NanoWeb: Search, Access and Explore Life Science Nanopublications on the Web*

(Discussion Paper)

Fabio Giachelle[1], Dennis Dosso[1] and Gianmaria Silvello[1]

[1]*Department of Information Engineering, University of Padua, Padua, Italy*

Abstract

Nanopublications are scientific statements represented in the Resource Description Framework (RDF), a brief machine-readable form representing data. Nanopublications consist of scientific facts extracted from the literature and contextualized with provenance and attribution information.

Nanopublications are designed to enhance knowledge spreading, support the re-use of scientific facts, and provide credit to the corresponding authors. Despite these promising features, nanopublications are not widely adopted, and their use is still quite limited to experts. We believe this is partly due to the lack of services for searching, retrieving, and understanding nanopublications.

To mitigate this, we propose NanoWeb, a Web-based system designed to allow general users to search, access, explore, and re-use nanopublications publicly available on the Web. Currently, NanoWeb is tailored for the life science domain, where plenty of nanopublications are available.

Keywords

Nanopublications, Semantic Web, Data Citation

1. Introduction

In the last decades we assisted to a change of paradigm in the current world of research, more precisely towards a more data-intensive approach, following the so-called *fourth paradigm of science* [2]. As a result, data acquired a prominent role, becoming the center of scientific discovery as well as of scholarship and scholarly communication [3]. Accordingly, data science has experienced an unprecedented growth that involved also other linked research fields regarding the search [4], provenance [5], citation [6], re-use [7], and exploration [8] of data.

Despite this premises, most of the huge volume of data available in the Web, is provided in an unstructured format, which is *human-readable* but not machine-readable. Hence, the heterogeneity of data and their representations lead to issues concerning interoperability; data access, search and re-use; and domain-dependent requirements.

*This is an extended abstract of [1].

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy

fabio.giachelle@unipd.it (F. Giachelle); dennis.dosso@unipd.it (D. Dosso); dennis.dosso@unipd.it (G. Silvello)
http://www.dei.unipd.it/~giachell/ (F. Giachelle); http://www.dei.unipd.it/~dosso/ (D. Dosso); http://www.dei.unipd.it/~silvello/ (G. Silvello)
0000-0001-5015-5498 (F. Giachelle); 0000-0001-7307-4607 (D. Dosso); 0000-0003-4970-4554 (G. Silvello)

Semantic Web technologies, such as the *nanopublication model* [9], aim to tackle some of these issues according to the Findable, Accessible, Interoperable, Reusable (FAIR) guiding principles [10]. Nanopublications represent statements (i.e. *assertions*) leveraging on the Linked Open Data (LOD) principles [11] to convey scientific facts in an efficient, succinct, and machine-readable form. The LOD features make nanopublications suitable to handle the connections between related scientific facts, encoded in RDF. These facts are connected in a network that can be explored, so to discover connections among them.

The nanopublications model is designed to make scientific claims accessible, so that the scientific knowledge can automatically be explored by agents. Several scientific fields adopted the nanopublications model, such as the Life Science domain, where more than ten million openly available nanopublications [12] were created. There is important evidence in the biomedical domain about the benefits of using nanopublications for expanding scientific insight [13]. However, to the best of our knowledge, there is no tool to visualize nanopublications and explore their underlying network of scientific facts. BioKB [14] provides functionalities similar to NanoWeb in terms of semantic search and graph visualization. In particular, it provides access to the semantic content of biomedical articles through a SPARQL endpoint and a web interface. Thus, allowing users to search for biomedical entities and visualize their graph of relations. However, BioKB does not account for nanopublications and does not support a multi-level exploration of the graph, enabling an in-depth exploration of the entities relation network. Currently, searching for nanopublications is possible only through sparse SPARQL endpoints. NanoWeb contributes to tackle these issues by providing: (a) a unified entry-point to access all the publicly available nanopublications from the Life Science domain; (b) a user-friendly web interface that enables users without a prior knowledge of SPARQL and related technologies to search, access and explore millions of nanopublications in a human-readable form.

The paper is organized as follows: Section 2 presents some background, Section 3 provides a brief overview of NanoWeb functionalities and describes one prominent feature, and Section 4 draws some final remarks.

2. Background

A nanopublication consists of a Resource Description Framework (RDF) graph based on an assertion, that represents a scientific statement extracted, manually or automatically, from a scientific publication. Nanopublications are representated using an extended RDF syntax supporting *quads* in addition to triples, where an identifier (an IRI) is added. Hence, groups of triples may be characterized as belonging to the same subgraph, i.e. to the same *named graph* [15, 16], if they share the same extra URI. More specifically, the structure of nanopublications consists of four *named graphs*: (a) the *head*, that connects the other three sub-graphs; (b) the *assertion* graph (in blue in Figure 1) that reports the statement extracted from the scientific publications as one RDF triple (subject-predicate-object), where the the two concepts of the assertion (subject and object) are linked by a specific relationship defined by the predicate. The predicate relationship may refer to external ontologies and scientific databases storing related data. (c) The provenance graph (in orange in Figure 1) reports the assertion metadata such as the assertion's generation methods and its creators; and, (d) the publication info graph (yellow)

that contains the metadata about when the nanopublication was created, its authors and the license terms for its reuse.

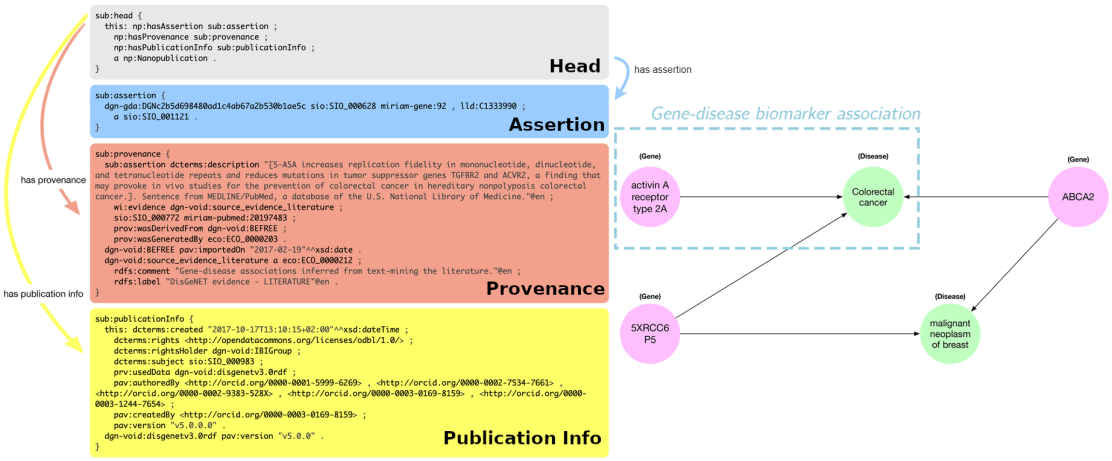

Figure 1: (A) RDF trig serialization of the nanopublication with assertion: ⟨*activin A receptor type 2A - gene-disease biomarker association - Colorectal Cancer*⟩; (B) network of gene-disease associations created by five nanopublications.

Figure 1.A shows an example of nanopublication regarding the biomedical domain. The assertion contains information about the gene-disease association ⟨*activin A receptor type 2A – gene-disease biomarker association – colorectal cancer*⟩, where *activin A receptor type 2A* is the subject, *gene-disease biomarker association* is the predicate and *colorectal cancer* is the object of the triple. The assertion is extracted from a paper [17], which states that a drug – *Mesalazine* – can reduce the mutations in a tumor suppressor gene – *activin A receptor type 2A (ACVR2)* – related to *colorectal cancer*. This assertion is represented as a RDF nanopublication with the TriG [1] serialization in Figure 1.A. TriG is a compact RDF syntax, characterized by the use of *prefixes* to avoid the repetition of IRIs. Figure 1.B illustrates an example of a five nodes network concerning gene-disease associations. If we consider for instance the *colorectal cancer* disease, it is initially connected with the gene *ACVR2*, through the gene-disease biomarker association. If we expand the *colorectal cancer* relation network, we discover two other genes *5XRCC6P5* and *ABCA2*. These genes are directly connected to the *colorectal cancer* disease and, after a further expansion, we discover that they are linked also to the *malignant neoplasm of breast* disease. These RDF triples are all inferred from five distinct papers distinguished by venue and time of publication, that also do not cite each other. The triple in Figure 1.B contained in the blue rectangle is also the one used as assertion in Figure 1.A. This shows how, by using RDF graphs, triples connected among them, the LOD principles, and nanopublications about these triples, it is possible to explore and discover new concepts and links among them not otherwise explicitly visible through the traditional scientific literature. Despite these promising features, the adoption of nanopublications by scientists is still quite limited to specific domains [18]. To

[1] https://www.w3.org/TR/trig/

date, there is no easy way to find, re-use and cite them, as well as no service that allows users to access them in a human-readable form, and search them using keywords or natural language queries. Moreover, there is no tool designed to allow the progressive exploration of the nanopublications relation network, which can be useful for the discovery of new meaningful connections between nanopublications' assertions and scientific facts.

SPARQL is a powerful but complex query language to access and interrogate RDF graphs [19]. Due to its complex syntax, SPARQL is not intuitive for end-users without a prior expertise and knowledge of the underlying schema of database, and of the IRIs used in it. To address the issues related to the use of SPARQL and enable users querying the RDF datasets via natural language queries, the *keyword search* paradigm has been introduced. Keyword-based methods are designed to ease the access to structured data [20, 21]. In contrast to SPARQL, keyword search returns a ranking of answers, ordered according to their *relevance* for the user-provided keyword query. Keyword query search systems over structured data consider both relational databases (RDB) [21] and graph-like databases such as RDF datasets [22, 20]. Keyword-based systems may be divided into three categories: (a) *schema-based* systems [23, 24] which exploit the information concerning the schema of the database to formulate structured-language queries (e.g. SQL or SPARQL queries) designed from the user-provided keyword query; (b) *graph-based* systems [25, 26] which rely on the transformation of relational databases into graphs; and (c) *virtual-document based* systems [27] which generate a *virtual document* [28] for a given graph, where the lexical content is preserved. Hence, virtual-document based systems can leverage on efficient state-of-the-art IR methods for indexing and ranking. However, there is no keyword search system for nanopublications, which are always searched via SPARQL endpoints. To allow end-users specify queries in natural language, NanoWeb exploits a very recent advancement in virtual-document based systems [29], thus enabling fast and effective keyword-search over RDF and nanopublications.

3. NanoWeb

NanoWeb[2] is a public and open-source[3] web application designed to provide intuitive search, exploration, citation [30] and re-use of nanopublications.

As of now, NanoWeb is tailored for the life science domain, and it is conceived to support life science experts in their research work. Nevertheless, NanoWeb can be applied, after domain-specific customization, to other scientific domains.

NanoWeb provides a unified access to the world of nanopublications, even to users without a specific expertise, which can search for nanopublications using natural language queries. Its main features are:

1. A crawler collecting the public nanopublications available on the web;
2. Keyword search and advanced search. The latter provides a structured guided search based on the boolean search paradigm;
3. A user-centric visual interface to search and consult nanopublications, enriched with information gathered from external authoritative ontologies;

[2] https://w3id.org/nanoweb/

[3] NanoWeb source code is available at https://github.com/giachell/nanoweb

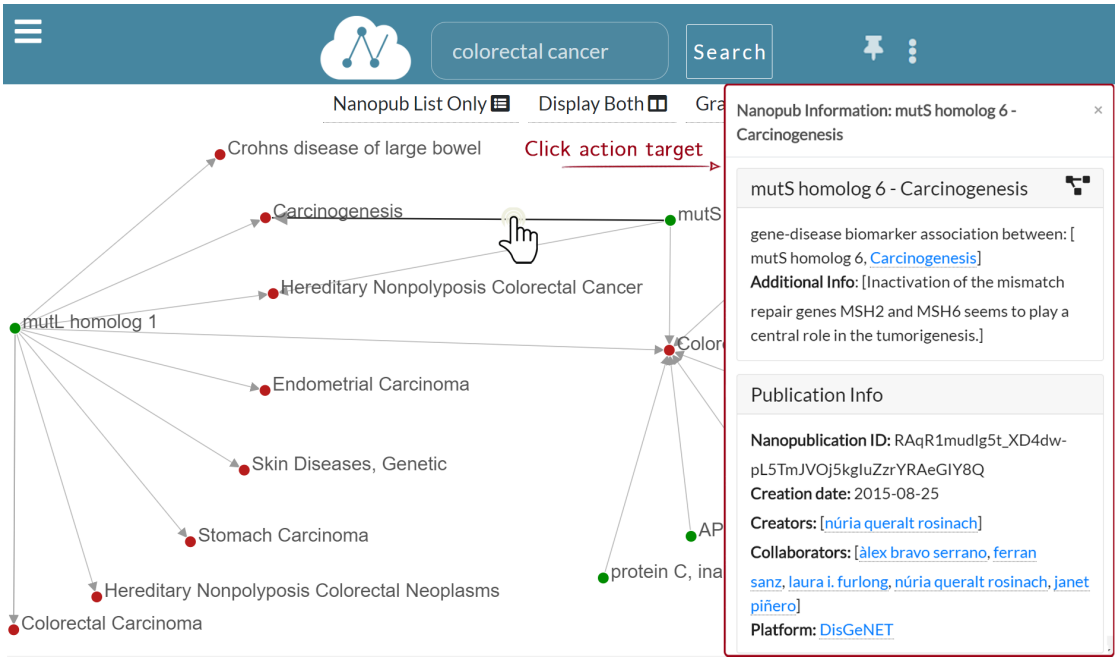

Figure 2: Graph exploration: users can inspect nanopublications and the information concerning entity connections by clicking on the edges. As a result, an information window is shown reporting the nanopublication's content, which refers to the entities linked by the edge clicked. Inside the red box (right side) is reported the information content of the nanopublication with assertion: ⟨*mutS homolog 6 - gene-disease biomarker association - Carcinogenesis*⟩.

4. A visual exploration of the nanopublications' relation network built around the scientific facts encoded in the assertions;
5. Data search capabilities providing direct connections to evidence papers and external scientific curated databases.

One of the prominent NanoWeb features is the exploration of the nanopublications relation network, i.e. the RDF graph composed by the triples extracted from the scientific evidence papers, as described above. NanoWeb represents the nanopublications relation network as a graph where the nodes are the subjects and the objects of the nanopublications' triples. The nodes are connected together with directed edges from subjects to objects.

Figure 2 shows a multi-level graph exploration of the relation network concerning the nanopublication with title *mutL homolog 1 - Colorectal Carcinoma*, which expresses a gene-disease association. The graph exploration functionality allows the user to understand the relationships between different nanopublications. The graph exploration functionality is conceived to let the user progressively expand the graph, focusing on both the nodes and connections of interest, without any limit on the depth of the exploration, i.e., to the graph's dimension visualized. Hence, the user can potentially expand the whole graph, having all the nodes of the relation network displayed. The exploration of the relation network is one of the biggest contribution, since it provides a deeper understanding of the connections between entities, such as gene-disease associations.

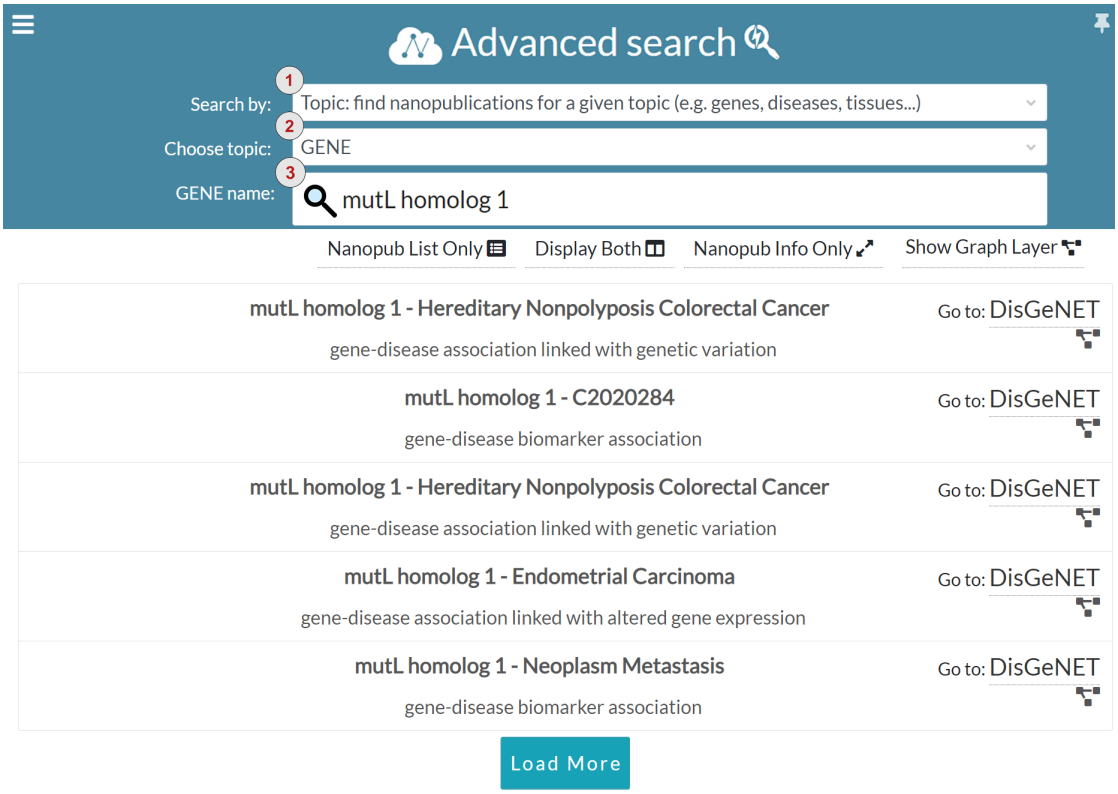

Figure 3: Advanced search: users are guided through the search process by means of sequential filters, that progressively refine the query domain and restrict the retrieval output accordingly. The figure shows an example of a topic-based search for nanopublications concerning *genes* and in particular the *mutL homolog 1* gene.

4. Conclusions

Nanopublications are concise, noise-free RDF resources designed to efficiently convey information and concepts.

The full adoption of nanopublications is held back by the lack of services to search, access, explore, and cite them. As of today, data search is possible using only SPARQL queries, that are only within the capabilities of experienced users. There is no service to search over all the publicly available nanopublications, using either keyword or natural language queries. To target these issues we designed and developed *NanoWeb*.

NanoWeb provides unified access to Life Science nanopublications so that users can search, access, explore, and re-use them on the Web. NanoWeb allows the users to (i) search for domain-specific nanopublications using keyword queries; (ii) explore their relation network to discover new nanopublications and meaningful connections; (iii) access and consult their information content; (iv) access the evidence paper information (e.g. abstract) and the entry points to related data record in external curated scientific databases; and, (v) easily cite nanopublications.

NanoWeb enables a serendipity-oriented perspective in the Life Science domain. Users can

benefit from this perspective through the exploration of nanopublication graphs, which could lead to a deeper understanding of the connections between entities (e.g. genes and diseases) and enrich the domain knowledge.

Acknowledgments

This work is supported by the ExaMode project, as part of the European Union Horizon 2020 program under Grant Agreement no. 825292.

References

[1] F. Giachelle, D. Dosso, G. Silvello, Search, access, and explore life science nanopublications on the web, PeerJ Computer Science 7 (2021) e335. doi:10.7717/peerj-cs.335.

[2] T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, USA, 2009.

[3] C. L. Borgman, Big Data, Little Data, No Data, MIT Press, 2015.

[4] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L. D. Ibáñez, E. Kacprzak, P. Groth, Dataset search: a survey, VLDB J. 29 (2020) 251–272. doi:10.1007/s00778-019-00564-x.

[5] J. Cheney, L. Chiticariu, W. Tan, Provenance in databases: Why, how, and where, Foundations and Trends in Databases 1 (2009) 379–474.

[6] G. Silvello, Theory and Practice of Data Citation, Journal of the American Society for Information Science and Technology (JASIST) 69 (2018) 6–20.

[7] L. A. Wynholds, J. C. Wallis, C. L. Borgman, A. Sands, S. Traweek, Data, Data Use, and Scientific Inquiry: Two Case Studies of Data Practices, 2012, pp. 19–22. doi:10.1145/2232817.2232822.

[8] P. Rahman, L. Jiang, A. Nandi, Evaluating Interactive Data Systems, VLDB J. 29 (2020) 119–146. doi:10.1007/s00778-019-00589-2.

[9] P. Groth, A. Gibson, J. Velterop, The Anatomy of a Nanopublication, Inf. Serv. Use 30 (2010) 51–56.

[10] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[11] C. Bizer, T. Heath, T. Berners-Lee, Linked Data – The Story So Far, Int. J. Semantic Web Inf. Syst. 5 (2009) 1–22. doi:10.4018/jswis.2009081901.

[12] T. Kuhn, A. Meroño-Peñuela, A. Malic, J. H. Poelen, A. H. Hurlbert, E. C. Ortiz, L. I. Furlong, N. Queralt-Rosinach, C. Chichester, J. M. Banda, E. L. Willighagen, F. Ehrhart, C. T. A. Evelo, T. B. Malas, M. Dumontier, Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data, in: 14th IEEE International Conference on e-Science, e-Science 2018, IEEE Computer Society, 2018, pp. 83–92. doi:10.1109/eScience.2018.00024.

[13] C. Chichester, P. Gaudet, O. Karch, P. T. Groth, L. Lane, A. Bairoch, B. Mons, A. Loizou, Querying neXtProt nanopublications and their value for insights on sequence variants

and tissue expression, J. Web Semant. 29 (2014) 3–11. doi:10.1016/j.websem.2014.05.001.
[14] M. Biryukov, V. Groues, V. Satagopam, R. Schneider, Biokb-text mining and semantic technologies for biomedical content discovery (2018).
[15] J. J. Carroll, P. Stickler, RDF triples in XML, in: Proc. of the WWW 2004 Conference (Alternate Track Papers & Posters), ACM Press, 2004, pp. 412–413.
[16] J. J. Carroll, C. Bizer, P. Hayes, P. Stickler, Named graphs, provenance and trust, in: Proceedings of the 14th international conference on World Wide Web, ACM Press, 2005, pp. 613–622.
[17] C. Campregher, C. Honeder, H. Chung, J. M. Carethers, C. Gasche, Mesalazine reduces mutations in transforming growth factor β receptor ii and activin type ii receptor by improvement of replication fidelity in mononucleotide repeats, Clin Cancer Res 16 (2010) 1950–1956.
[18] R. Page, Liberating links between datasets using lightweight data publishing: an example using plant names and the taxonomic literature, Biodiversity Data Journal 6 (2018) e27539.
[19] J. Pérez, M. Arenas, C. Gutierrez, Semantics and Complexity of SPARQL, ACM Trans. Database Syst. 34 (2009) 1–45.
[20] H. Bast, B. Buchhold, H. Haussmann, Semantic search on text and knowledge bases, Foundations and Trends in Information Retrieval (FnTIR) 10 (2016) 119–271.
[21] J. X. Yu, L. Qin, L. Chang, Keyword Search in Relational Databases: A Survey, IEEE Data Eng. Bull. 33 (2010) 67–78.
[22] H. Wang, C. C. Aggarwal, A survey of algorithms for keyword search on graph data, in: Managing and Mining Graph Data, Springer, 2010, pp. 249–273.
[23] A. Balmin, V. Hristidis, N. Koudas, Y. Papakonstantinou, D. Srivastava, T. Wang, A system for keyword proximity search on XML databases, in: Proceedings of 29th International Conference on Very Large Data Bases, VLDB, Morgan Kaufmann, 2003, pp. 1069–1072.
[24] Y. Luo, W. Wang, X. Lin, X. Zhou, J. Wang, K. Li, SPARK2: top-k keyword query in relational databases, IEEE Trans. Knowl. Data Eng. 23 (2011) 1763–1780.
[25] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, S. Sudarshan, Keyword searching and browsing in databases using BANKS, in: Proceedings of the 18th International Conference on Data Engineering, IEEE Computer Society, 2002, pp. 431–440.
[26] A. Simitsis, G. Koutrika, Y. E. Ioannidis, Précis: from unstructured keywords as queries to structured databases as answers, VLDB J. 17 (2008) 117–149.
[27] G. Kadilierakis, P. Fafalios, P. Papadakos, Y. Tzitzikas, Keyword Search over RDF Using Document-Centric Information Retrieval Systems, in: The Semantic Web, Springer International Publishing, Cham, 2020, pp. 121–137.
[28] J. I. Lopez-Veyna, V. J. S. Sosa, I. López-Arévalo, A Virtual Document Approach for Keyword Search in Databases, in: DATA, SciTePress, 2012, pp. 39–48.
[29] D. Dosso, G. Silvello, Search text to retrieve graphs: A scalable RDF keyword-based search system, IEEE Access 8 (2020) 14089–14111.
[30] E. Fabris, T. Kuhn, G. Silvello, A framework for citing nanopublications, in: Proc. of the 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, 2019, pp. 70–83. doi:10.1007/978-3-030-30760-8_6.