A model for capturing provenance of assertions about chemical substances

Kody Moodley[1], Amrapali Zaveri[1], Chunlei Wu[2], and Michel Dumontier[1][0000−0003−4727−9435]

[1] Institute of Data Science, Maastricht University, Universiteitsingel 60, 6229 ER, Maastricht, The Netherlands
firstname.lastname@maastrichtuniversity.nl
[2] Department of Molecular and Experimental Medicine, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA
cwu@scripps.edu

Abstract. Chemical substance resources on the Web are often made accessible to researchers through public APIs (Application Programming Interfaces). A significant problem of missing provenance information arises when extracting and integrating data in such APIs. Even when provenance is stated, it is usually not done with any prescribed templates or terminology. This creates a burden on data producers and makes it challenging for API developers to automatically extract and analyse this information. Downstream, these consequences hinder efforts to automatically determine the veracity and quality of extracted data, critical for proving the integrity of associated research findings. In this paper, we propose a model for capturing provenance of assertions about chemical substances by systematically analyzing three sources: (i) Nanopublications, (ii) Wikidata and (iii) selected Minimal Information Standards (MISTS) for reporting biomedical studies[3]. We analyse provenance terms used in these sources along with their frequency of use and synthesize our findings into a preliminary model for capturing provenance.

Keywords: API · provenance · evidence · data model · chemical substance

1 Introduction

The increasing number of chemical substance databases on the Web are often made accessible through public APIs (Application Programming Interfaces), queryable by researchers to enrich their computational analyses and scientific workflows. One such API is the BioThings API suite[4]. "BioThings" refer to objects of any biomedical entity-type represented in the biological knowledge space, such as genes, drugs, chemicals, diseases, etc. The popular `MyChem.info` (chemicals), `MyGene.info` (genes) and `MyVariant.info` (gene variants) APIs are demonstrable examples built and maintained using the SDK.

[3] Reported in FAIRsharing.org `https://fairsharing.org`
[4] `http://biothings.io/`

The problem of *provenance* for scientific *assertions* arises for stakeholders of such APIs. We adopt the W3C's definition of *provenance*: "Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness"[5]. An assertion refers to an individual statement about a particular entity, in this case, a chemical entity. For example, in PubChem it states "Acetaminophen has a melting point of 168 degrees celsius"[6]. In practice, this assertion might be encoded in a concrete data format, e.g. in JSON as a key-value pair, or in RDF (Resource Description Framework)[7] as a subject-predicate-object triple. In PubChem, two references are listed for this assertion: the publication reporting it, and a database from which it was retrieved. Both items indicate provenance, though provenance is not limited to these types. Provenance also includes *evidence* supporting the assertion (in the form of specific data elements, media, graphs etc.) and the *experimental methodologies* that generate it.

Data producers, for various sources that an API integrates, often do not submit sufficient provenance information for assertions, or, when they do, they often use inconsistent terminology. This inconsistent capturing is not helped by the lack of a standardized specification for doing so. The problem also makes it challenging for API developers to represent provenance in the results of submitted queries, in a machine-interpretable way. While there have been attempts to develop vocabularies for capturing provenance, there is no accepted guideline for identifying which provenance items are essential (*must* be specified), recommended (*should* be specified) and optional. Another missing feature is a universally accepted standard for which vocabularies to use when specifying provenance about chemical assertions. For example, Wikidata [14] uses "stated in" (encoded as `P248`) to refer to a database from which an assertion was extracted. However, Nanopublications (Nanopubs) [10] have a variety of possible terms to describe the same item, including: "hasSource", "references" and "cites". Downstream, this complicates automatic identification, extraction and processing of provenance by developers, which is crucial for validating research findings associated with given assertions.

To address these problems, we systematically analyse three sources: (i) Nanopublications, (ii) Wikidata and (iii) selected MISTS for reporting biomedical studies reported in FAIRsharing.org, to examine how they capture provenance. We then synthesise our findings to propose a preliminary model for capturing provenance of assertions about chemical substances. Our motivation for studying Nanopubs and Wikidata in particular is that they are the only large-scale databases of scientific assertions providing mechanisms for specifying machine-processable provenance information that is mapped to standard ontology terms. MISTS are studied because they are the only known recommendations of minimal information (including provenance of study assertions) required for describing biomedical studies.

[5] `https://www.w3.org/TR/prov-overview`
[6] `https://pubchem.ncbi.nlm.nih.gov/compound/1983#section=Melting-Point`
[7] `https://www.w3.org/RDF/`

2 Related Work

There have been many efforts at standardizing general data provenance on the Web [6,13,5]. Provenance of datasets in life sciences has also received attention with the BioSchemas [4], DATA tag suite [11] and HCLS [3] initiatives. However, there are currently no specialized models for chemical substance assertions.

In terms of scientific assertions, the most relevant initiative providing a partial specification, is Wikidata[8]. Wikidata is an online open knowledge repository storing structured data about information on a wide variety of topics, including chemicals. In Wikidata's provenance model, an assertion is called a "claim". An example of a claim is"Acetaminophen is a subclass of non-opioid analgesic". The entities "Acetaminophen" and "non-opioid analgesic", and the "subclass" relation, are referred to by Wikidata-specific identifiers `Q57055`, `Q1747785` and `P279` respectively. Each claim has a list of "references" which are each a separate record of provenance for the claim. Each reference can specify an arbitrary number of provenance items supporting the claim (e.g. database, publication or date retrieved). Wikidata's provenance model is not sufficient for our goals since it does not indicate which provenance items are essential, recommended and optional. It also uses Wikidata-specific terms for provenance which are not mapped to standard (bio-)ontology terms to increase their interpretability.

There have also been various efforts to standardize terminology that data producers can use to capture provenance [7,2,9,12,1]. We briefly describe three prominent terminologies below.

Open PHACTS & The W3C Provenance working group. Open PHACTS [15] is an online drug discovery platform integrating, linking, and providing access to data across numerous biomedical resources. To establish a standardized and interoperable way for exchanging provenance information, Open PHACTS participated in the activities of the W3C Provenance working group to influence the development of such a standard. A major output of the W3C Provenance working group is the Provenance Ontology (PROV-O) [8], a domain-independent RDF terminology for describing provenance information. PROV-O consists of terms to denote either physical or conceptual entities of interest, as well as property terms to denote provenance-related relationships between such entities [8, Figure 1]. For example, the assertion: "Rifampin is effective in the treatment of Pulmonary Tuberculosis" is represented in RDF as: `<http://purl.obolibrary.org/obo/CHEBI_28077>` `<http://purl.obolibrary.org/obo/RO_0002606>` `<http://purl.bioontology.org/ontology/SNOMEDCT/154283005>`. To associate provenance with this statement, such as the publication in which it was proposed, one can use `prov:hadPrimarySource`[9] or `prov:wasQuotedFrom` to specify the URL or DOI (Digital Object Identifier) for the article in which the assertion originates.

Provenance, Authoring & Versioning (PAV) ontology. The PAV ontology[10] addresses some limitations of PROV-O with regards to authoring and version-

[8] `http://Wikidata.org`

[9] prov is prefix for http://www.w3.org/ns/prov#

[10] `https://pav-ontology.github.io/pav`

ing of digital resources on the Web. In particular, PROV-O is not expressive enough to capture specialised authoring roles such as "contributor" and "curator". This finer-grained approach also extends to properties associated with these roles. The ontology was designed to be light-weight (containing as few terms as possible for satisfying the requirements) which makes it useful only as a complementary terminology, in combination with others like PROV-O, to fit use-cases in specialized knowledge domains.

Chemical Information (CHEMINF) Ontology. The Chemical Information Ontology[11] was established to structure information and standardize terminology in modern chemical research. This kind of research regularly requires various simulations and calculations to be performed on chemical data. CHEMINF provides a standardized vocabulary for annotating the various calculations obtained by the software, which is crucial for diagnosing errors and determining their veracity. The terms in CHEMINF are broadly classified into those related to: software, the algorithms implemented by them, and properties of chemical substances.

3 Provenance usage in Nanopubs, Wikidata and MISTS

3.1 Nanopublications

Building on the RDF standard for representing assertions on the Web, Barend Mons and Jan Velterop proposed the concept of Nanopublication in 2009 [10]. The motivation was to enable researchers to publish small structured snippets of knowledge (Nanopublications or nanopubs for short) from research data complementing the publishing of traditional full-length research texts. Indeed Mons and Velterop argue that: "Published contributions to science can be as short as single statements that interpret data, and yet be valuable to scientific progress and understanding" [10, Section 2.4]. A nanopub consists of a core assertion represented as an RDF triple, richly annotated with qualifying metadata and provenance for the assertion. Since all language features required to represent nanopubs are included in RDF, the nanopub specification is essentially an RDF design pattern for expressing assertions and their associated metadata. We queried all Nanopublications[12] (the latest dump on 5 April 2018 which includes 10,803,231 nanopubs) to retrieve all properties associated with them, along with their usage frequencies. A total of 333 properties were extracted. We then manually pruned this list of properties to retain only those that were related to provenance of an assertion, which resulted in 37 unique properties. From this pruned list, we found that the most frequently used properties are "hasPublicationInfo", "hasEvidence", "date of assertion" and "author". Figure 1 plots usage frequencies (logarithmic scale) of each provenance item.

We further classified the properties into five major dimensions:

[11] `https://www.ebi.ac.uk/ols/ontologies/cheminf`
[12] https://zenodo.org/record/1213293#.W6-WwxMzaAw

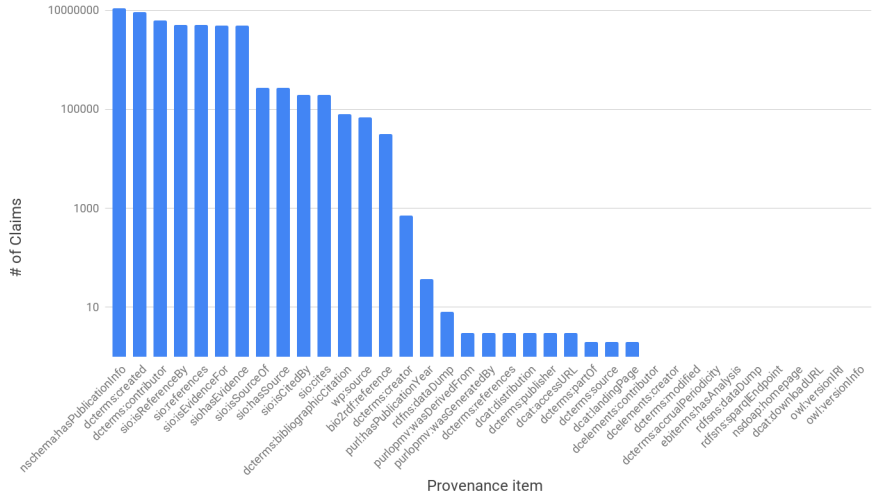

Fig. 1. Usage frequency of each provenance item in Nanopubs.

- *Authorship* refers to information about the person(s) or organisational entities responsible for generating the assertion
- *Temporal aspects* refer to information concerning *when* the assertion was made, modified etc.
- *Source* refers to information about the publication or database from which the assertion was extracted.
- *Generating process* concerns information about the experiments, studies and analytical procedures which led to the proposal of the assertion.
- *Evidence* refers to individual datasets, pieces of data and other assertions which support or prove the given assertion.

Figure 3 plots the usage frequency (logarithmic scale) of Nanopublication provenance properties in these different dimensions.

3.2 Wikidata

We performed a similar analysis of claims in Wikidata. We extracted c.a. 150,000 records of type "Chemical Compound" from Wikidata. For all claims about these compounds that had references, we counted the usage frequency of each type of provenance property (see Figure 2).

As a result of our analysis, we retrieved 76 unique metadata items (Figure 2). We further analysed this list to eliminate domain-specific provenance metadata items and pruned the list to 37 metadata items. Some examples of non-domain specific items we found are "statedin", "publicationdate", "retrieved", "software version", "DOI" etc. By examining the frequency across the provenance dimensions (Figure 3), we noticed that Nanopub authors submit much more evidence, authorship and temporal information than Wikidata contributors.

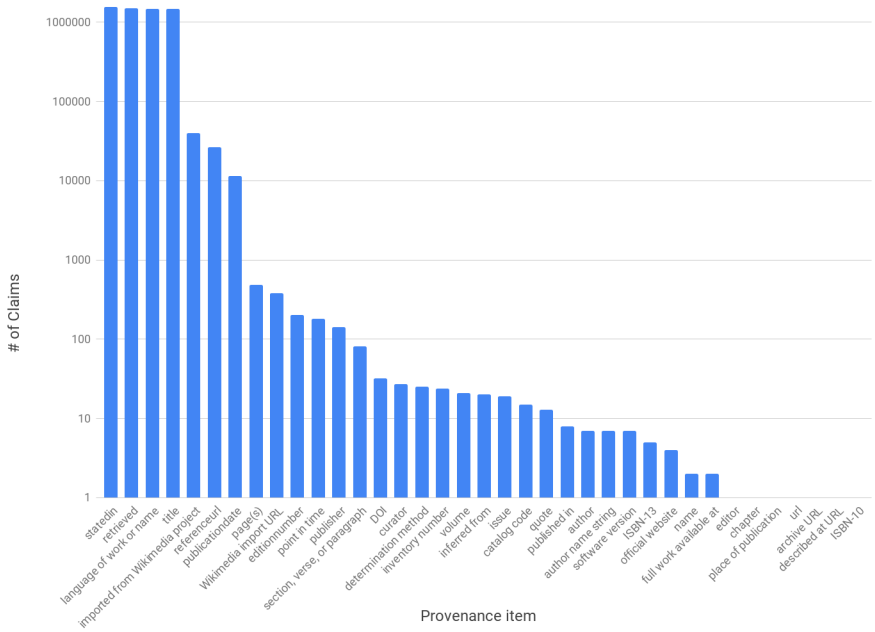

Fig. 2. Usage frequency of each provenance item in Wikidata.

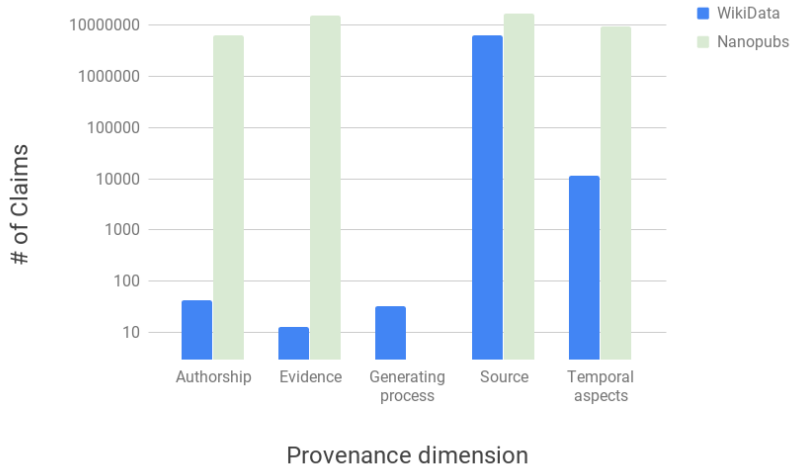

Fig. 3. Wikidata and Nanopubs provenance usage frequency per dimension.

Another finding is that for both Nanopubs and Wikidata, Generating process information is largely not specified.

3.3 MISTS

We additionally analyzed selected MISTS for reporting biomedical studies that were registered at `FAIRsharing.org`, which are 'Recommended' and have a publication ('Has Publication'). As a result of this query, we retrieved a list of 16 reporting guidelines, of which 2 were duplicates and 1 was unavailable. Thus, we analysed the following 14 reporting guidelines' metadata elements:

- Animals in Research: Reporting In Vivo Experiments
- STAndards for the Reporting of Diagnostic accuracy
- STrengthening the Reporting of OBservational studies in Epidemiology
- Preferred Reporting Items for Systematic reviews and Meta-Analyses
- CONSOlidated standards of Reporting Trials
- Minimum Information About a Microarray Experiment
- Minimal Information Required In the Annotation of Models
- Minimum Information about a Molecular Interaction Experiment
- Minimum Information About a Proteomics Experiment
- Minimum Information about any (x) Sequence
- Recommended reporting guidelines for life science resources
- Consolidated criteria for reporting qualitative research
- Case Reports
- Consolidated Health Economic Evaluation Reporting Standards

A total of 347 metadata elements were extracted from all of these reported guidelines. We analyzed each of them and pruned the lists to 44 elements such as "ethics statement", "apparatus", "duration", "location".

4 Proposed provenance model

The analysis in Section 3 demonstrates that all extracted provenance properties can be classified into the 5 dimensions of provenance information: Authorship, Temporal aspects, Evidence, Generating Process and Source. While there may be other interesting dimensions of provenance information that data producers do not yet record, our primary goal is to capture *existing* information, in a structured way. Therefore, we use the 5 dimensions as a guide for selecting the relevant provenance items for our model. For each dimension we also identified certain subcategories of provenance items relevant to that dimension. The complete hierarchy is depicted in Figure 4.

The procedure for selecting the provenance items (and corresponding ontology terms) to use for each provenance dimension, was based on two criteria: 1) the frequency of their use by data publishers in Wikidata and Nanopubs, and 2) their relative importance in determining the veracity of an assertion (as judged by two postdoctoral researchers with 12 years of combined experience

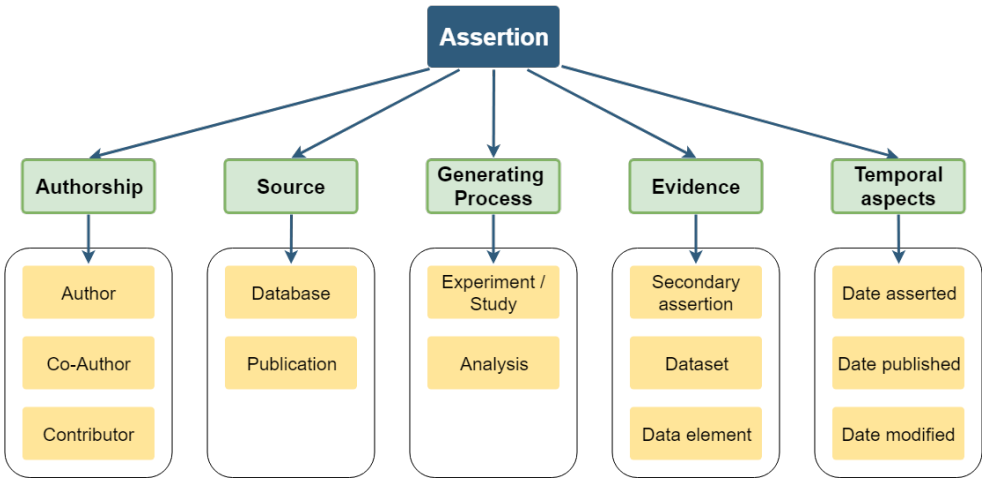

Fig. 4. Provenance dimensions for a scientific assertion about chemicals.

in bioinformatics research). For authorship of an assertion we identified three potentially important pieces of information: the name of the primary author of the assertion, names of any co-authors of the assertion, as well as names of any persons who contributed to the discovery or generation of the assertion. The Source dimension can be broadly separated into properties pertaining to the scientific publication in which the assertion was reported, and those concerning a database (potentially an indirect source) from which the assertion was retrieved. For Generating process, we divided the properties into those related to clinical trial studies, and those related to computational analyses. Evidence can be derived from another (secondary) assertion, a specific data element (such as images, audio/visual media, graphs, calculations etc.), or a specialized dataset. Finally, the main temporal properties we associate with an assertion are the dates on which it was conceived, published and last modified.

Thus, our proposed provenance model consists of (i) a listing of provenance properties to be used for assertions about chemical substances, (ii) a precise definition of each property (through mappings to standard ontology terms), and (iii) a recommendation of which properties are essential, recommended, and optional, respectively. These are fully detailed in the file "BioThingsProvenanceModel.xslx" within our GitHub repository[13]. Table 1 summarises the essential properties in each dimension. Each property is assigned an ontology term using prefixes (resolvable at `http://prefix.cc`).

The model is independent of any concrete data format, and is implementable in the JSON-LD and RDF standards, for example. We provide example instantiations of our model for these two formats in our GitHub repository, located in files "jsonldexample.json" and "rdfexample.ttl", respectively.

[13] `https://github.com/MaastrichtU-IDS/biothingsprovenancemodel`

#	Name	(Sub)Dimension	Ontology term
1	assertedBy	Author	dcterms:creator
2	coAssertedBy	Co-author	obo:MS_1002036
3	assertedOn	Date asserted	prov:generatedAtTime
4	publishedOn	Date published	dbpedia:publicationDate
5	supportedByDataSet	Dataset	prov:wasDerivedFrom
6	supportingDatasetVersion	Dataset	schema:version
7	supportingDatasetLicense	Dataset	schema:license
8	supportingDatasetURL	Dataset	schema:url
9	wasDerivedFrom	Data element	sio:SIO_000772
10	wasInferredFrom	Secondary assertion	prov:wasDerivedFrom
11	supportingExperimentID	Experiment / Study	schema:identifier
12	supportingAnalysisSoftware	Analysis	swo:SWO_0000001
13	supportingAnalysisMethod	Analysis	prov:wasGeneratedBy
14	supportingAnalysisSoftwareVersion	Analysis	sio:SIO_000654
15	supportingSoftwareLicense	Analysis	schema:license
16	publishedIn	Publication	prov:hadPrimarySource
17	publisher	Publication	schema:publisher
18	publicationTitle	Publication	dbpedia:publicationTitle
19	retrievedFrom	Database	pav:retrievedFrom
20	retrievalURL	Database	pav:retrievedFrom
21	databaseURL	Database	schema:url
22	databaseLicense	Database	schema:license
23	databaseVersion	Database	schema:version

Table 1. Summary of the prescribed essential provenance properties in our model.

5 Conclusions, limitations and future work

In this paper, we proposed a model for capturing provenance about chemical substances when retrieving information via an API across disparate data sources. We reported on a systematic analysis of three sources concerning how they report provenance and specifically which non-domain specific provenance items are advocated based on their frequency of usage. The three sources analyzed were (i) Nanopublications, (ii) Wikidata, (iii) Selected MISTS for reporting biomedical studies (from FAIRsharing.org). Our provenance model consists of 90 unique properties. As future work, we will implement the model on 10 prominent data sources amalgamated by the BioThings API (specifically `MyChem.info`), and evaluate it's utility. As next steps, we envision the relevance of selected elements to be discussed widely within the biomedical community, and that a future version will be recommended for widespread uptake.

6 Acknowledgements

Support for this work was provided by NCATS, through the Biomedical Data Translator program (NIH awards OT3TR002027 [Red]). Any opinions expressed in this document are those of the Translator community writ large and do not

necessarily reflect the views of NCATS, individual Translator team members, or affiliated organizations and institutions.

References

1. Brush, M.H., Shefchek, K., Haendel, M.: SEPIO: A semantic model for the integration and analysis of scientific evidence. In: Proceedings of the Joint International Conference on Biological Ontology and BioCreative (2016)
2. Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A.J., Goble, C., Clark, T.: Pav ontology: provenance, authoring and versioning. J Biomed Semantics **4**(1), 37 (2013)
3. Dumontier, M., Gray, A.J., Marshall, M.S., Alexiev, V., Ansell, P., Bader, G., Baran, J., Bolleman, J.T., Callahan, A., Cruz-Toledo, J., et al.: The health care and life sciences community profile for dataset descriptions. PeerJ **4**, e2331 (2016)
4. Garcia, L., Giraldo, O., Garcia, A., Dumontier, M.: Bioschemas: schema. org for the life sciences. Proceedings of SWAT4LS (2017)
5. Glavic, B., Dittrich, K.R.: Data provenance: A categorization of existing approaches. In: BTW. vol. 7, pp. 227–241 (2007)
6. Hartig, O.: Provenance information in the web of data. In: Proceedings of the Linked Data on the Web Workshop (2009)
7. Hastings, J., Chepelev, L., Willighagen, E., Adams, N., Steinbeck, C., Dumontier, M.: The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. PloS one **6**(10), e25513 (2011)
8. Lebo, T., Sahoo, S., McGuinness, D., Khalid Belhajjame, J.C., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: PROV-O: The PROV ontology. W3C Recommendation (2013), `http://www.w3.org/TR/prov-o`, (last accessed 28 August 2018)
9. Missier, P., Belhajjame, K., Cheney, J.: The w3c prov family of specifications for modelling provenance metadata. In: Proceedings of EDBT. pp. 773–776. ACM (2013)
10. Mons, B., Velterop, J.: Nano-publication in the e-science era. In: Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (2009)
11. Sansone, S.A., Gonzalez-Beltran, A., Rocca-Serra, P., Alter, G., Grethe, J.S., Xu, H., Fore, I.M., Lyle, J., Gururaj, A.E., Chen, X., et al.: Dats, the data tag suite to enable discoverability of datasets. Scientific data **4**, 170059 (2017)
12. Sarntivijai, S., Vasant, D., Jupp, S., Saunders, G., Bento, A.P., Gonzalez, D., Betts, J., Hasan, S., Koscielny, G., Dunham, I., Parkinson, H., Malone, J.: Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation. Journal of Biomedical Semantics **7**(1), 8 (2016)
13. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. SIGMOD Record **34**(3), 31–36 (2005)
14. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
15. 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), 1188 – 1198 (2012)