowledge was first published. Again the usage of `pav:authoredBy` is incorrect.

From the above analysis, we conclude that the existing nanopublications do not provide high quality information about the provenance. 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 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 inorder to retrieve the authorship information. Thus, from the triples contained in the published nanopublications we cannot see the complete picture in one nanopublication.

5. Conclusion

The Nanopublication is useful for the research community to store the claim in a meaningful way. More than 10 million nanopublications have been published in life sciences domain. They use PROV ontology, PAV ontology and SIO ontology in provenance graph for supporting the claim. Our analysis shows that these nanopublications have not all been created following a suitable methodology that pragmatic approaches may have been taken - perhaps given the data or limited expertise available to them. This approach allows valuable data to be recorded, however, it may compromise the quality of captured data. These nanopublications provide the authoring information, but these “authors” actually seem to be the curators or creators of the nanopublication but not the actual author of the claim. Some nanopublications use methodology that overcome the large overhead of triples using indexes such as WikiPathways’ nanopublications. Indexes are stored in the assertion graph of nanopublication that arguably not a common or good practice.

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. As a result, we

[10] https://web.expasy.org/groups/calipho/

[11] http://www.selventa.com/

[12] https://github.com/tkuhn/bel2nanopub

argue that standard approaches and frameworks towards the creation of nanopublications are necessary.

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] A. Jena, "Apache jena fuseki," *The Apache Software Foundation*, 2014.

[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] 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.

[17] 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.