Automated Nanopublications Generation from Biomedical Literature

Pedro Sernadela and José L. Oliveira

Abstract—**The continuous growth of unstructured information resulting from biomedical research is a trending challenge for the scientific community. In this way, novel methods for information management are emerging to improve knowledge distribution and access. The concept of nanopublications illustrates one of these recent strategies to implement machine-readable knowledge assertions. It tries to overcome inconsistency, ambiguity and redundancy of traditional publications. The purpose is that they are more suited than traditional papers to represent relationships that exist between research data, providing an efficient mechanism for knowledge exchange. Although the evident benefits of these RDF-based snippets, its applicability stills challenging due to the inexistence of extraction and publications methods. To solve that issue, we propose an automated workflow for nanopublications generation from biomedical literature. The proposed method consists of exploring an automated information extraction tool for relevant information detection from published documents and respective standardization of the mined information through semantic web recommendations, for further exploration.**

I. Introduction

Accessing or analysing the vast amount of information resulting from biomedical research is a trending challenge for the scientific community. The continuous growth of this unstructured data brings several concerns regarding information management. For that reason, it becomes essential to explore novel strategies to deal with scientific outcomes. In this way, the concept of nanopublications [1] is a natural response to the increasing number and complexity behind scientific communications. It tries to overcome inconsistency, ambiguity and redundancy of traditional publications, improving scientific results distribution and access. By exploiting semantic web features, nanopublications are more suitable to represent associations between research data than traditional papers. Furthermore, they include provenance details, offering an effective mechanism for knowledge exchange. Currently, the applicability of this concept has been well accepted by the scientific community. However, there stills exists challenges for the successful extraction and publication of these concise RDF-based snippets [2, 3]. In this paper, we discuss these translation issues and propose an automated workflow for nanopublications generation from biomedical literature. This workflow aims to support the integration of information extraction systems using nanopublication schema. Thus, enabling data distribution and exchange. The proposed method consists of exploring an automated information extraction tool for relevant information detection (e.g. concepts and respective relationships) from published documents. The method also includes the standardization of the mined information through semantic web recommendations, for further exploration.

II. Background

A. Information Extraction

Over the last decade, the number of biomedical information extraction tools has been growing steadily. Latest strategies use computational solutions to aid in the extraction and storage of relevant concepts, as well their respective attributes and relationships. The outcome of these complex workflows provides valuable insights over the overwhelming amount of biomedical information being produced.

Text-mining solutions have been increasingly applied to assist bio-curators, allowing the extraction of biomedical concepts such as genes, proteins, chemical compounds or diseases [4], and thus reducing curation times and cost [5]. Usually, state-of-the-art solutions follow a combination of pre-defined and sequential processes to apply and perform biomedical information extraction. Natural Language Processing (NLP) techniques [6] are commonly applied as pre-processing tasks to split documents text into meaningful components, such as sentences and tokens, assign grammatical categories (a process named part-of-speech tagging), or even apply linguistic parsing to identify the structure of each sentence. Next, concept recognition methods are employed, which involve Named Entity Recognition (NER) [4] to detect the concept mentions, and normalization processes [7] to distinguish and attribute unique identifiers to each detected entity name. Complete biomedical text-mining solutions also apply relation-mining techniques to identify the events and entity relations that make up complex biological networks. Conventional solutions are focused on investigating and extracting direct associations between two concepts (e.g. genes, proteins, drugs, etc.) [8]. The study of these associations has generated plenty of interest, especially in respect to protein-protein interactions (PPIs) [9] [10], drug-drug interactions (DDIs) [11], and relations between chemicals and target genes. Other solutions, such as FACTA [12], are targeted at uncovering implicit heterogeneous connections between different types of concepts.

Pedro Sernadela is with DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal (phone: +351 234370500, email: sernadela@ua.pt).

José L. Oliveira is with DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal (email: jlo@ua.pt).

B. *Semantic Web*

In recent years, semantic web became the *de facto* paradigm for data integration at a web-scale, focused on the semantics and the context of data [13]. It enables the creation of rich networks of linked data, establishing new possibilities to retrieve and discover knowledge (e.g. reasoning). Moreover, semantic web makes data integration and interoperability a standard feature, enabling the representation of data elements in the web as real-world entities and links, capable of creating large and complex networks for both human and machines consumption. Supported by technologies such as RDF (Resource Description Framework), OWL (Web Ontology Language) and SPARQL (SPARQL Protocol and RDF Query Language), technical recommendations of the World Wide Web Consortium, it facilitates the deployment of advanced algorithms for searching and mining large integrated datasets [14].

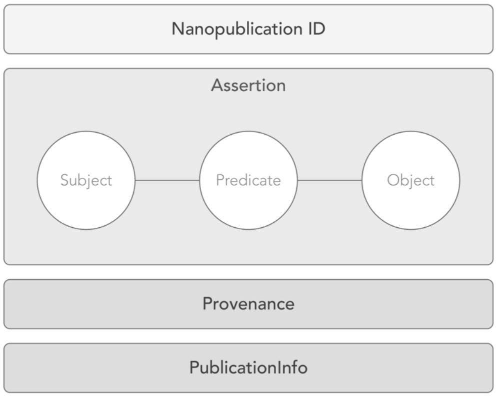
Strategies that combine the benefits of information extraction methods with semantic web features represent a growing trend that allows the establishment of curated databases with improved availability [15]. In this way, combining different text-mining results with semantic web technologies allow curation outcomes to be discoverable and shared across multiple research institutions. Coulet et al. [16] provide an overview of such strategies.

Some open-source applications have also been released to support the publication of text-mined information through semantic web standards. One approach is provided through the PubAnnotation [17] prototype repository. The idea was to build a sharable repository, where various corpora and annotations can be stored together and queried through SPARQL. However, the service uses specific JSON annotation files as its default input format (submitted through a REST API) producing limitations or incompatibilities when it is necessary to include annotations generated from others text-mining tools. A more recent approach is provided by the Ann2RDF tool [18], an interoperable semantic layer that unifies text-mining results originated from different tools, information extracted by curators, and baseline data already available in reference knowledge bases, enabling a proper exploration using semantic web technologies. The python-based tool automatically accepts the most common text-mining data formats, such as the BioC and standoff format and outputs a RDF/XML Knowledge Base with all processed annotations.

C. *Nanopublications*

Nanopublications are a natural response to the increasing number and complexity behind scientific communications. Thus, they attempt to overcome the inconsistency, ambiguity and redundancy of traditional publications, improving information extraction and analysis. With this strategy, researchers can access more quickly not only the more relevant information but also the supporting data and related metadata. Furthermore, it intends to reduce the publishing time and the search period, streamlining the investigation procedure. For these reasons, deploying data as nanopublications will benefit similar studies, saving time and unnecessary costs. The basic semantic web knowledge unit is built through the union of two concepts (subject and object) via a predicate, a triple statement, which formulates the assertion about something that can be uniquely identified. Nanopublications are also built on this semantic web strategy, allowing knowledge summarization. It standardizes how provenance, authorship, publication information and further relationships can be attributed, 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 retrieval and use. Moreover, with universal nanopublication identifiers, each one 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 [1]. Nowadays, the nanopublication community (http://nanopub.org) is developing this standard through an incremental process. Figure 1 represents the basic model according to nanopublication schema (http://nanopub.org/nschema). The unique nanopublication identifier is connected to Assertion, Provenance and Publication Information. Each of these contains a set of triples representing the nanopublication metadata. The Assertion graph must contain at least one assertion comprised by one or more RDF triples. Supporting information about these assertions is included in the Provenance graph, where DOIs, URLs, timestamps or associated information can be described. Additional information, such as attribution, generated time, keywords or tags can be added to a Publication Information graph to offer provenance information regarding the nanopublication itself.

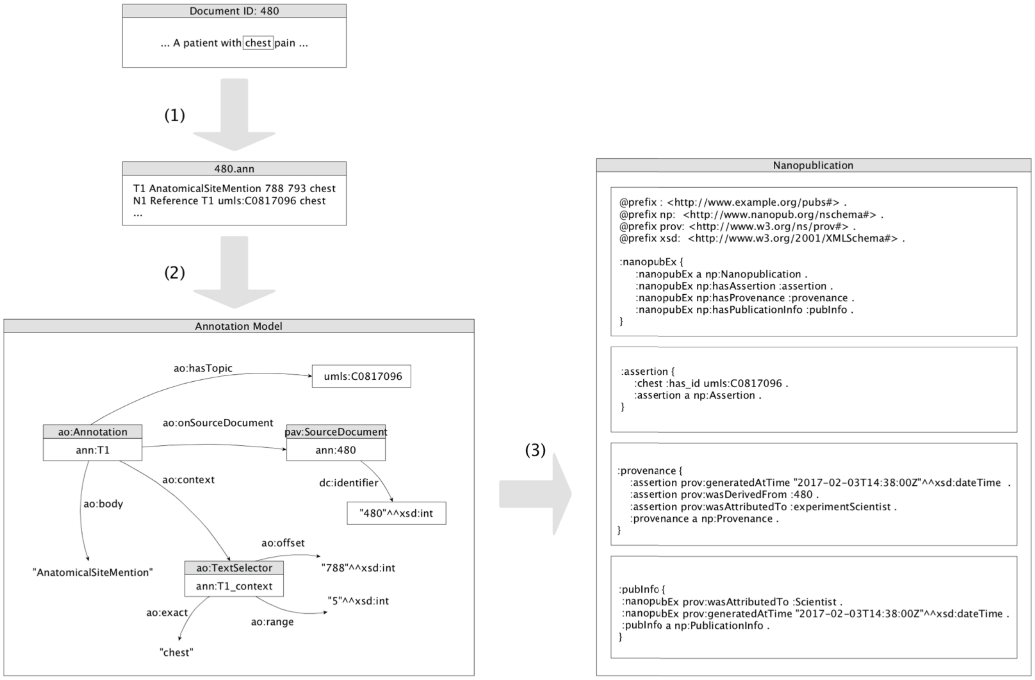
Figure 1. Nanopublication model overview.



III. Methods

This article proposes a pipeline for automated generation of nanopublications from unstructured information. To enable an automated generation, our method involves 3 main steps: information extraction, semantic web translation and nanopublication conversion. Figure 2 shows the overall pipeline from textual information to the nanopublication schema, including all transformation steps of the workflow. An example of the concept recognition *chest* is provided to illustrate how a simple and concise nanopublication can be created by our method.

Figure 2. Automated pipeline for nanopublications conception from textual information. (1) Information extraction tools such as Neji services [19] extracts concepts and relationships; (2) Annotation files are consumed by the Ann2RDF tool [18], converting the extrated annotations into a unifyed and semantic web-based annotation model; (3) Nanopublications are created by mapping the semantic web-based annotations to the nanopublication schema.



```
T1 AnatomicalSiteMention 788 793 chest
N1 Reference T1 umls:C0817096 chest
...
```

```
@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 {
    :chest :has_id umls:C0817096 .
    :assertion a np:Assertion .
}
```

```
:provenance {
   :assertion prov:generatedAtTime "2017-02-03T14:38:00Z"^^xsd:dateTime  .
   :assertion prov:wasDerivedFrom :480 .
   :assertion prov:wasAttributedTo :experimentScientist .
   :provenance a np:Provenance .
}
```

```
:pubInfo {
 :nanopubEx prov:wasAttributedTo :Scientist .
 :nanopubEx prov:generatedAtTime "2017-02-03T14:38:00Z"^^xsd:dateTime .
 :pubInfo a np:PublicationInfo .
}
```

A. Information Extraction

For the information extraction process (Figure 2, step 1), our solution is based on a modular framework for biomedical natural language processing called Neji [19]. This open-source framework was developed around four crucial characteristics: modularity, scalability, speed and usability. It follows several state-of-the-art methods for biomedical natural language processing, such as sentence splitting, tokenization, lemmatization, part-of-speech, chunking and dependency parsing. The concept recognition tasks are performed using dictionary matching and machine learning techniques with normalization. This framework implements a very flexible and efficient concept tree, where the recognized concepts are stored, which supports nested and intersected concepts with one or more identifiers. It supports several input and output formats including the most popular ones in biomedical text mining, namely IeXML, Pubmed XML, A1, CONLL and BioC. The architecture of Neji allows users to configure the processing of documents according to their specific objectives and goals, providing a very rich and complete concepts information.

The main component of Neji is the processing pipeline, which can be composed by several modules that will be executed following a FIFO strategy. Therefore, a pipeline is nothing more than a list of independent modules, each of them responsible for a specific processing task, that are executed sequentially.

B. Semantic Web Translation

Computational information extraction tools produce several annotation formats, creating a barrier to efficiently combine and exchange this information. The migration of this mined information into semantic web format and services, provides an additional value to share that knowledge. To assist our method (Figure 2, step 2), we take benefit of Ann2RDF [18] features to convert text-mined results into an open representation model. This short-term model is based on the Annotation Ontology (AO) [20] for the pre-representation of annotated biomedical scientific documents to formal ontological elements. In the Figure 2 (step 2), we show the core annotation model adopted. The central point of the annotation representation includes the annotation URI (e.g. *T1*), the document source (e.g. ID: *480*) and the respective annotated data (e.g. *chest*). The text selectors are used to identify the string detected on the document: the *ao:exact* data property represents the linear sequence of characters, i.e. the subject of the annotation, the *ao:offset* data property an integer indicating the distance from the beginning of the document up to a given element or position, and the *ao:range* data property an integer indicating the number of characters starting from the offset. Information about the annotation itself is connected through two different properties: the *ao:body* representing the annotated resource and the *ao:hasTopic* indicating that the annotated resource can be identified as indicated topic, i.e. the semantic identifier of the detected resource (e.g. *C0817096*). Moreover, the annotations are linked to the respective document source through the object property *ao:onSourceDocument* providing a provenance interchange mechanism. By using this simplified model, entity annotations can be easily mapped to RDF. The use of Ann2RDF tool results in a more suitable transition process, in which annotations are automated enriched and transformed following semantic web recommendations.

C. Nanopublications Generation

The nanopublication creation process (Figure 2, step 3) is supported entirely by an automated mapping process that extracts relevant data from the semantic web-based annotations. These data are harmonized to be included in the several fields of the nanopublication structure as mentioned in Figure 1. In the example of Figure 2, we have only represented a sample generated assertion for better visualization purposes. However, many others assertions are extracted from the curated data. Provenance information is also included due to the easy extraction of authors information from the published documents.

IV. Conclusion

The continued exponential growth of biomedical data has made it harder than ever for researchers to find and assimilate all the information relevant to their research. For those reasons, various efforts and strategies have been promoted to try summarize the knowledge scattered across multiple publications and store it in structured form, for further use. The nanopublication strategy aims to solve these issues and current problems with the extraction of relevant information, redundant data, lack of associations and provenance information. Therefore, novel and automated strategies are needed to explore the evident value of nanopublications and to enable data attribution mechanisms, an important feature for data owners. However, creating these valuable resources is a time-demanding task that requires the support of computerized solutions. In this paper, we have explored that issue by proposing an automated workflow to generate summarized and concise assertions from published documents, according to the nanopublication schema. The pipeline proposed uses information extraction techniques to identify susceptible concepts from biomedical literature and of a suitable semantic web-powered strategy for represent essential pieces of publishable information. In this way, the proposed method allows summarization of findings through state-of-the-art knowledge sharing mechanisms, allowing publishing and integration of text-mined information results into the semantic web ecosystem in a suitable exchange form. The described pipeline can also be applied to others research areas, in which, the information can be reduced to a handful facts for further exploration.

Acknowledgment

The research leading to these results has received funding from the European Community (FP7/2007-2013) under ref. no. 305444 – the RD-Connect project. Pedro Sernadela is funded by Fundação para a Ciência e Tecnologia (FCT) under the grant agreement SFRH/BD/52484/2014.

References

[1] P. Groth, A. Gibson, and J. Velterop, “The anatomy of a nanopublication,” *Inf. Serv. Use*, 2010.

[2] J. Velterop, “Nanopublications*: the future of coping with information overload.,” *LOGOS J. World B. Community*, 2010.

[3] P. Sernadela, E. Van der Horst, M. Thompson, P. Lopes, M. Roos, and J. Oliveira, “A Nanopublishing Architecture for Biomedical Data,” in *8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)*, vol. 294, Springer International Publishing, 2014, pp. 277–284.

[4] D. Campos, S. Matos, and J. Oliveira, “Current Methodologies for Biomedical Named Entity Recognition,” *Biol. Knowl. Discov. Handb. Preprocessing, Mining, Postprocessing Biol. Data*, pp. 839–868, 2012.

[5] B. Alex, C. Grover, and B. Haddow, “Assisted Curation: Does Text Mining Really Help?.,” *Pacific Symp. Biocomput.*, vol. 13, 2008.

[6] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, “Natural language processing: an introduction.,” *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 544–51, Jan. 2011.

[7] M. Schuemie, J. Kors, and B. Mons, “Word sense disambiguation in the biomedical domain: an overview,” *J. Comput. Biol.*, 2005.

[8] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak, and B. Shen, “Biomedical text mining and its applications in cancer research.,” *J. Biomed. Inform.*, vol. 46, no. 2, pp. 200–11, Apr. 2013.

[9] Q.-C. Bui, S. Katrenko, and P. M. A. Sloot, “A hybrid approach to extract protein-protein interactions.,” *Bioinformatics*, vol. 27, no. 2, pp. 259–65, Jan. 2011.

[10] R. Hoffmann and A. Valencia, “Implementing the iHOP concept for navigation of biomedical literature.,” *Bioinformatics*, vol. 21 Suppl 2, no. suppl_2, p. ii252-8, Sep. 2005.

[11] L. Tari, S. Anwar, S. Liang, J. Cai, and C. Baral, “Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism.,” *Bioinformatics*, vol. 26, no. 18, pp. i547-53, Sep. 2010.

[12] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, “FACTA: a text search engine for finding associated biomedical concepts.,” *Bioinformatics*, vol. 24, no. 21, pp. 2559–60, Nov. 2008.

[13] C. M. Machado, D. Rebholz-Schuhmann, A. T. Freitas, and F. M. Couto, “The semantic web in translational medicine: current applications and future directions.,” *Brief. Bioinform.*, Nov. 2013.

[14] D. J. Wild, Y. Ding, A. P. Sheth, L. Harland, E. M. Gifford, and M. S. Lajiness, “Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research.,” *Drug Discov. Today*, vol. 17, no. 9–10, pp. 469–74, May 2012.

[15] J. Laurila, N. Naderi, and R. Witte, “Algorithms and semantic infrastructure for mutation impact extraction and grounding,” *BMC Genomics*, vol. S24, 2010.

[16] A. Coulet, Y. Garten, M. Dumontier, R. B. Altman, M. A. Musen, and N. H. Shah, “Integration and publication of heterogeneous text-mined relationships on the Semantic Web.,” *J. Biomed. Semantics*, vol. 2 Suppl 2, p. S10, Jan. 2011.

[17] J. Kim and Y. Wang, “PubAnnotation: a persistent and sharable corpus and annotation repository,” *Proc. 2012 Work. Biomed. Nat. Lang. Process. Assoc. Comput. Linguist.*, 2012.

[18] P. Sernadela, S. Matos, and J. L. Oliveira, “Ann2RDF: moving annotations to semantic web,” in *Proceedings of the 17th International Conference on Information Integration and Web-based Applications &Services - iiWAS ’15*, 2015, pp. 1–5.

[19] D. Campos, S. Matos, and J. Oliveira, “A modular framework for biomedical concept recognition,” *BMC Bioinformatics*, vol. 14, no. 1, p. 281, 2013.

[20] P. Ciccarese, M. Ocana, L. J. Garcia Castro, S. Das, and T. Clark, “An open annotation ontology for science on web 3.0.,” *J. Biomed. Semantics*, vol. 2 Suppl 2, no. Suppl 2, p. S4, Jan. 2011.