A Nanopublishing Architecture for Biomedical Data

Pedro Sernadela[1,*], Eelke van der Horst[2], Mark Thompson[2], Pedro Lopes[1],
Marco Roos[2], and José Luís Oliveira[1]

[1] DETI/IEETA, University of Aveiro, Aveiro, Portugal
{sernadela,pedrolopes,jlo}@ua.pt
[2] Human Genetics Department, Leiden University Medical Center, Leiden, Netherlands
{e.van_der_horst,m.thompson,m.roos}@lumc.nl

Abstract. The massive production of data in the biomedical domain soon triggered a phenomenon known as information overload. The majority of these data are redundant without linked statements or associations, which hinders research methods. In this work, we describe an innovative and automated approach to integrate scientific results into small RDF-based data snippets called nanopublications. A nanopublication enhances attribution and ownership of specific data elements, representing the smallest unit of publishable information. It is particularly relevant for the scientific domain, where controlled publication, validation and ownership of data are essential. This proposal extends an existing semantic data integration framework by enabling the generation of nanopublications. Furthermore, we explore a streamlined integration and retrieval pipeline, empowered by current Semantic Web standards.

Keywords: Nanopublications, COEUS, Semantic Web, Data Integration.

1 Introduction

Peer-reviewed publications remain the main means for exchanging biomedical research information. However, there are several ways, apart from publishing and sharing scientific articles, in which researchers can contribute to scientific community, for example, the submission or curation of biological databases [1]. In both cases, most part of the information is actually growing at high levels [2] and it is increasingly difficult to find scientific data that are linked or associated, including provenance details. For example, if one researcher decides to investigate if the gene APP has some specific association with Alzheimer's disease, he may spent several days searching and analyzing the current scientific information available on the Web. This scenario will worsen if he wants to analyze multiple gene combinations in complex diseases, one of the most challenging domains of biomedical research.

In addition, few initiatives specify how academic credit is established for biomedical data sharing. Traditionally, the evaluation measure of a researcher's scientific career relies on his publication record in international peer-reviewed scientific journals. As stated above, there is a multitude of ways to contribute to the scientific community such as the submission and curation of databases entries and

J. Sáez-Rodríguez et al. (eds.), *8th International Conference on Practical Appl. of Comput. Biol. & Bioinform. (PACBB 2014)*, Advances in Intelligent Systems and Computing 294,
DOI: 10.1007/978-3-319-07581-5_33, © Springer International Publishing Switzerland 2014

records. In these specific cases, there is no successful way to attribute the credits of this work.

In an effort to tackle these challenges and with the dawn of the Semantic Web, a new strategy arises to interconnect and share data – nanopublications. With this approach for relating atomic data with its authors, accessing and exchanging knowledge becomes a streamlined process. The idea is that nanopublications are suitable to represent relationships between research data and efficient exchange of knowledge [3]. With the nanopublications format, most of experimental data or negative studies can be published in a standard format, such as RDF triples, instead of archived as supplemental information in an arbitrary format or independent databases. Researchers also need to access supporting data to make progress with their investigation. Analyzing only data content is not enough to fulfill most research studies requirements, becoming essential to analyze all the associated metadata. For these reasons, publishing this type of information as nanopublications will benefit similar studies saving time and unnecessary costs.

Additionally, even with the adoption of standards, some data sharing problems persist. The main reason for this is the lack of expertise by institutions or authors to transform local data into accepted data standards [4]. In this way, it is evident that researchers need an easy-setup mechanism that allows them to publish and share their scientific results through a reliable system.

In this paper, we propose to follow this idea presenting an innovative architecture to integrate automatically several data studies into a reusable format - the nanopublication. With this system, we make the transition from several common data formats to the Semantic Web paradigm, "triplifying" the data and making it publicly available as nanopublications. The main goal is to exploit the nanopublication format to efficiently share the information produced by the research community, assigning the respective academic credit to its authors. The proposed approach is provided as an extension of the COEUS[1] framework [5], a semantic data integration system. This framework includes advanced data integration and triplication tools, base ontologies, a web-oriented engine with interoperability features, such as REST (Representational State Transfer) services, a SPARQL (SPARQL Protocol and RDF Query Language) endpoint and LinkedData publication. Moreover, the resources can be integrated from heterogeneous data sources, including CSV and XML files or SQL and SPARQL query results, which will benefit our final solution.

This document is organized in 4 sections. Section 2 introduces the nanopublications standard and some related projects. Section 3 describes the system architecture. Finally, Section 4 discusses ongoing work and future research perspectives.

2 Background

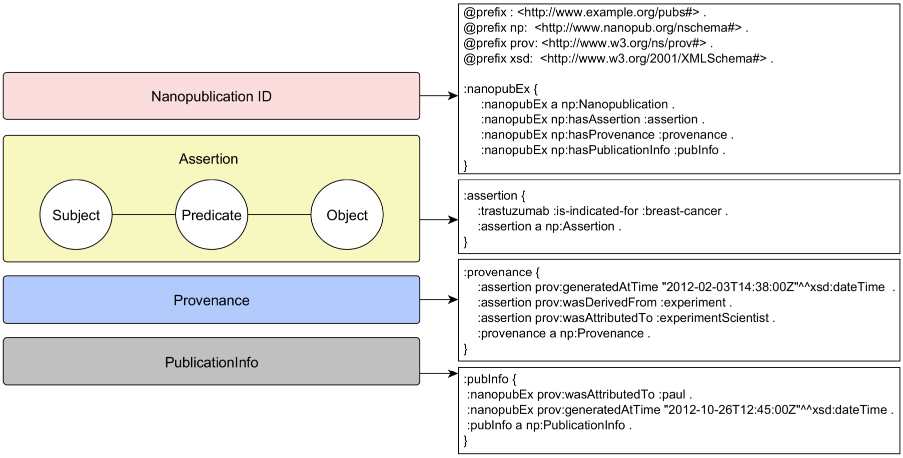
Nanopublications make it possible to report individualized knowledge assertions in a more efficient way. Due to the schema extensibility, it allows a useful aggregation

[1] http://bioinformatics.ua.pt/coeus

alternative to manage disparate data. Next, supported by the actual model, we analyze some uses cases demonstrating nanopublications' current potential.

2.1 The Nanopublication Model

The basic Semantic Web (SW) knowledge unit is built through the union of two concepts (subject and object) through a predicate, a triple statement, which formulates the assertion about something that can be uniquely identified. Nanopublications are also built on this SW strategy, allowing knowledge summarization to a set of thoroughly individualized list of assertions - the nanopublication [6]. It standardizes how one can attribute provenance, authorship, publication information and further relationships, always with the intention to stimulate information reuse. It is serializable through the interoperable RDF format, opening the door to new knowledge exchange possibilities and fostering their retrieval and use. Moreover, with universal nanopublications identifiers, each nanopublication can be cited and their impact tracked, encouraging compliance with open SW standards. Various efforts are under way to create guidelines and recommendations for the final schema [6]. Nowadays, the standard is being developed as an incremental process by the nanopublication community[2]. Figure 1 represents the basic model according to nanopublications schema[3].



```
@prefix : <http://www.example.org/pubs#> .
@prefix np:  <http://www.nanopub.org/nschema#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .

:nanopubEx {
    :nanopubEx a np:Nanopublication .
    :nanopubEx np:hasAssertion :assertion .
    :nanopubEx np:hasProvenance :provenance .
    :nanopubEx np:hasPublicationInfo :pubInfo .
}
```

```
:assertion {
   :trastuzumab :is-indicated-for :breast-cancer .
   :assertion a np:Assertion .
}
```

```
:provenance {
   :assertion prov:generatedAtTime "2012-02-03T14:38:00Z"^^xsd:dateTime  .
   :assertion prov:wasDerivedFrom :experiment .
   :assertion prov:wasAttributedTo :experimentScientist .
   :provenance a np:Provenance .
}
```

```
:pubInfo {
 :nanopubEx prov:wasAttributedTo :paul .
 :nanopubEx prov:generatedAtTime "2012-10-26T12:45:00Z"^^xsd:dateTime .
 :pubInfo a np:PublicationInfo .
}
```

Fig. 1. Basic nanopublication structure (left) with corresponding example (right)

The unique nanopublication identifier is connected to Assertion, Provenance and Publication Information objects. Each of these contains a set of axioms representing the nanopublication metadata. The Assertion graph must contain, at least, one assertion comprised by one or more RDF triples. Supporting information about these

[2] http://nanopub.org

[3] http://nanopub.org/nschema

assertions is included in the Provenance scope, where DOIs, URLs, timestamps or associated information can be described. Additional information, such as attribution, generated time, keywords or tags can be added too in the Publication Information graph to offer provenance information regarding the nanopublication itself.

In a sense, nanopublications are a natural response to the exploding number and complexity behind scientific communications. In this way, it offers not only a great opportunity to improve and publish conventional papers' research data, but also to explore experimental or negative data studies. Studies of this type are rarely published. Moreover, deploying data as nanopublications allows authors to receive the rightful credit for the shared content.

2.2 Related Projects

In 2008, the scientific journal Nature Genetics, was the first to introduce the concept of "microattribution" to enable an alternative reward system for scientific contributions [1]. Nevertheless, the first practical demonstration was only achieved in 2011, with a series of interrelated locus-specific databases to store all published and unpublished genetic variation related to hemoglobinopathies and thalassemia [7]. At the same time, some approaches emerged to deposit scientific results as nanopublications due to recent SW initiatives empowerment. Some of them are outlined next.

The Open PHACTS project[4] provides a nanopublications use case in their semantic knowledge infrastructure for public and commercial drug discovery research [8]. The nanopublications are used to store information as individual assertions from drug databases and from generated individuals through annotation tools. With the nanopublication format they promote data citation and provide credit to those producing important scientific assertions.

The LOVD nanopublication tool [1] encourages the submission of human genomic variant data for scientific community sharing. This application enables first-generation nanopublications from the Leiden Open-Access Variation Database[5] [9]. From the local database, the system populates a triple store and aggregates all different triples into nanopublications. The content can also be retrieved in XML format. Another module has also been developed for this tool to specify allele frequency data [5]. In this case, the data is submitted by uploading a pre-formatted Excel spreadsheet template in order to extract the data to the system, creating a nanopublication per record. To attribute work recognition the system uses the ResearcherID[6] user unique identity.

The Prizms approach [4] enables the creation of nanopublication data by providing an automated RDF conversion tool. The input data can be in any format, including CSV, XML, JSON and others formats. Making use of an extension of a data management infrastructure (Comprehensive Knowledge Archive Network – CKAN) it can cite derived datasets using the nanopublication provenance standards.

[4] http://openphacts.org/

[5] http://lovd.nl

[6] http://researcherid.com

Essentially, it generates RDF data to describe the datasets as a “datapub”, a nanopublication model for describing datasets, according to the authors. A public demonstration with 330 melanoma datasets is publicly available[7].

Other approach is related with exposing experimental data in life sciences. Mina *et al.* make use of the nanopublication model to create scientific assertions from the workflow analysis of Huntington’s Disease data, making it machine-readable, interoperable, and citable [10]. Mainly, they present how the results of a specific case study can be represented as nanopublications and how this integration could facilitate the data search by means of SPARQL queries. Also, they include and connect the nanopublications provenance graph to Research Objects (RO) [11], an aggregation object that bundles experimental studies resources. This linkage allows a context description of the workflow process. In contrast to nanopublications, RO encapsulate elements for an entire investigation, as opposed to individual claims [12].

The Nanobrowser portal[8] is a different approach that lets users create interactive and manual statements through the nanopublication concept. The tool uses English sentences to represent informal and underspecified scientific claims [13]. These sentences follow a syntactic and semantic scheme that the authors call AIDA (Atomic, Independent, Declarative, Absolute), which provides a uniform and succinct representation of scientific assertions [14]. Essentially, authors and curators manually write AIDA sentences, and text mining approaches automatically extract the content to create nanopublications assertions.

3 Nanopublishing Architecture

The previous projects show how nanopublications can be used in real world scenarios. However, there are several issues and challenges that still have to be addressed. Most of the available solutions target a specific domain (e.g. Open PHACTS, LOVD, etc.), which limits the creation of nanopublications by researchers. Others miss the main functionality that is actually needed: an automated transition from several data formats to nanopublications. In this way, we believe that certain features must be employed for a nanopublication transition solution to be successfully implemented in practice, which are described next:

1. The solution must accept common input data types (databases, delimited or structured files, etc.) and be capable of generating new nanopublications automatically, assigning the respective credit to its authors.
2. The application content, i.e. all nanopublications, must be created with the goal to be publicly available, promoting data sharing.
3. A search engine, supported by a SPARQL endpoint for instance, must be developed to provide a mechanism to query nanopublications.
4. A query federation solution for users’ information exchange must be available, according to LinkedData principles [15].
5. The solution must be easy to setup by researchers.

[7] http://data.melagrid.org

[8] http://nanobrowser.inn.ac

To tackle these requirements, we propose an extension to the COEUS' architecture to allow easy integration from several data formats to the nanopublications ontology graph. In the next section, we describe the main changes in the core system architecture to enable nanopublishing.

3.1 Integrating Nanopublications

The COEUS framework offers a good starting point to develop an architecture to support generic data loading and integration. Its main handicap is the data transformation process that must match the internal model ontology. Changing this strategy, COEUS' architecture will allow publishing universal scientific results automatically as nanopublications.

The COEUS engine provides a variety of connectors (CSV, XML, JSON, SQL, SPARQL, RDF, and TTL) to aggregate data from different sources. However, the "triplification" process is made through an organized ontology model. In this model, the data relationships are in an "Entity-Concept-Item" structure (e.g. Protein-Uniprot-P51587) to enable the integration of generic data into the Knowledge Base (KB). However, in the scenario addressed in this work, we know in advance the model of the data to be stored. Hence, we facilitate the user setup by completing automatically the nanopublications structure model. In this specific case, the user must only configure each "Resource" (data source properties) to integrate data as nanopublications. This introduces the first change to COEUS' internal setup.

The COEUS ontology translates the data elements into "Items" (coeus:Item), a basic representation of the produced data. As we are creating nanopublications, a specific data model, we associate a new predicate property: "Nanopublication" (np:Nanopublication). By adding this property to the internal ontology of the application, we split the core data transformation in 2 ways: the creation of COEUS concept data and the creation of nanopublication data. Attending to these modifications, the core application proceeds differently if the user wants to integrate data into COEUS' original model or into COEUS' nanopublication model.

To publish data as nanopublications, we also change the triples generation process. Due to the nanopublications schema, each nanopublication produced includes at least one Assertion along with the Provenance and PublicationInfo field. The creation of each field is an automatic and incremental process. The mapping between the data source and each nanopublication field content remains a manual user interaction process. This procedure is conducted through a specific COEUS web interface, facilitating the user interaction.

Figure 2 shows the new workflow diagram. The user starts by associating the data, creating one or more resources and their data endpoints (Figure 2, block 1). Linking the selected data to the nanopublications different fields will depend on the data type. This mapping process allows the engine to complete the nanopublication structure. Based on advanced Extract-Transform-and-Load (ETL) features, the engine generates, for each entry, a nanopublication record (Figure 2, block 2). Every nanopublication created is stored and made publicly accessible by several services (Figure 2, block 3).

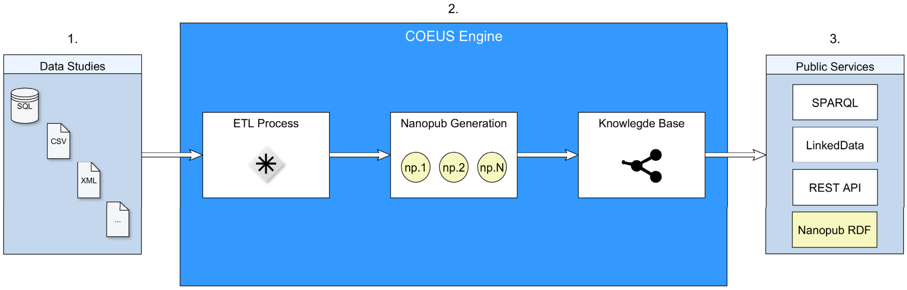


Fig. 2. Nanopublishing workflow: from generic data to nanopublications

To explore the data, COEUS has several interoperability features including REST services, a SPARQL endpoint and LinkedData interfaces. The creation of a nanopublication store forces this platform to adopt a new strategy to retrieve data. In this way, the system includes a RDF/XML exporting format option (represented in Figure 2 as “Nanopub RDF”), concordant with the nanopublication schema and accessible by a Uniform Resource Identifier (URI). We are also making collaborative efforts to maintain a compatibility interface with the nanopublication API, currently in development by the nanopublications community.

4 Discussion and Conclusions

The massive growth of scientific data generated year by year, including experimental data, begs for new strategies to grasp novel scientific outcomes. The nanopublication standard arises as a Semantic Web solution to this problem enabling researchers to synthesize and interconnect their results data. The appearance of this prominent solution, quickly triggered the provision of some tools. The majority are prototype solutions, each one targeting a specific domain. The approach described in this paper, intends to incentivize researches to publish and integrate their data as nanopublications in an easy way. Studies results can be generated in common formats to be submitted later to this framework. According to the workflow described, the user has the option to include the desired data into the engine, selecting and mapping the essential structured fields. Through the new design of ETL features on COEUS, we can integrate and deliver the data as nanopublications. This approach allows the reduction of redundancy and ambiguity of scientific statements contained in integrated studies. Also, it provides an attribution system with proper recognition to their authors, enabling appropriate data sharing mechanisms, according to LinkedData principles. With our expertise, we believe that such open source framework will benefit the research community and promote data sharing standards. We are also making efforts to offer an easy setup solution. By designing a new setup web interface we plan to make this framework more user-friendly and increase the researcher’s application range. In a near future, this work in progress will deliver, in a package, all tools needed to help researchers publish, store and retrieve all their outcomes as nanopublications.

Acknowledgments. The research leading to these results has received funding from the European Community (FP7/2007-2013) under ref. no. 305444 – the RD-Connect project, and from the QREN "MaisCentro" program, ref. CENTRO-07-ST24-FEDER-00203 – the CloudThinking project.

References

1. Patrinos, G.P., Cooper, D.N., van Mulligen, E., Gkantouna, V., Tzimas, G., Tatum, Z., Schultes, E., Roos, M., Mons, B.: Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. Hum. Mutat. 33, 1503–1512 (2012)
2. Velterop, J.: Nanopublications*: The future of coping with information overload. LOGOS J. World B. Community (2010)
3. Mons, B., Velterop, J.: Nano-Publication in the e-science era. Work. Semant. Web Appl. Sci. Discourse (2009)
4. McCusker, J., Lebo, T.: Next Generation Cancer Data Discovery, Access, and Integration Using Prizms and Nanopublications. Data Integr. Life Sci. 105–112 (2013)
5. Lopes, P., Oliveira, J.L.: COEUS: "semantic web in a box" for biomedical applications. J. Biomed. Semantics. 3, 11 (2012)
6. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Inf. Serv. Use (2010)
7. Giardine, B., Borg, J., Higgs, D.R., Peterson, K.R., Philipsen, S., et al.: Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. Nat. Genet. 43, 295–301 (2011)
8. Harland, L.: Open PHACTS: A semantic knowledge infrastructure for public and commercial drug discovery research. Knowl. Eng. Knowl. Manag. (2012)
9. Fokkema, I.F.A.C., Taschner, P.E.M., Schaafsma, G.C.P., Celli, J., Laros, J.F.J., den Dunnen, J.T.: LOVD v.2.0: The next generation in gene variant databases. Hum. Mutat. 32, 557–563 (2011)
10. Mina, E., Thompson, M.: Nanopublications for exposing experimental data in the life-sciences: A Huntington's Disease case study. In: Proc. 6th Int. Semant. Web Appl. Tools Life Sci. Work 2013 (2013)
11. Belhajjame, K., Corcho, O., Garijo, D.: Workflow-centric research objects: First class citizens in scholarly discourse. In: Proc. ESWC 2012 Work. Futur. Sch. Commun. Semant. Web (2012)
12. Belhajjame, K., Zhao, J., Garijo, D., Hettne, K., Palma, R., Corcho, Ó., Gómez-Pérez, J.-M., Bechhofer, S., Klyne, G., Goble, C.: The Research Object Suite of Ontologies: Sharing and Exchanging Research Data and Methods on the Open Web. arXiv Prepr. arXiv 1401.4307. 20 (2014)
13. Kuhn, T., Krauthammer, M.: Underspecified scientific claims in nanopublications. arXiv Prepr. arXiv1209.1483 (2012)
14. Kuhn, T., Barbano, P.: Broadening the scope of nanopublications. Semant. Web Semant. Big Data. 487–501 (2013)
15. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Int. J. Semant. Web Inf. Syst. (2009)