International Journal of Metadata, Semantics and Ontologies

ISSN online: 1744-263X - ISSN print: 1744-2621
https://www.inderscience.com/ijmso

Nano-PROV: FAIRification workflow for generating nanopublications based on provenance and semantic enrichment

Matheus Pedra Puime Feijoó, Rodrigo Jardim, Sergio Manuel Serra da Cruz, Maria Luiza Machado Campos

DOI: 10.1504/IJMSO.2023.10060808

Article History:

Received:	29 July 2022
Last revised:	28 March 2023
Accepted:	09 April 2023
Published online:	05 December 2023

Nano-PROV: FAIRification workflow for generating nanopublications based on provenance and semantic enrichment

Matheus Pedra Puime Feijoó*

Instituto de Computação,
Universidade Federal do Rio de Janeiro,
Rio de Janeiro, Brazil
and
Tecnun, Engineering School,
University of Navarra,
San Sebastian, Pamplone, Spain
Email: mpuime@unav.es
*Corresponding author

Rodrigo Jardim

Laboratório de Biologia Computacional e Sistemas,
Instituto Oswaldo Cruz (FIOCRUZ),
Rio de Janeiro, Brazil
Email: jardim@ioc.fiocruz.br

Sergio Manuel Serra da Cruz

Department of Computer Science,
Federal Rural University of Rio de Janeiro,
Rio de Janeiro, Brazil
and
Instituto de Computação,
Universidade Federal do Rio de Janeiro,
Rio de Janeiro, Brazil
Email: serra@ppgi.ufrj.br

Maria Luiza Machado Campos

Instituto de Computação,
Universidade Federal do Rio de Janeiro,
Rio de Janeiro, Brazil
Email: mluiza@ppgi.ufrj.br

Abstract: Providing research data to be readable, accurate and understandable by human and autonomous computational agents is challenging, primarily if published on the web. We present Nano-PROV, a workflow-based approach that aims at semantic enrichment of data and provenance control of published research data sets. The workflow uses the nanopublications for data transformation, a reliable format for dynamically publishing research outputs. Further, Nano-PROV adopts UN-PROV, a unified provenance guideline centred on nanopublication for identifying and controlling data and workflow provenance. In this paper, we developed computational experiments to evaluate the workflow by generating a nano-pub data model based on the genomic scenario, showing how the proposal may circumvent various issues regarded with data reusability, interoperability, and discoverability issues. Compared with related works, our results demonstrated the feasibility of Nano-PROV to enhance the semantic expressivity of research data and its metadata annotations.

Keywords: nanopublication; FAIR principles; FAIRification; data provenance; semantic web; metadata; research data management; ontologies; semantic enrichment.

Reference to this paper should be made as follows: Feijoó, M.P.P., Jardim, R., Serra da Cruz, S.M. and Campos, M.L.M. (2023) 'Nano-PROV: FAIRification workflow for generating nanopublications based on provenance and semantic enrichment', *Int. J. Metadata Semantics and Ontologies*, Vol. 16, No. 2, pp.138–151.

Biographical notes: Matheus Pedra Puime Feijoó received his MSc degree in Computer Science from the Federal Universidade Federal do Rio de Janeiro. He is also a PhD candidate at Tecnun – Engineering School University of Navarra. He works as a Researcher at the Knowledge Engineering (GRECO Group) and the Virus Outbreak Data Network Brazil (VODAN-BR Group). His research interests include data science, focusing on conceptual modelling, machine learning, databases, semantic web, linked data, FAIR data principles, bioinformatics and data provenance.

Rodrigo Jardim received his MSc degree in Geomatics from the University of the State of Rio de Janeiro and PhD degree in Computational and Systems Biology from the Oswaldo Cruz Foundation. Currently, he is working at the Oswaldo Cruz Foundation, Laboratory of Computational and Systems Biology, Instituto Oswaldo Cruz. Has experience in the area of Software Engineering with an emphasis in Bioinformatics, mainly in the following topics: open source, databases, cloud computing, comparative genomics and drug repositioning.

Sergio Manuel Serra da Cruz received his MSc degree in Computer Science from the Universidade Federal do Rio de Janeiro and PhD degree in Computer Science from the Universidade Federal do Rio de Janeiro. Currently, he is working as a Professor and Researcher at the Federal University of Rio de Janeiro and the Federal Rural University of Rio de Janeiro. Has experience in Computer Science, acting on the following subjects: database, scientific workflows, consultancy, provenance and educational data mining.

Maria Luiza Machado Campos received her MSc degree in Computer Engineering from the Federal University of Rio de Janeiro and PhD degree in Information Systems from the University of East Anglia. Currently, she is working as a Professor and Researcher at the Department of Computer Science, Federal University of Rio de Janeiro, co-coordinates the Knowledge Engineering Group (GRECO Group), and supervising Master's and Doctorate students at the Informatics Post-graduate Program at the same university. Her research interests include heterogeneous information integration, focusing on metadata, ontologies, conceptual modelling, databases, data warehousing and the semantic web.

This paper is a revised and expanded version of a paper entitled 'GAP: Enhancing Semantic Interoperability of Genomic Datasets and Provenance Through Nanopublications' presented at the '15th International Conference on Metadata and Semantic Research', Madrid, 2021.

1 Introduction

Reusing research data presents opportunities and challenges for individual researchers, their organisations, and the research community. Interdisciplinary scientific data-centric environments generate vast amounts of data, providing new possibilities for analysis. However, research faces barriers when extracting knowledge from large data sets.

These barriers are related to cross-domain complexity and the shortage of common governance models (Demchenko et al., 2012). Further, when individual researchers attempt to reuse data, the obstacles reach new levels. For instance, when searching and retrieving data sets, they face semantic issues related to redundant, inconsistent, inaccurate, incomplete and even obsolete records (Fan, 2015).

Poor (or absent) semantics models, non-interoperable (or non-integrated) research data environments and the lack of curation, transparency and provenance trails are the dilemmas experienced when trying to reuse published data (Demchenko et al., 2012). Providing accurate annotations on how data and metadata can be interoperated or analysed with other sources is crucial for expanding knowledge and data discoverability. The drawbacks increase if the reusers are autonomous agents who require an upper level of data management, chiefly reusable and understandable by these agents (Wilkinson et al., 2016).

If we consider semantic aware scenarios, Linked Data (LD) is one of the most accepted formalisms for defining data sets as Digital Objects (DOs); the data science community has defended LD as a foundation for providing machine-understandable (meta)data (Bizer et al., 2009). Similarly, the FAIR data principles proposed by Wilkinson et al. (2016) improve data findability, accessibility, interoperability and reusability. The principles increase the research data value, making them more easily discoverable, accessible, interoperable and reusable by any agent.

The research community has been developing several approaches to comply with these principles independently of the scientific domain. We highlight that the process of making research data FAIR (a.k.a FAIRification) is quickly gaining momentum in the research community. It is a workflow-based transformation process to control data and metadata management and assist data owners in transforming digital research outputs into FAIR DOs (Jacobsen et al., 2020). A broad spectrum of FAIRification workflows has singular characteristics related to specific research domains, tools, and techniques (Feijoó et al., 2022).

Nanopublication (NP) also generates DOs following FAIR data principles (Sustkova et al., 2020). NP can store minimal statements in a well-defined DO. These statements, also called assertions, can represent any scientific output information. NP provides a dynamic and reliable mechanism to store data in multipurpose domains.

Another benefit of NP is provenance control. Part of the NP schema stores the provenance of the assertion and the provenance of nanopublication DO generation. Despite that, Asif et al. (2019) investigated the published NP and noticed a shortage of data provenance in the NPs. This issue relates to data owners' lack of provenance awareness and the absence of specifications or guidelines for achieving an ideal NP provenance environment.

To mitigate the concerns, we propose the Nano-PROV FAIRification workflow. Our workflow-based method considers a boost in data reusability and knowledge discoverability to assist the development of novel semantic data models. The Nano-PROV FAIRification provides a guideline for data owners to achieve their FAIR objectives based on well-known converging formalisms such as LD, DO, NP, FAIR data principles, data provenance and semantic interoperability.

Unlike previous related works, our innovative study uses the NP definition to convert unstructured (meta)data (with poor semantics) into DOs, following the FAIR data principles. Additionally, we provide a unified NP Provenance (UN-PROV) guideline for ensuring a systematic and reliable mechanism to gather retrospective provenance metadata (Da Serra et al., 2009) by implementing provenance controls and ontologies like the W3C PROV data model (Moreau et al., 2015) and other standardised recommendations.

This manuscript is an extended version of our previous work (Feijoó et al., 2022). We introduce the FAIRification workflow and show a real-world use case to compose the Nanopublication Genome data model (NGen-DM). Furthermore, we present a guideline for controlling and generating machine-readable retrospective provenance metadata following the NP definition.

The paper is organised as follows. Section 2 includes the background and related work. Section 3 presents the Nano-PROV FAIRification, whose 12 steps are grouped into three categories (*pre-FAIRification*, *FAIRification* and *post-FAIRification steps*). Further, Section 3 also presents the UN-PROV. Section 4 presents the experiments and discusses the FAIRification application in a data-centric genomic scenario. Finally, Section 5 presents the conclusion, limitations and future works.

2 Background

2.1 Understandable data semantics environment

As highlighted by Roche et al. (2015) and Mons et al. (2017), reusers make a strenuous effort to find, access, interoperate and reuse data and metadata. According to the authors, about 80% of data are considered re-useless due to the absence of a Persistent Identifier (PID), descriptors, provenance annotations, licences and explicit semantic terms.

The FAIR guiding principles are a general guideline for improving (meta)data management and curation practices, focusing either on human-driven or machine-driven activities (Wilkinson et al., 2016). The GO FAIR working group introduced the FAIRification workflow-based methodology to promote a deeper understanding of Dos' unstructured and poor semantic data transformations (GO FAIR, 2019).

The FAIRification is a multistep process that data owners can adopt for controlling and adapting (meta)data to be FAIRer. The process considers the owners' premises, the data and the technologies. Today, there are several FAIRification proposals (Jacobsen et al., 2020; Sinaci et al., 2020; Kersloot et al., 2021; Groenen et al., 2021).

Nanopublication (NP) is another innovative (meta)data formalism for tackling data reusability and interoperability issues. Even though its development is not directly related to the FAIR principles, Kuhn et al. (2016) pointed out the similarities between these principles and NP objectives. For instance, NP adopts the LD structure for storing information in a minimal, dynamic and consistent DO, based on the RDF/TriG notation.

NP comprises four RDF/TriG graphs: head, assertion, provenance and publication info. The head identifies the NP and connects the graphs. The assertion stores the statements. The provenance highlights the statements' provenance, and the publication info stores the NP provenance. Owing its dynamic and machine-readable way of disseminating information in fragments, the NP publishes data with attached metadata and provenance (Groth et al., 2010).

Data provenance is crucial if we consider research reusability and knowledge discoverability. Data provenance is the metadata that assists the DO ancestry representation. In a data reusability scenario, retrospective data provenance is crucial in providing documentation for controlling data quality, authorship and curation (Da Serra et al., 2009).

W3C PROV Data Model (PROV-DM) is a well-founded formalism for tracking DO provenance. PROV-DM provides specifications to control data provenance, which any agent can use to model, serialise, exchange, access, merge and translate the provenance metadata (Moreau et al., 2015). Furthermore, the provenance can be semantically tracked in a machine-readable format by the PROV Ontology (PROV-O) (Moreau et al., 2015).

2.2 Related works

The technical literature has described various FAIRification workflows to assist (meta)data creation or evolution. For instance, the GO FAIR proposal is a template to support the development of novel FAIRification processes (GO FAIR, 2019). The template is a general process for implementing a FAIR environment in seven steps: (1) retrieve non-FAIR data; (2) analyse the retrieved data; (3) define the semantic model; (4) make data linkable; (5) assign licence; (6) define metadata for the data set and (7) deploy the FAIR data resource. Owing its generality, this FAIRification process can be adopted in distinct scientific domains. We stress that these workflow steps are 'technology agnostic'. They do not specify computational techniques or specific technologies to solve scientific problems. Unlike the FAIR guidelines that emphasise the intrinsic data and metadata association in almost all its principles, this workflow only deals with metadata in step 6, without a deeper examination of the (meta)data relationships.

Similarly, the FAIRification workflow described by Jacobsen et al. (2020) also focuses on general processes. Jacobsen's workflow steps are classified into three major phases: Pre-FAIRification, FAIRification and Post-FAIRification. The phases are decomposed into nine steps: (1) FAIRification objective identification; (2) data analysis; (3) metadata analysis; (4a) data semantic model definition and (4b) metadata semantic model definition; (5a) make data linkable and (5b) make metadata linkable; (6) host FAIR data and (7) assess FAIR data.

Jacobsen's workflow suggests mechanisms for achieving each step. Besides, the workflow performs similar steps for data and metadata, highlighting both artefacts' importance. Oliveira et al. (2022) derived a generic workflow-based approach based on Jacobsen. This work establishes well-defined actions for each step to assist data modellers during (meta)data transformation.

On the other hand, several authors emphasised the role of specialisation steps in research domains. For instance, Sinaci et al. (2020) developed a FAIRification workflow for data-centric health research. It incorporates data validation, de-identification and pseudonymisation steps necessary to this research domain. Additionally, this workflow includes steps for data versioning and indexing that the domain requires.

The *de novo* FAIRification process proposed by Groenen et al. (2021) also focused on health research to transform raw (meta)data of vascular anomalies. This workflow comprises five phases and fifteen steps related to adopting specific domain tools and techniques for transforming (meta)data.

As far as we know, these works mentioned above provide innovative ways of generating FAIRer environments. However, none of the previous works considers the necessary steps to analyse and represent a FAIR DM concentrating on both data and metadata, especially considering the metadata already attached to the published raw data.

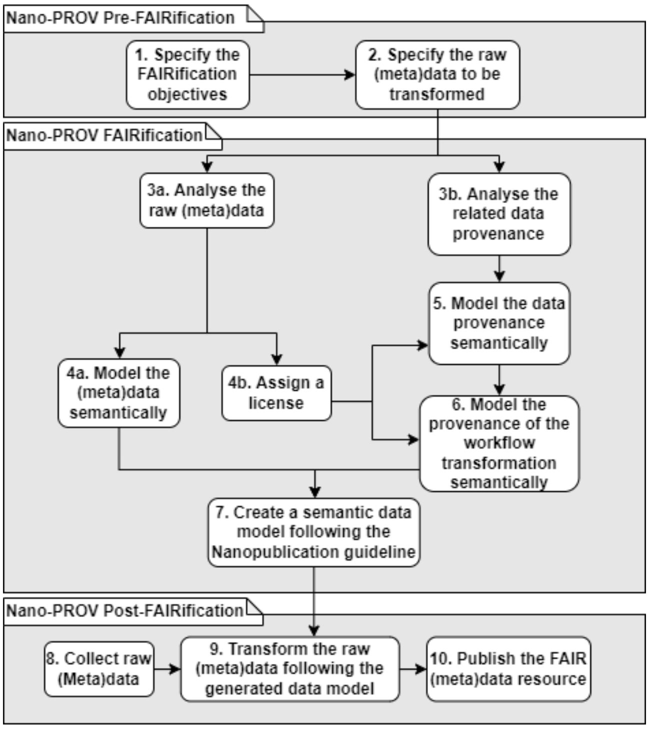
Moreover, these works do not contemplate the roles of provenance identification, control and modelling. Finally, none of the previous workflows considers NP a possible solution for FAIR DO creation.

3 Nano-PROV FAIRification workflows

This section describes the Nano-PROV FAIRification workflow (illustrated in Figure 1). Our method borrows some features described in GO FAIR and Jacobsen et al. (2020) FAIRification processes. However, unlike the GO FAIR FAIRification workflow, the Nano-PROV workflow focuses on creating a DM with particular attention to metadata analyses, provenance management and NP adoption.

As Wilkinson et al. (2016) pointed out, metadata is crucial in increasing data reusability, principally in data-centric and Big Data scenarios. Our workflow conducts all the steps with proper attention to data and metadata, following the FAIR data principles and formalisms of LD and the Semantic Web (Heath and Bizer, 2011).

Figure 1 Nano-PROV FAIRification workflow


Nano-PROV Pre-FAIRification
1. Specify the FAIRification objectives
2. Specify the raw (meta)data to be transformed
Nano-PROV FAIRification
3a. Analyse the raw (meta)data
3b. Analyse the related data provenance
5. Model the data provenance semantically
4a. Model the (meta)data semantically
4b. Assign a license
6. Model the provenance of the workflow transformation semantically
7. Create a semantic data model following the Nanopublication guideline
Nano-PROV Post-FAIRification
8. Collect raw (Meta)data
9. Transform the raw (meta)data following the generated data model
10. Publish the FAIR (meta)data resource

The Nano-PROV FAIRification is customised to data provenance. The reusable FAIR principle (R1.2) explicitly cites the detailed provenance necessity. Thus, considering this perspective, Nano-PROV includes three steps for controlling the data and workflow provenance by the Unified Nanopublication Provenance (UN-PROV) approach.

Contrasting the GO FAIR workflow, we propose the creation of Nanopublication DO. Other works have used NP as a feasible solution to a dynamical, minimal, machine-readable and semantic-supported DO. Further, Nano-PROV presents an overview for data producers to make (meta)data FAIR using the described concepts and technologies.

To achieve these goals, researchers must have confidence in the data when adopting the workflow and not only in defining the technologies. For instance, researchers must perform Competency Questions (CQ) like 'How to enrich the (meta)data?', 'Which are the essential provenance metadata?', 'Which ontologies does the domain community adopts?' or 'Where will the FAIR data be published?'. The CQ are examples of how the Nano-PROV FAIRification workflow may aid researchers.

The Nano-PROV workflow phases – Pre-FAIRification, FAIRification and Post-FAIRification – are further divided into twelve steps. Each step in the FAIRification process is fully described in the following sub-sections.

3.1 (Step 1) Specify the FAIRification objectives

The first step is understanding the role of (meta)data and developing FAIRification objectives. Data scientists must have an essential background in LD and Semantic Web and be familiar with the FAIR principles. Modellers must consider raw data sets, relevant metadata, reusers, stakeholders and data standards and guidelines.

Another crucial factor is to perform technical analysis, like conducting a FAIR evaluation framework (Herczog et al., 2020). This evaluation can identify the (meta)data reuse scenario, data storage characteristics, reusers agents, data domain specifications and additional metadata attributes.

As Wilkinson et al. (2016) described, a data set does not need to obey all the FAIR principles simultaneously but should achieve at least a minimal level, upgradable when possible. The workflow may support some principles depending on the data reusers premisses, FAIR status and other related characteristics.

We underline that some of the FAIR principles include technological factors. For example, the adoption of globally unique identifiers and the specification of communication protocols. The objectives must reflect characteristics that influence the objectives and stakeholders' resource limitations for implementing Nano-PROV.

The step 1 output is a checklist with metrics and CQ related to the stakeholders' needs. Moreover, the step identifies two user groups (data modellers and domain stakeholders). The groups discuss the issues experienced by reusers, investigate solutions for complying with the FAIR based on domain-relevant standards and propose the workflow outputs.

3.2 (Step 2) Specify the raw (meta)data to be transformed

The second step delineates the definition of data sets and their metadata to be transformed, considering the data type diversity. The analysed raw data used are divided into test and real data sets.

The test data sets are samples used during the FAIRification workflow to evaluate the semantic data and the provenance models. These data sets must reflect the characteristics of the original scenario and address the FAIRification objectives. The variety of data formats, the connections between internal and external data sets, the range of different data types, the versioning, the attached metadata licences and data provenance are analysed during the workflow. If the researcher controls these characteristics, he/she will prevent (meta)data misinterpretations or shortages.

The real data sets represent all the data sets in the analysed environment to be transformed into NP. In this case, the test data sets are included in the real data set. This real data set is used only in the Post-FAIRification phase after creating the NP model. In step 2, the output considers the definition of the data sets. Besides, a (meta)data collecting script is desirable to obtain the data sets and standardise the data transformation steps at the end of the FAIRification cycle.

3.3 (Step 3a) Analyse the raw (meta)data

Step 3a analyses the raw (meta)data in the test data sets. The raw (meta)data analysis contemplates an overview of the technical characteristics and an interpretative semantic investigation of the data sets. From a technical point of view, data researchers need to investigate the data structures, types, and representations (Jacobsen et al., 2020). The investigation assists the Semantic Model (SM) composition and introduces data semantical analysis.

The interpretative semantic investigation of the data set must regard both data and metadata. Some FAIRification workflows take this analysis separately or do not represent the metadata analysis (GO FAIR, 2019; Sinaci et al., 2020; Kersloot et al., 2021). Additionally, Mons et al. (2017) referred to the metadata in two different sections: intrinsic metadata (added during the generation of the data) and user-defined metadata (metadata related to the data provenance).

However, due to the different applied domains, ontologies and controlled vocabularies, the Nano-PROV workflow conducts the metadata analysis in different steps. The intrinsic metadata are analysed with the data, and the user-defined metadata are examined in the next step, given well-defined data provenance ontologies and guidelines.

The semantical analysis investigates if the published data and metadata are clear, unambiguous and pave the way for the workflow. The analysis of the data can be subdivided into: (i) searching for possible semantic standardisation; (ii) describing the semantic meaning of the data and (iii) highlighting the relations between data elements (Ganz et al., 2016; Shotton et al., 2009; Wilkinson et al., 2016).

At first, modellers need to search for a semantic standardisation related to the data sets, which can attenuate the effort during the analysis and composition of the SM (Jacobsen et al., 2020). After that, modellers must semantically describe all the data elements with the domain reusers' assistance. Lastly, researchers need to identify the connections between the data elements. The possible relationships can be intrinsic in the data, referring to other data elements or data from other databases, although not semantically informed. Identifying these relationships can increase data interoperability (Wilkinson et al., 2016).

As previously presented, metadata are essential for data reusability. Along with that, metadata must follow the same three topics of semantic data analyses. Additionally, it is necessary to identify the attached metadata, the metadata that are not strictly linked with the data, and related metadata that are not present in the environment (Jacobsen et al., 2020).

In addition, the modellers must try to capture as much metadata as possible to provide a favourable reusable scenario. The richer the metadata, the more reusers can utilise them to verify if the data are good enough for their objectives (Mons et al., 2017). Finally, this step must consider the semantical analysis of the data and metadata relationship to facilitate the development of the semantic models.

The outputs of step 3a need to represent a description of the (meta)data environment. Modellers can develop the semantic model effortlessly with the generated semantic description and the discovered technical information.

3.4 (Step 3b) Analyse data provenance

Step 3b identifies and analyses the data provenance. As discussed, provenance is crucial for data interpretability, discoverability, and reproducibility. Tracking provenance is challenging, principally considering published data.

W3C PROV supports data researchers in managing and searching provenance and preparing it to be semantically modelled. The modellers need to understand how the data were generated, the generation purposes, their attributes during the generation, filter and cleaning processes, and identify the owners, creators and curators (Jacobsen et al., 2020). Further, when reusers collect information from the data provenance, they face distinct provenance perspectives, motivated by their objectives and data requirements. Owing these perspectives, the richer the data provenance can be, the better it can be reused by agents (Gil et al., 2010).

The W3C PROV focuses on mapping the data provenance by the agent, activity and entity roles. Provenance describes the creation and use of entities by activities that agents may influence differently (Groth and Moreau, 2013). For example, four provenance roles are identified when mapping the provenance related to the action of generating a data set by scrapping an article. In this scenario, the creator is the agent, the scrapping is the activity and the data set and article are the entities. From this perspective, the FAIRification actors can track the provenance relationships. Additionally, PROV template adoption will support the provenance SM in future steps by applying the PROV ontology, additional provenance ontologies and controlled vocabularies.

The outputs of step 3b generate data provenance artefacts for increasing data understandability and reusability. Adopting the W3C PROV can simplify mapping the data lineage and assist in identifying the data provenance characteristics.

3.5 (Step 4a) Model the (meta)data semantically

Step 4a focuses on modelling the (meta)data semantically. Modellers must define a model that characterises the data set entities and relationships to be semantically accurate in a machine-readable environment.

Creating an SM requires effort to translate the (meta)data scenario to be understood by machines. At first, modellers must investigate if there are any SMs related to the applied domain. The GO FAIR community defends SM reuse as a possible advantage of the (meta)data domain standardisation (GO FAIR, 2019).

If there is no related SM, modellers may: (i) generate a conceptual model following the (meta)data analyses; (ii) search for controlled vocabularies, ontologies and thesaurus referring to the domain and (iii) establish an SM following the conceptual model and the ontologies.

The conceptual model represents the environment as a graph with terms and relations. The subject-predicate-object format can assist modellers in generating the conceptual model and preparing it to be translated using ontological terms. After that, modellers must find domain ontologies and other supportive ontologies. Searching domain ontologies that fit the created model is quite challenging. Ontology search engines like BioPortal (Whetzel et al., 2011), OLS Ontology Search (Cote et al., 2008) and Ontobee (Ong et al., 2016) can assist in this identification.

The last action in this step 4a is to conceive the SM that complies with the selected ontologies. Modellers must translate the terms and relations based on the chosen ontologies. Additionally, the SM may represent other external data sets from different sources by semantically linking them. Thus, the output of this step is the generated SM that represents the actual (meta)data environment to be readable by machines.

3.6 (Step 4b) Assign a licence

Step 4b spotlights the licence attribution for the Data Model (DM). Research data are copyrighted material, and data licences lack or incorrect use generates a non-reusable data scenario (Wilkinson et al., 2016). Providing clear and acceptable data licences assists reusers in understanding the requirements and permissions granted by data holders to correctly reuse data (Hyland et al., 2014).

Further, licensing research data can contribute to data preservation and protect researchers from possible conflicts related to the conclusions obtained from the data. LD licences can promote benefits such as deterring data fraud, reducing data duplication, encouraging interdisciplinary research and recognising data holders (Ball, 2014). Choosing a licence that best fits the data environment and data owners' requirements is challenging. Different sources, like Creative Commons (CC), Open Data Commons (ODC) and GNU Public Licences can support this decision.

In some cases, the data sets already have an attached licence. Therefore, the modellers must follow the attached licence and cannot provide a new licence that changes the original definitions. Moreover, the NP provides (meta)data information in a free and open form to be reusable without obstacles (Mons and Velterop, 2009). Modellers must define a licence that complies with the NP approach, the FAIRification objectives and the data sets licences. These step outputs must consider the licence identification and definition related to the (meta)data and the NP data model.

3.7 (Step 5) Model data provenance (UN-PROV Data)

Step 5 focuses on modelling the data provenance. By the data provenance artefacts generated in the 3b step, data modellers must create the conceptual model following the subject-predicate-object notation, although adopting provenance-related ontologies. Nanopublication establishes a clear provenance definition. However, there are minimal details for modelling provenance in NP, and the published NPs suffer from a lack of provenance (Asif et al., 2019).

The Unified Nanopublication Provenance (UN-PROV) approach was created to support provenance management in NPs. UN-PROV reuses existing provenance ontologies to propose an environment for controlling and publishing provenance in NP. The UN-PROV provides triples representing the agent-entity-activity notation defended by the W3C PROV.

By the NP definition, the provenance and publication info graphs store provenance information. Owing the graphs objectives, the UN-PROV guideline is divided into two steps. At first, the provenance graph is modelled during this step. After that, the SM referring to the NP publication info graph is developed in FAIRification step 6.

The NP guideline specifies CQs for controlling the assertion provenance graph, like 'how the assertion was generated', 'who generated it', 'when it was generated', 'where the assertion was obtained from' and any other additional information (Groth et al., 2021). These guidelines, though, do not provide related information about the CQs, the provenance ontologies, and a possible DM for storing assertion provenance. By that, the UN-PROV supports modellers in identifying and developing a data provenance model to be attached to the final NP data model.

The UN-PROV SM for the NP provenance graph (a.k.a UN-PROV Data) follows the triple notation and is represented by seven subgraphs. These subgraphs identify the assertion graph as an entity, the performed activity to generate this entity and the attributed agents. Further, the SM provides additional information like data description, citation requirements, and storage location. The UN-PROV data provenance SM is highlighted in the red box in Figure 2.

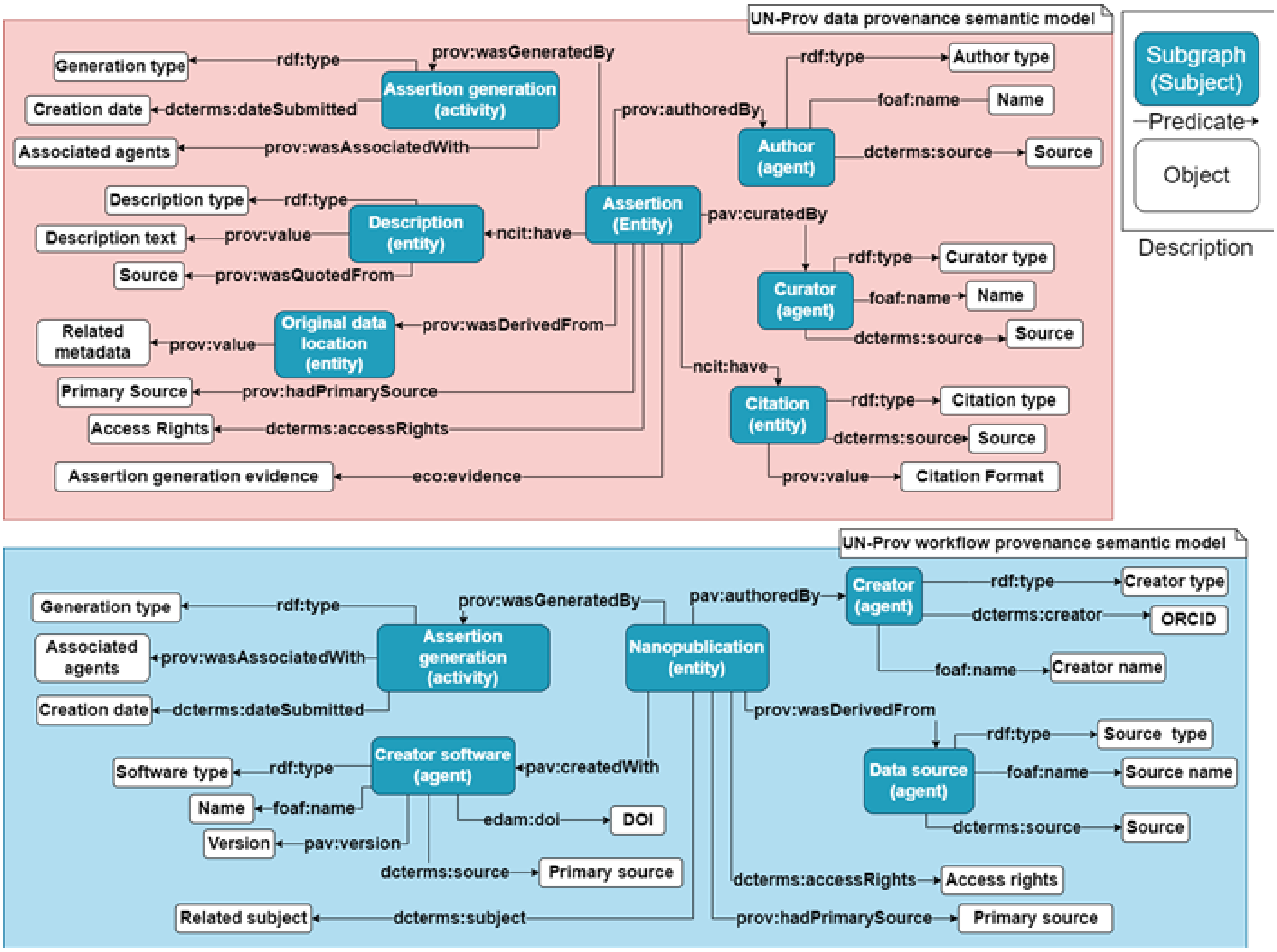
The assertion entity is the core attribute in the provenance graph. Nine provenance metadata are related to the assertion: two agents, three entities, one activity and three additional metadata. These metadata are not mandatory, and additional provenance can be attached to the UN-PROV Data.

At first, triples related to the primary source identification, the access rights definition and the assertion generation evidence are identified. The *prov:hadPrimarySource* property identifies the source related to the data in the assertion entity. The data access rights can be identified by *dcterms:accessRights*. The eco:ECO0000501 term from the Evidence & Conclusion Ontology (Chibucos et al., 2016) highlights the evidence for the data provenance.

After that, the author and curators are identified. The UN-PROV Data designs these agents using the same metadata information. The PROV-O and PAV properties *pav:authoredBy*, *pav:curatedBy* and *prov:wasAttributedTo* differentiate the agents' roles. The agent's name can be identified using the Friend of A Friend (FOAF) vocabulary or a semantic term related to the agent. The agent's related source can also be identified.

The (meta)data generation activity is modelled by the generation type, the activity agents, and the creation date. The generation type considers if an automatic or manual process originated the data. The *prov:wasAssociatedWith* property identifies the agents. Further, the term *dcterms:dateSubmitted* is used to track the creation date.

Figure 2 UN-PROV semantic models templates (see online version for colours)



The UN-PROV Data identifies the storage location by the term *prov:wasDerivedFrom*. Related metadata, such as the data location characteristics, the related agents, and the source URL, may be tracked. The description is identified by *dcterms:description* term, the description text by the *prov:value* property and the description source by *prov:wasQuotedFrom*. Moreover, the provided citation is highlighted by the *dcterms:bibliographicCitation* property, *prov:value* quotes the citation, and *dcterms:source* identifies the related sources.

The output of step 5 is the provenance SM based on the (meta)data environment and following the UN-PROV Data. The created data provenance SM is further utilised in the composition of the final Nanopublication DM.

3.8 *(Step 6) Model the provenance of the (meta)data transformation process semantically (UN-PROV Workflow)*

Step 6 follows the same process as the previous one. However, the difference is related to tracking the workflow provenance and the generated NP data model. This step was created to comply with the NP publication info graph. Analogous to the NP provenance graph, no formalised provenance guideline is provided. The only condition for the publication info graph is to store the creator's identification and the NP creation timestamp (Groth et al., 2021).

As presented in the blue coloured box in Figure 2, five subgraphs comprise the UN-PROV Workflow. Three agents, one activity and three additional metadata, compose the SM. The NP creator can be identified by *pav:authoredBy* in the NP entity triple block. The author's Open Research and Contributor ID (ORCID), name and type are also provided. Simultaneously, the *pav:retrievedFrom* property identifies the data source location. The data source performs the role of an agent, and its type can be identified using the FOAF vocabulary, additional semantic terms or the data source name. The URL for the data source is also stored.

A particular agent for this graph is the creation software. Owing the Big Data characteristics, a coded script or software could generate the NP data sets. The *pav:createdWith* property relates the software to the SM. The software type is tracked by the term *rdf:type*, the software name by the *foaf:name* term and the version by the *pav:version* property. The ORCID identifies the software authors. Further, the software deposit location is presented by the *dcterms:source*, and the related Digital Object Identifier (DOI) by the *edam:data_1188* property.

Like the UN-PROV Data SM, three metadata describe the activity to generate the model. At first, the generation type is identified by *rdf:type*. After that, the related agents are highlighted by the *prov:wasAssociatedWith* property. The last metadata activity is the creation date, tracked by *dcterms:dateSubmitted* property.

Lastly, the provenance workflow is distinguished by the primary source, subject and access rights. Adopting the term *prov:hadPrimarySource,* modellers identify the location used to compose the NP. Subsequently, the subject identifies the associated terms related to the (meta)data environment, acting like the scientific paper keywords. To identify the subject, modellers must use *dcterms:subject* term followed by the semantic terms. Furthermore, the assigned NP access rights are highlighted by the *dcterms:accessRights* term.

The output of this step is the workflow provenance semantic model following the UN-PROV Workflow by mapping and controlling the (meta)data transformation provenance.

3.9 *(Step 7) Create a semantic data model following the Nanopublication guideline*

Nano-PROV step 7 generates the final model following the NP guideline. The created model must follow a data framework and a machine-understandable notation. The NP approach is adopted due to the dynamic and effortless way of generating DOs according to the FAIR principles and Semantic Web definitions (Feijoó et al., 2022).

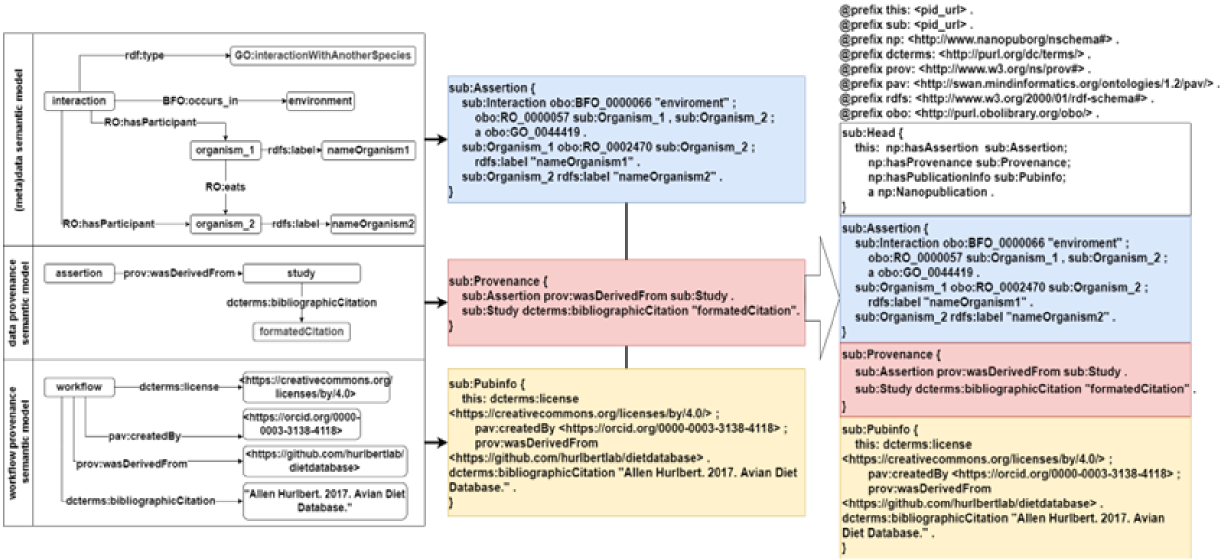
In this step, modellers assemble the final DM by aggregating the previously developed SMs in previous steps. The assertion graph must receive the (meta)data SM generated in the 4a step. The provenance graph acquires the UN-PROV Data SM from step 5. Lastly, the publication info graph meets the UN-PROV Workflow SM generated in step 6.

The first part of this step is to translate the created semantic models to the RDF/TriG notation. Following the previous Nano-PROV steps, the generated SMs must follow the subject-predicate-object. By that, modellers translate the SM more effortlessly, without re-interacting with the created models and defining each semantic term role in this syntax. Figure 3 represents a possible translation from the generated semantic models to a final NP model.

The example in Figure 3 was based on a real NP related to the organism's interaction. The figure left part presents the developed SMs. The modellers develop the NP data model following the defined ontologies starting from the semantic models. Although different from the created human-readable SMs, the NP model must follow the URI of the ontological terms. For example, the term *BFO:occurs_in* property in the (meta)data SM example refers to the same *obo:BFO0000066* ontology property used in the NP assertion graph but focuses on the machine readability. Further, data modellers must identify the used prefix in the ontological classes and properties. The right part of Figure 3 presents the final NP model with all the necessary graphs and prefixes.

Another part of the transformation from the SMs to the NP model is related to data literals. Data literals are (meta)data that were not transformed following a related semantic term, e.g., the bibliographic citation presented in Figure 3. By the non-use of a controlled vocabulary, data modellers must avoid data literals simply because machines cannot primarily understand them. Depending on the case, some data literals can be transformed into semantically related terms, and one example of that is related to date. As Kuhn et al. (2021) discussed, the date information is not semantically tracked in most DOs. Although comparing the effort of machine understanding date as a literal and date with semantic terms, the second approach is more efficient.

Figure 3 Transformation example from the created SM to the final NP data model (see online version for colours)


(meta)data semantic model
interaction
rdf:type
GO:interactionWithAnotherSpecies
BFO:occurs_in
environment
RO:hasParticipant
organism_1
rdfs:label
nameOrganism1
RO:eats
RO:hasParticipant
organism_2
rdfs:label
nameOrganism2
data provenance semantic model
assertion
prov:wasDerivedFrom
study
dcterms:bibliographicCitation
formatedCitation
workflow provenance semantic model
workflow
dcterms:license
<https://creativecommons.org/licenses/by/4.0>
pav:createdBy
<https://orcid.org/0000-0003-3138-4118>
prov:wasDerivedFrom
<https://github.com/hurlbertlab/dietdatabase>
dcterms:bibliographicCitation
"Allen Hurlbert. 2017. Avian Diet Database."

```
sub:Assertion {
   sub:Interaction obo:BFO_0000066 "enviroment" ;
      obo:RO_0000057 sub:Organism_1 , sub:Organism_2 ;
      a obo:GO_0044419 .
   sub:Organism_1 obo:RO_0002470 sub:Organism_2 ;
      rdfs:label "nameOrganism1" .
   sub:Organism_2 rdfs:label "nameOrganism2" .
}
```

```
sub:Provenance {
   sub:Assertion prov:wasDerivedFrom sub:Study .
   sub:Study dcterms:bibliographicCitation "formatedCitation".
}
```

```
sub:Pubinfo {
   this: dcterms:license
<https://creativecommons.org/licenses/by/4.0/> ;
      pav:createdBy <https://orcid.org/0000-0003-3138-4118> ;
      prov:wasDerivedFrom
<https://github.com/hurlbertlab/dietdatabase> .
dcterms:bibliographicCitation "Allen Hurlbert. 2017. Avian Diet
Database." .
}
```

```
@prefix this: <pid_url> .
@prefix sub: <pid_url> .
@prefix np: <http://www.nanopuborg/nschema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix pav: <http://swan.mindinformatics.org/ontologies/1.2/pav/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix obo: <http://purl.obolibrary.org/obo/> .
sub:Head {
   this:  np:hasAssertion  sub:Assertion;
      np:hasProvenance sub:Provenance;
      np:hasPublicationInfo sub:Pubinfo;
      a np:Nanopublication .
}
sub:Assertion {
   sub:Interaction obo:BFO_0000066 "enviroment" ;
      obo:RO_0000057 sub:Organism_1 , sub:Organism_2 ;
      a obo:GO_0044419 .
   sub:Organism_1 obo:RO_0002470 sub:Organism_2 ;
      rdfs:label "nameOrganism1" .
   sub:Organism_2 rdfs:label "nameOrganism2" .
}
sub:Provenance {
   sub:Assertion prov:wasDerivedFrom sub:Study .
   sub:Study dcterms:bibliographicCitation "formatedCitation" .
}
sub:Pubinfo {
   this: dcterms:license
<https://creativecommons.org/licenses/by/4.0/> ;
      pav:createdBy <https://orcid.org/0000-0003-3138-4118> ;
      prov:wasDerivedFrom
<https://github.com/hurlbertlab/dietdatabase> .
dcterms:bibliographicCitation "Allen Hurlbert. 2017. Avian
Diet Database." .
}
```

With the created model, data modellers can test the written model's syntax with the NP community-developed Validator for Nanopublication,[1] a tool to check if the created DO follows the RDF and NP requirements (Kuhn, 2015).

The generated output of step 7 is the final Nanopublication DM representing all the created SMs based on NP guidelines and RDF/TriG notation. The created NP model must pass the data modellers' validations. Further, the validated model can be used in the following steps for generating the NP DOs based on the defined real data sets.

3.10 (Step 8) Collect the raw (meta)data following the generated data model

Step 8 concentrates on collecting the real data sets based on the NP model. Data modellers can apply different techniques to obtain real data sets. Owning or having access to these data sets is the ideal scenario. Data scientists, though, sometimes cannot directly retrieve these data sets using a query language. Data modellers may develop scripts, tools or services to circumvent this obstacle.

Modellers can adopt techniques to extract real data sets. One of these examples is Selenium,[2] an open-source tool for testing web applications. Selenium has been widely adopted to collect (meta)data from web pages. Another example is Scrapy,[3] an open-source and collaborative framework for extracting published (meta)data, mainly used in Big Data due to its flexibility and adaptability in collecting (meta)data and generating structured data sets. Besides the technique used, the final output of this step must represent the real data sets following the established Pre-FAIRification requirements and the created NP data model.

3.11 (Step 9) Post-FAIRification: Transform the raw (meta)data following the generated data model

In Step 9, modellers must adopt or develop a script to transform the real data sets into NP DOs based on the created Nanopublication data model. During this transformation step, data modellers must investigate the best approach to deal with the present characteristics, like the data set total size and the semantic term proportion. These characteristics may influence the transformation performance.

One possibility adopted by data modellers is the OpenRefine[4] tool. According to Thompson et al. (2020), OpenRefine is a versatile tool focusing on exploring, cleaning, transforming, and extending data sets according to the data scientists' objectives. Although the DO created by OpenRefine does not follow the NP guidelines, additional steps must be developed to transform the data sets into the NP.

Moreover, modellers can develop a script for converting the real data sets, especially when data collection scripts are developed. The nano-pub-java library can support the creation of the transformation script. This library follows the NP guidelines, and one of the features is the definition of NP DOs in RDF TriG notation based on the delimited (meta)data semantic terms (Kuhn, 2015).

An example is the NP Python library which provides ways to create, search and publish NPs (Van der Burg et al., 2021). Similar to the nano-pub-java, this library assists data scientists in creating NP DOs following the RDF TriG notation. Further, it focuses on controlling the ownership and provenance of the NP by delimiting the owner's ORCID and defining custom provenance and publication info triples.

Further, the FAIR data principles highlight that the DO must be assigned with a global and unique PID. A series of services assign PIDs for DOs. Archival Resource Keys (ARK), Identifiers.org and Persistent Uniform Resource Locators (PURL) are some examples (Juty et al., 2020).

The output of step 9 is the nanopublication data set referring to the developed model and the attached PID. Additionally, the outputs must consider adopting or developing an approach for transforming the real data sets into DOs.

3.12 (Step 10) Publish the FAIR (meta)data resource

The nanopublication DOs dissemination is the final step of the Nano-PROV. Depending on the established objectives, modellers can adopt different techniques for publishing their DOs. The nanopublication Server (NServer) is the primary location to publish NPs. The NServer is a decentralised network where the active servers can store, retrieve, and replicate NPs (Kuhn, 2022). The submission of NPs in this network occurs by creating a new NServer, integrating it into the network and publishing the NPs. After the NPs publication, the servers in the network can replicate them. The cited libraries in the previous step have features for publishing nanopublications in an NServer. Almost eleven million NP were stored using the NServer (Kuhn, 2022).

Storing the NPs in an NServer takes advantage of most of them being published in the same network, increases NP DOs integration, and unifies access to nanopublications (Kuhn et al., 2016). At the same time, the NServer suffers from a lack of services for searching, querying, analysing and using data. In addition, the network focuses on machine readability, principally for identifying the NPs by their PIDs and decreases the readability for humans (Giachelle et al., 2021).

Besides publishing NPs in an NServer environment, data modellers can adopt other consolidated approaches like web applications, SPARQL endpoints or RDF triple stores. Nevertheless, it is necessary to clarify that the selected method must follow the FAIR principles.

The output of this step is the publication of the created nanopublication DOs in a reliable and trustworthy way for reusers. Furthermore, data modellers can adopt the FAIR Data Point to provide metadata about the stored DOs in a transparent and controlled access way (Hooft, 2020).

As previously considered, the end of the FAIRification workflow is delimited by the publication of the DOs. After that, new objectives, requirements, revision of the semantic models or other characteristics contribute to the FAIRification process reinteraction. Furthermore, disseminating the objectives, decisions, semantic models or any other generated output during the FAIRification execution can assist other data modellers in their FAIRification process and consolidate the convergence of the used approaches during the workflow.

4 Usage and discussion of the Nano-PROV FAIRification workflow

This section presents an overview of the Nano-PROV adoption for generating the nanopublication Genome data model (NGen-DM). The created DM provides a semantic enhancement of the published (meta)data in the US National Centre for Biotechnology Information (NCBI) databases, referring to information on biological organisms, genomic assemblies and scientific articles domains.

Genomic researchers face distinct drawbacks related to un-reusable (meta)data in these databases. These issues are related to the unstructured and unsemantic (meta)data published in NCBI. The NGen-DM was first introduced in our previous work (Feijoó et al., 2022).

During the Pre-FAIRification steps 1 and 2, we identified that the environment has a diverse range of (meta)data characteristics. Further, the literature highlights that reusers commonly know that the published data have a high percentage of reusability and interoperability shortages related to the non-adoption of machine-readable approaches (Feijoó et al., 2020). These drawbacks are found in several highly used and recognised biological databases, such as the GenBank Assembly database. We started the development of NGen DM based on the GenBank database faced issues.

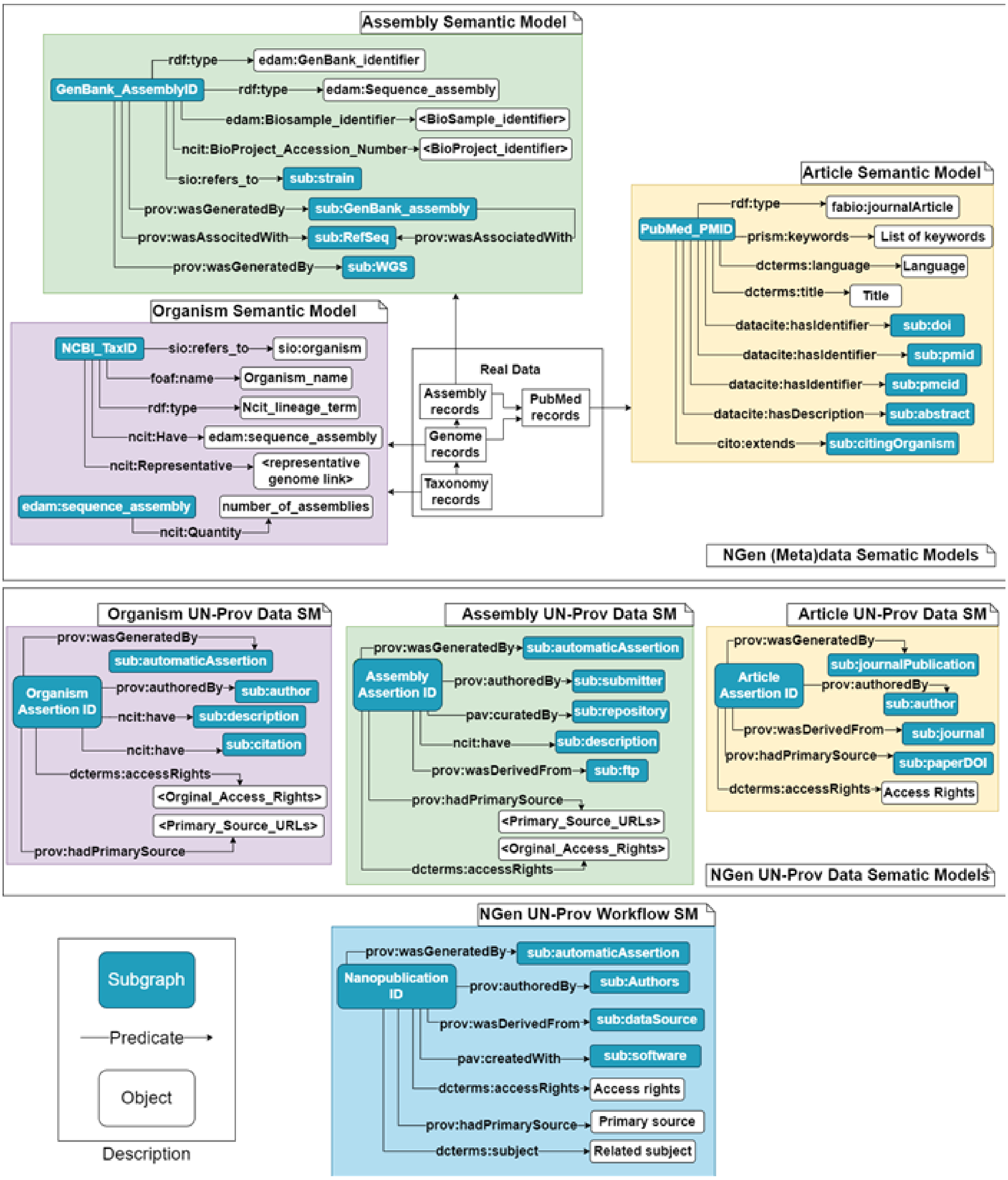
Adopting the outputs from the Pre-FAIRification steps, the raw (meta)data and related provenance were analysed in the Nano-PROV steps 3a and 3b. Our analyses revealed that the GenBank records connect with other databases related to biological and scientific publication domains. With that in mind, three distinct SMs were created following the NCBI Taxonomy and NCBI Genome, GenBank Assembly and PubMed databases.

The upper box in Figure 4 represents the actual connections between the databases and the generated outputs from the semantic (meta)data modelling performed in the Nano-PROV step 4a (NGen (meta)data semantic models). Similar to what was proposed in the step 4a Nano-PROV FAIRification step, general and domain ontologies were adopted to express the real semantic meaning. However, we perceived that it was impracticable to semantically transform them due to user-defined (meta)data. These terms are related to the text descriptions, the article authors' names and the organisations' names. To solve these untracked terms, we propose a series of semantic triples for highlighting them semantically, like the *NCBI_TaxID foaf:name Organism_name* triple in the Organism Semantic Model.

The 4b licence assignment step was performed after concluding step 4a. The NCBI environment already has a description of the data reusable policy,[5] so there was no need to choose licences for the (meta)data. On the other hand, we adopted the CCBY4.0[6] licence for the NPs. This licence follows the NCBI policy and defines the attributions that the reusers need to execute.

With the outputs of the provenance analyses in the step 3b, we started semantically modelling the data and workflow provenance based on the UN-PROV. Following FAIRification step 5, we developed three data provenance SMs related to each analysed domain. The NGen UN-PROV Data Semantic Model box in Figure 4 highlights the provenance of the (meta)data. The models were developed effortlessly by adopting the UN-PROV Data template, considering each database record-related provenance. Further, the related (meta)data licence analysed in step 4b was attached.

Figure 4 NGen semantic models (see online version for colours)

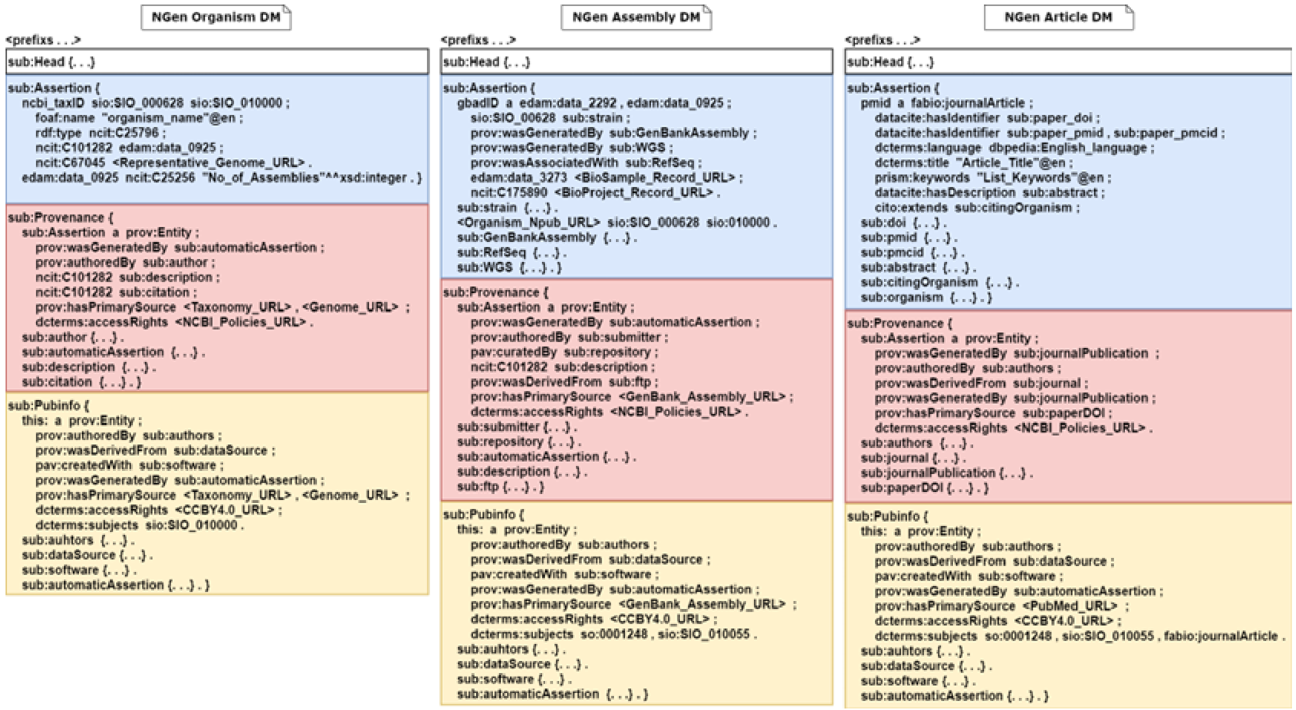


After UN-PROV Data SM generation, we conducted the Nano-PROV step 6 for generating the workflow provenance semantic models based, represented in the UN-PROV Workflow SM box in Figure 4. A unique SM was generated due to the similarities in performing the workflow in the three domains simultaneously. The SM distinctions are the NP PIDs, the related domain subject and the original resources.

After generating all the semantic models, we performed the last DM generation in step 7. The data model was based on the chosen ontologies and the subject-predicate-object notation. Firstly, we translated the human-oriented terms and relationships by their original identifiers. Furthermore, each generated SM was related to its triple graph following the NP scheme. From the semantic models designation, three NP models were developed following their data domain: Organism NP, Assembly NP and Article NP. Figure 5 presents the final NGen DM.

After generating the NGen-DM, the next steps in the Nano-PROV workflow were collecting the real data sets in step 8 and transforming them into NP following the created Data Model in step 9. We generated a Python script[7] to perform these two Post-FAIRification steps. The script scraped the four databases and extracted the necessary data and metadata. The script generates JSON files with the extracted (meta)data and converts the outputs to NPs following the generated DM.

Figure 5 Final NGen data model (see online version for colours)



NGen Organism DM

```
<prefixs . . .>
sub:Head {. . .}
sub:Assertion {
   ncbi_taxID  sio:SIO_000628  sio:SIO_010000 ;
      foaf:name  "organism_name"@en ;
      rdf:type  ncit:C25796 ;
      ncit:C101282  edam:data_0925 ;
      ncit:C67045  <Representative_Genome_URL> .
   edam:data_0925  ncit:C25256  "No_of_Assemblies"^^xsd:integer . }
sub:Provenance {
   sub:Assertion  a  prov:Entity ;
      prov:wasGeneratedBy  sub:automaticAssertion ;
      prov:authoredBy  sub:author ;
      ncit:C101282  sub:description ;
      ncit:C101282  sub:citation ;
      prov:hasPrimarySource  <Taxonomy_URL> , <Genome_URL>  ;
      dcterms:accessRights  <NCBI_Policies_URL> .
   sub:author {. . .} .
   sub:automaticAssertion  {. . .} .
   sub:description  {. . .} .
   sub:citation  {. . .} . }
sub:Pubinfo {
   this:  a  prov:Entity ;
      prov:authoredBy  sub:authors ;
      prov:wasDerivedFrom  sub:dataSource ;
      pav:createdWith  sub:software ;
      prov:wasGeneratedBy  sub:automaticAssertion ;
      prov:hasPrimarySource  <Taxonomy_URL> , <Genome_URL>  ;
      dcterms:accessRights  <CCBY4.0_URL> ;
      dcterms:subjects  sio:SIO_010000 .
   sub:auhtors  {. . .} .
   sub:dataSource {. . .} .
   sub:software {. . .} .
   sub:automaticAssertion {. . .} . }
```

NGen Assembly DM

```
<prefixs . . .>
sub:Head {. . .}
sub:Assertion {
   gbadID  a  edam:data_2292 , edam:data_0925 ;
      sio:SIO_00628  sub:strain ;
      prov:wasGeneratedBy  sub:GenBankAssembly ;
      prov:wasGeneratedBy  sub:WGS ;
      prov:wasAssociatedWith  sub:RefSeq ;
      edam:data_3273  <BioSample_Record_URL> ;
      ncit:C175890  <BioProject_Record_URL> .
   sub:strain  {. . .} .
   <Organism_Npub_URL>  sio:SIO_000628  sio:010000 .
   sub:GenBankAssembly  {. . .} .
   sub:RefSeq  {. . .} .
   sub:WGS  {. . .} . }
sub:Provenance {
   sub:Assertion  a  prov:Entity ;
      prov:wasGeneratedBy  sub:automaticAssertion ;
      prov:authoredBy  sub:submitter ;
      pav:curatedBy  sub:repository ;
      ncit:C101282  sub:description ;
      prov:wasDerivedFrom  sub:ftp ;
      prov:hasPrimarySource  <GenBank_Assembly_URL> ;
      dcterms:accessRights  <NCBI_Policies_URL> .
   sub:submitter {. . .} .
   sub:repository {. . .} .
   sub:automaticAssertion {. . .} .
   sub:description {. . .} .
   sub:ftp {. . .} . }
sub:Pubinfo {
   this:  a  prov:Entity ;
      prov:authoredBy  sub:authors ;
      prov:wasDerivedFrom  sub:dataSource ;
      pav:createdWith  sub:software ;
      prov:wasGeneratedBy  sub:automaticAssertion ;
      prov:hasPrimarySource  <GenBank_Assembly_URL>  ;
      dcterms:accessRights  <CCBY4.0_URL> ;
      dcterms:subjects  so:0001248 , sio:SIO_010055 .
   sub:auhtors {. . .} .
   sub:dataSource {. . .} .
   sub:software {. . .} .
   sub:automaticAssertion  {. . .} . }
```

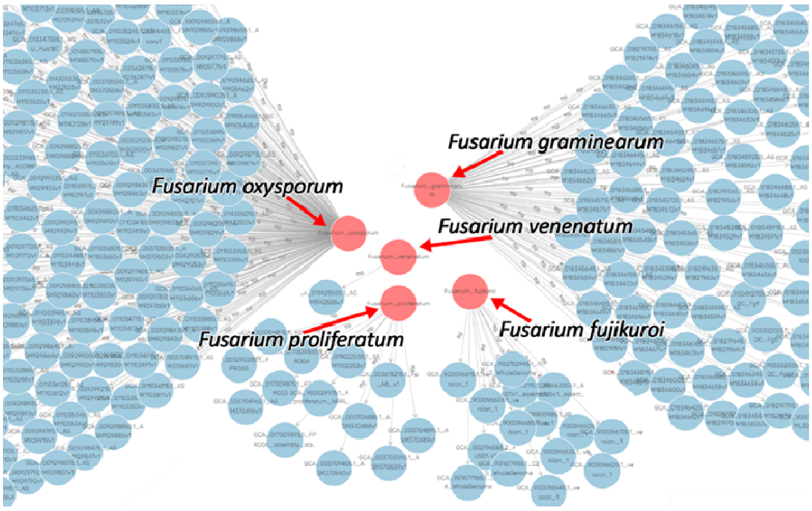
NGen Article DM

```
<prefixs . . .>
sub:Head {. . .}
sub:Assertion {
   pmid  a  fabio:journalArticle ;
      datacite:hasIdentifier  sub:paper_doi ;
      datacite:hasIdentifier  sub:paper_pmid , sub:paper_pmcid ;
      dcterms:language  dbpedia:English_language ;
      dcterms:title  "Article_Title"@en ;
      prism:keywords  "List_Keywords"@en ;
      datacite:hasDescription  sub:abstract ;
      cito:extends  sub:citingOrganism ;
   sub:doi  {. . .} .
   sub:pmid  {. . .} .
   sub:pmcid  {. . .} .
   sub:abstract  {. . .} .
   sub:citingOrganism  {. . .} .
   sub:organism  {. . .} . }
sub:Provenance {
   sub:Assertion  a  prov:Entity ;
      prov:wasGeneratedBy  sub:journalPublication  ;
      prov:authoredBy  sub:authors ;
      prov:wasDerivedFrom  sub:journal ;
      prov:wasGeneratedBy  sub:journalPublication ;
      prov:hasPrimarySource  sub:paperDOI ;
      dcterms:accessRights  <NCBI_Policies_URL> .
   sub:authors  {. . .} .
   sub:journal {. . .} .
   sub:journalPublication {. . .} .
   sub:paperDOI {. . .} . }
sub:Pubinfo {
   this:  a  prov:Entity ;
      prov:authoredBy  sub:authors ;
      prov:wasDerivedFrom  sub:dataSource ;
      pav:createdWith  sub:software ;
      prov:wasGeneratedBy  sub:automaticAssertion ;
      prov:hasPrimarySource  <PubMed_URL>  ;
      dcterms:accessRights  <CCBY4.0_URL> ;
      dcterms:subjects  so:0001248 , sio:SIO_010055 , fabio:journalArticle .
   sub:auhtors {. . .} .
   sub:dataSource {. . .} .
   sub:software {. . .} .
   sub:automaticAssertion {. . .} . }
```

As a test of the execution of the script, 6688 Nanopublication DOs were created based on the Organism-NP, Assembly-NP and Article-NP data models. In this test, 1399 organisms, 4199 genomic assemblies and 1090 articles records were semantically converted to NP following the NGen-DM. Step 10 was the last step of the FAIRification workflow regarding the publication of the generated FAIR DOs in a FAIR repository.

The NPs were published in two different repositories. These repositories were locally installed on a Linux Ubuntu machine. We carried out tests to prove if there were errors during the (meta)data transformation and the identification of semantical increase compared with the original data. The chosen repositories were the NServer and GraphDB. To run the tests, only GraphDB was used due to its search features based on the SPARQL language. The publication of the generated data set in the Nserver was carried out to test the functioning of the NP repository. Finally, no errors were found in the generated NP. Figure 6 presents a query result that can be made using the generated NPs. In that case, all GenBank FTP links referring to distinct Fusarium organisms stored in the NCBI databases can be retrieved simultaneously. In the original scenario, the GenBank FTP links are constantly used by researchers, although the common access presented in the GenBank assembly database can only be conducted manually. By using the NPs, the data retrieval can be easily done by ruing a single SPARQL query. Further, compared with the (meta)data stored in the GenBank databases, an increase in identifying distinct records from the traceable semantic terms was noticeable when using the GraphDB records. The generated NP data set can be retrieved in the presented repository.

Figure 6 Fusarium organisms and their associated GenBank FTP links (see online version for colours)



5 Conclusions

Nano-PROV FAIRification workflow increases the control and semantic enrichment of raw data sets. Three points can be highlighted about the workflow. Firstly, the workflow provides a more straightforward and reliable environment for generating data models that follow FAIR principles. Secondly, the generation of DO following the NP notation enhances the management of published data and metadata. Thirdly, the definition of the UN-PROV guidelines provides greater control and accuracy regarding the changes that occur with the stored data and the data transformation process aimed at the FAIR environment.

When comparing the Nano-PROV FAIRification with the related works, we identified some similarities with the leading GO FAIR (2019) and Jacobsen et al. (2020) FAIRification processes. However, it is understandable that most related works focus on applying essential concepts and techniques to the specific domain.

Besides that, the Nano-PROV FAIRification workflow provides general concepts and techniques for improving data and metadata while providing an understandable environment for humans and machines.

Additionally, the workflow spotlights the data provenance, which is essential for NP generation. From the UN-PROV proposal, the data and workflow provenance can be better managed and defined with ontologies and methods to be understood by agents.

By generating the NGen-DM based on our FAIRification approach, we perceived a (meta)data semantical increase and better provenance control in the unsemantic records published in research databases. However, some limitations during the application of the FAIRification were identified:

- It is still necessary to improve the used techniques, such as NServer, which in its current version can only be used by researchers with NServers expertise;
- NP management tools are scarce, causing a more significant effort to recover NPs by reusers;
- Depending on the purpose and size of the analysed literals, it becomes difficult to identify equivalent semantic terms so that machines can understand them.

In future works, we intend to improve the FAIRification process with the evolution of the used tools, provide a possible DM focused on the storage of common concepts among NP and better investigate the applicability of the FAIRification process in the other domains like digital agriculture (Da Cruz et al., 2009, 2018).

References

Asif, I., Chen-Burger, J. and Gray, A.J.G. (2019) 'Data quality issues in current nanopublications', *Proceedings of the 15th International Conference on EScience*, IEEE.

Ball, A. (2014) *How to License Research Data*, DCC How-to Guides, Digital Curation Centre.

Bizer, C., Heath, T. and Berners-Lee, T. (2009) 'Linked data – the story so far', *International Journal on Semantic Web and Information Systems*, Vol. 5, No. 3, pp.1–22.

Chibucos, M.C., Siegele, D.A. and Giglio, M. (2016) 'The evidence and conclusion ontology (ECO): supporting GO annotations', *Methods in Molecular Biology*, Springer, New York, NY, pp.245–259.

Cote, R.G., Jones, P., Martens, L., Apweiler, R. and Hermjakob, H. (2008) 'The ontology lookup service: more data and better tools for controlled vocabulary queries', *Nucleic Acids Research*, Vol. 36, pp.W372–W376.

Da Cruz, S.M.S. et al. (2018) 'Data provenance in agriculture', in Belhajjame, K., Gehani, A. and Alper, P. (Eds): *International Provenance and Annotation of Data and Processes*, Springer, pp.257–261.

Da Cruz, S.M.S., Campos, M.L.M. and Mattoso, M. (2009) 'Towards a taxonomy of provenance in scientific workflow management systems', *Congress on Services – I*, IEEE, USA. Doi: 10.1109/services-i.2009.18.

Demchenko, Y., Zhao, Z., Grosso, P., Wibisono, A. and De Laat, C. (2012) 'Addressing big data challenges for scientific data infrastructure', *Proceedings of the 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, IEEE, USA.

Fan, W. (2015) 'Data quality', *ACM SIGMOD Record*, Vol. 44, No. 3, pp.7–18.

Feijoó, M.P.P., Jardim, R., Da Cruz, S.M.S. and Campos, M.L.M. (2020) *Evaluating FAIRness of Genomic Databases*, Springer International Publishing, Cham.

Feijoó, M.P.P., Jardim, R., Da Cruz, S.M.S. and Campos, M.L.M. (2022) 'GAP: enhancing semantic interoperability of genomic datasets and provenance through nanopublications', *Metadata and Semantic Research*, Springer International Publishing, Cham, pp.336–348.

Ganz, F., Barnaghi, P. and Carrez, F. (2016) 'Automated semantic knowledge acquisition from sensor data', *IEEE Systems Journal*, Vol. 10, No. 3, pp.1214–1225.

Giachelle, F., Dosso, D. and Silvello, G. (2021) 'Search, access, and explore life science nanopublications on the web', *PeerJ Computer Science*, Vol. 7. Doi: 0.7717/peerj-cs.335.

Gil, Y., Cheney, J., Groth, P., Hartig, O., Miles, S., Moreau, L. and Da Silva, P.P. (2010) *Provenance XG Final Report*, W3C.

GO FAIR, G-F. (2019) *FAIRification Process – GO FAIR*, GO FAIR. Available online at: https://www.GO FAIR.org/fair-principles/fairification-process/

Groenen, K.H.J., Jacobsen, A., Kersloot, M.G., Dos Santos Vieira, B., van Enckevort, E., Kaliyaperumal, R. and Arts, D.L. et al. (2021) 'The de novo FAIRification process of a registry for vascular anomalies', *Orphanet Journal of Rare Diseases*, Vol. 16, No. 1. Doi: 10.1186/s13023-021-02004-y.

Groth, P. and Moreau, L. (2013) *PROV-Overview*, W3C.

Groth, P., Gibson, A. and Velterop, J. (2010) 'The anatomy of a nanopublication', *Information Services and Use*, Vol. 30, Nos. 1/2, pp.51–56.

Groth, P., Schultes, E., Thompson, M. and Tatum, Z. (2021) *Nanopublication Guidelines*. Available online at: https://nanopub.org/guidelines/working_draft/

Heath, T. and Bizer, C. (2011) 'Linked Data: Evolving the Web into a Global Data Space', *Synthesis Lectures on Data, Semantics, and Knowledge*, Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-79432-2

Herczog, E., Russell, K. and Stall, S. (2020) 'FAIR Data Maturity Model. Specification and Guidelines', *Research Data Alliance*. Doi: 10.15497/RDA00050.

Hooft, R. (2020) *FAIR Data Point*.

Hyland, B., Atemezing, G. and Villazón-Terrazas, B. (2014) *Best Practices for Publishing Linked Data*, W3C.

Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L.O.B., Mons, B., Schultes, E., Roos, M. and Thompson, M. (2020) 'A generic workflow for the data FAIRification process', *Data Intelligence*, Vol. 2, Nos. 1/2, pp.56–65.

Juty, N., Wimalaratne, S.M., Soiland-Reyes, S., Kunze, J., Goble, C.A. and Clark, T. (2020) 'Unique, persistent, resolvable: identifiers as the foundation of FAIR', *Data Intelligence*, Vol. 2, Nos. 1/2, pp.30–39.

Kersloot, M.G., Jacobsen, A., Groenen, K.H.J., dos Santos Vieira, B., Kaliyaperumal, R., Abu-Hanna, A. and Cornet, R. et al. (2021) 'De-novo FAIRification via an electronic data capture system by automated transformation of filled electronic case report Forms into machine-readable data', *Journal of Biomedical Informatics*, Vol. 122. Doi: 10.1016/j.jbi.2021.103897.

Kuhn, T. (2015) 'nanopub-java: a java library for nanopublications', *Proceedings of 5th Workshop on Linked Science*, Vol. 1572, No. 1, pp.1–7.

Kuhn, T. (2022) 'GitHub – tkuhn/nanopub-server: a simple server to publish nanopublications', *GitHub*. Available online at: https://github.com/tkuhn/nanopub-server (accessed on 7 June 2022).

Kuhn, T., Chichester, C., Krauthammer, M. and Dumontier, M. (2016) 'Publishing without publishers: a decentralised approach to dissemination, retrieval, and archiving of data', *International Semantic Web Conference*, Springer International Publishing, Cham, pp.656–672.

Kuhn, T., Taelman, R., Emonet, V., Antonatos, H., Soiland-Reyes, S. and Dumontier, M. (2021) 'Semantic micro-contributions with decentralised nanopublication services', *PeerJ Computer Science*, Vol. 7. Doi: 10.7717/peerj-cs.387. eCollection 2021.

Mons, B. and Velterop, J. (2009) 'Nano-Publication in the e-science era', *Workshop on Semantic Web Applications in Scientific Discourse*, Vol. 2, No. 2.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L.O.B. and Wilkinson, M.D. (2017) 'Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science Cloud', *Information Services; Use*, Vol. 37, No. 1, pp.49–56.

Moreau, L., Groth, P., Cheney, J., Lebo, T. and Miles, S. (2015) 'The rationale of PROV', *Journal of Web Semantics*, Vol. 35, pp.235–257.

Oliveira, N.Q., Borges, V., Rodrigues, H.F., Campos, M.L.M. and Lopes, G.R. (2022) 'A practical approach of actions for FAIRification workflows: metadata and semantic research', *Communications in Computer and Information Science*, Springer, Cham. Doi: 10.1007/978-3-030-98876-0_8.

Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J. and Mungall, C. et al. (2016) 'Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration', *Nucleic Acids Research*, Vol. 45, No. D1, pp.D347–D352.

Roche, D.G., Kruuk, L.E.B., Lanfear, R. and Binning, S.A. (2015) 'Public data archiving in ecology and evolution: how well are we doing?', *PLOS Biology*, Vol. 13, No. 11. Doi: 10.1371/journal.pbio.1002295.

Shotton, D., Portwin, K., Klyne, G. and Miles, A. (2009) 'Adventures in semantic publishing: exemplar semantic enhancements of a research article', *PLoS Computational Biology*, Vol. 5, No. 4. Doi: 10.1371/journal.pcbi.1000361.

Sinaci, A.A., Núñez-Benjumea, F.J., Gencturk, M., Jauer, M-L., Deserno, T., Chronaki, C. and Cangioli, G. et al. (2020) 'From raw data to FAIR data: the FAIRification workflow for health research', *Methods of Information in Medicine*, Vol. 59, No. S01, pp.e21–e32.

Sustkova, H.P., Hettne, K.M., Wittenburg, P., Jacobsen, A., Kuhn, T., Pergl, R. and Slifka, J. et al. (2020) 'FAIR convergence matrix: optimising the reuse of existing FAIR-related resources', *Data Intelligence*, Vol. 2 No. 1–2, pp.158–170.

Thompson, M., Burger, K., Kaliyaperumal, R., Roos, M. and Da Silva Santos, L.O.B. (2020) 'Making FAIR easy with FAIR tools: from creolization to convergence', *Data Intelligence*, Vol. 2, Nos. 1/2, pp.87–95.

Van der Burg, S., Richardson, R. and Smits, D. (2021) *Nanopub: A Python Library for Searching, Publishing and Modifying Nanopublications*, GitHub.

Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T. and Musen, M.A. (2011) 'BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications', *Nucleic Acids Research*, Vol. 39, pp.W541–W545.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A. and Blomberg, N. et al. (2016) 'The FAIR guiding principles for scientific data management and stewardship', *Scientific Data*, Vol. 3, No. 1.

Websites

1 https://nanopub.petapico.org/

2 https://www.selenium.dev/

3 https://scrapy.org/

4 https://openrefine.org/

5 https://www.ncbi.nlm.nih.gov/home/about/policies/

6 https://creativecommons.org/licenses/by/4.0/deed

7 Omitted due to ongoing blind review; available on request.