CEUR-WS.org/Vol-2275/short2.pdf

Increasing the nanopublication recall with a BridgeDb Identifier Mapping Service

Egon Willighagen[1][0000−0001−7542−0286]

Dept. Bioinformatics - BiGCaT, NUTRIM, Maastricht University, NL
egon.willighagen@maastrichtuniversity.nl

Abstract. The volume of literature in the life sciences is continuously growing and keeping up with it is a problem. While review articles and databases help us by summarizing vast amounts of research, dissemination of core research outcomes is still mostly restricted to scholarly journal. Nanopublications have been proposed as a solution to capture scientific statements. This led to a 2010 proposal to serialize nanopubs in the Resource Description Framework (RDF) and in 2016 to an international network of nanopublication servers. However, RDF has a limitation that the Internationalized Resource Identifier (IRI) for resources does not have to be normalized and unique. To overcome this issue, the Open PHACTS project developed an Identifier Mapping Service and an approach called *scientific lenses* for mapping of equivalent IRIs. We here demonstrate the application of this approach to improve the recall of nanopublications from the international network.

Keywords: nanopublications · identifier · BridgeDb.

1 Introduction

The volume of literature in the life sciences is continuously growing and keeping up with literature is a problem [1]. While literature reviews and databases summarize vast amounts of research, dissemination of core research outcomes is still mostly restricted to the written word. Nanopublications have been proposed to capture scientific statements with provenance to the origin of that statement [2]. This led to a 2010 proposal to serialize nanopubs in the Resource Description Framework (RDF) [3]. Kuhn *et al.* introduced in 2016 an international network of connected servers to host nanopublications, with initial data sets from DisGeNET [4], neXtProt [5], and others [6]. Nanopublications for WikiPathways were added later [7]. The network currently hosts about 10 million nanopublications [8].

However, RDF has a limitation that the Internationalized Resource Identifier (IRI) for resources does not have meaning and do not have to be unique. This causes an infinite number of possible IRIs for the same resource. And because when creating RDF content, one is not meant to reuse domain names not under your control and one often wishes to make resource IRIs dereferenceable, this is exactly what we see in practise: different data sets use different IRIs for the same gene, protein, or metabolite.

The Open PHACTS project has had the same problem when starting to link different pharmacology data sets [9], including ChEMBL [10] and WikiPathways [11]. To overcome this issue, it developed an Identifier Mapping Service (IMS) based on BridgeDb [12] and the *scientific lenses* approach that allowed mapping of equivalent IRIs [13,14]. The alternative, of course, is to normalize IRIs in data sets before the integration, e.g. with identifiers.org [15], but removes flexibility to change the level of equivalence depending on the data analysis done [13,14]. The IMS implemented the identifier mapping as integral part of the Open PHACTS Linked Data API, hiding this need to map equivalent IRIs when querying the underlying data sets.

Therefore, the assumption is that if we use an IMS as outlined here with appropriate loaded scientific lenses, we will find more nanopubs for a particular biological entity. To test this hypothesis, we searched nanopubs with information about a set of genes on the international network of nanopublication servers.

2 Methods

To implement our workflow, an R Markdown document was developed to perform the various steps detailed below. The full document is available at `https://github.com/egonw/swat4hcls2018/`. It uses a few R packages and introduces a helper function to simplify searching nanopublications.

Genes As representative data set we took two pathways from the list of most viewed WikiPathways (`https://www.wikipathways.org/index.php/Special:PopularPathwaysPage`). Pathway WP241 was selected, about the human one carbon metabolism [16], with mostly NCBI Gene identifiers. WP2059 was selected as a second example, with predominantly Ensembl gene identifiers [17]. The *rrdf* package was used to query all genes in this pathways from the SPARQL endpoint [18].

Nanopublication Server As source of nanopubs we took a nearby server from the international network of mirroring servers, hosted by the Institute for Data Sciences at Maastricht University (`http://graphdb.dumontierlab.com/repositories/nanopubs`) [6]. To simplify the interaction with the server, we took advantage of an online running instance of grlc (grlc.io) [19] that wraps the nanopublication server API with an OpenAPI [20]: `https://github.com/peta-pico/nanopub-api`.

BridgeDb Identifier Mapping Server For the IRI mapping, we used the Docker image of the IMS developed by the Open PHACTS project (`https://hub.docker.com/r/openphacts/identitymappingservice/`) [12,13,9]. This was recently repurposed by Ehrhart *et al.* for gene-variant mappings [21]. The IMS is started and loaded with IRI mapping data as explained in [21]: 1. the user starts to Docker image, and then 2. loads the identifier mapping files into the IMS instance using a loading script. The actual mapping files were generated by J. Mélius based on data from Ensembl 87 (see `http://bridgedb.org/data/`

`linksets/current/HomoSapiens/`. The source code can be found at `https://github.com/BiGCAT-UM/EnsemblLinksetsCreator`). These linksets map Ensembl identifiers with NCBI Gene, HGNC, and others.

Data Analysis The R Markdown notebook integrates the three aforementioned approaches into a single analysis. To test the hypothesis, it first retrieves the genes from two WikiPathways and for each gene it searches for nanopubs. This is done by looking up equivalent gene IRIs using the IMS server and then for each gene IRI search for nanopublications. It then counts only the original IRI and for all equivalent IRIs and reports the differences.

3 Results

With the R Markdown notebook we searched for nanopublications for two popular WikiPathways, WP241 and WP2059. The first has mostly NCBI Gene identifiers and the second mostly Ensembl Gene identifiers. The script reports that NCBI Gene identifiers return the most nanopublications when searching the full nanopublication network, indicating nanopublication data sets prefer to use NCBI Gene identifier-based IRIs. This observation affects the number of additional nanopubs found via equivalent IRIs: for WP241 we indeed find a high number of found pathways when using only the IRI returned by the WikiPathways SPARQL endpoint and a lower number for WP2059. For WP241 we get on average 464 nanopublications (min: 10, max: 1000, median: 288), while for WP2059 we retrieve on average 21 nanopublications (min: 0, max:1000, median: 1). The count is currently capped at 1000 nanopublications, imposed by the grlc API wrapping around the nanopublication server, which explains this artifact.

As hypothesized, using equivalent IRIs, as returned by the IMS, will retrieve additional nanopublications. The number of equivalent IRIs by the IMS is different for both pathways. For WP241 it returns on average 9 IRIs (min: 7, max: 23, median: 7) and for WP2059 it returns on average 11 IRIs (min: 7, max: 29, median: 10). The returned IRIs are a mix of mappings to other database sources and different IRI patterns for the same database identifier.

Indeed, with these additional IRIs, more nanopublications are found on the international nanopublication network. Furthermore, when the additional IRIs include NCBI Gene identifier-based IRIs, we find a higher number of additional nanopublications. This observation is similar for that observed when only using the original NCBI Gene identifier-based IRI, as explained above. Therefore, we find fewer additional nanopublications for WP241 which used predominantly NCBI Gene identifiers: on average it find 17 additional nanopublications (min: 0, max: 230, median: 3). However, for WP2059 which used predominantly Ensembl identifiers we find a much higher number of additional nanopublications, on average 44 (min: 0, max: 1099, median: 0).

4 Discussion

It goes without discussion that the results show that we indeed find more nanopublications about a certain gene. However, there are some aspects that must be noted. First, the enrichment is only as good as the completeness and quality of the gene-gene identifier mapping link sets. The link sets used in this study are biased towards equivalent IRIs based on Ensembl and NCBI Gene identifiers. If nanopublication data sets use other gene identifiers, these will still not be found. Besides this completeness issue, the author had the impression that some mappings were missing, something will be explored. A more elaborate analysis is planned, involving more pathways and more identifier sources.

Another effect of this completeness aspect is that the link sets used only cover gene-gene identifier mappings. However, the scientific lenses approach also allows gene-RNA and gene-protein mappings, under a lens that equates genes and proteins. This is particularly relevant for WikiPathways, where gene and proteins are frequently used as equivalent. The ability to load such additional link sets and the feature of the IMS to turn on and off the lenses, would allow returning even more nanopublications.

5 Conclusion

The results show that using an IRI mapping service increases the recall when searching nanopublication, overcoming the problem that nanopublications do not (and should not need to) normalize IRIs. This paper demonstrates this with the application of a locally installed BridgeDb IMS service and an R script in combination with online nanopublication services.

References

1. Pain, E.: How to keep up with the scientific literature. Science (November 2016)
2. Mons, B., Velterop, J.: Nano-Publication in the e-Science Era. In: Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009), collocated with the 8th International Semantic Web Conference (ISWC-2009), Washington DC, USA, October 26, 2009. Volume 523., CEUR Workshop Proceedings (October 2009)
3. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use **30**(1) (January 2010) 51–56
4. Queralt-Rosinach, N., Kuhn, T., Chichester, C., Dumontier, M., Sanz, F., Furlong, L.I.: Publishing DisGeNET as nanopublications. Semantic Web **7**(5) (June 2016) 519–528
5. Chichester, C., Gaudet, P., Karch, O., Groth, P., Lane, L., Bairoch, A., Mons, B., Loizou, A.: Querying neXtProt nanopublications and their value for insights on sequence variants and tissue expression. Web Semantics: Science, Services and Agents on the World Wide Web **29** (December 2014) 3–11

6. Kuhn, T., Chichester, C., Krauthammer, M., Queralt-Rosinach, N., Verborgh, R., Giannakopoulos, G., Ngonga Ngomo, A.C., Viglianti, R., Dumontier, M.: Decentralized provenance-aware publishing with nanopublications. PeerJ Computer Science **2** (August 2016) e78+
7. 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., Fernandez, M., Tamma, V., Lecue, F., Cudr-Mauroux, P., Sequeda, J., Lange, C., Heflin, J., eds.: The Semantic Web ISWC 2017. Volume 10587. Springer International Publishing, Cham (2017) 436–451
8. Kuhn, T., Meroo-Peuela, A., Malix, A., Poelen, J., Hurlbert, A., Centeno, E., Furlong, L.I., Queralt-Rosinach, N., Chichester, C., Banda, J., Willighagen, E., Ehrhart, F., Evelo, C., Malas, T., Dumontier, M.: Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data. In: Proceedings of IEEE eScience 2018, arXiv.org (September 2018)
9. Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., Mons, B.: Open PHACTS: semantic interoperability for drug discovery. Drug Discovery Today **17**(21-22) (November 2012) 1188–1198
10. Willighagen, E.L., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A.J., Tkachenko, V., Hastings, J., Chen, B., Wild, D.J.: The ChEMBL database as linked open data. Journal of Cheminformatics **5**(1) (2013) 23
11. Waagmeester, A., Kutmon, M., Riutta, A., Miller, R., Willighagen, E.L., Evelo, C.T., Pico, A.R.: Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. PLOS Computational Biology **12**(6) (June 2016) e1004989
12. Van Iersel, M.P., Pico, A.R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B.R., Evelo, C.T.: The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics **11**(1) (January 2010) 5+
13. Brenninkmeijer, C., Evelo, C., Goble, C., Gray, A.J.G., Groth, P., Pettifer, S., Stevens, R., William, A.J., Willighagen, E.L.: Scientific Lenses over Linked Data: An Approach to Support Task Specific Views of the Data. A Vision. In: Linked Science 2012 - Tackling Big Data. (2012)
14. Batchelor, C., Brenninkmeijer, C.Y.A., Chichester, C., Davies, M., Digles, D., Dunlop, I., Evelo, C.T., Gaulton, A., Goble, C., Gray, A.J.G., Groth, P., Harland, L., Karapetyan, K., Loizou, A., Overington, J.P., Pettifer, S., Steele, J., Stevens, R., Tkachenko, V., Waagmeester, A., Williams, A., Willighagen, E.L.: Scientific Lenses to Support Multiple Views over Linked Chemistry Data. In Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandei, D., Groth, P., Noy, N., Janowicz, K., Goble, C., eds.: The Semantic Web ISWC 2014. Volume 8796. Springer International Publishing, Cham (2014) 98–113
15. Juty, N., Le Novre, N., Laibe, C.: Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Research **40**(D1) (January 2012) D580–D586
16. Adriaens, M., Evelo, E., Willighagen, E., Marianthi, F., Kalafati, Pico, A., Kelder, T., Slenter, D., Hanspers, K., Txr24, Egoyenechea, Fehrhart, Thakur, G., Jeff: One Carbon Metabolism (Homo sapiens)
17. Salomonis, N., Hanspers, K., Fehrhart, Pico, A., Willighagen, E., Kelder, T., Mlius, J.: Alzheimers Disease (Homo sapiens)
18. Willighagen, E.: Accessing biological data in R with semantic web technologies. Technical report, PeerJ Inc. (March 2014)

19. Meroo-Peuela, A., Hoekstra, R.: grlc Makes GitHub Taste Like Linked Data APIs. In Sack, H., Rizzo, G., Steinmetz, N., Mladeni, D., Auer, S., Lange, C., eds.: The Semantic Web. Volume 9989. Springer International Publishing, Cham (2016) 342–353
20. Sferruzza, D., Rocheteau, J., Attiogb, C., Lanoix, A.: Extending OpenAPI 3.0 to Build Web Services from their Specification. In: Proceedings of the 14th International Conference on Web Information Systems and Technologies, Seville, Spain, SCITEPRESS - Science and Technology Publications (2018) 412–419
21. Ehrhart, F., Melius, J., Cirillo, E., Kutmon, M., Willighagen, E.L., Coort, S.L., Curfs, L.M., Evelo, C.T.: Providing gene-to-variant and variant-to-gene database identifier mappings to use with BridgeDb mapping services. F1000Research **7** (September 2018) 1390