Proc. International Conference on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2015)
(electronic proc.: http://www.swat4ls.org/wp-content/uploads/2015/2010/SWAT2014LS_2015_paper_2011.pdf)

Exposing Provenance Metadata Using Different RDF Models

Gang Fu[1], Evan Bolton[1], Núria Queralt-Rosinach[2], Laura I. Furlong[2], Vinh Nguyen[3], Amit Sheth[3], Olivier Bodenreider[4], Michel Dumontier[5]

[1] gang.fu@nih.gov, bolton@ncbi.nih.gov
National Center of Biotechnology Information, National Institute of Health,
[2] nqueralt,lfurlong@imim.es
Universitat Pompeu Fabra, Barcelona, Spain
[3] vinh,amit@knoesis.org
Kno.e.sis Center, Wright State University, Ohio, USA
[4] olivier@nlm.nih.gov
National Library of Medicine, National Institute of Health, Maryland, USA
[5] michel.dumontier@stanford.edu
Stanford Center for Biomedical Informatics, Stanford University, Stanford, USA

Abstract. A standard model for exposing structured provenance metadata of scientific assertions on the Semantic Web would increase interoperability, discoverability, reliability, as well as reproducibility for scientific discourse and evidence-based knowledge discovery. Several Resource Description Framework (RDF) models have been proposed to track provenance. However, provenance metadata may not only be verbose, but also significantly redundant. Therefore, an appropriate RDF provenance model should be efficient for publishing, querying, and reasoning over Linked Data. In the present work, we have collected millions of pairwise relations between chemicals, genes, and diseases from multiple data sources, and demonstrated the extent of redundancy of provenance information in the life science domain. We also evaluated the suitability of several RDF provenance models for this crowdsourced data set, including the N-ary model, the Singleton Property model, and the Nanopublication model. We examined query performance against three commonly used large RDF stores, including Virtuoso, Stardog, and Blazegraph. Our experiments demonstrate that query performance depends on both RDF store as well as the RDF provenance model.

1 Introduction

Evidence and provenance are key aspects of a healthy scientific discourse. A standard model to provide structured and interoperable metadata linked to scientific assertions is of increasing interest [22,16]. The Resource Description Framework (RDF), the lingua franca for the Semantic Web, offers the building blocks by which statements can be provided along with their metadata. Structured metadata, such as whether the resource was manually curated or automatically text

mined from scientific literature, is key to assessing quality of information. Hence, a scalable and well-designed RDF-based metadata model is crucial for knowledge integration.

Specifying the provenance of a single entity can be easily achieved using existing RDF terminologies such as PROV. However, it is the specification of the provenance of a binary or n-ary relation which remains non-standard. Several models for exposing the provenance metadata of the relations have been proposed including adding provenance annotations to i) an instance of a class that represents the n-ary relation (N-ary model) [2]; ii) an instantiated property, i.e. Singleton property (SP) model [18]; and iii) a graph that contains the relational assertions, i.e. Nanopublication model [12]. In the life sciences, the N-ary model has been used to capture the provenance information for protein-protein interactions (i.e. iRefIndex database [20]) and text-mined gene-disease interactions (i.e. DisGeNET [6]), while the recently proposed SP model [18] has been used across elements of biomedical and material sciences. Despite their use to represent various data, no study has yet been performed to examine the advantages and disadvantages of all these models using a common dataset.

In the present study, we aim to evaluate the consequence of using different RDF models to capture provenance metadata for life science data. We examine the number of triples generated and query performance on three RDF stores: Virtuoso [4], StarDog [3], and BlazeGraph [1]. Regarding to the provenance metadata of the relational assertions, we consider the data source, the supporting scientific publication, and the biological species where the given assertion holds true. In addition to the three basic RDF models described above, we also examine the implementions of the so-called *cardinal assertion model* that was first introduced by Nanopublications [5] on the N-ary and SP models, to create a non-redundant network of assertions. This consideration is particularly important as there exists substantive overlap in the assertions from multiple databases. For instance, the asserted relation between dexamethasone (PubChem Compound 5743) and glucocorticoid receptor (GR) (NCBI Gene 2908) was mentioned by four different data sources, but each data source cites an entirely different set of scientific publications in support of the assertion. This work is crucial for the efficient implementation of scalable, interoperable, and extensible knowledge models for open data sources including PubChemRDF [10], Bio2RDF [7], and DisGeNET-RDF[19].

2 Methods

2.1 Dataset preparation

We generated a reference dataset of pairwise relations between chemicals, genes, and diseases from multiple data sources across life science domain. The chemical-disease relations were obtained from National Drug File Reference Terminology (NDFRT) [8], CTD [9], KEGG [13], and SIDER [15]; chemical-gene relations were obtained from CTD [9], DrugBank [14], KEGG [13], IUPHAR-DB [23], and ChEMBL [11]; protein-protein relations were obtained from iRefIndex [20] and

BioGRID [24]; gene-disease were contributed by DisGeNET [6]. All chemicals were represented using PubChem Compound identifiers (CIDs), all genes were represented using National Center for Biotechnology Information (NCBI) Gene identifiers (GIDs), and all diseases were represented using the Unified Medical Language System (UMLS) Concept Unique Identifiers. The pairwise relations were normalized using the modified Semantic Network standard vocabulary [21]. The interrelations between biomedical entities (chemicals, genes, and diseases) constitute a semantic network, and SPARQL queries were used to explore the network topology on behalf of evidence-based hypothesis generation. However, it is fairly common to collect the identical assertion from multiple sources, in particular, for such a consolidated knowledge base. Hence, additional constraints were applied in the searching strategies.

2.2 RDF model construction

Five RDF models were studied, including N-ary model with and without cardinal assertion (Fig. 1), SP model with and without cardinal assertion (Fig. 2), and the Nanopublication model (Fig. 3). Only the assertion graphs and the provenance graphs were considered in the Nanopublication model. In both N-ary and SP cardinal assertion variants, a predicate `cito:providesAssertionFor` is used to link the cardinal assertion of the pairwise relation to the multiple evidence (Fig. 1A and 2A). Without cardinal assertion, the pairwise relation would be asserted redundantly by multiple data sources (Fig. 1B and 2B). In the Nanopublication model variant A, one assertion graph may correspond with one or more than one provenance graphs (Fig. 3). In the following comparative analysis, Model **I** refers to the N-ary model with cardinal assertion, Model **II** refers to the N-ary model without cardinal assertion, Model **III** refers to the SP model with cardinal assertion, Model **IV** refers to the SP model without cardinal assertion, and Model **V** refers to the Nanopublication model.

2.3 Query formulation

An interesting research topic in drug discovery is to determine which proteins are responsible for eliciting particular drug side effects. We formulated SPARQL queries to examine this question using different levels of complexity (**Q1**, **Q2**) and provenance constraints (**Q3**, **Q4**). **Q1** explores the hypothesis that if chemical **A** inhibits gene **B**, and gene **B** interacts with gene **C**, and gene **C** is linked to disease **D**, then the above path can be used to explain the disease/adverse side effect **D** caused by chemical **A**. It should be noted that the observed side effect can be explained in several ways: either the aforementioned three-step indirect paths, or the two-step indirect path involving only the chemical-gene interaction and gene-disease associations. Therefore, we have constructed another query, i.e. **Q2**, to filter out the diseases that are associated with genes that directly interact with the given chemical. The first two queries do not take into account the provenance metadata, and it is usually the case that only the integrated

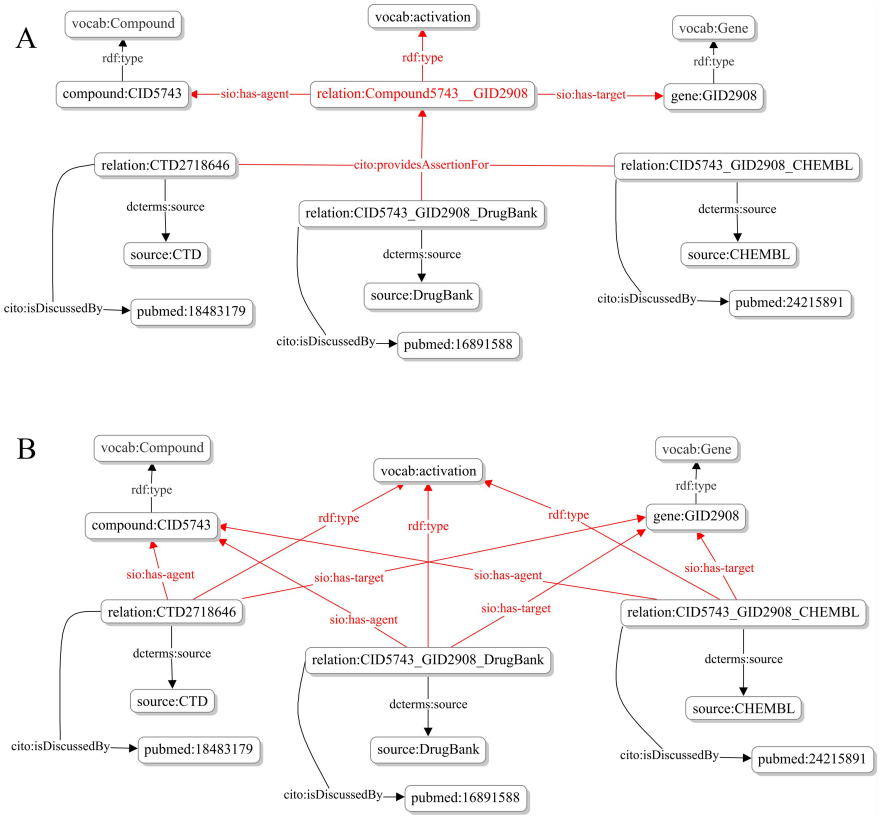

Fig. 1. Graphical representation of N-ary model for the relation between compound CID5743 and gene 2908: (a) with cardinal assertion (Model I); (b) without cardinal assertion (Model II).

assertions are considered on behalf of hypothesis generation and knowledge discovery. **Q3** narrows down the search results by applying data source constraints. **Q4** restricts by number of aggregated evidence on **Q1**: such that the query only considers the pairwise relations in the indirect path that have more than one supporting literature references.

We carried out **Q1** through **Q4** on six chemicals that have extensive biomedical annotations from multiple data sources: propranolol (CID4946), clotrimazole (CID2812), mitoxantrone (CID4212), risperidone (CID5073), chlorpromazine (CID2726), and haloperidol (CID3559). There are hundreds of similar compounds in the integrated dataset and they are of key interest in the context of drug repurposing and development.

All queries were performed against three RDF stores without further tuning: open source Virtuoso 7.1, Stardog 2.2, and Blazegraph 1.5. The configuration allowed up to 16 GB memory for each RDF store to run queries, which were

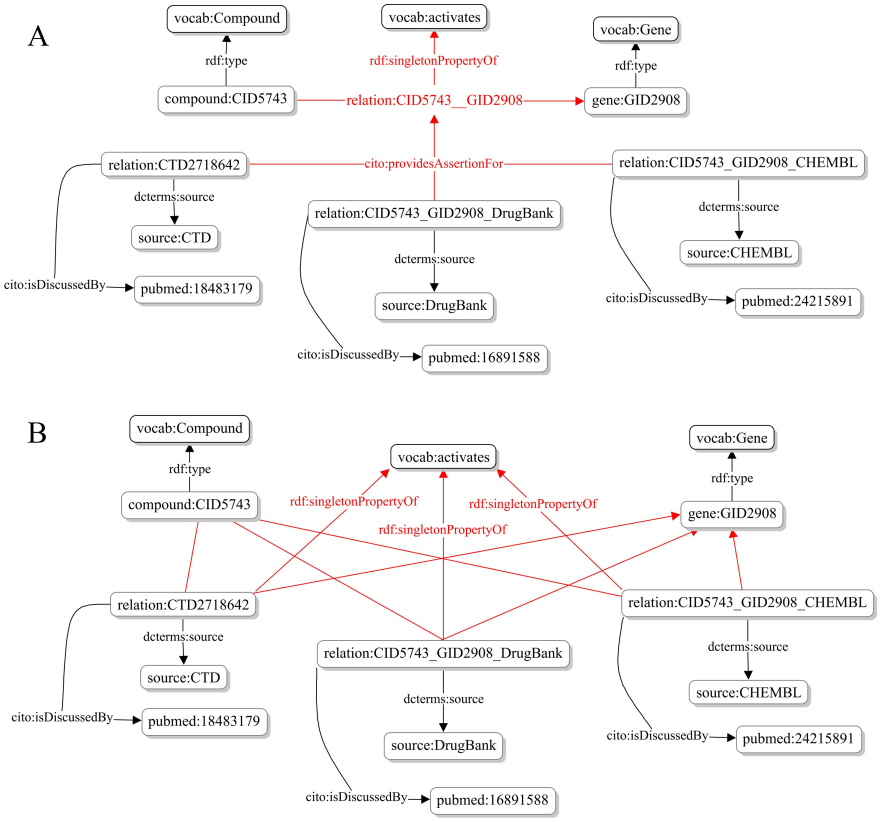

Fig. 2. Graphical representation of SP model for the relation between compound CID5743 and gene 2908: (a) with cardinal assertion (Model III); (b) without cardinal assertion (Model IV).

performed on cold cache. The Log10 transformations of the execution time in millisecond were illustrated in boxplot; the averages and standard deviations of the execution time in seconds were summarized as well in the comparative analysis.

The data sets and the SPARQL queries are available at: `http://figshare.com/articles/Provenance_RDF_Models/1399197`.

3 Results and Discussions

3.1 Data set statistics

We first compared the total number of triples that each RDF model contains. The most efficient RDF model is SP model without cardinal assertion (Model **IV**), which contains 17,239,427 triples, and the cardinal assertion of SP model (Model

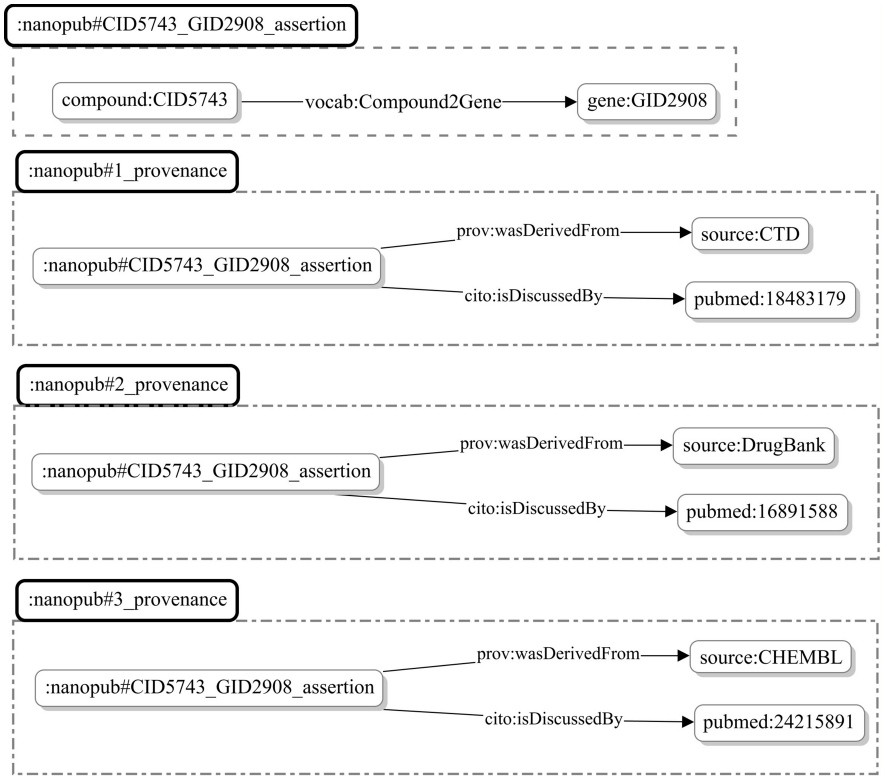

Fig. 3. Graphical representation of Nanopublication model (Model V) for the relation between compound CID5743 and gene 2908.

III) increased the total number of triples by about 14% to 19,575,298. For N-ary models, the cardinal assertion also increased the total number of triples by about 6%, from 21,445,348 (Model **II**) to 22,787,218 (Model **I**). The N-ary model requires two triples (predicates `sio:has-agent` and `sio:has-target`) to represent the agent and target in a biological process, while the SP model maintains the previous binary relation structure in only one triple. Hence, with the cardinal assertion, the N-ary model (Model **I**) contains 3,211,920 more triples ($\sim$ 16%) in comparison with the SP model (Model **III**), and without the cardinal assertion, the N-ary model (Model **II**) contains even more triples (4,205,921 triples) in contrast to the SP model (Model **IV**). The Nanopublication model is the most verbose model in this regard, which contains 27,605,782 triples distributed in 8,251,238 graphs.

We also studied the amount of evidence associated with each relational assertion to illustrate the degree of redundancy with respect to the identical pairwise relations in the life science domain. We only examine object property instances representing the pairwise relations that were created in the SP models (Model **III** and Model **IV**), as the degree of redundancy is same across other RDF mod-

els. The total number of unique subjects in the SP models with and without cardinal assertion are 7,654,605 (Model **III**) and 4,442,685 (Model **IV**), respectively. The difference between the two numbers accounts for the total number of object property instances arbitrarily created for the cardinal assertions. If there are multiple cases of evidence for a given assertion, the cardinal assertion variant may reduce the total number of triples to express the same information, however, if there is only one case of evidence for a given assertion, the cardinal assertion will increase the total number of triples. Hence, whether the cardinal assertion can reduce the total number of triples depends on the extent of redundancy of the identical pairwise relations in the data set. Among 3,211,920 cardinal assertions, 2,800,124 (∼87%) of them are only associated with one evidence, 238,558 (∼7%) of them are associated with two cases of evidence, 67,088 (∼2%) of them are associated with three cases of evidence, and 98,625 (∼3%) of them are associated with more than three cases of evidence. The pairwise relations between PubChem compound CID5694 and NCBI gene GID5465 is associated with the most number of cases of evidence (3,096). Although there were many redundant assertions from multiple data sources, the majority have only one supporting evidence. Hence, the increase in the total number of triples were largely attributable to publication assertions.

3.2 Query performance evaluation

We undertook a performance evaluation using three RDF databases (see Table 1). With Virtuoso, the SP models with and without cardinal assertion (Model **III** and **IV**) largely outperformed the other models. **Q1** and **Q2** executed roughly 100 times faster on the SP models as compared to the N-ary models. Although Model **V** yielded comparable performance with Model **III** and I**V** in **Q1**, the additional filtering constraint made it much slower in **Q2**. In **Q4**, Model **III**, **IV**, and **V** performed similarly, which are 10 times faster than Model **II** and 100 times faster than Model **I**. In general, Virtuoso performed best using the SP models. With the Stardog RDF store, the N-ary models and the SP models were comparable in performance, but they always outperformed Nanopublication model. In particular, when the aggregated evidence was considered in **Q4**, both N-ary and SP models with and without cardinal assertion were carried out over 10 times faster than the Nanopublication model. Using Blazegraph, the Nanopublication model generally outperformed other models. In particular, **Q1** and **Q2** were carried out over 10 times faster in Model **V** rather than in other models.

Without querying the provenance metadata, the models with cardinal assertion (Model **I** and **III**) always yielded better performance in comparison with the models without cardinal assertion (Model **II** and **IV** accordingly). Hence, if we remove the redundant identical assertions from various data sources in both N-ary and SP models, the graph traversal-like queries can be executed much faster. If we think of conjunctive queries (i.e. graph traversal or inner join) as performing Cartesian products, the computational costs go up exponentially as the number of data items increase. Hence, the redundant pairwise relations cost

much more time rather than cardinal assertions in **Q1** and **Q2**. However, if the provenance restrictions were considered, the model without cardinal assertion (Model **II** and **IV**) usually outperformed, except the **Q3** of the SP models executed in Stardog and **Q4** of both N-ary and SP models executed in Blazegraph. But the difference of query performance were usually small, except for the **Q4** of the N-ary models executed in Virtuoso, and the **Q3** of both N-ary and SP models executed in Blazegraph. So in general, if the provenance restrictions were considered, the models with and without cardinal assertion were comparable.

Table 1. The average execution time and standard deviation in seconds

		Model I[a]	**Model II**[a]	**Model III**[a]	**Model IV**[a]	**Model V**[a]
Virtuoso	Q1	44.827 (±15.918)	269.854 (±99.266)	**0.665** (±0.263)	2.398 (±1.082)	1.283 (±0.212)
	Q2	260.337 (±253.588)	369.635 (±120.482)	**0.535** (±0.301)	2.375 (±1.083)	585.52 (±193.382)
	Q3	4.04 (±0.294)	3.069 (±0.161)	3.075 (±0.243)	**2.248** (±0.718)	2.287 (±0.049)
	Q4	352.312 (±204.483)	14.994 (±11.587)	2.201 (±0.028)	**1.953** (±0.331)	2.531 (±0.054)
StarDog	Q1	**1.906** (±0.214)	5.499 (±0.191)	3.354 (±0.783)	16.805 (±13.315)	21.291 (±0.621)
	Q2	**2.45** (±0.262)	6.366 (±0.208)	4.072 (±0.492)	18.398 (±14.383)	184.96 (±90.820)
	Q3	2.738 (±0.240)	**1.291** (±0.068)	3.834 (±0.500)	16.463 (±14.277)	27.537 (±1.602)
	Q4	14.344 (±1.576)	9.084 (±0.147)	9.575 (±0.751)	**8.698** (±0.467)	45.959 (±1.263)
Blazegraph	Q1	11.087 (±5.129)	54.74 (±32.618)	33.597 (±10.515)	41.491 (±11.862)	**0.732** (±0.133)
	Q2	10.853 (±4.494)	56.599 (±34.590)	33.469 (±10.292)	41.099 (±12.697)	**4.215** (±1.635)
	Q3	10.944 (±4.915)	1.56 (±0.533)	6.05 (±1.284)	**0.465** (±0.054)	0.581 (±0.071)
	Q4	83.384 (±1.612)	117.131 (±2.721)	**74.505** (±1.570)	89.054 (±1.119)	80.729 (±1.436)

[a] The average execution times are in the first line, and the standard deviations are in the second line within parenthesis; the best performance has been highlighted in bold.

4 Conclusion

In this study, we evaluated three existing RDF models and two cardinal assertion models for representing relations and exposing their provenance metadata. We examined the effect of each model on overall graph size and query time execution across three different RDF databases. Since our integrated life science dataset contained many duplicate assertions, graph traversal can be accomplished in a much more efficient way using the cardinal assertion. The redundant assertions add up a lot of computational overhead when searching through the integrated knowledge base for evidence-based hypothesis exploration. Surprisingly, we found that each RDF store performed the best using a different provenance model. It has been demonstrated that SPARQL queries may be executed in a RDF store specific manner in a previous analysis [17]. Our results drew a similar conclusion and may have contentious implications for the standardization of a provenance model, which should ideally be software/platform/system agnostic. A more extensive analysis with larger benchmark datasets and more query patterns would be helpful in the future study.

Acknowledgements This work was initiated at the 2014 BioHackathon in Fukashima; This research was supported [in part] by the Intramural Research Program of the National Library of Medicine; The research leading to these results has received support from Instituto de Salud Carlos III-Fondo Europeo de Desarrollo Regional (PI13/00082 and CP10/00524), the Innovative Medicines Initiative Joint Undertaking under grant agreements n 115191 (Open PHACTS)], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution. The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

References

1. Blazegraph. `http://www.systap.com/rdf/`
2. Rdf n-ary. `http://www.w3.org/TR/swbp-n-aryRelations/`
3. Stardog. `http://stardog.com/`
4. Virtuoso. `http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/`
5. A, G., JCJ, v.D., EA, S., M, R., B, M.: Towards computational evaluation of evidence for scientific assertions with nanopublications and cardinal assertions. In: 5th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS). pp. 28–30
6. Bauer-Mehren, A., Rautschka, M., Sanz, F., Furlong, L.I.: Disgenet: a cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics 26(22), 2924–6 (2010)
7. Belleau, F., Nolin, M., Tourigny, N., Rigault, P., Morissette, J.: Bio2rdf: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics 41(5), 706–716 (2008)
8. Brown, S.H., Elkin, P.L., Rosenbloom, S.T., Husser, C., Bauer, B.A., Lincoln, M.J., Carter, J., Erlbaum, M., Tuttle, M.S.: Va national drug file reference terminology: a cross-institutional content coverage study. Stud Health Technol Inform 107(Pt 1), 477–81 (2004)

9. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B.L., Wiegers, T.C., Mattingly, C.J.: The comparative toxicogenomics database's 10th year anniversary: update 2015. Nucleic Acids Res 43(Database issue), D914–20 (2015)
10. Fu, G., Batchelor, C., Dumontier, M., Hastings, J., Willighagen, E., Bolton, E.: Pubchemrdf: towards the semantic annotation of pubchem compound and substance databases. J Cheminform 7, 34 (2015)
11. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P.: Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40(Database issue), D1100–7 (2012)
12. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Inf. Serv. Use 30(1-2), 51–56 (2010)
13. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: Kegg for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40(Database issue), D109–14 (2012)
14. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A.C., Wishart, D.S.: Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39(Database issue), D1035–41 (2011)
15. Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J., Bork, P.: A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 6, 343 (2010)
16. Machado, C.M., Rebholz-Schuhmann, D., Freitas, A.T., Couto, F.M.: The semantic web in translational medicine: current applications and future directions. Brief Bioinform 16(1), 89–103 (2015)
17. Mironov, V., Seethappan, N., Blonde, W., Antezana, E., Splendiani, A., Kuiper, M.: Gauging triple stores with actual biological data. BMC Bioinformatics 13 Suppl 1, S3 (2012)
18. Nguyen, V., Bodenreider, O., Sheth, A.: Don't like rdf reification?: Making statements about statements using singleton property. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 759–770. WWW '14, ACM, Republic and Canton of Geneva, Switzerland (2014), `http://dx.doi.org/10.1145/2566486.2567973`
19. Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., Furlong, L.I.: Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford) 2015 (2015)
20. Razick, S., Magklaras, G., Donaldson, I.M.: irefindex: a consolidated protein interaction database with provenance. BMC Bioinformatics 9, 405 (2008)
21. Rosemblat, G., Shin, D., Kilicoglu, H., Sneiderman, C., Rindflesch, T.C.: A methodology for extending domain coverage in semrep. J Biomed Inform 46(6), 1099–107 (2013)
22. Sahoo, S., Nguyen, V., Bodenreider, O., Parikh, P., Minning, T., Sheth, A.: A unified framework for managing provenance information in translational research. BMC Bioinformatics (2011)
23. Sharman, J.L., Benson, H.E., Pawson, A.J., Lukito, V., Mpamhanga, C.P., Bombail, V., Davenport, A.P., Peters, J.A., Spedding, M., Harmar, A.J.: Iuphar-db: updated database content and new features. Nucleic Acids Res 41(Database issue), D1083–8 (2013)
24. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: Biogrid: a general repository for interaction datasets. Nucleic Acids Res 34(Database issue), D535–9 (2006)