Nanopublications: A growing resource of provenance-centric scientific linked data

Kuhn, Tobias; Merono-Penuela, Albert; Malic, Alexander; Poelen, Jorrit H.; Hurlbert, Allen H.; Ortiz, Emilio Centeno; Furlong, Laura I.; Queralt-Rosinach, Nuria; Chichester, Christine; Banda, Juan M.; Willighagen, Egon; Ehrhart, Friederike; Evelo, Chris; Malas, Tareq B.; Dumontier, Michel

published in
2018 IEEE 14th International Conference on eScience (e-Science)
2018

DOI (link to publisher)
10.1109/eScience.2018.00024

document version
Publisher's PDF, also known as Version of record

document license
Article 25fa Dutch Copyright Act

citation for published version (APA)
Kuhn, T., Merono-Penuela, A., Malic, A., Poelen, J. H., Hurlbert, A. H., Ortiz, E. C., Furlong, L. I., Queralt-Rosinach, N., Chichester, C., Banda, J. M., Willighagen, E., Ehrhart, F., Evelo, C., Malas, T. B., & Dumontier, M. (2018). Nanopublications: A growing resource of provenance-centric scientific linked data. In *2018 IEEE 14th International Conference on eScience (e-Science): [Proceedings]* (pp. 83-92). Article 8588643 Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/eScience.2018.00024

Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data

Tobias Kuhn
Department of Computer Science
Vrije Universiteit Amsterdam
The Netherlands
t.kuhn@vu.nl

Albert Meroño-Peñuela
Department of Computer Science
Vrije Universiteit Amsterdam
The Netherlands

Alexander Malic
Institute of Data Science
Maastricht University
The Netherlands

Jorrit H. Poelen
400 Perkins St
Oakland, California
USA

Allen H. Hurlbert
Department of Biology
University of North Carolina
Chapel Hill, USA

Emilio Centeno Ortiz
Research Programme on
Biomedical Informatics, Hospital del
Mar Medical Research Institute
Universitat Pompeu Fabra
Barcelona, Spain

Laura I. Furlong
Research Programme on
Biomedical Informatics, Hospital del
Mar Medical Research Institute
Universitat Pompeu Fabra
Barcelona, Spain

Núria Queralt-Rosinach
Department of Integrative Structural
and Computational Biology
The Scripps Research Institute
La Jolla, USA

Christine Chichester
Datafair.xyz
Geneva, Switzerland

Juan M. Banda
Department of Computer Science
Georgia State University
USA

Egon Willighagen
Department of Bioinformatics
BiGCaT, NUTRIM
Maastricht University
The Netherlands

Friederike Ehrhart
Department of Bioinformatics
BiGCaT, NUTRIM
Maastricht University
The Netherlands

Chris Evelo
Department of Bioinformatics
BiGCaT, NUTRIM
Maastricht University
The Netherlands

Tareq B. Malas
Department of Human Genetics
Leiden University Medical Center
The Netherlands

Michel Dumontier
Institute of Data Science
Maastricht University
The Netherlands

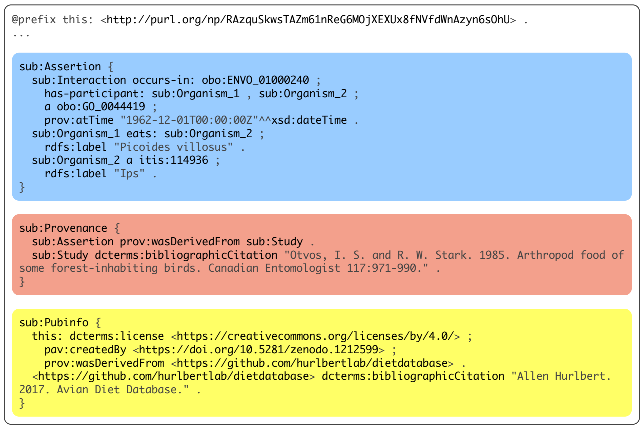
***Abstract*—Nanopublications are a Linked Data format for scholarly data publishing that has received considerable uptake in the last few years. In contrast to the common Linked Data publishing practice, nanopublications work at the granular level of atomic information snippets and provide a consistent container format to attach provenance and metadata at this atomic level. While the nanopublications format is domain-independent, the datasets that have become available in this format are mostly from Life Science domains, including data about diseases, genes, proteins, drugs, biological pathways, and biotic interactions. More than 10 million such nanopublications have been published, which now form a valuable resource for studies on the domain level of the given Life Science domains as well as on the more technical levels of provenance modeling and heterogeneous Linked Data. We provide here an overview of this combined nanopublication dataset, show the results of some overarching analyses, and describe how it can be accessed and queried.**

I. Introduction

Provenance has been identified as a crucial aspect to enable trust in Linked Data environments in general and eScience in particular [28], and has led to the now widely adopted PROV ontology [23]. However, there is still a lack of proven methods on how to represent and publish such provenance information for scientific data in a general, reliable, and agreed-upon manner. Nanopublications [14], [27] have been proposed as a solution to this problem by providing a granular and principled way of publishing scientific (and other types of) data in a provenance-centric manner. Such a nanopublication consists of an atomic snippet of a formal statement (called "assertion") that comes with information about where this knowledge came from (the provenance of the assertion, i.e. how it was discovered) and with metadata about the nanopublication as a whole (the provenance of the nanopublication, i.e. how it was created; this part is called "publication info"). All these three parts are represented as Linked Data (in RDF) and together they constitute a self-contained entity that we call a nanopublication. On the technical level, nanopublications are implemented with the help of named RDF graphs [8], with one graph for each of assertion, provenance, and publication info, plus an additional head graph that holds everything together. Here, we present a growing dataset of such nanopublications, currently consisting of more than 10 million nanopublications containing diverse data (mostly from the Life Sciences so far) coming from a variety of contributors.

Figure 1 shows an example of a nanopublication from a recently added dataset on bird diets. It shows the three

978-1-5386-9156-4/18/$31.00 ©2018 IEEE
DOI 10.1109/eScience.2018.00024



```
@prefix this: <http://purl.org/np/RAzquSkwsTAZm61nReG6MOjXEXUx8fNVfdWnAzyn6sOhU> .
...

sub:Assertion {
  sub:Interaction occurs-in: obo:ENVO_01000240 ;
    has-participant: sub:Organism_1 , sub:Organism_2 ;
    a obo:GO_0044419 ;
    prov:atTime "1962-12-01T00:00:00Z"^^xsd:dateTime .
  sub:Organism_1 eats: sub:Organism_2 ;
    rdfs:label "Picoides villosus" .
  sub:Organism_2 a itis:114936 ;
    rdfs:label "Ips" .
}

sub:Provenance {
  sub:Assertion prov:wasDerivedFrom sub:Study .
  sub:Study dcterms:bibliographicCitation "Otvos, I. S. and R. W. Stark. 1985. Arthropod food of
some forest-inhabiting birds. Canadian Entomologist 117:971-990." .
}

sub:Pubinfo {
  this: dcterms:license <https://creativecommons.org/licenses/by/4.0/> ;
    pav:createdBy <https://doi.org/10.5281/zenodo.1212599> ;
    prov:wasDerivedFrom <https://github.com/hurlbertlab/dietdatabase> .
  <https://github.com/hurlbertlab/dietdatabase> dcterms:bibliographicCitation "Allen Hurlbert.
2017. Avian Diet Database." .
}
```

Fig. 1. An example of a nanopublication on bird diets

main named graphs (the head graph is not shown), and the prefix of the first line shows the nanopublication's identifier (not shown are all other prefixes). The assertion states that there is an interaction of type "interspecies interaction between organisms" (`obo:GO_0044419`) with two participating organisms occurring in a place of "conifer woodland" (`obo:ENVO_01000240`) on a given day in 1962. We are further told that the interaction was in fact that one of the organisms ate the other, where the eating organism is described with the species label "Picoides villosus" and the eaten organism is classified under the genus with a certain taxonomic serial number (`itis:114936`) labeled as "Ips". If we are wondering about how we happen to know about this particular "interaction", we can look at the provenance graph, where we read that this assertion was derived from a study reported in a paper published in 1985. While the provenance part tells us where the knowledge encoded in the assertion came from, the publication info part tells us more about how the triples that make up this nanopublication were created. We see that it was created by a thing identified with a DOI (https://doi.org/10.5281/zenodo.1212599, which happens to be a software tool), derived from a GitHub repository (`hurlbertlab/dietdatabase`) attributed to Allen Hurlbert in 2017. We see that the recorded interaction, the recording of the interaction, and the publication in the current form as a nanopublication happened in very different contexts spread over many decades. It thereby also demonstrates the importance of differentiating between them in an unambiguous and systematic manner, which is exactly the strength and purpose of the nanopublication format.

II. Background

In previous work, we addressed the problems of the reliability and persistence of such nanopublications and of principled and efficient versioning for evolving datasets. Nanopublications were from the start defined as being immutable, but only with our work on trusty URIs [20], [21] we could provide technical guarantees on immutability with the use of cryptographic hashes. Nanopublications are thereby identified by URIs that contain a hash value that is calculated on their entire content, whereby even minimal changes can be detected and users can formally verify nanopublication data against their identifiers, to ensure the data perfectly corresponds to what they were looking for. An example of such a trusty URI is shown in the first line of Figure 1.

In order to allow users to reliably retrieve nanopublications given their URIs, we built upon this work to develop a decentralized publishing network [18], [19], currently consisting of fifteen server instances in nine countries[1]. We showed that nanopublication datasets can be efficiently and reliably archived and retrieved, without depending on the uptimes of individual servers. This server network is fully open, in the sense that everybody can publish nanopublications through it and even set up their own server to become a node in the network.

In order to account for the fact that we are often interested in entire datasets and not just individual data entries, and that such datasets change over time, we developed the concept of nanopublication indexes, which are themselves nanopublications. Such indexes point to other nanopublications, thereby defining sets of them of arbitrary size, which can also be used to define incremental dataset versions that reuse as much as possible from the previous versions [22]. We could demonstrate that this kind of dataset representation and versioning is indeed efficient and effective, and that the overheads implied by nanopublications are offset by the benefits of referencing and using specific well-defined subsets.

On a more practical level, we developed a Java nanopublication library [16] and a command-line utility tool called npop[2] to support the development of more high-level applications.

III. Related Work

Scientific data publishing has lately been a very popular topic and led to the proposal and adoption of the FAIR guiding principles for scientific data management and stewardship [34], mandating scientific data to be Findable, Accessible, Interoperable, and Reusable. Linked Data approaches directly tackle the interoperability challenge, and the Linked Open Data cloud [6] has come to contain large amounts of scientifically relevant structured data. However, we are still lacking accepted technologies to directly publish evolving Linked Data resources in a reliable manner, and much of the existing Linked Open Data is in fact extracted with custom scripts from different heterogeneous data formats, e.g. via the Bio2RDF initiative [12].

Apart from the nanopublication approach on which we focus in this paper, a number of other proposals address scientific data publishing. Solutions based on HTML for scholarly communication, such as RASH [11] and Dokieli [7], start from scientific articles and annotate them with Linked Data via HTML, RDFa, and related technologies. Other approaches

[1] http://purl.org/nanopub/monitor
[2] https://github.com/tkuhn/npop

like Research Objects [4] embrace a broader variety of types of digital resources, such as data tables, metadata, source code, presentation slides, log files, and workflow definitions, and provide a linked format for bundling them in well-defined packages. As a further notable approach, micropublications [10] combine elements of the approaches above, while putting an emphasis on the structure of scientific arguments and their relation to pieces of evidence. Nanopublications differ from these approaches in their exclusive focus on Linked Data (thereby not directly covering things like source code or diagrams) and in their emphasis of provenance at the most fine-grained level.

IV. Data

The data presented here is a wild collection of various datasets, consisting in total of 10 803 231 nanopublications. It contains new datasets extracted and modeled from existing resources, including nanopublication datasets on DrugBank (via Bio2RDF) and GloBI, and two smaller new datasets that were modeled as nanopublications from the start ("nano-born") on polycystic kidney disease and monogenic rare diseases, respectively. Furthermore, the collection also includes datasets derived from OpenBEL and the Human Protein Atlas, which are already a few years old but have until now not been properly introduced, as well as syntactically improved versions (including minting of trusty URIs) of previously introduced datasets on GeneRIF/AIDA and neXtProt. Moreover, the collection also contains new versions of the previously announced nanopublication datasets on DisGeNET (4 versions), WikiPathways (21 versions), and LIDDI (2 versions), the latter of which is also nano-born. Lastly, there is a small number of loose nanopublications that are not part of any dataset. We will now briefly describe all of these datasets.

The *DrugBank/Bio2RDF* dataset was extracted from Bio2RDF resources [12] originating from the DrugBank database [35], including drug–drug interactions, drug targets, and food interactions. These nanopublications were automatically extracted with a detailed meta-model to describe (in the publication info part) the processes, installations, versions, and codebases that produced each individual nanopublication[3].

Global Biotic Interactions (GloBI) is an initiative that aims to make existing species interaction datasets more easily accessible to help address the Eltonian shortfall in openly accessible biodiversity data. Instead of acting as a centralized data repository, GloBI provides tools to continuously discover, integrate and aggregate existing datasets across general purpose repositories like GitHub and Zenodo in addition to supporting various specialized infrastructures like iNaturalist[4], Web of Life[5], and the Interaction Bank of the Natural History Museum, London [2]. Resulting aggregate datasets can now be indexed, linked, transformed and made available as data archives, APIs and, more recently, as nanopublications covering a subset of the data. By providing tools to find and link datasets that describe how organisms interact (e.g., parasite-host, pollinator-plant, pathogen-host) with each other and across other biodiversity data platforms, GloBI helps to better integrate valuable knowledge acquired in recent and not-so recent past. Apart from the iNaturalist database mentioned above, the current nanopublication subset also includes datasets on turfgrass diseases[6], bees[7], interactions of ocean species[8], and the *Avian Diet Database*[9]. We have already seen an example of the latter in Figure 1. The Avian Diet Database is a compilation of quantitative diet data for bird species of North America extracted from the primary literature. Individual records describe a trophic link between a bird species and a prey item, and the strength of that link based on the fraction of the diet by weight, number of items, or occurrence. In addition, each trophic link is described by contextual information about the time and place of the diet study, sample size, and original citation.

A small dataset on a *meta-analysis of polycystic kidney disease expression profiles* was derived from sequenced male mouse RNA data [24]. The data is split into three phases of disease progression: early, moderate, and advanced. The mice of the different phase groups were sacrificed at different weeks after gene disruption and had different cystic phenotypes (mild, moderate, and severe). The results of this meta-analysis were then represented in 1657 nanopublications.

Another small dataset on *monogenic rare diseases and their genes* was recently published in the nanopublication format. Starting from diseases and their causative genes in the OMIM database [1], a subset of disease-gene associations were selected for monogenic (disease is caused by a mutation in one gene), rare (occuring in less than 1:2000) diseases. For each of them, the reference stating for the first time that this disease is causally associated with by this gene was manually curated. These associations are now exposed as nanopublications, following the DisGeNET model (see [30] and below). The nanopublications represents the diseases with their OMIM identifiers, genes with their HGNC symbols (provided by Wikidata [26], [29]), and the Ensembl gene identifier (which was added using BioMart [15]) for the gene, and the PubMed identifier for the provenance.

The *OpenBEL* dataset consists of nanopublications converted from the Biological Expression Language (BEL) format from resources provided by the OpenBEL initiative[10]. The conversion process maps BEL-specific vocabularies to existing accepted ontologies, as far as possible[11].

The *Human Protein Atlas* (HPA) [32] is a dataset that concentrates on the genome-wide analysis of human proteins. The nanopublications are a subset of the entire dataset, presently covering the immunohistochemistry results. They were built

[3] https://github.com/tkuhn/bio2rdf2nanopub

[4] https://inaturalist.org

[5] http://www.web-of-life.es/

[6] https://github.com/globalbioticinteractions/aps-turfgrasses

[7] https://github.com/globalbioticinteractions/Catalogue-of-Afrotropical-Bees

[8] https://github.com/globalbioticinteractions/raymond

[9] https://github.com/hurlbertlab/dietdatabase

[10] http://openbel.org/

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

on the neXtProt tissue expression nanopublication model (see below) with some minor modifications. For example, the Neuroscience Information Framework (NIF) Standard Ontology is used in the assertion graph to specify the quality of the immunohistochemical staining.

The *neXtProt* database is a high-quality corpus of data describing human proteins. Three categories of nanopublications had been derived and published from the a subset of the neXtProt data [9]. They detail information on the tissue expression of proteins, protein posttranslational modifications (PTM), and single amino acid polymorphisms, respectively. For the tissue expression nanopublications, the assertion graph contains the protein, tissue, and quality of the expression result, whereas the provenance graph contains the method of detection and an assessment of the evidence. The PTM nanopublication assertion delineates the specific modification in one specific protein isoform. The nanopublications for the amino acid polymorphisms describe the codon or codons variations that result in a protein change. Since their first publication, these neXtProt nanopublications have been syntactically improved by minting trusty URIs for them.

The *GeneRIF/AIDA* dataset contains sentences in a controlled natural language that were generated from input of the GeneRIF dataset on gene functions [17]. As with the neXtProt dataset, these nanopublications have been improved since their original publication by including trusty URIs.

Starting from version v2.1.0.0, *DisGeNET* has released its linked dataset on human gene-disease associations (GDAs) also in the form of nanopublications [30]. That first nanopublication version consisted of 940 034 nanopublications, representing the same number of scientific assertions for 381 056 different GDAs with their detailed provenance, levels of evidence and publication information descriptions. Since then, three more versions have been released: v3.0.0.0, v4.0.0.0, and v5.0.0.0. The assertion part of these nanopublications contains the description for a specific single GDA, and the provenance graph includes provenance, evidence and attribution statements that were directly mapped from the VoID description of DisGeNET's RDF representation. With each DisGeNET database release, the data collected from the existing data sources are updated and new data sources are added. BeFree is an important data source, providing a growing dataset generated by text-mining millions of biomedical abstracts from Medline. In the last version v5.0.0.0, there are 1 469 541 nanopublications for 561 119 GDAs between 17 074 genes and 20 370 diseases, disorders, traits, and clinical or abnormal human phenotypes.

WikiPathways is a database for biological pathways, including metabolic, signaling and genetic pathways [31]. Since 2016 WikiPathways publishes their monthly data releases also as in the nanopublication format, which has led to than twenty new versions since the initial release [22]. These nanopublications still describe only a subset of the knowledge in the database, focusing and exposing facts that are explicitly backed by literature, identified with PubMed identifiers. The assertions cover participation of genes, proteins, and metabolites in pathways, complexes, and interactions between these entities in pathways.

The *LIDDI* dataset on Linked Drug-Drug Interactions, finally, was described in an earlier publication [3], but a new version has since been released. In this updated version, the drug mappings have been cleaned and the drug-drug-adverse event triples properly linked. Multiple research groups have started using the dataset and some of their suggestions for improvements have been incorporated. For an upcoming more substantial release, coverage of more drugs and more adverse events is planned.

All these datasets and their versions are defined by nanopublication indexes, which link to the respective sets of nanopublications. As nanopublication indexes are nanopublications themselves, they can be published via the same server network and form part of the nanopublication collection as regular data entries. Tables III and IV at the end of this paper show all indexes that have been published until now, 129 in total, including some experimental unnamed indexes. *Date* denotes the creation date of the index, and *Sub* shows the number of sub-indexes. The WikiPathways and OpenBEL datasets make use of such sub-indexes to partition their data into several referenceable subsets. For example, index number 6 (first version of OpenBEL dataset) has indexes number 1 and 2 as sub-indexes and therefore contains the union of their nanopublications. Such subsets can also be defined post-hoc on and across existing datasets, for example index number 11 in Table III, which includes nanopublications from the OpenBEL and neXtProt datasets. The column *Size*, finally, shows the number of contained nanopublications (including sub-indexes).

V. Analysis

The unifying format of nanopublications now allows us to make an overarching analysis over all these heterogeneous datasets in a complete and uniform manner. The 10 803 231 nanopublications are in made up of 378 654 287 triples in total, 61 184 484 in the assertion graphs, 122 229 003 in the provenance graphs, and 136 738 995 in the publication info graphs. We see that these datasets are indeed "provenance-centric" with an average of 11.3 provenance triples per nanopublication. The average size of a nanopublication is 35.1 triples.

Next, we wanted to get an idea of who created these nanopublications. For that, we ran a SPARQL query to find all identifiers that were listed as creators or authors of a nanopublication in its publication info part. Specifically, we considered the predicates `dct:creator`, `dce:creator`, `pav:createdBy`, `pav:authoredBy`, and `prov:wasAttributedTo`. The summary of the result is shown in Table I. In total there are more than 47 million such creator mentions, i.e. an average of 4.4 creators per nanopublication. This set is however dominated by a much smaller number of "power-users" with just 41 unique identifiers. By far the most widely adopted identification scheme is ORCID (86% of all mentions), followed by a somewhat irregular use of literal strings (14%), and URIs identifying software tools (0.17%).

TABLE I
Creators and authors of nanopublications

type	total	unique	example (with max. frequency)
ORCID	40 964 679	26	http://orcid.org/0000-0003-0169-8159
Literal string	6 534 917	3	"CALIPHO project"
Tool URI	79 617	4	https://doi.org/10.5281/zenodo.1212599
Google Scholar URI	3	3	https://scholar.google.it/citations?user=9aI21r8AAAAJ&hl=en
ResearcherID	3	3	http://www.researcherid.com/rid/B-6035-2012
Other URI	16	2	http://sorry.vse.cz/~xhudj19
Total	47 579 235	41	

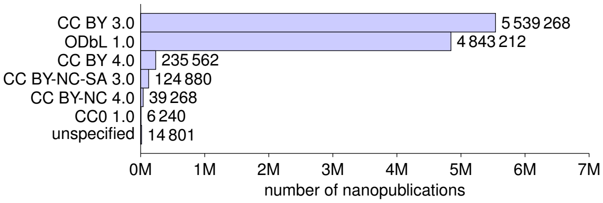


Fig. 2. License distribution

Next, we can turn to the topic of licensing. Figure 2 shows the distribution of licenses as reported in the nanopublications, considering `dct:license` and `dct:rights`. Apart from DisGeNET using the Open Database License (ODbL), Creative Commons (CC) licenses dominate the dataset, mostly in its CC BY flavor in version 3.0 and (to a much lesser extent) 4.0. This effect is mainly driven by a few large datasets that use CC BY 3.0, namely WikiPathways, neXtProt, and the Human Protein Atlas. (WikiPathways has just switched to CC0 1.0 for future versions.)

Next, we can look a bit more deeply to find out what these nanopublications are really about. We can do this by counting the used namespaces (i.e. the shared first part of URIs that group identifiers into vocabularies and collections) in the four different graphs (head, assertion, provenance, and publication info) and the three different triple positions (subject, predicate, and object). Table II shows the most relevant namespaces in the nanopublications datasets, according to their frequency of occurrence in the subject, predicate and object position for all head, assertion, provenance, and publication info graphs. The shown percentages denote the ratio of nanopublications where the given namespace appears at least once in the given position.

In the head graph, we always find predicates in the `rdf` and `np` namespaces, which is unsurprising since otherwise these would not be valid nanopublications. Subjects and objects in the head mostly denote the nanopublication itself and its graphs, respectively, and thereby tend to come from dataset-specific namespaces. In the assertion graph, `rdf` is the most frequent namespace in predicate position, mostly due to instantiations (`rdf:type`), but otherwise this category is — unsurprisingly — dominated by domain vocabularies like `sio`, `obo`, `oborel`, and `obox`. This is even more the case for the object position with `sio`, `ncbi` and `umls` being the most frequent namespaces, whereas the subject position is a mix of domain-specific and dataset-specific vocabularies. In the provenance graphs, we see the classical provenance and metadata vocabularies in predicate position, including `prov`, `rdf`, `rdfs`, `dc`, `pav`, and `owl`. It is notable that `prov` reaches near-universal acceptance by occurring in 98.95% of all nanopublications in that position. The Weighted Interests Vocabulary (`wi`) is also surprisingly popular, appearing in more than 93% of all nanopublications. This provenance information is supposed to be attached to the assertion graph, which explains the domain-specific namespaces in subject position. In object position we see a mix of domain-specific and dataset-specific namespaces, which can be explained by provenance pointing to external domain entities as well as internal activities or objects. The publication info graph is about the provenance of the nanopublication as a whole, and it is therefore not surprising to find domain-specific namespaces in subject position, and the classical provenance ontologies again in predicate positions, like `prov` and `pav`. DC Terms (`dc`) even achieves perfect adoption by appearing in 100% of nanopublications in a predicate of this graph. In the object position, we find near-universal adoption of `orcid` at 98.67%. Creative Commons license URIs (`cc`) and plain domain names of websites (like http://nextprot.org, which are summarized by the prefix `http://`) are popular as well. This kind of analysis provides us with interesting insights into the content of such a large number of nanopublications, but we also have to be aware that these numbers are often driven by a few large datasets.

Finally, we can investigate a bit more the variety of content found in these nanopublications. For this we filtered all triples from the assertion graphs that have `rdf:type` as predicate, thereby assigning an individual to a type. Overall, we found 50 384 007 such individual–type assignments, involving 14 941 unique types. The most frequent type, `eco:ECO_0000218` representing “manual assertion”, occurs 8 828 067 times, whereas on the other end of the scale many types such as `https://www.inaturalist.org/taxa/104422` (standing for Spotted Spreadwing, a species of damselflies) appear just once. Figure 3 visualizes the frequency distribution of these types, showing the continuous occurrence of classes along the whole size spectrum, which can be seen as an indication that the overall dataset is indeed varied and broad. The plateau in the middle is caused by the Human Protein Atlas via its reporting of the occurrence of proteins in a larger

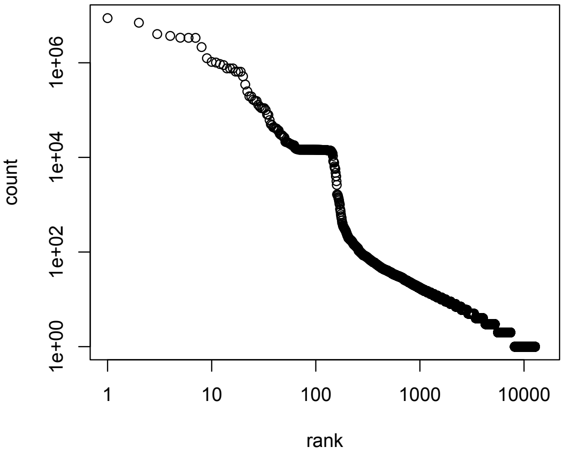


Fig. 3. Frequency distribution of types

number of different tissues, including the cases where no information about the occurrence is available. Thereby, each of the tissues produces a group of nanopublications of a size that matches the number of covered proteins (around 14 500), thereby producing this plateau effect in the diagram.

VI. Availability

All nanopublications of the data we presented here can be directly retrieved from the decentralized server network. Most of the nanopublication URIs directly resolve to a node of the network, and in case this does not work they can all be found on various servers in the network via a well-defined procedure [18]. To retrieve individual nanopublications or entire datasets in a more efficient way, the nanopublication library or its command-line interface can be used [16]. A nanopublication dump of the complete nanopublication content of the server network is furthermore made available on Zenodo[12].

In order to facilitate easier and more powerful access to nanopublications, we also developed a Linked Data API to access the full set of nanopublications available on the network. This API is powered by a `grlc` server [25] at the front and a GraphDB triple store instance with a SPARQL endpoint at the back.[13] This API offers a standard entry point to the data in the nanopublications network, which any Linked Data client can consume via HTTP without specific knowledge of SPARQL or RDF.[14] Concretely, our API provides the following methods:

- **find_latest_nanopubs_with_pattern**. Find all nanopublications for a given triple pattern, sorted by recency
- **find_nanopubs_with_pattern**. Find all nanopublications for a given triple pattern, in undefined order (faster than the method above)
- **find_latest_nanopubs_with_uri**. Find all nanopublications that contain the given URI, sorted by recency
- **find_nanopubs_with_uri**. Find all nanopublications that contain the given URI, in undefined order (faster than the method above)
- **get_all_indexes**. Get all nanopublication indexes (excluding incomplete ones); this gives a result similar to Tables III and IV
- **get_index_elements**. Get all nanopublications that the given index directly contains as elements
- **get_nanopub**. Get nanopublication by URI identifier (alternatively, this can also be achieved by directly calling the servers of the nanopublication network)

The GraphDB instance automatically loads nanopublications published to the network, normally within a few minutes or less.

Lastly, we also published the source code used for the shown analyses and processes on GitHub[15] and archived it on Zenodo[16].

VII. Discussion

The nanopublication approach can be seen as an instance of "containerization" analogous to the industrial containerization for efficient transport of goods in standardized, physical containers (and to the more recent containerization of software with solutions like Docker). Efficiency is improved with a standardized form and size of the containers, which allows for large-scale automatic and reliable processing of the content (data and physical goods, respectively). The diverse dataset presented here illustrates the recent uptake and expected benefits of this nanopublication-based containerization approach for scientific data publishing.

The current collection of nanopublications can be a valuable resource on the domain level, in particular for biomedical studies where reliability, reproducibility, and trust are important. Our dataset provides good coverage on the biomedical domains of genes, proteins, diseases, biological pathways, and drugs.

Furthermore, our data may also prove to be valuable to study the modeling and processes related to provenance in general and the PROV ontology in particular, with more than 122 million provenance triples, and with more than 10 million nanopublications using the PROV ontology.

Lastly, the dataset can also be useful on a lower technical level, as a dataset with a very large number of graphs. The LIDDI dataset has already been used for benchmarking named graph handling [13], highlighting LIDDI's "extremely large number of graphs, 392,340". The combined nanopublication dataset presented here is in this sense more than 100 times more extreme. LOD Laundromat [5], as another point of comparison, contains about 100 times more triples (currently more than 38 billion), but only about 1.5% the number of named graphs of our nanopublication dataset (658 045 documents with a named graph each in LOD Laundromat, compared to more than 43 million named graphs in our nanopublication dataset).

[12] https://doi.org/10.5281/zenodo.1213293

[13] The API is available at http://purl.org/nanopub/api

[14] The underlying parametrized queries can be found at https://github.com/peta-pico/nanopub-api/

[15] https://github.com/tkuhn/nanoresource

[16] https://doi.org/10.5281/zenodo.1213690

As future work, we have concrete plans to attract further datasets, including the entire Bio2RDF database and larger subsets of resources such as GloBI. We are also working to improve the API and to establish a whole ecosystem of decentralized services feeding from the data backbone of the nanopublication network. In order make the execution of API calls more efficient and scalable, we will also look into the techniques of HDT-Quads [13] and the quad version of Triple Pattern Fragments [33], as soon as these technologies become stable enough for this kind of application.

References

[1] Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., Hamosh, A.: OMIM.org: Online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic acids research **43**(D1), D789–D798 (2014)

[2] Baker, E., Kitching, I.J., Beccaloni, G.W., Whitaker, A., Dupont, S., Smith, V.S., Noyes, J.S.: NHM interactions bank (2016), https://doi.org/10.5519/0060767

[3] Banda, J.M., Kuhn, T., Shah, N.H., Dumontier, M.: Provenance-centered dataset of drug-drug interactions. In: International Semantic Web Conference. pp. 293–300. Springer (2015)

[4] Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., et al.: Why linked data is not enough for scientists. Future Generation Computer Systems **29**(2), 599–611 (2013). https://doi.org/10.1016/j.future.2011.08.004

[5] Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: Lod laundromat: a uniform way of publishing other peoples dirty data. In: International Semantic Web Conference. pp. 213–228. Springer (2014)

[6] Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. International journal on semantic web and information systems **5**(3), 1–22 (2009)

[7] Capadisli, S., Guy, A., Verborgh, R., Lange, C., Auer, S., Berners-Lee, T.: Decentralised authoring, annotations and notifications for a read-write web with dokieli. In: International Conference on Web Engineering. pp. 469–481. Springer (2017). https://doi.org/10.1007/978-3-319-60131-1_33

[8] Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. In: Proceedings of the 14th international conference on World Wide Web. pp. 613–622. ACM (2005)

[9] Chichester, C., Karch, O., Gaudet, P., Lane, L., Mons, B., Bairoch, A.: Converting neXtProt into Linked Data and nanopublications. Semantic Web **6**(2), 147–153 (2015)

[10] Clark, T., Ciccarese, P., Goble, C.: Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. Journal of Biomedical Semantics **5**(1), 28 (2014). https://doi.org/10.1186/2041-1480-5-28

[11] Di Iorio, A., Nuzzolese, A.G., Osborne, F., Peroni, S., Poggi, F., Smith, M., Vitali, F., Zhao, J.: The RASH framework: enabling HTML + RDF submissions in scholarly venues. In: In Proceedings of the ISWC 2015 Posters & Demonstrations Track (2015)

[12] Dumontier, M., Callahan, A., Cruz-Toledo, J., Ansell, P., Emonet, V., Belleau, F., Droit, A.: Bio2RDF release 3: a larger connected network of linked data for the life sciences. In: Proceedings of the 2014 International Conference on Posters & Demonstrations Track. vol. 1272, pp. 401–404 (2014)

[13] Fernández, J.D., Martínez-Prieto, M.A., Polleres, A., Reindorf, J.: HDTQ: managing RDF datasets in compressed space. In: European Semantic Web Conference. pp. 191–208. Springer (2018)

[14] Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services & Use **30**(1-2), 51–56 (2010)

[15] Kinsella, R.J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al.: Ensembl biomarts: a hub for data retrieval across taxonomic space. Database **2011** (2011)

[16] Kuhn, T.: nanopub-java: A java library for nanopublications. In: Proceedings of the 5th Workshop on Linked Science (LISC 2015) (2015)

[17] Kuhn, T., Barbano, P.E., Nagy, M.L., Krauthammer, M.: Broadening the scope of nanopublications. In: Extended Semantic Web Conference. pp. 487–501. Springer (2013)

[18] Kuhn, T., Chichester, C., Krauthammer, M., Dumontier, M.: Publishing without publishers: a decentralized approach to dissemination, retrieval, and archiving of data. In: International Semantic Web Conference. pp. 656–672. Springer (2015)

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

[20] Kuhn, T., Dumontier, M.: Trusty URIs: Verifiable, immutable, and permanent digital artifacts for linked data. In: European Semantic Web Conference. pp. 395–410. Springer (2014)

[21] Kuhn, T., Dumontier, M.: Making digital artifacts on the web verifiable and reliable. IEEE Transactions on Knowledge and Data Engineering **27**(9), 2390–2400 (2015)

[22] Kuhn, T., Willighagen, E., Evelo, C., Queralt-Rosinach, N., Centeno, E., Furlong, L.I.: Reliable granular references to changing linked data. In: International Semantic Web Conference. pp. 436–451. Springer (2017)

[23] Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: PROV-O: The PROV ontology. W3C recommendation **30** (2013)

[24] Malas, T.B., Formica, C., Leonhard, W.N., Rao, P., Granchi, Z., Roos, M., Peters, D.J., 't Hoen, P.A.: Meta-analysis of polycystic kidney disease expression profiles defines strong involvement of injury repair processes. American Journal of Physiology-Renal Physiology **312**(4), F806–F817 (2017)

[25] Meroño-Peñuela, A., Hoekstra, R.: Automatic Query-centric API for Routine Access to Linked Data. In: The Semantic Web ISWC 2017, 16th International Semantic Web Conference. vol. 10587, pp. 334–339. Springer LNCS (2017)

[26] Mietchen, D., Hagedorn, G., Willighagen, E., Rico, M., Gmez-Prez, A., Aibar, E., Rafes, K., Germain, C., Dunning, A., Pintscher, L., et al.: Enabling open science: Wikidata for research (wiki4r). Research Ideas and Outcomes **1**, e7573 (Dec 2015)

[27] Mons, B., van Haagen, H., Chichester, C., den Dunnen, J.T., van Ommen, G., van Mulligen, E., Singh, B., Hooft, R., Roos, M., Hammond, J., et al.: The value of data. Nature genetics **43**(4), 281–283 (2011)

[28] Moreau, L.: Provenance-based reproducibility in the semantic web. Web semantics: science, services and agents on the World Wide Web **9**(2), 202–221 (2011)

[29] Putman, T.E., Lelong, S., Burgstaller-Muehlbacher, S., Waagmeester, A., Diesh, C., Dunn, N., Munoz-Torres, M., Stupp, G.S., Wu, C., Su, A.I., et al.: Wikigenomes: an open web application for community consumption and curation of gene annotation data in wikidata. Database **2017**(1) (Jan 2017)

[30] Queralt-Rosinach, N., Kuhn, T., Chichester, C., Dumontier, M., Sanz, F., Furlong, L.I.: Publishing DisGeNET as nanopublications. Semantic Web **7**(5), 519–528 (2016)

[31] Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., et al.: WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic acids research **46**(D1), D661–D667 (2017)

[32] Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al.: Towards a knowledge-based human protein atlas. Nature biotechnology **28**(12), 1248 (2010)

[33] Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., Haesendonck, G., Colpaert, P.: Triple pattern fragments: a low-cost knowledge graph interface for the web. Web Semantics: Science, Services and Agents on the World Wide Web **37**, 184–206 (2016)

[34] Wilkinson, M.D., Dumontier, M., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific data **3** (2016)

[35] Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research **36**(suppl_1), D901–D906 (2007)

TABLE II

Namespaces and their frequency in the different positions in nanopublications. The full URIs for the shown prefixes are shown at the bottom.

Graph	subj		pred		obj	
head	nxpt	37.27%	rdf	100.00%	np	100.00%
	disgen	36.13%	np	100.00%	nxpt	37.27%
	panp	11.61%	rdfs	11.61%	disgen	22.53%
	disgent	8.70%			disgent	8.70%
	liddi-r	1.82%				
	klab	1.44%				
	nnp	1.41%				
	bel2-r	1.16%				
	dbv	0.25%				
	dbr	0.11%				
assertion	gda-r	36.13%	rdf	69.61%	sio	44.87%
	ncbi	31.27%	sio	52.43%	ncbi	44.85%
	umls	31.22%	obo	36.24%	umls	44.83%
	nxpts	30.24%	oborel	30.24%	evs	31.27%
	pa	11.61%	obox	13.50%	caloha	30.24%
	gda	8.70%	nif	11.61%	pa	11.61%
	nxpt	7.02%	rdfs	11.15%	caloha-l	11.61%
	liddi-r	1.82%	dc	2.44%	nxpt	7.06%
	liddi	1.81%	dbv	2.17%	nxpts	7.02%
	nxptt	1.02%	npx	1.58%	rsp	7.02%
provenance	nxpt	37.27%	prov	98.95%	eco	45.97%
	nxptq	36.25%	rdf	95.66%	pubmed-id	44.09%
	disgen	22.53%	wi	93.71%	nxpt	37.27%
	disgen-void5	13.60%	rdfs	83.25%	nxptq	37.27%
	disgen-void4	13.10%	dc	48.38%	obox	36.13%
	panp	11.61%	pprov	47.86%	efo	21.49%
	disgen-void3	9.43%	sio	47.55%	bao	14.76%
	bao	8.75%	pav	37.36%	disgen-void5	13.60%
	disgen-void2	8.70%	owl	8.75%	disgen-void4	13.10%
	disgent	8.70%	pav2	8.70%	pas	11.61%
pubinfo	nxpt	37.27%	dc	100.00%	orcid	98.67%
	disgen	36.13%	pprov	93.71%	cc	50.69%
	disgen-void5	13.60%	swan	49.61%	http://	48.93%
	disgen-void4	13.10%	prov	40.81%	sio	44.83%
	panp	11.61%	pav	39.28%	odbl	44.83%
	disgen-void3	9.43%	rdf	38.02%	nxpt	37.27%
	disgen-void2	8.70%	pav2	8.70%	eco	37.27%
	disgent	8.70%	rdfs	1.54%	disgen-void5	13.60%
	liddi-r	1.82%	npx	0.56%	disgen-void4	13.10%
	klab	1.44%	dce	0.13%	rid	11.61%

Prefix	Namespace
bel2-r	http://www.tkuhn.ch/bel2nanopub/
bao	http://www.bioassayontology.org/bao#
caloha	ftp://ftp.nextprot.org/pub/current_release/controlled_vocabularies/caloha.obo#
caloha-l	http://purl.obolibrary.org/obo/caloha.obo#
cc	http://creativecommons.org/licenses/by/
dbv	http://bio2rdf.org/drugbank_vocabulary
dbr	http://bio2rdf.org/drugbank:
dc	http://purl.org/dc/terms/
dce	http://purl.org/dc/elements/1.1/
disgen	http://rdf.disgenet.org/resource/nanopub/
disgen-void2	http://rdf.disgenet.org/v2.1.0/void.ttl#
disgen-void3	http://rdf.disgenet.org/v3.0.0/void/
disgen-void4	http://rdf.disgenet.org/v4.0.0/void/
disgen-void5	http://rdf.disgenet.org/v5.0.0/void/
disgent	http://rdf.disgenet.org/nanopublications.trig#
eco	http://purl.obolibrary.org/obo/eco.owl#
efo	http://www.ebi.ac.uk/efo/
evs	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#
gda	http://rdf.disgenet.org/gene-disease-association.ttl#
gda-r	http://rdf.disgenet.org/resource/gda/
obo	http://purl.obolibrary.org/obo/#
oborel	http://purl.org/obo/owl/OBO_REL#
obox	http://purl.obolibrary.org/obo/
klab	http://krauthammerlab.med.yale.edu/nanopub/
liddi	http://liddi.stanford.edu/LIDDI
liddi-r	http://liddi.stanford.edu/LIDDI_resource:
ncbi	http://identifiers.org/ncbigene/

Prefix	Namespace
nif	http://ontology.neuinfo.org/NIF/Backend/NIF-Quality.owl#
np	http://www.nanopub.org/nschema#
nnp	http://purl.org/np/
npx	http://purl.org/nanopub/x/
nxpt	http://www.nextprot.org/nanopubs#
nxptq	http://www.nextprot.org/help/quality_criteria/
nxpts	http://www.nextprot.org/db/search#
nxptt	http://www.nextprot.org/db/term/
odbl	http://opendatacommons.org/licenses/odbl/
orcid	http://orcid.org/
owl	http://www.w3.org/2002/07/owl#
pa	http://www.proteinatlas.org/
pas	http://www.proteinatlas.org/search/
pav	http://purl.org/pav/
pav2	http://purl.org/pav/2.0/
panp	http://www.proteinatlas.org/about/nanopubs/
prov	http://www.w3.org/ns/prov#
pprov	http://purl.org/net/provenance/ns#
pubmed-id	http://identifiers.org/pubmed/
rid	http://www.researcherid.com/rid/
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
rsp	http://sadiframework.org/ontologies/GMOD/RangedSequencePosition.owl#
sio	http://semanticscience.org/resource/
umls	http://linkedlifedata.com/resource/umls/id/
swan	http://swan.mindinformatics.org/ontologies/1.2/pav/
wi	http://purl.org/ontology/wi/core#

TABLE III

Overview of nanopublication indexes (continued in Table IV)

#	Index Title (some abbreviated)	Date	Sub	Size
1	OpenBEL's Small Corpus 1.0	2015-03-02		2033
2	OpenBEL's Large Corpus 1.0	2015-03-02		48674
3	OpenBEL's Small Corpus 20131211	2015-03-02		1288
4	OpenBEL's Large Corpus 20131211	2015-03-02		72885
5	AIDA GeneRIF	2015-03-04		156026
6	OpenBEL's Small and Large Corpus 1.0	2015-03-04	2	50707
7	OpenBEL's Small and Large Corpus 20131211	2015-03-04	2	74173
8	DisGeNET v2.1.0.0	2015-03-04		940034
9	neXtProt protein data (preliminary)	2015-03-09		4025981
10	*(unnamed)*	2015-04-09		3
11	Data about CDKN2A from BEL2nanopub & neXtProt	2015-04-14		5
12	Linked Drug-Drug Interactions (LIDDI) dataset v1.01	2015-07-17		98085
13	*(unnamed)*	2015-08-18		3
14	DisGeNET v3.0.0.0	2015-11-15		1018735
15	*(unnamed)*	2016-01-09		2
16	Regional goverments of Czech Republic	2016-01-09		13
17	DisGeNET v4.0.0.0 (Nanopub Index)	2016-05-13		1414902
18	Rett Syndrome data from DisGeNET v3.0.0.0	2016-11-01		1284
19	Rett Syndrome data from DisGeNET v4.0.0	2016-11-02		968
20	*(unnamed)*	2016-11-02		1642
21	*(unnamed)*	2016-11-02		2566
22	*(unnamed)*	2016-11-02		692
23	*(unnamed)*	2016-11-02		233
24	Human Protein Atlas data	2016-11-02		1254468
25	My example dataset	2016-11-28		3
26	*(unnamed)*	2017-05-04		1512
27	DisGeNET v2.1.0.0, incremental dataset	2017-05-09		940034
28	DisGeNET v3.0.0.0, incremental dataset	2017-05-09		1018735
29	DisGeNET v4.0.0.0, incremental dataset	2017-05-09		1414902
30	complexes extracted from WikiPathways version 20160610	2017-05-11		41
31	complexes extracted from WikiPathways version 20160710	2017-05-11		37
32	complexes extracted from WikiPathways version 20160810	2017-05-11		37
33	complexes extracted from WikiPathways version 20160910	2017-05-11		37
34	complexes extracted from WikiPathways version 20161010	2017-05-11		37
35	complexes extracted from WikiPathways version 20161110	2017-05-11		36
36	complexes extracted from WikiPathways version 20161210	2017-05-11		37
37	complexes extracted from WikiPathways version 20170210	2017-05-11		42
38	complexes extracted from WikiPathways version 20170310	2017-05-11		42
39	complexes extracted from WikiPathways version 20170410	2017-05-11		42
40	complexes extracted from WikiPathways version 20170510	2017-05-11		42
41	interactions extracted from WikiPathways version 20160610	2017-05-11		8977
42	interactions extracted from WikiPathways version 20160710	2017-05-11		10136
43	interactions extracted from WikiPathways version 20160810	2017-05-11		10086
44	interactions extracted from WikiPathways version 20160910	2017-05-11		10087
45	interactions extracted from WikiPathways version 20161010	2017-05-11		10090
46	interactions extracted from WikiPathways version 20161110	2017-05-11		10092
47	interactions extracted from WikiPathways version 20161210	2017-05-11		10100
48	interactions extracted from WikiPathways version 20170210	2017-05-11		10375
49	interactions extracted from WikiPathways version 20170310	2017-05-11		10371
50	interactions extracted from WikiPathways version 20170410	2017-05-11		10371
51	interactions extracted from WikiPathways version 20170510	2017-05-11		10371
52	pathwayParticipation extracted from WikiPathways version 20161110	2017-05-11		3830
53	pathwayParticipation extracted from WikiPathways version 20161210	2017-05-11		3838
54	pathwayParticipation extracted from WikiPathways version 20170210	2017-05-11		3906
55	pathwayParticipation extracted from WikiPathways version 20170310	2017-05-11		3906
56	pathwayParticipation extracted from WikiPathways version 20170410	2017-05-11		3910
57	pathwayParticipation extracted from WikiPathways version 20170510	2017-05-11		3910
58	WikiPathways, incremental dataset, 20160610	2017-05-11	2	9018
59	WikiPathways, incremental dataset, 20160710	2017-05-11	2	10173
60	WikiPathways, incremental dataset, 20160810	2017-05-11	2	10123
61	WikiPathways, incremental dataset, 20160910	2017-05-11	2	10124
62	WikiPathways, incremental dataset, 20161010	2017-05-11	2	10127
63	WikiPathways, incremental dataset, 20161110	2017-05-11	3	13958
64	WikiPathways, incremental dataset, 20161210	2017-05-11	3	13975
65	WikiPathways, incremental dataset, 20170210	2017-05-11	3	14323

TABLE IV
OVERVIEW OF NANOPUBLICATION INDEXES (CONTINUED FROM TABLE III)

#	Index Title (some abbreviated)	Date	Sub	Size
66	WikiPathways, incremental dataset, 20170310	2017-05-11	3	14319
67	WikiPathways, incremental dataset, 20170410	2017-05-11	3	14323
68	WikiPathways, incremental dataset, 20170510	2017-05-11	3	14323
69	*(unnamed)*	2017-05-11		83771
70	*(unnamed)*	2017-05-11		4859
71	*(unnamed)*	2017-05-11		18098
72	Data for meta-analysis of polycystic kidney disease expression profiles	2017-05-18		1657
73	DisGeNET v5.0.0.0	2017-10-17		1469541
74	complexes extracted from WikiPathways version 20170610	2017-12-18		42
75	interactions extracted from WikiPathways version 20170610	2017-12-18		10371
76	pathwayParticipation extracted from WikiPathways version 20170610	2017-12-18		3910
77	WikiPathways, incremental dataset, 20170610	2017-12-18	3	14323
78	complexes extracted from WikiPathways version 20170710	2017-12-18		42
79	interactions extracted from WikiPathways version 20170710	2017-12-18		10371
80	pathwayParticipation extracted from WikiPathways version 20170710	2017-12-18		3910
81	WikiPathways, incremental dataset, 20170710	2017-12-18	3	14323
82	complexes extracted from WikiPathways version 20170810	2017-12-18		43
83	interactions extracted from WikiPathways version 20170810	2017-12-18		13208
84	pathwayParticipation extracted from WikiPathways version 20170810	2017-12-18		4362
85	WikiPathways, incremental dataset, 20170810	2017-12-18	3	17613
86	complexes extracted from WikiPathways version 20170910	2017-12-18		43
87	interactions extracted from WikiPathways version 20170910	2017-12-18		13080
88	pathwayParticipation extracted from WikiPathways version 20170910	2017-12-18		4364
89	WikiPathways, incremental dataset, 20170910	2017-12-18	3	17487
90	complexes extracted from WikiPathways version 20171010	2017-12-18		43
91	interactions extracted from WikiPathways version 20171010	2017-12-18		12620
92	pathwayParticipation extracted from WikiPathways version 20171010	2017-12-18		4386
93	WikiPathways, incremental dataset, 20171010	2017-12-18	3	17049
94	complexes extracted from WikiPathways version 20171116	2017-12-18		44
95	interactions extracted from WikiPathways version 20171116	2017-12-18		12632
96	pathwayParticipation extracted from WikiPathways version 20171116	2017-12-18		4403
97	WikiPathways, incremental dataset, 20171116	2017-12-18	3	17079
98	complexes extracted from WikiPathways version 20171210	2017-12-18		44
99	interactions extracted from WikiPathways version 20171210	2017-12-18		12638
100	pathwayParticipation extracted from WikiPathways version 20171210	2017-12-18		4407
101	WikiPathways, incremental dataset, 20171210	2017-12-18	3	17089
102	complexes extracted from WikiPathways version 20180110	2018-03-30		44
103	interactions extracted from WikiPathways version 20180110	2018-03-30		12645
104	pathwayParticipation extracted from WikiPathways version 20180110	2018-03-30		4408
105	WikiPathways, incremental dataset, 20180110	2018-03-30	3	17097
106	complexes extracted from WikiPathways version 20180210	2018-03-30		45
107	interactions extracted from WikiPathways version 20180210	2018-03-30		12683
108	pathwayParticipation extracted from WikiPathways version 20180210	2018-03-30		4413
109	WikiPathways, incremental dataset, 20180210	2018-03-30	3	17141
110	complexes extracted from WikiPathways version 20180310	2018-03-30		45
111	interactions extracted from WikiPathways version 20180310	2018-03-30		12685
112	pathwayParticipation extracted from WikiPathways version 20180310	2018-03-30		4417
113	WikiPathways, incremental dataset, 20180310	2018-03-30	3	17147
114	Core drug data extracted from Drugbank via Bio2RDF	2018-03-30		7740
115	Drug-drug interaction data extracted from Drugbank via Bio2RDF	2018-03-30		12104
116	Food interaction data for drugs extracted from Drugbank via Bio2RDF	2018-03-30		310
117	Drug target data extracted from Drugbank via Bio2RDF	2018-03-30		4007
118	Drug target relations extracted from Drugbank via Bio2RDF	2018-03-30		15107
119	Drug data extracted from Drugbank via Bio2RDF	2018-03-30	5	39268
120	Linked Drug-Drug Interactions (LIDDI) dataset v1.02	2018-03-30		98085
121	Dataset of loose nanopublications	2018-04-04		42
122	Dataset of loose example nanopublications	2018-04-04		183
123	Nanopublications for the Avian Diet Database	2018-04-05		25962
124	Nanopublications for Turfgrass Diseases	2018-04-05		3430
125	Nanopublications for iNaturalist.org Species Interaction Observations	2018-04-05		29882
126	Nanopublications for Catalogue of Afrotropical Bees	2018-04-05		5267
127	Nanopublications for Species Interaction Dataset by Raymond et al.	2018-04-05		14961
128	Nanopublication Collection of Global Biotic Interactions, Version 1	2018-04-05	5	79502
129	Monogenic rare diseases: gene associations with specific literature references	2018-04-05		4583