Resource Description Framework (RDF) Modeling of Named Entity Co-occurrences Derived from Biomedical Literature in the PubChemRDF

Qingliang Li¹, Sunghwan Kim¹, Leonid Zaslavsky¹, Tiejun Cheng¹, Bo Yu¹ and Evan E. Bolton¹

¹ National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Abstract

Named entities, such as chemicals/drugs, diseases, and genes/proteins, and their associations are not only important components of biomedical literature, but also the foundation of creating biomedical knowledge bases/graphs. This work details the expression of co-occurrence associations among chemicals, genes, proteins, and diseases in the Resource Description Framework (RDF) format within the PubChemRDF resource, which is freely accessible and publicly available. The co-occurrence model is populated into a triplestore with named entities and their associations that are derived from text mining of about 35 million biomedical references in PubMed. Use cases are provided to demonstrate the utility of the model. Together with meta-data modeling of the references including the information about the author, journal, grant, and funding agency, this data model can address pertinent biomedical questions through SPARQL queries and help exploit bio-medical knowledge in various user perspectives and use cases.

Keywords

PubChem, RDF, Semantic Web, Linked Data, Triplestore, SPARQL, Modeling, Drug, Chemical, Disease, Gene, Protein, PubChemRDF, Co-Occurrence.

1. Introduction

PubChem (<u>https://pubchem.ncbi.nlm.nih.gov</u>) [1–3] is an open chemical database at the National Center for Biotechnology Information (NCBI), the National Library of Medicine (NLM), the U.S. National Institutes of Health (NIH). Among many tools and services provided by PubChem are the literature knowledge panels [4], which assist users in quickly finding important relationships between chemicals, genes, proteins, and diseases. The literature knowledge panels for a given entity (e.g., a chemical, gene, or protein) display a selected set of "co-occurrence neighbors", which are defined as any chemicals, genes, proteins, and diseases mentioned together with that entity in the biomedical literature. In addition, the literature knowledge panel provides a sample of PubMed records that comention the entity and those selected co-occurrence neighbors. The list of the co-occurrence neighbors and relevant PubMed records can be downloaded for further analysis through the download button in the panel. Note that the use of the term "co-occurrence neighbors" avoids confusion with the existing 2-dimensional (2-D) and 3-dimensional (3-D) chemical structure-based neighbor relationships [5–7].

ORCID: 0000-0002-6453-236X (Q. Li); 0000-0001-9828-2074 (S. Kim); 0000-0001-5873-4873 (L. Zaslavsky); 0000-0002-4486-3356 (T. Cheng); 0000-0003-3952-8921 (B. Yu); 0000-0002-5959-6190 (E. E. Bolton)



^{© 2023} Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The 14th International Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS) Conference, February 13–16, 2023, Basel, Switzerland

EMAIL: qingliang.li@nih.gov (Q. Li); kimsungh@ncbi.nlm.nih.gov (S. Kim); zaslavsk@ncbi.nlm.nih.gov (L. Zaslavsky); chengt2@ncbi.nlm.nih.gov (T. Cheng) ; bo.yu2@nih.gov (B. Yu); bolton@ncbi.nlm.nih.gov (E. E. Bolton)

CEUR Workshop Proceedings (CEUR-WS.org)

The underlying data for the knowledge panel is derived from text mining the 35 million biomedical references in PubMed, using the LeadMine software [8]. Chemicals, genes, proteins, and diseases are extracted from the titles and abstracts of the PubMed records and the most relevant co-occurrence neighbors are identified using statistical analysis and relevance-based sampling. A detailed explanation of the method used to develop the knowledge panel is given in our previous paper [4].

The present paper describes a data model that expresses the named entities and their co-occurrence associations in the Resource Description Framework (RDF) format [9], along with the meta-data modeling of the reference information (including the author, journal, grant, and funding agency). The data model augments the existing RDF-formatted data (also known as PubChemRDF) [10] and helps find answers to biomedical questions (as demonstrated in a recent study [11]) through SPARQL Protocol and RDF Query Language queries [12].

2. Model design

The purpose of the model design is (1) to semantically describe the co-occurrence relations of the named entities (i.e., chemicals, genes/proteins, and diseases), which are derived from the biomedical literature, and (2) to take advantage of the existing PubChemRDF graph to address sophisticated biomedical questions in a semantic way through RDF. In this model, the term "compound" is used to indicate a chemical named entity to keep consistency in the naming convention within PubChemRDF. Because gene and protein names are often used interchangeably in literature, the data model designed in this study does not distinguish genes and proteins and, instead, treats all of them as genes, unless explicitly specified.

Figure 1 shows the design of the co-occurrence RDF model. There are five types of nodes in this model:

- Green ovals corresponding to named entities (i.e., compounds, genes, and diseases)
- Blue ovals representing references and their metadata (e.g., author, journal, grant, funding agency)
- Purple ovals for the co-occurrence of named entities (e.g., compound-compound, compound-gene, and compound-disease associations)
- White rectangles describing the characteristic of an association (e.g., text mining and type of associations) used for linking to external ontologies.
- White hexagons representing the co-occurrence score of an association (see below).

The relationships between nodes are defined using the PubChem-specific vocabulary as well as several external ontologies: the Semanticscience Integrated Ontology (SIO) [13], Funding, Research Administration and Projects Ontology (FRAPO) [14], Dublin Core Metadata Initiative (DCMI) metadata terms [15], Publishing Requirements for Industry Standard Metadata (PRISM) vocabularies [16] and EMBRACE Data And Methods (EDAM) ontology [17].

Named entities identified from text mining of 35 million PubMed records in our previous study [4] are matched to compounds, genes/proteins, and diseases in PubChemRDF. The associations between the PubMed records and the named entities found in them are linked with a predicate, pcvocab:discussesAsDerivedByTextMining.

When two named entities are identified in a PubMed record, a co-occurrence node is created and linked to the corresponding named entity nodes, with the rdf:subject and rdf:object predicates. The co-occurrence node is also linked to the nodes specifying the nature of the co-occurrence (e.g., the types of co-mentioned named entities and whether they are text-mined or not).

Some entities have thousands of co-occurrence neighbors. For example, the term "neoplasms" cooccurs with over 47,000 compounds, 28,000 genes, and 6,000 diseases. The importance of individual entity relationships is quantified using the co-occurrence score (S_{CO}), as explained in our previous study [4]. It can be considered as a variant of the term frequency-inverse document frequency (TF-IDF) score [18–21]. The co-occurrence score values are used to identify up to 1,000 co-occurrence neighbors of each neighbor type (i.e., compounds, genes, and diseases) for a given entity, and the corresponding cooccurrence nodes are created in the model.



Figure 1: Diagram of RDF co-occurrence modeling of compounds, genes, and diseases, derived from biomedical literature. Blue ovals correspond to biomedical literature (references) and its metadata. Green ovals represent named entities (i.e., compounds, genes, and diseases). Purple ovals indicate the co-occurrence associations between named entities (e.g., compound-disease association, disease-gene association). White rectangles specify the types of individual associations in terms of external ontologies. White hexagons indicate the co-occurrence scores.

It is important to note that the co-occurrence neighboring relationships between entities (e.g., compound-disease) are not symmetrical due to the asymmetry of the co-occurrence scores and the truncation of the neighbor lists. For example, when a disease is one of the top 1,000 co-occurrence neighbors of a chemical, there is no guarantee that the chemical is among the top 1,000 co-occurrence neighbors of the disease. This asymmetry is reflected in the directed graph in the RDF by means of designating the subject and the object.

Literature data is modeled in reference nodes, which are linked to the nodes representing their metadata (e.g., journal, author, publication date, grant number, funding agency, and MeSH terms [22]), using the relevant terms from FRAPO [14], DCMI [15], and PRISM [16] as predicates.

3. Applications

In this section, we present three use cases of the co-occurrence RDF resource. Here, these use cases assume that the co-occurrence RDF resource and PubChemRDF data are loaded into the triplestore of Virtuoso [23]. It is possible to load these data into other triplestores or RDF-aware graph databases, such as Apache Jena.

3.1. Use Case 1: Diseases Co-occurring with a Chemical

The simplest use case of the co-occurrence RDF is to retrieve named entities commonly mentioned with a query entity in PubMed articles (e.g., diseases or genes/proteins co-occurring with a chemical). As an example, Figure 2 shows the SPARQL query that retrieves the top 25 diseases mentioned together with indomethacin (CID 3715) in terms of their co-occurrence scores [4]. The diseases returned from the query are listed in Table 1, along with their co-occurrence scores and preferred disease names. The preferred disease name was retrieved from the PubChemRDF disease subdomain (with the term prefLabel in the Simple Knowledge Organization System (SKOS) [24] as a predicate).

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX compound: <http://rdf.ncbi.nlm.nih.gov/pubchem/compound/>
SELECT ?disease ?score ?disease_prefLabel
    FROM <http://rdf.ncbi.nlm.nih.gov/pubchem/cooccurrence>
    FROM <http://rdf.ncbi.nlm.nih.gov/pubchem/disease>
WHERE {
        ?cooccurrence rdf:subject compound:CID3715 .
        ?cooccurrence rdf:object ?disease .
        ?cooccurrence rdf:type sio:SIO_000993 .
        ?cooccurrence sio:SIO_000300 ?score .
        ?disease skos:prefLabel ?disease_prefLabel .
}
ORDER BY DESC(?score)
LIMIT 25
```

Figure 2: SPARQL query to retrieve the top 25 diseases commonly mentioned with indomethacin (CID 3715) in literature. The SIO terms SIO_000993 and SIO_000300 mean "chemical-disease association" and "has a value", respectively.

Table 1

S _{CO} ^a	Preferred disease name
58416	Inflammation
56017	Ulcer
51446	Edema
42250	Ductus Arteriosus, Patent
36563	Stomach Ulcer
34696	Pain
34306	Premature Birth
28989	Hypertension
22891	Arthritis
21757	Drug-Related Side Effects and Adverse Reactions
21500	Arthritis, Rheumatoid
20986	Infections
20916	Neoplasms
20628	Blood Platelet Disorders
20014	Hypotension
18336	Depressive Disorder
18081	Hemorrhage
17384	Arthritis, Experimental
16928	Headache
16868	Rheumatic Diseases
15627	Нурохіа
14744	Ischemia
14733	Fever
12385	Osteoarthritis
12070	Bartter Syndrome

Top 25 diseases most mentioned with indomethacin (CID 3715), along with their co-occurrence scores (S_{CO}), returned from the SPARQL query in Figure 2.

^{*a*} Derived from the values computed from Formula (3) in Reference [4] by multiplying by 100 and rounding to the nearest integer.

The most commonly occurring disease with indomethacin in literature is "inflammation", followed by "ulcer". It reflects the fact that indomethacin is a non-steroidal anti-inflammatory drug (NSAID) and that NSAIDs' common side effects include stomach ulcer. In essence, this query performs the same data retrieval task used to create the chemical-disease co-occurrence knowledge panel, available on the Summary page of indomethacin (<u>https://pubchem.ncbi.nlm.nih.gov/compound/3715</u> #section=Chemical-Disease-Co-Occurrences-in-Literature). It is noteworthy that the co-occurrence RDF allows the user to get an arbitrary number of co-occurrence neighbors (up to 1000, as explained in the Model Design section), while the knowledge panel in the Compound Summary page shows provides a maximum of 25 co-occurrence neighbors with the current settings.

3.2. Use Case 2: References that Co-mentions a Particular Chemical and Disease Pair

It is noteworthy that the diseases in Table 1 are related to the input chemical (indomethacin) in different contexts. For example, indomethacin is used to treat inflammatory diseases like arthritis, while it is also known to cause stomach ulcer (as a side effect). To understand the context of the relationship between two entities, it is often necessary to get relevant articles that mention them together. The

SPARQL query for this task is shown in Figure 3, with the indomethacin–inflammation pair as an example.

```
PREFIX prism: <http://prismstandard.org/namespaces/basic/3.0/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX pcvocab: <http://rdf.ncbi.nlm.nih.gov/pubchem/vocabulary#>
PREFIX disease: <http://rdf.ncbi.nlm.nih.gov/pubchem/disease/>
PREFIX compound: <http://rdf.ncbi.nlm.nih.gov/pubchem/compound/>
SELECT ?ref, ?date, ?journal, ?title
FROM <http://rdf.ncbi.nlm.nih.gov/pubchem/reference>
WHERE {
  ?ref pcvocab:discussesAsDerivedByTextMining compound:CID3715 .
  ?ref pcvocab:discussesAsDerivedByTextMining disease:DZID8173 .
  ?ref dcterms:date ?date .
  ?ref dcterms:title ?title .
  ?ref prism:publicationName ?journal .
}
ORDER BY DESC(?date)
LIMIT 8
```

Figure 3: SPARQL query to get eight most recent references that mention indomethacin (CID3715) and inflammation (DZID8173) together.

Table 2

Date	Journal	Title
2022-12-31	Pharm. Biol.	Gastroprotective effects of water extract of domesticated <i>Amauroderma rugosum</i> against several gastric ulcer models in rats
2022-07-20	Macromol. Biosci.	A Straightforward Approach Towards Antibacterial and Anti- Inflammatory Multifunctional Nanofiber Membranes with Sustained Drug Release Profiles
2022-07-01	Biomed. Pharmacother.	AMPK/mTOR-driven autophagy & Nrf2/HO-1 cascade modulation by amentoflavone ameliorates indomethacin- induced gastric ulcer
2022-07-01	Environ. Toxicol.	Protective effect of lupeol on arthritis induced by type II collagen via the suppression of P13K/AKT signaling pathway in Sprague dawley rats
2022-06-25	Int. J. Pharm.	Chitosan/sulfobutylether-β-cyclodextrin based nanoparticles coated with thiolated hyaluronic acid for indomethacin ophthalmic delivery
2022-06-14	Prostaglandins Other Lipid Mediat.	Post-mortem changes of prostanoid concentrations in tissues of mice: Impact of fast cervical dislocation and dissection delay
2022-06-07	J. Dairy Sci.	Induction of leaky gut by repeated intramuscular injections of indomethacin to preweaning Holstein calves
2022-06-03	Nat. Prod. Res.	<i>Ranunculus</i> species suppress nitric oxide production in LPS- stimulated RAW 264.7 macrophages

Eight recent PubMed records that mention indomethacin and inflammation together, returned from the SPARQL query in Figure 3.

In the query for Use Case 2, while the named entities are specified using the PubChem-specific vocabulary (pcvocab:discussesAsDerivedByTextMining), the external vocabularies from DMCI [15] and PRISM [16] are used to get the metadata for the reference (i.e., date, title, and journal). The result of the query is shown in Table 2.

3.3. Use Case 3: Diseases Implicitly Related to a Chemical via Genes

Use Case 3 intends to identify diseases related to a chemical via genes, by first identifying genes commonly mentioned with the query chemical and then retrieving diseases co-occurring with those genes. In this Use Case, while some of the resulting diseases may already be mentioned together with the query chemical in scientific articles, others may not. This implicit relationship can serve as a good starting point to formulate a new hypothesis to test in future studies. Figure 4 shows the SPARQL query with maribavir (CID 471161) as an example.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX compound: <http://rdf.ncbi.nlm.nih.gov/pubchem/compound/>
SELECT ?gene, ?disease, ?disease prefLabel
    FROM <http://rdf.ncbi.nlm.nih.gov/pubchem/cooccurrence>
    FROM <http://rdf.ncbi.nlm.nih.gov/pubchem/disease>
WHERE {
    {
        SELECT ?gene
            FROM <http://rdf.ncbi.nlm.nih.gov/pubchem/cooccurrence>
        WHERE {
            ?cooccurrence rdf:subject compound:CID471161 .
            ?cooccurrence rdf:object ?gene .
            ?cooccurrence rdf:type sio:SIO 001257 .
            ?cooccurrence sio:SIO 000300 ?score .
        }
        ORDER BY DESC(?score)
        LIMIT 1
    }
    ?cooccurrence2 rdf:subject ?gene .
    ?cooccurrence2 rdf:object ?disease .
    ?cooccurrence2 rdf:type sio:SIO 000983 .
    ?cooccurrence2 sio:SIO 000300 ?score2 .
    ?disease skos:prefLabel ?disease prefLabel .
}
ORDER BY DESC(?score2)
LIMIT 10
```

Figure 4: SPARQL query to get the top ten diseases associated with the gene most commonly mentioned with maribavir (CID471161). The SIO terms, SIO_001257, SIO_000983, and SIO_000300 mean "chemical-gene association", "gene-disease association", and "has a value", respectively.

Maribavir is an antiviral drug approved in 2021 by the U.S. Food and Drug Administration (FDA) for the treatment of posttransplant cytomegalovirus (CMV) infection. Because of its short history, this drug has not been mentioned in many papers, compared to old drugs introduced in the market decades ago (e.g., indomethacin). Table 3 shows the diseases retrieved from the query. The gene most mentioned together with maribavir is the protein kinase, X-linked (PRKX) gene. While some of the diseases co-occurring with PRKX are directly co-mentioned with maribavir in literature, other diseases, including "genetic translocation", "depressive disorder", "ischemia", and "neurodegenerative diseases", have not appeared with maribavir in PubMed records, implying implicit associations between Maribavir and these diseases via the PRKX gene.

Table 3

Top ten diseases associated with the protein kinase, X-linked (PRKX) gene, which is the most cooccurring gene with maribavir (CID 471161). The diseases were returned from the SPARQL query in Figure 4.

Preferred disease name	Co-occurring with maribavir
Neoplasms	Yes
Translocation, Genetic	No
Carcinogenesis	No
Polyploidy	Yes
Infections	Yes
Inflammation	Yes
Depressive Disorder	No
Drug-Related Side Effects and Adverse Reactions	Yes
Ischemia	No
Neurodegenerative Diseases	No

4. Conclusions

In this paper, we described the RDF data model that expresses the co-occurrence associations between chemicals, genes, and diseases derived from biomedical literature. This data model allows users to quickly identify chemicals, genes/proteins, and diseases mentioned together with a given named entity (Use Case 1). In addition, the model can be used to get references that mention two entities together, helping one to understand the context of the co-occurrence association between the entities (Use Case 2). It can also be used to find an implicit link between entities that are not mentioned together, through the common entities associated with them (Use Case 3).

The underlying data used in the co-occurrence RDF was derived from text mining of 35 million references available in PubMed as shown in our previous study [4]. While this data is also used to generate the PubChem literature knowledge panel, the co-occurrence RDF enables additional tasks. For example, with the co-occurrence RDF, the user can work with a large set of relevant co-occurrence neighbors (up to 1,000) for a given entity and automate this data retrieval task using a computer program or script. The co-occurrence RDF model is an enhancement to the PubChemRDF ecosystem that can facilitate exploring biomedical knowledge and seeking new discoveries in a semantic way. More importantly, it is naturally connecting to other linked data resources in various scientific communities to greatly enhance the usability and accessibility of biomedical data. The co-occurrence RDF data generated in this study is freely available at Zenodo [25].

5. Acknowledgements

This work was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health.

6. References

- [1] Sunghwan Kim, Exploring Chemical Information in PubChem, Current Protocols 1 (2021) e217. doi:10.1002/cpz1.217.
- [2] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton, PubChem in 2021: new data content and improved web interfaces, Nucleic Acids Research 49 (2021) D1388–D1395. doi:10.1093/nar/gkaa971.
- [3] Sunghwan Kim, Getting the most out of PubChem for virtual screening, Expert Opinion on Drug Discovery 11 (2016) 843–855. doi:10.1080/17460441.2016.1216967.
- [4] Leonid Zaslavsky, Tiejun Cheng, Asta Gindulyte, Siqian He, Sunghwan Kim, Qingliang Li, Paul Thiessen, Bo Yu, and Evan E. Bolton, Discovering and Summarizing Relationships Between Chemicals, Genes, Proteins, and Diseases in PubChem, Frontiers in Research Metrics and Analytics 6 (2021) 689059. doi:10.3389/frma.2021.689059.
- [5] Evan E. Bolton, Sunghwan Kim, and Stephen H. Bryant, PubChem3D: Similar conformers, Journal of Cheminformatics 3 (2011) 13. doi:10.1186/1758-2946-3-13.
- [6] Sunghwan Kim, Evan E. Bolton, and Stephen H. Bryant, Similar compounds versus similar conformers: complementarity between PubChem 2-D and 3-D neighboring sets, Journal of Cheminformatics 8 (2016) 62. doi:10.1186/s13321-016-0163-1.
- [7] PubChemRDF neighbor subdomain, https://pubchem.ncbi.nlm.nih.gov/docs/rdf-neighbor, last accessed 2023/01/11.
- [8] Daniel M. Lowe and Roger A. Sayle, LeadMine: a grammar and dictionary driven approach to entity recognition, J Cheminform 7 (2015) S5. doi:10.1186/1758-2946-7-S1-S5.
- [9] RDF 1.1 Concepts and Abstract Syntax, https://www.w3.org/TR/rdf11-concepts/, last accessed 2022/11/12.
- [10] Gang Fu, Colin Batchelor, Michel Dumontier, Janna Hastings, Egon Willighagen, and Evan Bolton, PubChemRDF: towards the semantic annotation of PubChem compound and substance databases, Journal of Cheminformatics 7 (2015) 34. doi:10.1186/s13321-015-0084-4.
- [11] Begoña Talavera Andújar, Dagny Aurich, Velma T. E. Aho, Randolph R. Singh, Tiejun Cheng, Leonid Zaslavsky, Evan E. Bolton, Brit Mollenhauer, Paul Wilmes, and Emma L. Schymanski, Studying the Parkinson's disease metabolome and exposome in biological samples through different analytical and cheminformatics approaches: a pilot study, Anal Bioanal Chem 414 (2022) 7399–7419. doi:10.1007/s00216-022-04207-z.
- [12] SPARQL 1.1 Query Language, https://www.w3.org/TR/sparql11-query/, last accessed 2022/11/12.
- [13] Michel Dumontier, Christopher JO Baker, Joachim Baran, Alison Callahan, Leonid Chepelev, José Cruz-Toledo, Nicholas R. Del Rio, Geraint Duck, Laura I. Furlong, Nichealla Keath, Dana Klassen, Jamie P. McCusker, Núria Queralt-Rosinach, Matthias Samwald, Natalia Villanueva-Rosales, Mark D. Wilkinson, and Robert Hoehndorf, The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery, J Biomed Semant 5 (2014) 14. doi:10.1186/2041-1480-5-14.
- [14] FRAPO, the Funding, Research Administration and Projects Ontology, https://sparontologies.github.io/frapo/current/frapo.html, last accessed 2022/11/12.
- [15] Dublin Core Metadata Initiative (DCMI) Metadata Terms, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, last accessed 2022/11/12.
- [16] Publishing Requirements For Industry Standard Metadata (PRISM) Specification Package, https://www.w3.org/Submission/prism/, last accessed 2022/11/12.
- [17] Jon Ison, Matúš Kalaš, Inge Jonassen, Dan Bolser, Mahmut Uludag, Hamish McWilliam, James Malone, Rodrigo Lopez, Steve Pettifer, and Peter Rice, EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats, Bioinformatics 29 (2013) 1325–1332. doi:10.1093/bioinformatics/btt113.
- [18] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008). doi:10.1017/CBO9780511809071.

- [19] Akiko Aizawa, An information-theoretic perspective of tf-idf measures, Information Processing & Management 39 (2003) 45–65. doi:10.1016/S0306-4573(02)00021-3.
- [20] Stephen Robertson, Understanding inverse document frequency: on theoretical arguments for IDF, Journal of Documentation 60 (2004) 503–520. doi:10.1108/00220410410560582.
- [21] Anand Rajaraman and Jeffrey David Ullman, Mining of Massive Datasets. Cambridge University Press, Cambridge (2011). doi:10.1017/CBO9781139058452.
- [22] Medical Subject Headings, https://www.nlm.nih.gov/mesh/, last accessed 2022/11/12.
- [23] OpenLink Software: Virtuoso Homepage, https://virtuoso.openlinksw.com/, last accessed 2022/11/12.
- [24] Thomas Baker, Sean Bechhofer, Antoine Isaac, Alistair Miles, Guus Schreiber, and Ed Summers, Key choices in the design of Simple Knowledge Organization System (SKOS), Journal of Web Semantics 20 (2013) 35–49. doi:10.1016/j.websem.2013.05.001.
- [25] Qingliang Li, Sunghwan Kim, Leonid Zaslavsky, Tiejun Cheng, Bo Yu, and Evan Bolton, Resource Description Framework (RDF) Modeling of Named Entity Co-occurrences Derived from Biomedical Literature in the PubChemRDF [Data set], Zenodo (2023). doi:10.5281/zenodo.7521846.