Exploring nanopublishing with COEUS

Pedro Lopes*, Pedro Sernadela, José Luís Oliveira

DETI/IEETA, University of Aveiro, Aveiro, Portugal
{pedrolopes,sernadela,jlo}@ua.pt

Abstract. A nanopublication represents the smallest unit of publishable information. This schema enhances attribution and ownership of specific data elements. With these guidelines for relating atomic data with its authors, accessing and exchanging knowledge becomes a more streamlined process. Nanopublications are particularly relevant in the scientific domain, where scientific publication, validation and ownership of data are essential.
The COEUS semantic web application framework delivers, in a single package, all the tools required to rapidly build a new semantic knowledge base from scratch, including multiple data integration algorithms and interoperability services.
This work introduces the combination of COEUS' integration and interoperability features with the nanopublications standard. This results in a unique nanopublishing pipeline, where collections of annotated data can be modeled and integrated, stored in a semantic knowledge base, and published through COEUS API. These improvements to the COEUS framework greatly benefit the scientific community, where creating and publishing nanopublications is still cumbersome.

Keywords: Semantic Web, nanopublications, data integration, application framework.

1 Introduction

The Semantic Web [1] paradigm introduces multiple technologies and strategies that are a perfect fit to represent real-world relationships in digital information systems, namely in the life sciences. Semantic Web standards tackle challenges in the most diverse domains, from data heterogeneity to service interoperability (or lack thereof) [2]. The cornerstone of this flexibility is the concept of nanopublications, a simple micro attribution strategy enabling the creation of machine-readable knowledge assertions, empowering a new structure level for the huge amounts of information flooding the scientific field [3, 4].

With this standard still its infancy, new tools are required to streamline the nanopublications generation and publishing process. Nowadays, this process is still manual, based on ad-hoc tools tailored to niche use cases.

This work introduces a strategy that exploits COEUS' [5] features to streamline the creation, storage and publishing of nanopublications. This new pipeline starts with a

translation process, integrating data from existing datasets into a semantic knowledge base, and modeling it according to the nanopublications standard. Aggregated knowledge is made available through COEUS' API, including REST services, a SPARQL endpoint, LinkedData interfaces and a nanopublications URI interface.

2 Background

Migrating systems to a Semantic Web environment is no different than the transition to previous paradigms. New technologies, algorithms and development strategies are introduced, making this transition a cumbersome task. The COEUS framework was built to overcome these challenges. The COEUS platform improves four key features in the development of new Semantic Web applications: (1) the transition from primitive to semantically enhanced systems; (2) the integration and triplification of data; (3) the sharing of knowledge through interoperable interfaces; (4) the deployment of a knowledge federation layer.

COEUS' flexible integration engine improves traditional data warehousing Extract-Transform-Load tasks, enabling the acquisition of data from heterogeneous resources (in CSV, JSON, XML, SQL, SPARQL, RDF and LinkedData) and its translation to a semantic data abstraction. The latter organizes knowledge in a cohesive structure, based on Entity-Concept-Item relationships. COEUS' API comprises various methods to access data (REST services, SPARQL endpoint, LinkedData interface, Java methods), making them easily available for querying and integration in external systems.

Nanopublications expand existing Semantic Web strategies to standardize how one can attribute provenance, authorship, publication information and further relationships, always with the intention to stimulate information reuse. In a sense, nanopublications are a natural response to the exploding number and complexity behind scientific data. With this standard, we can summarize published knowledge to a set of thoroughly individualized list of assertions - the nanopublication.

In summary, nanopublications are composed of three sections, each detailing assertion information, authorship and provenance, for elaborate knowledge statements [6] – Figure 1. Nanopublications are serializable through the interoperable RDF format,

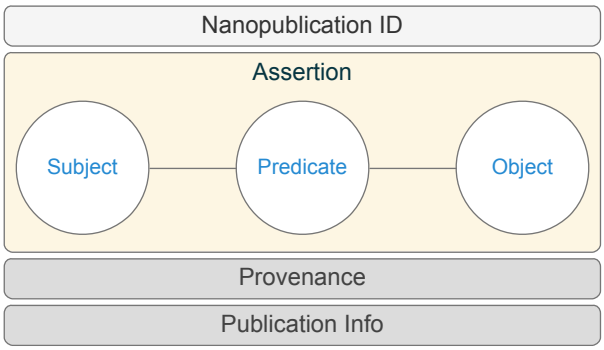

Fig. 1. Basic nanopublication structure: the unique nanopublication identifier is connected to Assertion, Provenance and Publication Information objects. Each of these contains a set of axioms representing the nanopublication metadata.

opening the door to many 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 semantic web standards. In addition to normalized positive assertions, Semantic Web's expressiveness can also be leveraged to expose negative knowledge assertions

3 Methods

The rational behind this work is to extend the base COEUS framework with support for nanopublications. This will enable a new nanopublishing pipeline where all tasks are automated. This process includes three general steps: (1) configuring the data abstraction to a semantic model - Figure 2-1; (2) integrating & translating data from external resources into the internal knowledge base - Figure 2-2; (3) sharing the nanopublications dataset - Figure 2-3. Adding nanopublishing support to COEUS leverages on its flexibility.

Before the actual integration process, we need to setup where the data comes from and how it will be represented according to the nanopublications format. COEUS setup allows organizing data according to a predefined hierarchy, based on an Entity-Concept-Item structure. With the nanopublication extension, two new properties were added to COEUS' configuration to enable the automated creation of nanopublications. COEUS integration engine was updated to detect these settings and proceed accordingly. The setup configuration changes are detailed next.

- **coeus:isNanopublication**. This predicate can be applied to the Concept configuration metadata, defining the base data imports for creating new nanopublications. Where this property is enabled, COEUS' integration engine automatically generates new nanopublications and their respective URIs autonomously.
- **coeus:np_element**. This predicate can be applied to the Resource configuration

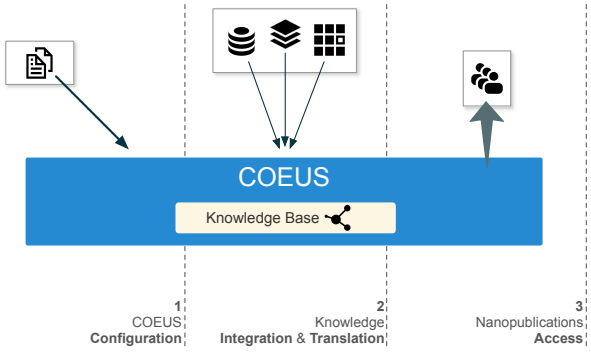

Fig. 2. Nanopublishing pipeline activities. (1) Configuration: data translations, from primitive formats to the nanopublications standard, are defined. (2) Integration and translation: COEUS access configured resources on-the-fly and automatically generates a nanopublications knowledge base. (3) Access: the nanopublications dataset can be queried through any of the available methods in the COEUS API.

metadata (for external resources associated with a nanopublication Concept), defining the type of data being loaded: a new assertion, provenance or publication information object.

In addition to these nanopublications-specific elements, the already existing dynamic properties can still be used to add any predicate from any ontology to the newly generated nanopublications.

4 Conclusion

Nanopublications arise as a new strategy to cope with knowledge provenance, ownership and sharing issues. The standard enables publishing comprehensive datasets, featuring large datasets, as a collection of rich individual assertions.

This work introduces an innovative nanopublishing pipeline. By extending COEUS, we enable a new semantic web framework to flexibly generate nanopublications dataset. These results streamline the translation of data in primitive formats to a semantic environment and, consequently, its delivery through open web interfaces. Furthermore, COEUS is now a "turn-key" nanopublishing solution, making the nanopublications dataset generation process much more agile.

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. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Sci Am 284, 34 - 43 (2001)
2. Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M.S., Ogbuji, C., Rees, J., Stephens, S., Wong, G.T., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.-H.H.: Advancing translational research with the Semantic Web. BMC Bioinformatics 8, S2-S2 (2007)
3. Velterop, J.: Nanopublications*: the future of coping with information overload. LOGOS: The Journal of the World Book Community 21, 3-4 (2010)
4. Mons, B., van Haagen, H., Chichester, C., den Dunnen, J.T., van Ommen, G., van Mulligen, E., Singh, B., Hooft, R., Roos, M., Hammond, J.: The value of data. Nature genetics 43, 281-283 (2011)
5. Lopes, P., Oliveira, J.L.: COEUS:"semantic web in a box" for biomedical applications. Journal of Biomedical Semantics 3, 1-19 (2012)
6. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use 30, 51-56 (2010)