Extending Nanopublications with Knowledge Provenance for Multi-Source Scientific Assertions

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

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

Abstract

Nanopublications are RDF graphs that enable the possibility of sharing machine-readable assertions on the Web while tracking their provenance and publication information. However, the current nanopublication model focuses on the provenance of single-source assertions derived from a specific publication or database. This work proposes extending the nanopublication model to include a fourth component called *knowledge provenance*. Knowledge provenance captures the context where an assertion is not derived from a single publication but from a body of knowledge that can comprehend supporting and conflicting pieces of evidence that we need to track and refer to. We apply the defined model to the facts generated by the Collaborative Oriented Relation Extraction (CORE) and published 197, 511 assertions in the form of extended nanopublications, allowing the identification, representation, access, and citation of individual gene expression-cancer associations.

Keywords

Nanopublications, Knowledge Provenance, Data Provenance, Knowledge Bases, Gene-Cancer Associations.

1. Introduction

Given the high volume of publications, scientific evidence is often extracted automatically and organized into Knowledge Bases (KBs), which are widely applicable because they are understandable to humans and machines [1, 2]. Making sure each statement is accessible on its own is key to create a unified resource that contains all the relevant knowledge in a specific field. This approach allows for easy retrieval, access, and reference to specific data points we mention. To accomplish this goal, the nanopublication model proves particularly effective as it enables the identification, representation, access, and citation of individual assertions [3, 4]. This model sees extensive use in representing statements, especially in the life science domain [5, 6, 7]. The structure of the nanopublication model consists of three named graphs, each containing information about the assertion, its provenance, and details about the nanopublication itself. While the nanopublication model is well-suited for single assertions originating from a single

IRCDL 2025: 21st conference on Information and Research science Connecting to Digital and Library science, February 20-21, 2025, Udine, 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)

source of evidence, the provenance graph in the nanopublication model specifically addresses what is termed "Information Provenance." However, when dealing with information derived from the aggregation of multiple sources, challenges arise. In such cases, supporting and conflicting evidence exists, and each assertion may potentially be associated with a reliability score computed in diverse ways. Hence, there is a need to keep track of the provenance of each piece of information used to generate assertions by systems exploiting an aggregation of multiple sources of evidence.

This limitation of the nanopublication model is more evident when considering the assertions of a large-scale knowledge discovery platform storing more than 230K gene expression-cancer associations: CoreKB [8]. [1] CoreKB stores facts generated by the CORE system, which undertakes biomedical literature and extracts detailed aspects from various evidence sources to produce scientific facts suitable for publication each as a Gene Cancer Status (GCS). Creating a GCS involves extracting information about a gene expression related to a specific disease from numerous sentences in multiple articles. Thus, supporting and conflicting evidence is aggregated to determine the most probable scientific assertion related to the pool of sentences [9]. Each fact generated by CORE can be published as a nanopublication following the standard model [10]. However, the *provenance* graph of the current nanopublication model does not provide enough information to link the GCS and the role of each evidence sentence from the literature. In this work, we extend the nanopublication model to account for multi-source assertions by introducing a novel component to its structure called *knowledge provenance*. The *knowledge provenance* graph describes which pieces of information contributed to support or confute the considered assertion. In addition, it includes information about the reliability of each assertion based on each source of information and its extraction process. It is important to note that in the proposed model, the components of the original nanopublication model remain unchanged, allowing for backward compatibility. We model *knowledge provenance* as an additional named graph of a nanopublication, defined according to an appropriate ontology. In this regard, we define the PROV-K ontology [2], a general resource to represent the provenance information of assertions derived from multiple sources of evidence. The PROV-K ontology is an integration of the PROV Ontology (PROV-O) and is grounded in the literature defining knowledge provenance [11, 12, 13]. To show the applicability of the proposed model, we serialize all facts in CoreKB as extended nanopublications. We published 197,511 extended nanopublications representing the facts in CoreKB, which can be browsed in the CoreKB platform and downloaded separately from the same platform or in bulk in Zenodo [14]. We also release the source code [3] for building the extended nanopublications, which can be used as a template for future applications on different resources.

The rest of this work is organized as follows. Section 2 introduces the original nanopublication model and describes previous efforts in data, information, and knowledge provenance. Section 4 presents an in-use and large-scale knowledge discovery platform highlighting the limitations of the current nanopublication model when dealing with multi-source assertions. Section 3 defines the extended nanopublication model accounting for knowledge provenance. Section

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

[2] Publicly available here: https://prov-k.dei.unipd.it/ontology/

[3] https://github.com/mntlra/knowledgeProvenance

4 describes the serialization of extended nanopublications starting from the facts in CoreKB. Section 5 draws some final remarks.

2. Related Work

The nanopublication model aims to facilitate the integration, exchange, accessibility and comprehension of scientific statements, and enable citations at the granularity of individual claims [3, 4]. Following this framework, a scientific publication can be divided into single statements or assertions, with each assertion encapsulated in a distinct nanopublication containing all pertinent information about that specific claim. Using Semantic Web technologies, the nanopublication model represents scientific claims in a distinctive, identifiable, citable, and reusable format. Representing data as nanopublications enhances data-intensive science and allows for fact discovery exploiting machine-readable information [15]. From a technical viewpoint, a nanopublication is a named graph that comprises three basic components; each represented as a named graph itself: (i) the *assertion* graph, containing the scientific assertion; (ii) the *provenance* graph, containing information about where the assertion comes from and how it has been defined; (iii) the *publication info* graph, containing all the metadata of the nanopublication, such as who curated it and when it was created. The components of the nanopublication are interconnected using a fourth graph called the *head* graph.

The nanopublication model has been used to represent statements from different fields, especially in the life science domain. Chichester et al. [16] created nanopublications from scientific facts associated with more than 38K proteins stored in the neXtProt database. [4] This approach showed that using the nanopublication model for the neXtProt database eases access to its information and can be a useful tool for expanding biological research [5]. Queralt-Rosinach et al. [7] published the contents of the DisGeNET database [5] as nanopublications to provide a Linked Data resource. Waagmeester et al., in [6], described their endeavors in converting WikiPathways, an online collaborative pathway resource, into nanopublications. [6] Overall, there are more than 10M nanopublications publicly accessible worldwide [17]. Concerning the aggregation of multiple nanopublications, Bucur et al. [18] proposed an approach where nanopublications representing snippets of scientific articles related to the same publication are interlinked, utilizing properties like *refersTo*. Albeit the unifying model proposed in [18] is relevant to our study, it still does not consider the reliability of an assertion and the supporting or conflicting relationships between pieces of information. The concept of nanopublications has already been expanded in [19]. Here, the *assertion* graph has been extended to account for English sentences representing textual scientific claims following a semantic scheme called AIDA (Atomic, Independent, Declarative, Absolute). However, we are interested in machine-readable representations, like the nanopublication model.

In the era of truth discovery algorithms and automatic information extraction, the nanopublication model fails to represent data reliability and the provenance of assertions constituted by an ensemble of contrasting and supporting evidence. In this regard, Clark et al. [20] formalized the

[4] https://www.nextprot.org/
[5] https://www.disgenet.org/rdf
[6] https://github.com/wikipathways/nanopublications

micropublication model, which represents empirical evidence beyond statement-based models like nanopublications. The proposed model offers a representation of biomedical evidence with particular interest in building claim networks and their lineage. Although related, the work by Clark et al. only targets the modeling of the biomedical communications ecosystem, including reproducibility and verifiability in research – which is out of scope for this study. Besides, the micropublication model represents the claim of a statement in textual form, as it happens also for AIDA nanopublications [19].

The Data–Information–Knowledge–Wisdom (DIKW) pyramid is a widely recognized model for representing information and knowledge within management systems [21]. It describes the processes involved in the data transformation, from a piece of data to the wisdom embedded in it. Each step adds value to the final results, starting from raw *Data*, where one can extract *Information*, to *Wisdom*, that is the application of *Knowledge* acquired from the information block. We establish a connection between the DIKW pyramid and provenance. In earlier studies, the initial level, known as *Data Provenance*, has received extensive attention within databases. Its primary emphasis lies in tracing the data lineage in response to a query [22, 23]. In this context, Provenance encompasses the origin and the pathway through which a specific piece of data was introduced into the given database. Over the years, various conceptualizations of provenance have been proposed and explored, such as "why-provenance," "where-provenance," and "how-provenance" [22, 23].

The second level concerns *Information Provenance*, which represents the provenance of assertions inferred from data. This is embedded in the provenance graph of the nanopublication model and has been studied in the context of the Semantic Web. Provenance on the Semantic Web comprises metadata representing the creation and publication of resources. The PROV Ontology (PROV-O) [7] provides a formal language to encode provenance information in a machine-readable format. It is based on the PROV Data Model (PROV-DM) [8] and the Open Provenance Model (OPM) [24]. While extensive in scope, the PROV-O models provenance as in the provenance graph of the nanopublication model; therefore, it lacks the representation of supporting and contradicting evidence and reliability scores.

The third level, called *Knowledge Provenance*, is the focus of this work, and it has been studied in different works by Fox and Huang [25, 12, 13, 26]. Knowledge Provenance (KP) has been proposed to create an approach to annotate the reliability of information extracted from web sources based on who created the assertion, how much the creator can be trusted, and what the information depends on. Little work has been done towards this end, and it mostly focuses on providing a taxonomy of four levels of provenance based on the certainty degree of each assertion [25]: *Static KP* (Level 1) for assertions for which the truth value does not change over time [25]; *Dynamic KP* (Level 2) allowing the validity of information to change over time [12]; *Uncertainty-oriented KP* (Level 3) considering truth values and relationships that are uncertain [13]; *Judgment-based KP* (Level 4) for provenance supported by social processes, e.g., truth propagation in social networks [26]. The *Static Knowledge Provenance Ontology* defines a taxonomy of proposition types and a set of axioms allowing the development of a reasoner to assess truth values based on different cases [11]. Although the ontology has been formalized

[7] http://www.w3.org/TR/2013/REC-prov-o-20130430/
[8] https://www.w3.org/TR/prov-dm/

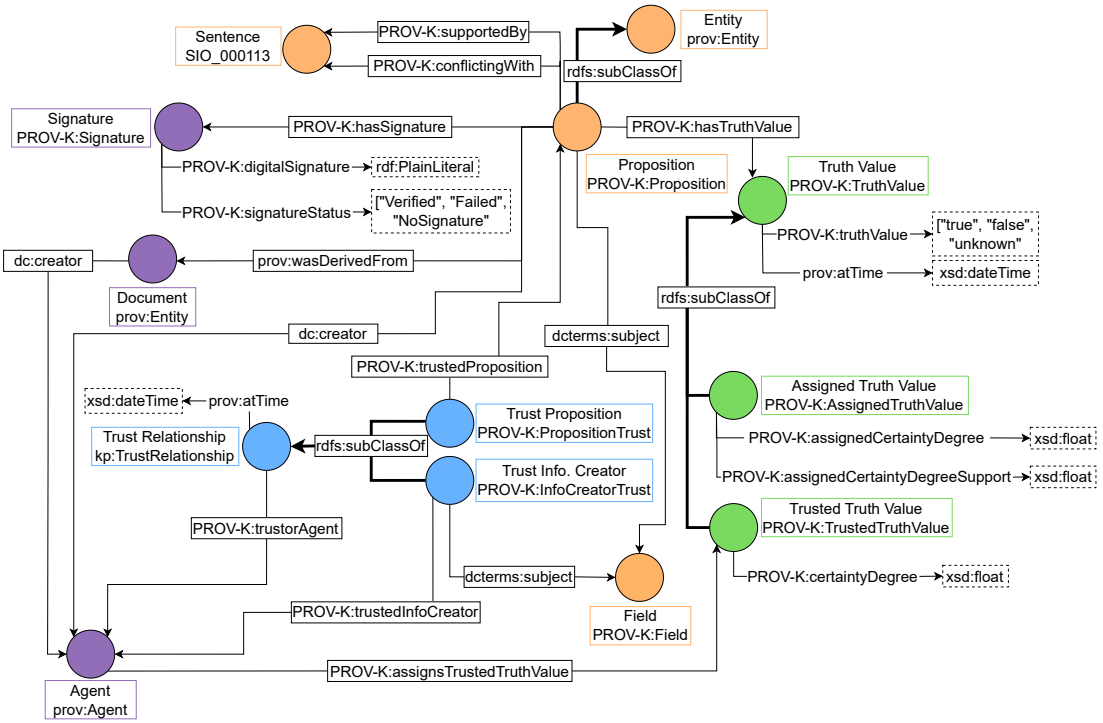

Figure 1: The PROV-K ontology. We divided the PROV-K ontology into four main areas: Propositions (displayed in orange), Digital Signature and Information Sources (depicted in purple), Trust Relationships (pictured in blue), and Truth Value (represented in green).

in [11], it is not available as a resource.

The fourth level, *Wisdom Provenance*, focuses on keeping track of the provenance of the wisdom inferred from knowledge or applications exploiting such knowledge, which is still unexplored in the literature.

3. Extended Nanopublication Model

We extend the original nanopublication model by introducing a novel component named *knowledge provenance*, which allows for the tracking of provenance for multi-sourced assertions. In this context, "knowledge provenance" is the nanopublication component describing which pieces of information contributed to support or confute the assertion. It can also include information about the reliability of the assertion and its assigned certainty degree. It is important to note that our approach does not change the other components of the original nanopublication model.

Since each module of a nanopublication is a named graph, we model knowledge provenance as a named graph according to an ontology called the PROV-K ontology. [9] We designed the

[9]The PROV-K ontology and its complete documentation are available at https://prov-k.dei.unipd.it/ontology/

PROV-K ontology by extending PROV-O [10] to represent provenance information of assertions derived from multiple sources using an aggregation algorithm. Note that the PROV-K ontology is a standalone resource, meaning it can also be used independently of nanopublications to represent provenance. We developed the PROV-K ontology following the guidelines provided by the Static KP ontology [11] with the addition of some elements from Dynamic KP, such as timestamps for truth values and trust relationships [12]. We also incorporate the concepts of *assigned certainty degree* and *certainty degree* from Uncertainty-oriented KP [13]. Representing truth values as probability distributions is more meaningful for our study than the static KP assumption, where the truth value is a categorical variable. We also expanded the concepts from Uncertainty-oriented KP to represent reliability testing. In this way, we can classify assertions into reliable or unreliable facts and include the conditions an assertion may fail to satisfy. We describe the PROV-K ontology based on four main areas: Propositions, Digital Signature and Information Sources, Trust Relationships, and Truth Value. Figure 1 reports the ontology schema.

Proposition. The central unit of the PROV-K ontology is the "Proposition", which is defined as "*the smallest piece of information to which provenance-related attributes may be ascribed*" and as "*a declarative sentence that is either true or false*" [11]. In our context, the nanopublication model's assertion graph can be considered the proposition. Since the PROV-K ontology extends PROV-O, we model propositions as subclasses of `prov:Entity`. We also define a taxonomy of propositions based on [11], which differentiates between independent and dependent propositions, i.e., assertions whose truth value depends upon other propositions. Each proposition can be supported by or conflicting with other knowledge sources. To encompass this situation, we defined two object properties called `supportedBy` and `conflictingWith`, both with range class "Sentence" from the Semanticscience Integrated Ontology (SIO). [11] Each proposition can be linked to one or more knowledge fields with the object property `dc:subject` from the Dublin Core (DC) Metadata Items. [12]

Information Source and Signature. To determine the provenance of a proposition, it may be useful to represent the document in which the considered proposition appears. We link the proposition to the "document" it belongs to with the object property `prov:wasDerivedFrom` from PROV-O. In this way, we allow a proposition to belong to any entity, e.g., a textual document or a dataset. For any document and any proposition, one can define its creator with the object property `dc:creator` from the DC Metadata Items, with range class `prov:Agent` from PROV-O. In this way, the creator of a proposition or a document may be any agent, e.g., a person or a digital artifact. We also represent the digital signature and signature status that can be assigned to a proposition. We apply the PROV-K ontology to model knowledge provenance in the context of nanopublications. However, the PROV-K ontology is a general resource designed to track provenance information for aggregated sources of evidence beyond the *knowledge provenance* graph of the nanopublication model. Thus, we defined the "Signature" class within

[10] https://www.w3.org/TR/prov-o/
[11] http://semanticscience.org/resource/SIO_000113
[12] http://purl.org/dc/terms/subject

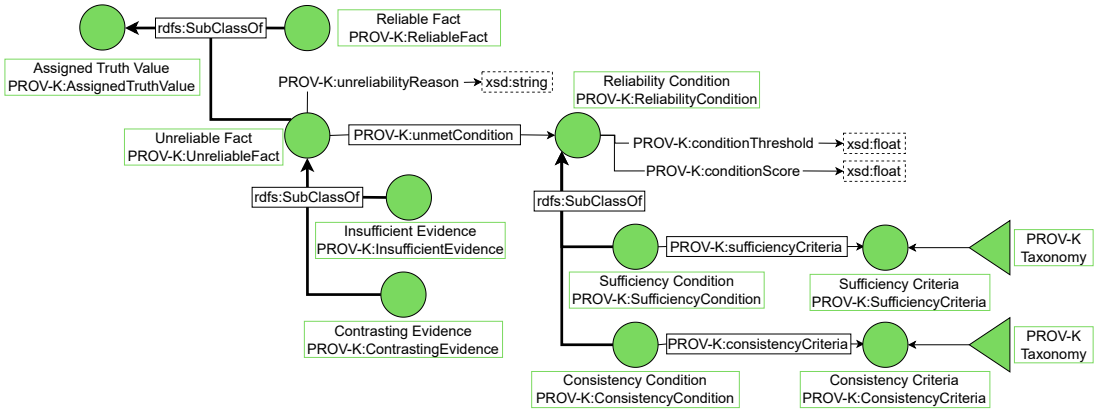

Figure 2: Assigned truth value modeling in the PROV-K ontology.

the ontology rather than solely relying on the modeling provided by the nanopublication model. Nevertheless, when we apply such an ontology to the nanopublication model, we can represent digital signatures with the "Nanopub Signature Element" [13] class from `nanopubx`.

Trust Relationships. We identify two trust relationships, one between two agents, which are referred to as provenance requester and information creator (class "InfoCreatorTrust"), and another between an agent and a proposition (class "PropositionTrust"). The former is defined as "*the provenance requester `a` "trusts" information creator `c` in a specific knowledge field `f`*", where "trust" means "*`a` believes any proposition created by `c` in field `f` to be true*". The latter takes the form of "*Proposition `x` is trusted by an agent `a`*" [11]. We also include the timestamp reporting when the trust relationship was issued using the data property `prov:atTime` from PROV-O. This work focuses on tracking the provenance of each piece of evidence and determining the reliability of a given fact. Nevertheless, the PROV-K ontology can be easily expanded to account for more complex trust relationships and decision processes.

Truth Values. Fox and Huang defined two types of truth values: the *assigned truth value* and the *trusted truth value* [11]. The former refers to the truth value assigned to the proposition by its creator. At the same time, the latter identifies the truth value evaluated by an external agent called "*provenance requester*". In Static KP, the truth value of a proposition can be "True", "False", or "Unknown". Thus, we link to each proposition the class `TruthValue` with object property `hasTruthValue`, where one can store both the assigned and trusted truth values defined in [11]. We also represent the timestamp reporting when the truth value was assigned or trusted with the data property `prov:atTime` from PROV-O. We define the *assigned certainty degree* as the probability that the proposition's creator assigns the truth value of "True" to the proposition and the *certainty degree* as the probability that an agent evaluates the trusted truth value as

[13] http://purl.org/nanopub/x/NanopubSignatureElement

"True" [13]. Based on the assigned certainty degree, we may classify propositions as reliable or unreliable. For this reason, we expand the concept of "assigned truth value" to account for reliability testing. We report the ontology schema for the assigned truth value in Figure 2. We classify the assigned truth value as "Reliable Fact" or "Unreliable Fact", where the latter identifies propositions failing some pre-defined reliability tests. One can represent whether the proposition is unreliable due to insufficient evidence (subclass `InsufficientEvidence`) or due to contrasting evidence (subclass `ContrastingEvidence`). We also include a data property called `unreliabilityReason` to describe why the proposition is deemed as "unreliable", as well as an object property called `unmetCondition` to specify the reliability conditions the proposition fails to respect. Each condition has a score, a threshold, and a criteria which describes how the reliability condition works. For instance, in CORE we have two sufficiency and one consistency criteria, which are modeled as named individuals of type `skos:Concept` and either `SufficiencyCriteria` or `ConsistencyCriteria`.

4. Real-World Use Case

CoreKB. CORE is a Knowledge Base Construction (KBC) system based on the combination of ML-based models and domain-expert feedback [9]. CORE harvests text from the literature, identifies sentences containing pairs of relevant entities, and extracts fine-grained aspects from them to generate gene expression-cancer associations that can be published as facts – i.e., GCS. For each fact, CORE combines three aspect probabilities to assign the gene class likelihood to the three mutually exclusive gene classes: `oncogene`, `tumor suppressor gene`, and `biomarker`. Then, the system performs a two-stage reliability test that, for each fact, first verifies that the fact has sufficient evidence and subsequently 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 fed back to domain experts for manual annotations in an active learning paradigm, making CORE suitable to iterative KB versioning. For technical details and the evaluation of the CORE system, we resort the interested reader to [9].

The data extracted by CORE, and then ingested by CoreKB, contains information about 23,879 genes and 11,530 diseases for a total of more than 230K fine-grained facts supported by 1,037,845 sentences from 251,038 research articles. Figure 3 shows a GCS card displayed by the CoreKB platform Each GCS comprises information about the gene and disease involved, which are identified by National Center for Biotechnology Information (NCBI) Gene IDs and Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) respectively, together with the assigned gene class. In addition, each GCS is linked to the sentences supporting the fact, i.e. identifying the same gene class, and those conflicting with it. For each sentence, provenance information includes the PubMed ID of the article from which the sentence has been extracted and the year of publication of such an article. CoreKB comprises three types of GCS: reliable facts, unreliable facts due to insufficient evidence, and unreliable facts due to low consensus (contrasting evidence). The former are facts that passed the reliability tests performed by CORE, while the others are facts that failed any of the two checks performed by the testing component.

Facts generated by CORE can be published as nanopublications since each GCS can be viewed as an assertion graph on its own. Giachelle et al. [10] showed the serialization of a

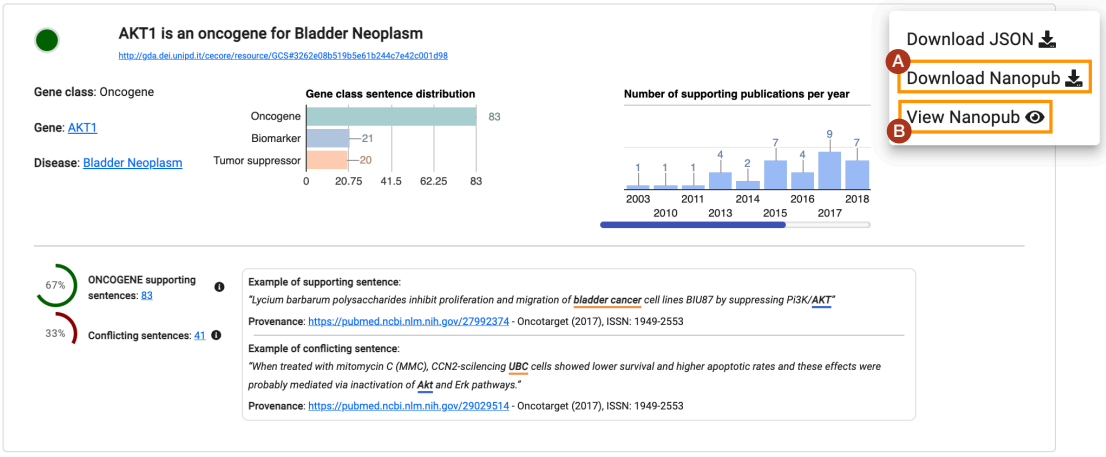

Figure 3: Landing page for the GCS `3262e08b519b5e61b244c7e42c001d98` displayed by CoreKB platform. Each GCS card comprises information about the gene and cancer labels, the gene class and its distribution across the associated publications, and statistics about the number of supporting and conflicting evidence. Each GCS can be downloaded in JSON format or serialized in TriG syntax as an extended nanopublication (A). The extended nanopublication representation a GCS can also be visualized (B).

GCS in CoreKB following the standard nanopublication model. In summary, the assertion graph comprises the GCS itself in the form of an Resource Description Framework (RDF) graph, the publication info graph includes metadata about the nanopublication, and the provenance graph describes how the assertion was derived. However, by publishing facts within CoreKB as classic nanopublications we cannot represent supporting and conflicting sentences and embed information about their reliability. For instance, consider the GCS in Figure 3, the standard nanopublication model fails to represent that the assertion is derived from the ensemble of 83 supporting sentences and 41 conflicting sentences. Moreover, it cannot include information about the reliability of such an assertion, i.e., it cannot convey the information that the GCS is reliable as the probability that gene "AKT1" is an oncogene for "bladder neoplasm" is 0.82.

Creation of the extended nanopublications. The extended nanopublication model has been applied to serialize all the facts in CoreKB as nanopublications with knowledge provenance. We followed the same methodology for the original nanopublication components used for the DisGeNET nanopublications [7]. Figure 4 shows a GCS modeled with the extended nanopublication model and serialized in TriG format. [14]

A nanopublication representing the facts generated by CORE comprises five named graphs: *head*, *assertion*, *provenance*, *publication information*, and *knowledge provenance*. The *head* graph connects all the components by linking the nanopublication URI to its subgraphs. The *assertion* graph contains the GCS in RDF format. The *provenance* graph includes the information prove-

[14] Access at:
https://gda.dei.unipd.it/cecore/resource/nanopub/3262e08b519b5e61b244c7e42c001d98/.

Figure 4: An extended nanopublication representing the fact from CoreKB. Due to space reasons, we only report the head graph (in grey) and the knowledge provenance, *knowledgeProv*, graph (in green). The classic components of the nanopublication are described in detail in [10].

nance and evidence used to build the GCS. In our case, all facts are derived from CoreKB and are generated automatically (class "Automatic Assertion"). [15] Since the facts generated by CORE often integrate more than one source of evidence (i.e., sentences), the source evidence is an instance of class "Combinatorial Evidence" from the Evidence and Conclusion Ontology (ECO). [16] The *publication information* graph includes the general topic of the nanopublications, information about the authors of the nanopublications, and the used dataset. Since CoreKB comprises fine-grained gene expression-cancer associations, the general topic for all nanopublications is "gene-disease association linked with altered gene expression" from SIO. [17] The *knowledge provenance* graph includes all supporting and conflicting sentences and information about the reliability of the GCS. For instance, the GCS in Figure 4 is a reliable fact. Hence, its truth value

[15] http://purl.obolibrary.org/obo/ECO_0000203
[16] http://purl.obolibrary.org/obo/ECO_0000212
[17] http://semanticscience.org/resource/SIO_001123

Table 1
Gene class distribution of the facts in CoreKB serialized as extended nanopublications.

Class	# of Nanopublications
Biomarker	107,830
Oncogene	35,821
Tumor Suppressor Gene	12,521
Contrasting Evidence	41,339

is an instance of class “ReliableFact" and we report its assigned certainty degree and support.

To represent the *assertion* graph, we rely on the ontology underlying KBs generated by CORE, [18] while for the *provenance* graph we rely on PROV-O. [19] For the authorship and versioning, we employ the Provenance, Authoring, and Versioning (PAV) vocabulary [27], and for the description of the used datasets we employ the Provenance Vocabulary Core ontology Specification (PRV) [28]. The evidence annotation is described using the Weighted Evidence (WI) vocabulary, [20] which comprises the object property `wi:evidence` to link the assertion to its evidence, and the ECO ontology. [21] For the description of the topic of the nanopublications and the process used to build the assertion, we use the SIO ontology. [22] For the knowledge provenance graph, we extended the PROV-K ontology to include the reliability condition defined by the CORE system. Specifically, we added two sufficiency criteria and one consistency criterion. A fact generated by CORE passes the sufficiency checks if the probability of Change of Cancer Status (CCS) and Gene-Cancer Interaction (GCI) being not informative is below a threshold value α set to 0.7. The consistency test instead checks whether the difference between the probabilities of the fact being classified with the two gene classes with the highest likelihood is bigger than a threshold value β set to 0.4.

To build the extended nanopublications, we extended the Python package `nanopub`. [23] We kept the provenance, publication information, and assertion graph unchanged to provide backward compatibility with the original nanopublication model. In addition, we developed a Python package to publish the facts in CoreKB as extended nanopublications serialized in TriG syntax. The code can take as input two CSV files comprising the facts and the sentences supporting or conflicting with it, or one can provide a Turtle (.ttl) file comprising the CoreKB dump available in Zenodo [29]. The code for serializing the facts in CoreKB as extended nanopublications can also serve as a template for future applications on different resources.

CoreKB comprises 231,099 GCS, which can be divided into reliable facts, unreliable due to insufficient evidence, and unreliable due to low consensus. We filter out unreliable facts due to insufficient evidence, as publishing them as independent publications provides little to no information. As a result, we published 197,511 facts from CoreKB as extended nanopublications, accounting for 156,172 reliable facts and 41,339 unreliable ones due to contrasting

[18] http://gda.dei.unipd.it/cecore/ontology/
[19] http://www.w3.org/TR/prov-o/
[20] http://www.evidenceontology.org/
[21] https://ontobee.org/ontology/ECO
[22] https://ontobee.org/ontology/SIO
[23] https://github.com/fair-workflows/nanopub

evidence. Table 1 shows the gene class distribution of the facts in CoreKB serialized as extended nanopublications. The extended nanopublications are also available in Zenodo [14].

We include serialized nanopublications into the CoreKB platform to ease facts visualization. For each GCS, one can explore the serialized nanopublication by clicking on the eye icon placed in the drop-down list on the right side of the claim (see point B in Figure 3). [24] The visualization depicts each component with a different color and displays URIs redirected to a functioning website containing the description of the considered element. One can also download the extended nanopublication representing a specific GCS thanks to the download button (see point A in Figure 3).

5. Final Remarks

This work extends the current nanopublication model to include a novel component called *knowledge provenance*, accounting for the provenance information of assertions derived from the aggregation of multiple sources of evidence. We described *knowledge provenance* as a named graph tracking the provenance of each piece of information that contributes to support or confute the assertion. To support the semantics of the *knowledge provenance* graph, we designed the PROV-K ontology, an integration of PROV-O representing provenance information of assertions derived from the aggregation of multiple sources. The PROV-K ontology is a general resource designed to track provenance information for aggregated sources of evidence beyond the context of nanopublications. We applied the proposed model by serializing more than 197K facts in CoreKB and publishing them as extended nanopublications. Such nanopublications can be easily browsed and downloaded through the CoreKB platform. The serialization of facts in CoreKB can also serve as a template for applying the extended nanopublication model on different 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] G. Weikum, X. L. Dong, S. Razniewski, F. M. Suchanek, Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases, Found. Trends Databases 10 (2021) 108–490. URL: https://doi.org/10.1561/1900000064.

[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.

[24] The serialized nanopublication representing the GCS used as an example throughout the paper can be visualized at:
https://gda.dei.unipd.it/cecore/resource/nanopub/3262e08b519b5e61b244c7e42c001d98/

[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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] F. Giachelle, S. Marchesin, L. Menotti, G. Silvello, Publishing CoreKB Facts as Nanopublications, in: Proc. of the 20th conference on Information and Research science Connecting to Digital and Library science (IRCDL 2024), volume 3643 of *CEUR-WS Proceedings*, CEUR, 2024, pp. 16–24. URL: https://ceur-ws.org/Vol-3643/paper2.pdf.

[11] M. S. Fox, J. Huang, An ontology for static knowledge provenance, in: Proc. of the Knowledge Sharing in the Integrated Enterprise - Interoperability Strategies for the Enterprise Architect, 2004 International Conference on Enterprise Integration and Modelling Technology, ICEIMT 2004, The 7th International Conference on Design of Information Infrastructure Systems for Manufacturing, DIISM 2004, 9-11 October 2004, University of Toronto, Canada, volume 183 of *IFIP*, Springer, 2004, pp. 203–213. URL: https://doi.org/10.1007/0-387-29766-9_17.

[12] J. Huang, M. S. Fox, Dynamic knowledge provenance, in: Proc. of the Business Agents and Semantic Web Workshop, National Research Council of Canada, 2004, pp. 372–387. URL: http://www.eil.utoronto.ca/wp-content/uploads/km/papers/huang-nrc04.pdf.

[13] J. Huang, M. S. Fox, Uncertainty in Knowledge Provenance, in: Proc. of The Semantic Web: Research and Applications, First European Semantic Web Symposium, ESWS 2004, Heraklion, Crete, Greece, May 10-12, 2004, volume 3053 of *Lecture Notes in Computer Science*, Springer, 2004, pp. 372–387. URL: https://doi.org/10.1007/978-3-540-25956-5_26.

[14] F. Giachelle, S. Marchesin, L. Menotti, G. Silvello, CORE Extended Nanopublications,

Zenodo, 2023. URL: https://zenodo.org/records/10392177.
[15] 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.
[16] 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.
[17] 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.
[18] C.-I. Bucur, T. Kuhn, D. Ceolin, A unified nanopublication model for effective and user-friendly access to the elements of scientific publishing, in: Proc. of Knowledge Engineering and Knowledge Management (EKAW 2020), volume 12387 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 104–119. URL: https://doi.org/10.1007/978-3-030-61244-3_7.
[19] 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.
[20] 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.
[21] J. E. Rowley, The wisdom hierarchy: representations of the DIKW hierarchy, J. Inf. Sci. 33 (2007) 163–180. URL: https://doi.org/10.1177/0165551506070706.
[22] P. Buneman, S. Khanna, W. C. Tan, Why and where: A characterization of data provenance, in: Proc. of the Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, volume 1973 of *Lecture Notes in Computer Science*, Springer, 2001, pp. 316–330. URL: https://doi.org/10.1007/3-540-44503-X_20.
[23] J. Cheney, L. Chiticariu, W. C. Tan, Provenance in databases: Why, how, and where, Found. Trends Databases 1 (2009) 379–474. URL: https://doi.org/10.1561/1900000006.
[24] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. G. Stephan, J. V. den Bussche, The Open Provenance Model core specification (v1.1), Future Gener. Comput. Syst. 27 (2011) 743–756. URL: https://doi.org/10.1016/j.future.2010.07.005.
[25] M. S. Fox, J. Huang, Knowledge Provenance, in: Proc. of the Advances in Artificial Intelligence, 17th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2004, London, Ontario, Canada, May 17-19, 2004, volume 3060 of *Lecture Notes in Computer Science*, Springer, 2004, pp. 517–523. URL: https://doi.org/10.1007/978-3-540-24840-8_47.
[26] J. Huang, M. S. Fox, Trust Judgment in Knowledge Provenance, in: Proc. of the 16th International Workshop on Database and Expert Systems Applications (DEXA 2005), 22-26

August 2005, Copenhagen, Denmark, IEEE Computer Society, 2005, pp. 524–528. URL: https://doi.org/10.1109/DEXA.2005.193.

[27] 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.

[28] 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.

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