Publishing CoreKB Facts as Nanopublications

Fabio Giachelle[1], Stefano Marchesin[1], Laura Menotti[1] and Gianmaria Silvello[1]

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

Abstract

The Collaborative Oriented Relation Extraction (CORE) system generates gene expression-cancer associations by combining scientific evidence from the literature. Such facts are then ingested into the CoreKB platform, where one can browse and search for associations. In this work, we publish 197, 511 assertions from CoreKB as nanopublications, allowing the sharing of machine-readable gene-cancer associations while tracking their provenance and publication information.

Keywords

Nanopublications, Information Provenance, Knowledge Bases, Gene-Cancer Associations

1. Introduction

Representing data in a machine-readable and interoperable format, i.e., Resource Description Framework (RDF), is fundamental for data-intensive science discovery [1]. This paradigm emphasizes data's crucial role for various purposes, ranging from research to developing innovative solutions and making informed decisions. In this context, Knowledge Bases (KBs) are pivotal as they are suitable for numerous applications thanks to their structure that is both human-readable and machine-interoperable [2]. KBs provide a novel publishing environment for scientific information, unveiling novel research questions related to the identification and referencing of individual assertions. In this context, the nanopublication model supports statement-based publishing by providing a framework to represent assertions as single publications [3]. In this way, each statement can be uniquely identified, accessed, and cited [4]. The nanopublication model relies on named graphs, providing both a human-readable and machine-understandable resource in line with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles for knowledge discovery [5]. The nanopublication model can be useful to represent facts generated by the Collaborative Oriented Relation Extraction (CORE) system, a Knowledge Base Construction (KBC) pipeline that combines scientific literature to provide fine-grained gene-cancer associations, called Gene Cancer Status (GCS) [6]. Such a resource provides information about genes whose expression level interacts with different types of cancer, either by enhancing its

IRCDL 2024: 20th conference on Information and Research science Connecting to Digital and Library science, February 22-23, 2024, Bressanone, Brixen, Italy

fabio.giachelle@unipd.it (F. Giachelle); stefano.marchesin@unipd.it (S. Marchesin); laura.menotti@unipd.it (L. Menotti); gianmaria.silvello@unipd.it (G. Silvello)

https://www.dei.unipd.it/~giachell/ (F. Giachelle); https://www.dei.unipd.it/~marches1/ (S. Marchesin); https://www.dei.unipd.it/~menottilau/ (L. Menotti); https://www.dei.unipd.it/~silvello/ (G. Silvello)

0000-0001-5015-5498 (F. Giachelle); 0000-0003-0362-5893 (S. Marchesin); 0000-0002-0676-682X (L. Menotti); 0000-0003-4970-4554 (G. Silvello)

CEUR Workshop Proceedings (CEUR-WS.org)

progression or regression [7]. Thus, having access to single statements instead of reading the whole literature is pivotal for advancing cancer research [8]. Facts generated by CORE were ingested into the CoreKB platform, allowing users to search for gene-cancer associations that can be browsed through entity landing pages [9]. [1]

In this work, we apply the nanopublication model to the facts generated by the CORE system [6], comprising gene-expression cancer associations extracted from the scientific literature. We publish $197,511$ facts in the form of nanopublications, which can be browsed and downloaded from the CoreKB platform [9]. We rely on the nanopublication model because it allows the publication of single statements that can be uniquely identified and referenced and has been widely used in the life-science domain [10, 11, 12, 13]. Moreover, other solutions for publishing single assertions represent claims in textual form [14, 15] and cannot be directly used in our use-case, where assertions are represented as RDF statements.

The remainder of this work is organized as follows. Section 2 outlines the nanopublication model and portrays previous endeavors in publishing information exploiting such a paradigm. Section 3 describes the CORE systems and the CoreKB platform, illustrating the structure of its facts. Section 4 presents the serialization of nanopublications representing assertions generated by CORE. Section 5 concludes the paper.

2. Background

The nanopublication model eases information access, allowing for representing statements at a finer granularity in a distinctive, identifiable, citable, and reusable way [3, 4]. In this way, scientific resources can be divided into single statements called *assertions*, resolving in a unique nanopublication comprising information relevant to that specific piece of knowledge. The nanopublication model provides a novel publishing environment, where authors share scientific claims in both human-readable and machine-actionable manner in line with FAIR principles. At a technical level, a nanopublication is a named graph consisting of three basic components, each represented as a named graph as well: (i) the *assertion* graph, representing the individual claim in RDF format; (ii) the *provenance* graph, containing where the assertion comes from and how it was generated; (iii) the *publication* graph, comprising all metadata describing the nanopublication. In addition, the nanopublication model contains a fourth graph called the *head* graph, which defines the nanopublication and connects all the components.

The nanopublication model has been employed to publish scientific information from different fields, especially in the life science domain. There are more than 10 million nanopublications publicly accessible worldwide [16], as representing data as nanopublications fosters data-intensive science and fact discovery using machine-readable information [17]. Chichester et al. [12] published nanopublications representing the neXtProt database [2], which contains scientific facts associated with more than 38K proteins. As a result, this work showed that using the nanopublication model provides easy access to information and can be a useful tool for advancing biological research [11]. Waagmeester et al. [13] converted WikiPathways, an

[1] https://gda.dei.unipd.it/

[2] https://www.nextprot.org/

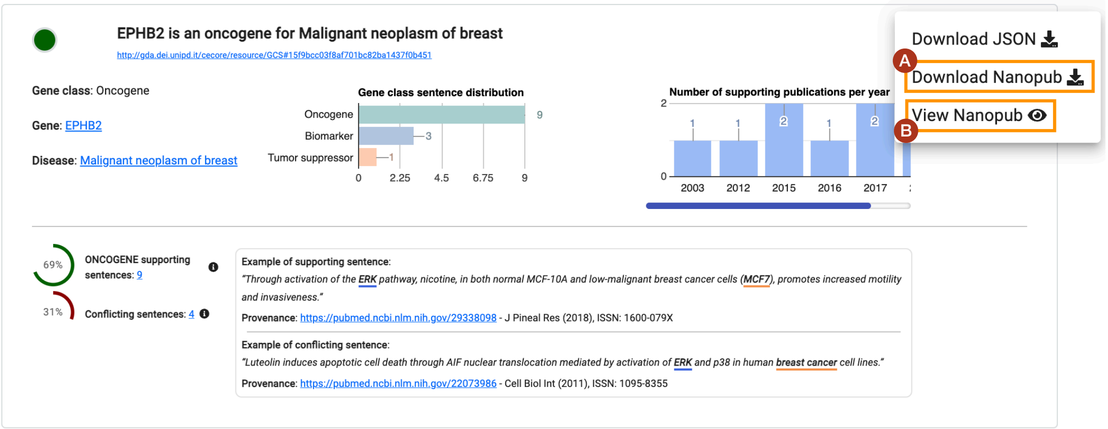
online collaborative pathway resource, into nanopublications. [3] Queralt-Rosinach et al. [10] created nanopublications from the DisgeNET database [4] to provide an alternative means to mine the Gene-Disease Associations (GDAs) contained in DisgeNET. Kuhn et al. [15] expanded the concept behind nanopublications to account for textual statements. In particular, the *assertion* graph has been extended to consider English sentences following a particular syntactic schema called AIDA (Atomic, Independent, Declarative, Absolute). In this matter, Clark et al. [14] proposed the micropublication model, where statements are in textual form as well. Differently from nanopublications, the micropublication model represents empirical evidence focusing on building claims networks and tracking their provenance, specifically for the biomedical communications ecosystem. However, both of these works are out of our scope, as both the micropublication model and the AIDA nanopublications represent claims in a textual form, while we focus on assertions in RDF format.

Nanopublications can be used by researchers to express research findings in a machine-understandable way. Such a practice is called *nanopublishing*. To this end, Knowledge Pixels [5] pioneered a new publishing environment based on nanopublications. This startup provides software and services for managing machine-readable scientific findings and recently beta-launched the Nanodash platform [6] to publish and search nanopublications. Concerning nanopublishing, Knowledge Pixels started three pilot projects with IOS Press and Pensoft [7] to develop and test a nanopublication-based publishing environment in three journals, namely the Data Science Journal [8], the Biodiversity Data Journal [9], and the RIO Journal [10].

3. Gene expression-cancer associations generated by CORE

CORE is a Knowledge Base Construction (KBC) system based on the combination of bootstrapping mechanisms and active learning paradigms [6]. CORE harvests articles from scientific literature and identifies sentences containing relevant fine-grained aspects to generate gene expression-cancer associations that can be published as facts, called Gene Cancer Status (GCS). The CORE architecture is based on several modules and processes, starting from Data acquisition and Named-Entity Recognition and Disambiguation (NERD), where the scientific literature is acquired and processed to provide entity-annotated sentences for manual annotations, to the Relation Extraction (RE) module. Each sentence provides information related to four different aspects: *i)* Change of Gene Expression (CGE): `up`, `down`, or `not informative`; *ii)* Change of Cancer Status (CCS): `progression`, `regression`, or `not informative`; *iii)* Gene-Cancer Interaction (GCI), indicating the interaction occurring between CGE and CCS: `causality`, `correlation`, or `not informative`; *iv)* Gene-Cancer Context (GCC), which evaluates the sentence utility in context and serves as a filter to differentiate between gene-cancer associa-

[3] https://github.com/wikipathways/nanopublications
[4] https://www.disgenet.org/rdf
[5] https://knowledgepixels.com/
[6] https://nanodash.petapico.org/
[7] https://blog.pensoft.net/tag/nanopublications/
[8] https://nanodash.petapico.org/connector/ios/ds
[9] https://nanodash.petapico.org/connector/pensoft/bdj
[10] https://nanodash.petapico.org/connector/pensoft/rio



EPHB2 is an oncogene for Malignant neoplasm of breast

http://gda.dei.unipd.it/cecore/resource/GCS#15f9bcc03f8af701bc82ba1437f0b451

Gene class: Oncogene

Gene: EPHB2

Disease: Malignant neoplasm of breast

ONCOGENE supporting sentences: 9

Conflicting sentences: 4

Example of supporting sentence:

"Through activation of the ERK pathway, nicotine, in both normal MCF-10A and low-malignant breast cancer cells (MCF7), promotes increased motility and invasiveness."

Provenance: https://pubmed.ncbi.nlm.nih.gov/29338098 - J Pineal Res (2018), ISSN: 1600-079X

Example of conflicting sentence:

"Luteolin induces apoptotic cell death through AIF nuclear translocation mediated by activation of ERK and p38 in human breast cancer cell lines."

Provenance: https://pubmed.ncbi.nlm.nih.gov/22073986 - Cell Biol Int (2011), ISSN: 1095-8355

Figure 1: Landing page for the GCS 15f9bcc03f8af701bc82ba1437f0b451 in the CoreKB platform. For each GCS, CoreKB displays information about the associated gene and cancer, the gene class, and its distribution across the involved publications. Each GCS represented as a nanopublication can be downloaded serialized in TriG format (A), or visualized (B).

tions related to changes in the gene expression level (`expression`) and those encompassing other types of gene-cancer relationship (`other`). Sentences with GCC value "`other`" are filtered out as they contain gene-cancer pairs not inherent to the task. Sentences related to the same gene-cancer pair are grouped and their CGE, CCS, and GCI are combined to generate facts – that is, GCSs – and assign to each of them one of the three mutually exclusive gene classes: *i)* `Oncogene`, representing genes favouring tumor growth; *ii)* `Tumour Suppressor Gene (TSG)`, denoting genes hindering cancer progression; *iii)* `Biomarker`, indicating genes that show altered expression levels in cancer but are not identified as oncogene or TSG. The generated facts undergo a reliability test, identifying reliable facts to populate the KB. Facts are subsequently classified as reliable or unreliable by a two-stage reliability test that, for each GCS, first verifies the fact has sufficient evidence and then checks that mutually exclusive classes are not similarly probable, i.e. it assesses the degree of contradicting evidence. In this way, unreliable facts can be provided to domain experts for manual annotations in an active learning paradigm, which makes CORE suited to iterative KB versioning.

The data extracted by CORE is available in Zenodo [18] and contains information about 23,853 genes and 1,767 diseases, for a total of 231,099 fine-grained facts supported by 1,037,45 sentences mined from 251,038 research papers. The facts generated by CORE were also ingested in the CoreKB platform [9]. CoreKB [11] allows users to search for gene-cancer associations with entity landing pages that summarize useful information of the considered GCS. Figure 1 shows an example of a GCS card. The fact in textual form is reported on the topside of the card and the coloured circle on the left side of the fact represents the reliability of the fact – green for "reliable fact", gray for facts with insufficient evidence, and red for unreliable facts due to contrasting information. Each card displays the assigned gene class (`Oncogene` in this

[11] https://gda.dei.unipd.it

case) and the gene and disease involved, whose labels are displayed. In addition, the number of sentences supporting the fact, i.e. identifying the same gene class, and those conflicting with it are reported below the gene and disease information. Each card shows an example of a supporting and conflicting sentence with provenance information of the PubMed article from which they were extracted. For each fact, the complete list of evidence can be accessed by clicking on the number of supporting or conflicting sentences. On the left side, the card displays a horizontal bar chart reporting the gene class distribution across the related sentences and bibliometrics about the number of supporting evidence over the years. In conclusion, a drop-down menu offers different functionalities, such as downloading the GCS in JSON format.

4. Implementation

Facts generated by CORE can be represented as nanopublications, since each GCS can be considered an assertion graph. To build nanopublications representing the facts in CoreKB, we adopted the same methodology used for DisgeNET nanopublications [10]. Figure 2 shows the serialization of a GCS from CoreKB [12] following the nanopublication model.

A nanopublication representing a GCS generated by CORE comprises four named graphs: *head*, *assertion*, *provenance*, and *publication information*. The *head* graph defines the nanopublication URI and links its subgraphs. The *assertion* graph is the GCS itself. The *provenance* graph describes how the assertion was derived. In this case, all facts are derived from the CoreKB platform, which is identified by the URL pointing at its homepage, and are generated by an automatic process, Class `Automatic Assertion` [13] from the Evidence and Conclusion Ontology (ECO). Since the CORE system generates facts by integrating multiple source of evidence, the source evidence is an instance of the class `Combinatorial Evidence` [14] from ECO. The *publication information* graph includes metadata about the nanopublication, such as its subject, license, and who created it. Since the CORE system generates fine-grained gene expression-cancer associations, the general topic for all nanopublications is `gene-disease association linked with altered gene expression` [15] from the Semanticscience Integrated Ontology (SIO). We also include information about the creators and authors of the nanopublications and a timestamp indicating when we serialized the nanopublication using the data property `created` from the Dublin Core (DC) metadata terms.

Concerning the used ontologies, to represent the *assertion* graph we employ the ontology supporting the KB creation process in the CORE system. [16] For the *provenance* graph, we mainly rely on the PROV Ontology (PROV-O) [17], except for the process behind the generation of the assertion, which is represented by the ECO ontology. [18] The type of evidence is represented using the Weighted Evidence (WI) vocabulary, [19] which includes the object property `evidence`

[12] http://gda.dei.unipd.it/cecore/resource/GCS#15f9bcc03f8af701bc82ba1437f0b451

[13] http://purl.obolibrary.org/obo/ECO_0000203

[14] http://purl.obolibrary.org/obo/ECO_0000212

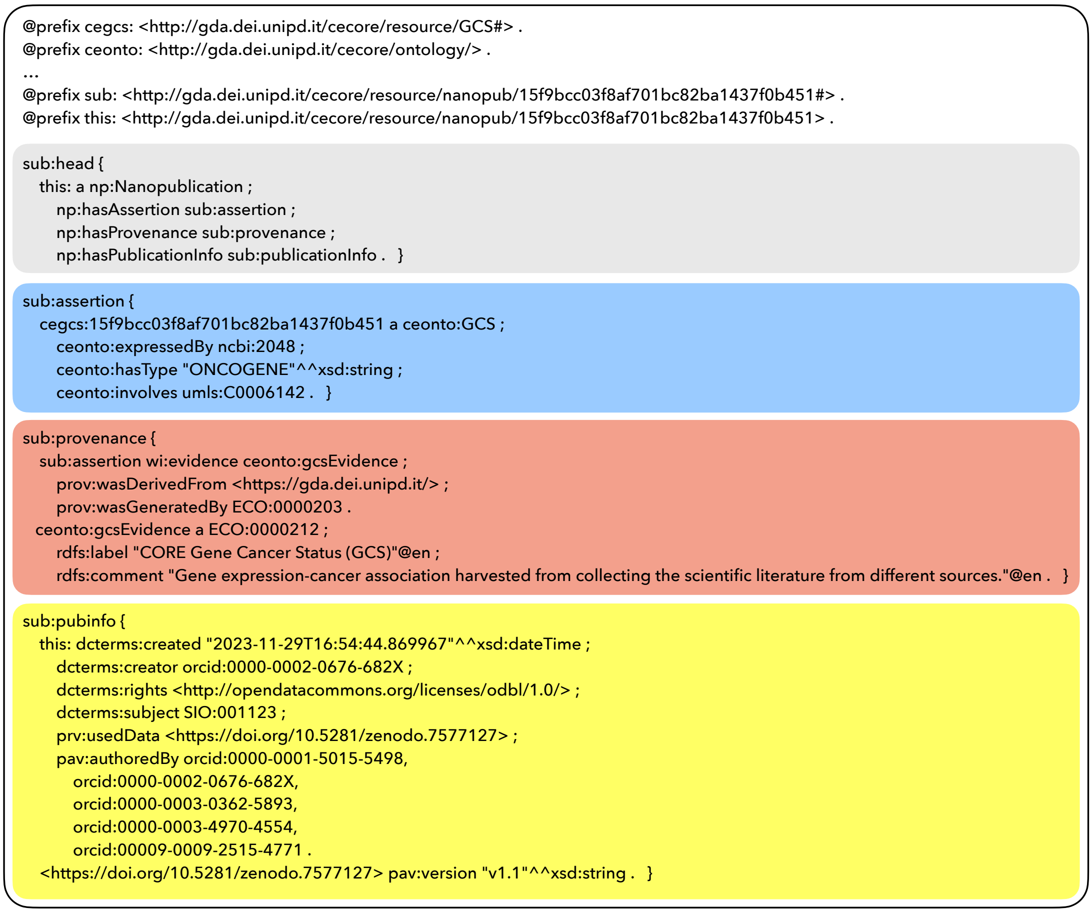
[15] http://semanticscience.org/resource/SIO_001123

[16] http://gda.dei.unipd.it/cecore/ontology/

[17] http://www.w3.org/TR/prov-o/

[18] https://ontobee.org/ontology/ECO

[19] http://www.evidenceontology.org/



```
@prefix cegcs: <http://gda.dei.unipd.it/cecore/resource/GCS#> .
@prefix ceonto: <http://gda.dei.unipd.it/cecore/ontology/> .
...
@prefix sub: <http://gda.dei.unipd.it/cecore/resource/nanopub/15f9bcc03f8af701bc82ba1437f0b451#> .
@prefix this: <http://gda.dei.unipd.it/cecore/resource/nanopub/15f9bcc03f8af701bc82ba1437f0b451> .

sub:head {
  this: a np:Nanopublication ;
    np:hasAssertion sub:assertion ;
    np:hasProvenance sub:provenance ;
    np:hasPublicationInfo sub:publicationInfo .  }

sub:assertion {
  cegcs:15f9bcc03f8af701bc82ba1437f0b451 a ceonto:GCS ;
    ceonto:expressedBy ncbi:2048 ;
    ceonto:hasType "ONCOGENE"^^xsd:string ;
    ceonto:involves umls:C0006142 .  }

sub:provenance {
  sub:assertion wi:evidence ceonto:gcsEvidence ;
    prov:wasDerivedFrom <https://gda.dei.unipd.it/> ;
    prov:wasGeneratedBy ECO:0000203 .
  ceonto:gcsEvidence a ECO:0000212 ;
    rdfs:label "CORE Gene Cancer Status (GCS)"@en ;
    rdfs:comment "Gene expression-cancer association harvested from collecting the scientific literature from different sources."@en .  }

sub:pubinfo {
  this: dcterms:created "2023-11-29T16:54:44.869967"^^xsd:dateTime ;
    dcterms:creator orcid:0000-0002-0676-682X ;
    dcterms:rights <http://opendatacommons.org/licenses/odbl/1.0/> ;
    dcterms:subject SIO:001123 ;
    prv:usedData <https://doi.org/10.5281/zenodo.7577127> ;
    pav:authoredBy orcid:0000-0001-5015-5498,
      orcid:0000-0002-0676-682X,
      orcid:0000-0003-0362-5893,
      orcid:0000-0003-4970-4554,
      orcid:00009-0009-2515-4771 .
  <https://doi.org/10.5281/zenodo.7577127> pav:version "v1.1"^^xsd:string .  }
```

Figure 2: A nanopublication representing GCS `15f9bcc03f8af701bc82ba1437f0b451` from CoreKB serialized in TriG format. Different colours identify the components of the nanopublication: grey for the head graph, blue for the assertion graph, red for the provenance graph, and yellow for the publication info graph.

to link the assertion, and ECO for the class of evidence. The *publication information* graph exploits the Provenance, Authoring, and Versioning (PAV) vocabulary [19] for authorship and versioning, the Provenance Vocabulary Core ontology Specification (PRV) [20] for the description of the used datasets, and the SIO ontology [20] for the topic of the nanopublications.

From a technical viewpoint, we developed a Python package that exploits the nanopub library [21] and serialize facts generated by CORE as named graphs in TriG syntax. To build the nanopublications, we retrieved all facts from CoreKB and excluded those deemed unreliable due to insufficient evidence. Indeed, publishing facts with insufficient evidence as independent publications provides little to no information. Thus, we publish 197, 511 facts from CoreKB as

[20] https://ontobee.org/ontology/SIO
[21] https://github.com/fair-workflows/nanopub

nanopublications, regarding $107,830$ biomarkers, $35,821$ oncogenes, $12,521$ TSGs, and $41,339$ unreliable facts due to contrasting evidence. The nanopublications are available in Zenodo [21]. To ease the visualization and access to facts, we integrated the nanopublications into the CoreKB platform. For each GCS, one can access the corresponding nanopublication[22] and browse it by selecting the eye icon from the drop-down menu beside the claim (see point B in Figure 1). We also provide a download button (see point A in Figure 1) to download the nanopublication representing a GCS. In this way, one can either download all the nanopublications via the Zenodo repository or select only the ones of interest in the CoreKB platform.

5. Conclusion and Future Works

In this work, we applied the nanopublication model to publish more than 197K facts generated by a large-scale knowledge discovery platform, namely CORE. We integrated the nanopublications in the CoreKB platform so that they can be easily browsed. One can also download the nanopublications separately from CoreKB, or in bulk in Zenodo [21]. As a result, each statement generated by CORE is considered a separate publication, allowing the identification, access, and citation of individual gene expression-cancer associations.

Representing each GCS as a nanopublication allows us to track its provenance. As we showed in this work, the *provenance* graph describes how the assertion was derived, i.e., facts are generated by an automatic process embodied in the CORE system. However, the *provenance* graph does not represent the supporting or conflicting evidence of each fact nor its reliability score, excluding pivotal information regarding the creation process of assertions in CORE. Thus, future works could delve into the extension of the nanopublication model for adequately keeping track of each piece of evidence used to generate assertions derived from an aggregation of multiple resources.

Acknowledgments

This project has received funding from the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074.

References

[1] T. Hey, S. Tansley, K. Tolle, J. Gray, The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009. URL: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

[2] X. L. Dong, Generations of knowledge graphs: The crazy ideas and the business impact, Proc. VLDB Endow. 16 (2023) 4130–4137. URL: https://doi.org/10.14778/3611540.3611636.

[22] The nanopublication representing the example GCS of the paper can be accessed at: https://gda.dei.unipd.it/cecore/resource/nanopub/15f9bcc03f8af701bc82ba1437f0b451

[3] P. Groth, A. Gibson, J. Velterop, The anatomy of a nanopublication, Inf. Serv. Use 30 (2010) 51–56. URL: https://doi.org/10.3233/ISU-2010-0613.

[4] E. Fabris, T. Kuhn, G. Silvello, A framework for citing nanopublications, in: Proc. of the Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, volume 11799 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 70–83. URL: https://doi.org/10.1007/978-3-030-30760-8_6.

[5] 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, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. ’t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. URL: https://doi.org/10.1038/sdata.2016.18.

[6] S. Marchesin, L. Menotti, F. Giachelle, G. Silvello, O. Alonso, Building a large gene expression-cancer knowledge base with limited human annotations, Database J. Biol. Databases Curation 2023 (2023). URL: https://doi.org/10.1093/database/baad061.

[7] X. Li, J. L. Warner, A Review of Precision Oncology Knowledgebases for Determining the Clinical Actionability of Genetic Variants, Front. Cell Dev. Biol. 8 (2020). doi:10.3389/fcell.2020.00048.

[8] B. Neary, J. Zhou, P. Qiu, Identifying gene expression patterns associated with drug-specific survival in cancer patients, Scientific Reports 11 (2021) 5004. URL: https://doi.org/10.1038/s41598-021-84211-y.

[9] F. Giachelle, S. Marchesin, G. Silvello, O. Alonso, Searching for reliable facts over a medical knowledge base, in: Proc. of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 3205–3209. URL: https://doi.org/10.1145/3539618.3591822.

[10] N. Queralt-Rosinach, T. Kuhn, C. Chichester, M. Dumontier, F. Sanz, L. I. Furlong, Publishing disgenet as nanopublications, Semantic Web 7 (2016) 519–528. URL: https://doi.org/10.3233/SW-150189. doi:10.3233/SW-150189.

[11] C. Chichester, P. Gaudet, O. Karch, P. 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. URL: https://doi.org/10.1016/j.websem.2014.05.001.

[12] C. Chichester, O. Karch, P. Gaudet, L. Lane, B. Mons, A. Bairoch, Converting neXtProt into Linked Data and nanopublications, Semantic Web 6 (2015) 147–153. URL: https://doi.org/10.3233/SW-140149.

[13] A. Waagmeester, M. Kutmon, A. Riutta, R. A. Miller, E. L. Willighagen, C. T. A. Evelo, A. R. Pico, Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources, PLoS Comput. Biol. 12 (2016). URL: https://doi.org/10.1371/journal.pcbi.1004989.

[14] T. Clark, P. Ciccarese, C. A. Goble, Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications, J. Biomed. Semant. 5 (2014) 28. URL: https://doi.org/10.1186/2041-1480-5-28.

[15] T. Kuhn, P. E. Barbano, M. L. Nagy, M. Krauthammer, Broadening the scope of nanopublications, in: Proc. of The Semantic Web: Semantics and Big Data (ESWC 2013), volume 7882 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 487–501. URL: https://doi.org/10.1007/978-3-642-38288-8_33.

[16] 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: Proc. of the 14th IEEE International Conference on e-Science, e-Science 2018, Amsterdam, The Netherlands, October 29 - November 1, 2018, IEEE Computer Society, 2018, pp. 83–92. URL: https://doi.org/10.1109/eScience.2018.00024.

[17] B. Mons, H. van Haagen, C. Chichester, P.-B. t. Hoen, J. T. den Dunnen, G. van Ommen, E. van Mulligen, B. Singh, R. Hooft, M. Roos, J. Hammond, B. Kiesel, B. Giardine, J. Velterop, P. Groth, E. Schultes, The value of data, Nature Genetics 43 (2011) 281–283. URL: https://doi.org/10.1038/ng0411-281.

[18] S. Marchesin, L. Menotti, G. Silvello, O. Alonso, CORE: Gene Expression-Cancer Knowledge Base, Zenodo, 2023. URL: https://doi.org/10.5281/zenodo.7577127.

[19] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A. J. G. Gray, C. A. Goble, T. Clark, PAV ontology: provenance, authoring and versioning, J. Biomed. Semant. 4 (2013) 37. URL: https://doi.org/10.1186/2041-1480-4-37.

[20] O. Hartig, J. Zhao, Publishing and consuming provenance metadata on the web of linked data, in: Proc. of the Provenance and Annotation of Data and Processes - Third International Provenance and Annotation Workshop, IPAW 2010, Troy, NY, USA, June 15-16, 2010. Revised Selected Papers, volume 6378 of *Lecture Notes in Computer Science*, Springer, 2010, pp. 78–90. URL: https://doi.org/10.1007/978-3-642-17819-1_10.

[21] F. Giachelle, S. Marchesin, L. Menotti, G. Silvello, CoreKB Facts Serialized as Nanopublications, Zenodo, 2023. URL: https://doi.org/10.5281/zenodo.10409288.