Representing NCBO Annotator results in standard RDF with the Annotation Ontology

Soumia Melzi, Clement Jonquet

Representing NCBO Annotator results in standard RDF with the Annotation Ontology

Soumia Melzi and Clement Jonquet

Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM)
& Computational Biology Institute (IBC) of Montpellier
University of Montpellier, France
soumia.melzi@lirmm.fr, jonquet@lirmm.fr

Abstract. Semantic annotation is part of the Semantic Web vision. The Annotation Ontology is a model that have been proposed to represent any annotations in standard RDF. The NCBO Annotator Web service is a broadly used service for annotations in the biomedical domain, offered within the BioPortal platform and giving access to more than 350+ ontologies. This paper presents a new output format to represent the NCBO Annotator results in RDF with the Annotation Ontology. We briefly present both technologies and describe the mappings to enable the representation. A Java library is available to parse the current JSON outputs to RDF/XML format. By rendering results in RDF, we make the annotations generated by the NCBO Annotator follow the Semantic Web standards making possible among other things to offer them as linked data.

Keywords: Semantic Web, biomedical ontologies, semantic annotation, NCBO Annotator, RDF, linked-data, Annotation Ontology.

1 Introduction

In the Semantic Web vision, semantic annotations enable data to be represented and tagged with ontologies thus facilitating data integration, interoperability, indexing and search [3]. In order to avoid the paradox where the variety of annotation tools and formats will become as large as the data to annotate, we need to develop common models for representing and sharing semantic annotations independently of the tools or experts that have generated them. The Annotation Ontology (AO) is such a model allowing to represent any annotations in standard RDF [1]. In the following, we present how we use AO to represent biomedical semantic annotations returned by the NCBO Annotator. Similar contribution has already been offered inside the DOMEO annotation tool [2], however our approach differs as it provides a (Java) library that could be plugged to any annotation postprocessing process. In addition, we explicitly describe the mappings (Table 1) also allowing anyone to extend the library with outputs in other format such as JSON-LD or other RDF syntaxes.

2 Background - the NCBO Annotator & the Annotation Ontology

The NCBO Annotator Web service [4] (http://bioportal.bioontology.org/annotator), is an annotation tool offered within the BioPortal platform [6] and giving access to

more than 350+ biomedical ontologies. The annotation workflow is based on a highly efficient syntactic concept recognition tool (using concept names and synonyms) and on a set of semantic expansion algorithms that leverage the semantics in ontologies (e.g., is-a relations and mappings). The Annotator is parameterizable to customize the annotation process (ontologies to include, use of semantic expansion, stopwords, longest match only, etc.) and when used as a web service through the REST API (http://data.bioontology.org/documentation#nav_annotator), the service returns JSON or XML outputs. The outputs do not include anything about the data being annotated (the inputs) nor the parameters, but it does offers for each annotating concept the following metadata:

- Description of the annotating concept (`annotatedClass`) with its URI, and references to its description, ontology, children, parents, descendants, ancestors, tree, notes, mappings and UI link in BioPortal;
- If applicable, description of the parents concepts (`hierarchy`) also annotating the data, with information about the ancestor level;
- If applicable, description of the mapped concepts (`mappings`) also annotating the data, with information about the inter-ontology mapping;
- The set of terms in the text data that have generated the `annotations` with for each: the character position within the text (from & to), the type of match (preferred name or synonym) and the exact piece of text that has been matched.

Figure 1 shows a portion of the results for a piece of text mentioning "cancer". The term *cancer* in the text has generated an annotation with the concept *Neoplasms* in MeSH: `http://purl.bioontology.org/ontology/MESH/D009369`

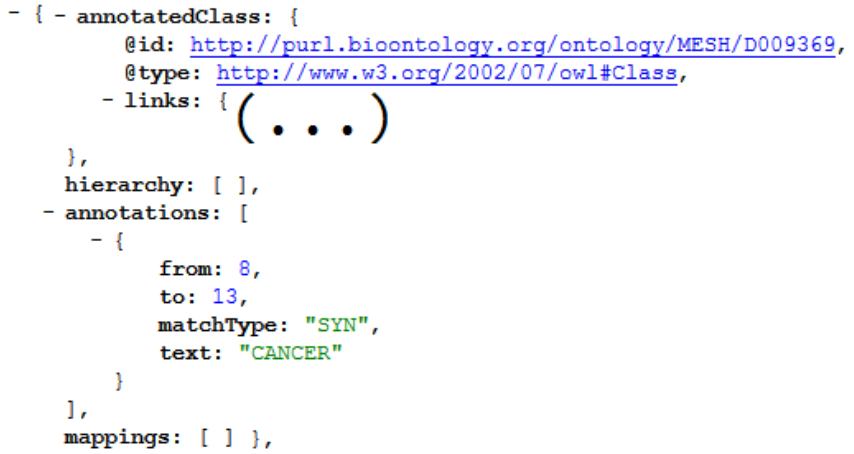

Fig. 1. Example of annotation returned by the Annotator Web service

The Annotation Ontology is a OWL ontology (http://purl.org/ao/) [1] proposed by Harvard Medical School researchers to represent any kind of semantic annotations either generated by automatic tool or by human experts, about any kind of data (text, image, video, etc.). It offers provenance information as well as metadata about the

annotations. AO helps to address the need for open standards to index and represent scientific data as linked data within the Semantic Web. It is a vocabulary, originally inspired from the Annotea initiative beginning of the 2000's [5], that actually reuses several defacto standards e.g., for provenance (PAV), metadata (Dublin Core), peoples (FOAF), communities (SIOC). Figure 2 shows and example of annotation represented using AO. We can notice different parts such as: (i) the *annotation provenance*, which includes the information about the tool or person that has created the annotation and when; (ii) the *document provenance*, which includes information about the data being annotated; (iii) the *annotation topic*, which represents the annotating concept with its URI; (iv) the *annotation type* and (v) the *annotation selector* as described after.

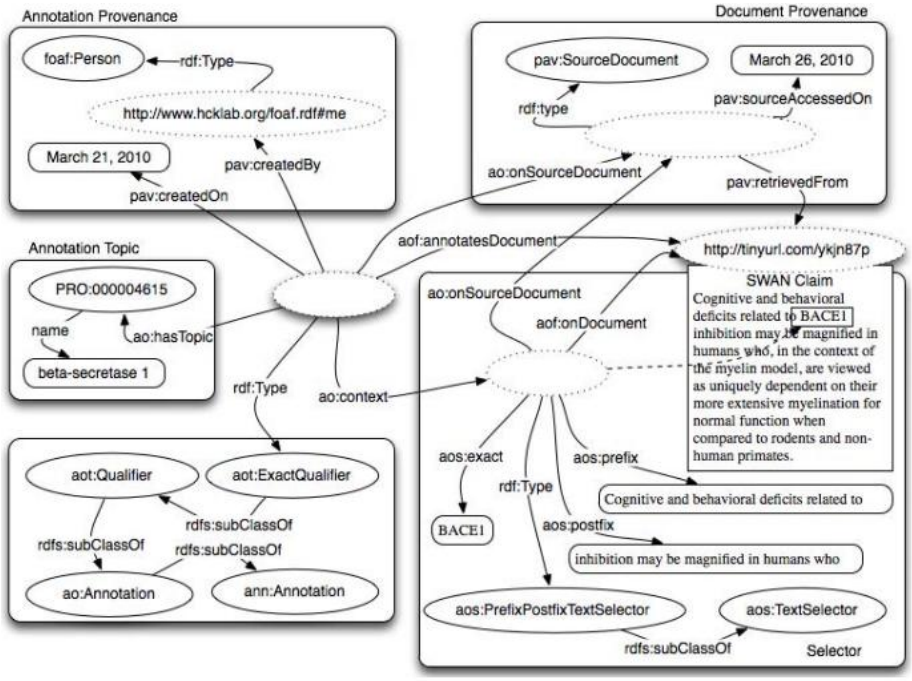

Fig. 2. Example of representation of an annotation with the Annotation Ontology (from [1]).

AO defines several types of selectors to identify part of the resource being annotated (part of a text document, section of an image, audio excerpt, etc.). In the following, we will use the type `aos:TextSelector` which allows to identify part of a text document either by: (i) specifying the number of characters from the beginning of the document to the part being annotated and the offset (OffsetRangeSelector) and (ii) by specifying a short prefix and postfix phrase (PrefixPostfixSelector), case illustrated in Figure 2. AO defines several types of annotations such as note, errata, example, definition or SKOS like qualifiers and it is possible to subclass or combine those main types. Especially, qualifiers are used when explicitly annotating with an RDF resource (with an URI) not just a tag. Figure 2 illustrates an `ExactQualifier` generally used when

the object of the relationship `ao:hasTopic` is representing exactly the portion of the annotated document.

3 NCBO Annotator results represented with AO

The Annotator only deals with free text data for the moment, therefore we have used the `aos:OffsetRangeSelector` to describe the context of annotations. Plus, we have used: `ExactQualifier` for direct annotations made with a preferred name and `Qualifier` otherwise.[1] Table 1 list the other alignements between the AO ontology and the Annotator annotation format.

Table 1. Mappings between NCBO annotation properties and AO properties.

Description	AO property	NCBO annotation property used or created
Exact matching term	*ao:exact*	*annotations:text*
Number of characters since the beginning of the document and the matching term	*ao:offset*	*annotations:from*
Size of the matching term	*ao:range*	*annotations:to* − *annotations:from*
Annotating concept	*ao:hasTopic*	*annotatedClass:id*
Annotation creation date	*pav:createdOn*	Not available - populated additionally when generating RDF
The annotation tool	*pav:createdBy*	Static value: http://bioportal.bioontology.org/annotator (cf. bottom of Fig. 3)

Because of missing information about the input data, we had to populate ourselves the `aof:annotatesDocument` and `ao:onSourceDocument` properties with an automatic generated id concatenated to the *pav:createdBy* property value. Figure 3 shows the RDF outputs generated for the annotation with *Neoplasms* used previously.

4 Conclusion

In this paper we have presented our approach to offer the NCBO Annotator results in RDF represented with a reference ontology for annotation of scientific data: the Annotation Ontology. In addition to the adoption of a standard RDF format facilitating the release of annotations as linked data, users have also access to all Semantic Web technologies to display or query their annotations. Considering the large number of annotations generated by different annotation tools, having them in RDF allows to query them semantically using for instance SPARQL or OWL descriptions. The parser to transform JSON to RDF outputs is available as a Java library that simply takes a JSON file

[1] In the next version of the parser, we will use `ExactQualifier` for both preferred name and synonym matches, `BroadQualifier` for is-a hierarchy annotations and `CloseQualifier` for mapping-based annotations.

```
<rdf:Description rdf:about="http://bioportal.bioontology.org/annotator/ann/4019/1">
    <rdf:type rdf:resource="http://purl.org/ao/types/Qualifier"/>
    <rdf:type rdf:resource="http://purl.org/ao/Annotation"/>
    <rdf:type rdf:resource="http://www.w3.org/2000/10/annotation-ns#Annotation"/>
    <aof:annotatesDocument rdf:resource="http://data.bioontology.org/annotator?4019"/>
    <ao:context rdf:resource="http://bioportal.bioontology.org/annotator/sel/4019/1"/>
    <ao:hasTopic rdf:resource="http://purl.bioontology.org/ontology/MESH/D009369"/>
    <pav:createdOn>23-08-14</pav:createdOn>
    <pav:createdBy rdf:resource="http://bioportal.bioontology.org/annotator"/>
</rdf:Description>
<rdf:Description rdf:about="http://bioportal.bioontology.org/annotator/sel/4019/1">
    <rdf:type rdf:resource="http://purl.org/ao/Selector"/>
    <rdf:type rdf:resource="http://purl.org/ao/selectors/TextSelector"/>
    <rdf:type rdf:resource="http://purl.org/ao/selectors/OffsetRangeSelector"/>
    <ao:exact>CANCER</ao:exact>
    <ao:offset>8</ao:offset>
    <ao:range>6</ao:range>
    <aof:onDocument rdf:resource="http://data.bioontology.org/annotator?4019"/>
</rdf:Description>
<rdf:Description rdf:about="http://bioportal.bioontology.org/annotator">
        <rdf:type rdf:resource="http://purl.org/swan/1.2/agents/Software"/>
        <foaf:name>NCBO annotator</foaf:name>
</rdf:Description>
```

Fig. 3. Example of RDF representation of an NCBO annotation.

returned by the Annotator and produces a RDF/XML file. This Java library is only available on request for now, but will be released in 2015. Our long term perspective, within the Semantic Indexing of French Biomedical Data Resources (SIFR) project (http://www.lirmm.fr/sifr) is to offer a service endpoint implementing several improvements (scoring, negation, disambiguation, new semantic expansion, new outputs formats, etc.) of the NCBO Annotator done with pre and post processing while still calling the Annotator service. Future versions will also include JSON-LD format and we will also follow the Open Annotation Data Model (http://www.openannotation.org/spec/core/ which is currently a W3C draft.

5 Acknowledgements

This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University of Montpellier, CNRS and the Computational Biology Institute (IBC) of Montpellier. We thanks the National Center for Biomedical Ontology (NCBO) for latest information about the Annotator.

References

1. Ciccarese, P., Ocana, M., Castro, L.J.G., Das, S., Clark, T.: An open annotation ontology for science on web 3.0. Biomedical Semantics 2(2:S4) (May 2011)
2. Ciccarese, P., Ocana, M., Clark, T.: Open semantic annotation of scientific publications using DOMEO. Biomedical Semantics 3(S1) (April 2012)
3. Handschuh, S., Staab, S. (eds.): Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications, vol. 96. IOS Press (2003)
4. Jonquet, C., Shah, N.H., Musen, M.A.: The Open Biomedical Annotator. In: American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09. pp. 56–60. San Francisco, CA, USA (March 2009)
5. Kahan, J., Koivunen, M.R., Prud'Hommeaux, E., Swick, R.R.: Annotea: an open RDF infrastructure for shared Web annotations. In: 10th Int.ernational World Wide Web conference, WWW'01. pp. 623–632. Hong Kong (May 2001)
6. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N.B., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research 37, 170–173 (May 2009)