KOBPSY, a Knowledge Base in Psychology and Behavioral Sciences

Alexander Garcia-Castro
Linking Data, LLC, USA/
Ontology Engineering Group,
Universidad Politecnica de
Madrid
Boadilla del Monte
28660, Spain
alexgarciac@gmail.com

Isabel Barth
Leibniz Institute for
Psychology Information
Universitätsring 15
54296 Trier, Germany
barth@zpid.de

Erich Weichselgartner
Leibniz Institute for
Psychology Information
Universitätsring 15
54296 Trier, Germany
wga@zpid.de

ABSTRACT

In this paper we present the Semantic Enhancement for the Open Journal Systems, SE4OJS, to generate semantically annotated documents. With this tool, we created a knowledgebase for psychology, KOBPSY. We have identified structural document elements such as authors, sections, and domain knowledge about diseases, population, treatments and more. Our corpus of documents comes from PsychOpen. The datasets we are building for each article include i) the RDF description of article metadata, e.g. information on title, keywords, authors and editors; ii) structural information such as sections, section types, paragraphs, in-text citations; and also, iii) RDF for content and annotations. We are structuring the annotations by using the Annotation Ontology. Our approach makes it possible to build concept-based queries; we are delivering a flexible, reusable and adaptable set of tools for metadata enrichment and semantic processing of scientific documents in psychology.

CCS Concepts

•**Information systems → Resource Description Framework (RDF);** *Web Ontology Language (OWL);*

Keywords

Semantic Web; scholarly communication; OWL; ontologies.

1. INTRODUCTION

Advances in technology have made it possible to adopt electronic dissemination channels for the scientific article, from paper-based journals to purely electronic formats. However, scientific literature remains locked up in discrete documents. Furthermore, although widely available over the WWW, it is not always machine-readable; discovering connections amongst papers remains largely a manual process. The connectivity tissue provided by RDF technology has not yet been widely used to support the generation of self-describing, machine-readable documents. In this paper, we present our contribution to the generation of self-describing documents for scientific literature in psychology; we have focused our efforts on delivering a semantically processed dataset for psychology as well as on building a modular reusable infrastructure.

We are extending the Open Journal Systems (OJS)[1] publication workflow. In addition to the output formats currently supported by the OJS, e.g. HTML, JATS-XML and PDF, we are also facilitating the generation of RDF. The RDF publication module makes use of existing biomedical infrastructure; for instance, Bioportal [6] for entity recognition and ontology management and MetaMap[2] [1] for Natural Language Processing (NLP). Our RDF model relies on existing ontologies; for instance, we are reusing a subset of the Semantic Publishing and Referencing Ontologies (SPAR)[3] [9]. This is a suite of OWL 2 DL ontologies that describes bibliographic records, citations, text structure and publishing roles. We are also reusing the Collections Ontology (CO) [4], a structural ontology that facilitates the management of ordered lists in OWL DL frameworks. In addition, we are annotating domain terminology with the UMLS Metathesaurus[4] and mapping to Bioportal identifiers whenever these are available. By using UMLS we are accessing the APA Thesaurus of Psychological Index Terms [5], Health Level Seven Reference Implementation Model, Version 3 as well as the NCIT[5] and other ontologies. We are also using the Neuroscience Information Framework (NIF) Standard Ontology, GALEN, the Bilingual Ontology of Alzheimer's Disease and Related Diseases and other ontologies available over Bioportal.

We are using the Annotation Ontology (AO) [3] to structure and describe the context of the annotation. Our approach defines an extended layer of metadata that complements the conventional bibliographic information available for scientific publications; this new layer of metadata is rooted within the content. As we are extending the OJS,

i-KNOW '15, October 21 - 23, 2015, Graz, Austria

ISBN 978-1-4503-3721-2/15/10…$15.00

DOI: http://dx.doi.org/10.1145/2809563.2809604

[1] https://pkp.sfu.ca/ojs/

[2] http://metamap.nlm.nih.gov

[3] http://sempublishing.sourceforge.net

[4] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

[5] National Cancer Institute Thesaurus; http://goo.gl/mtg1hO

our approach may easily be adopted by other publication workflows. By enacting the SE4OJS module, we are generating a Knowledge Base in Psychology, KOBPSY. We have applied our approach to PsychOpen[6]; this is a free full-text archive of seven scientific journals in psychology. KOBPSY delivers a dataset of interoperable, interlinked, semantically annotated and self-describing documents in psychology and behavioral sciences. It facilitates the execution of concept-based queries. In a similar fashion to efforts such as Biotea [2], we are bringing existing ontologies together in order to facilitate the semantic representation of sections in scientific literature as well as the identification of meaningful phrases. These are fragments of text corresponding to psychological disorders, attention deficit disorders, populations, chemicals, drugs, among other psychology scientific publications in psychology; by embedding scientific literature in psychology within the Web of Data (WoD) we are making it possible for developers and researchers to benefit from the advantages of the Linked Open Data (LOD) cloud.

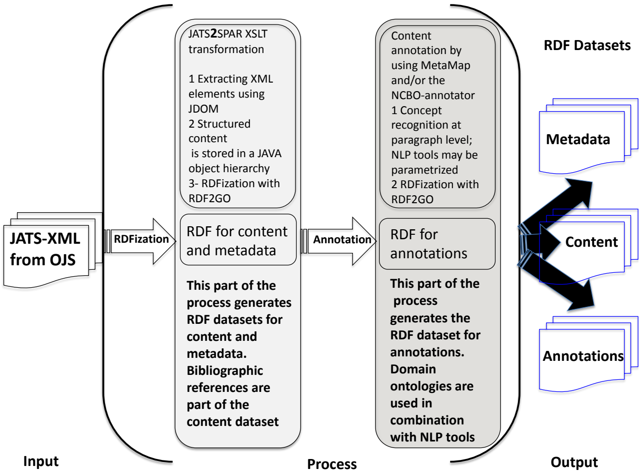
JATS2SPAR XSLT transformation
1 Extracting XML elements using JDOM
2 Structured content
is stored in a JAVA object hierarchy
3- RDFization with RDF2GO
Content annotation by using MetaMap and/or the NCBO-annotator
1 Concept recognition at paragraph level; NLP tools may be parametrized
2 RDFization with RDF2GO
RDF Datasets
Metadata
JATS-XML from OJS
RDFization
RDF for content and metadata
Annotation
RDF for annotations
Content
This part of the process generates RDF datasets for content and metadata. Bibliographic references are part of the content dataset
This part of the process generates the RDF dataset for annotations. Domain ontologies are used in combination with NLP tools
Annotations
Input
Process
Output

Figure 1: SE4OJS workflow. ©ZPID

2. WORKFLOW

Our workflow starts by consuming the JATS-XML generated from the OJS process. Firstly, an XSLT generates an RDF description for article metadata, e.g. information about contributors, article provenance and references. Then, the text-structure module extracts the nodes for the text from each XML input file; RDF triples describing the text structure and citations are thus generated. The content annotation module reuses the Java-text-structure representation, it sends the textual content to the external annotation tools, e.g. NCBO annotator, MetaMap, and creates a third RDF graph that stores the annotated concepts. A description of our workflow is illustrated in Figure 1. SE4OJS is currently using the NCBO-Annotator web service[7] as well as MetaMap for dictionary-based concept matching. The NCBO-Annotator does entity recognition; it also facilitates the use of synonyms available in the ontologies. MetaMap applies a number of natural language preprocessing steps to the input, such as tokenization, part-of-speech processing, shallow syntactic parsing and optionally, word-sense disambiguation. In addition, it makes it possible to identify negated concepts. Selecting the vocabularies to be used is possible for the NCBO-Annotator as well as for MetaMap.

3. RESULTS

3.1 Our Semantic Model

Our semantic model is based on the SPAR Ontologies and the Annotation Ontology. It comprises three RDF files for each JATS-XML article. There is a file describing the metadata for the article, contributors, and references; a separate RDF file describes the text-structure and in-text citations. There is also a third file representing the annotations derived from the content. Figure 2 provides a high-level overview of the type of information encoded by the generated RDF. The arrows present the relations between the entity to be annotated and the information items.

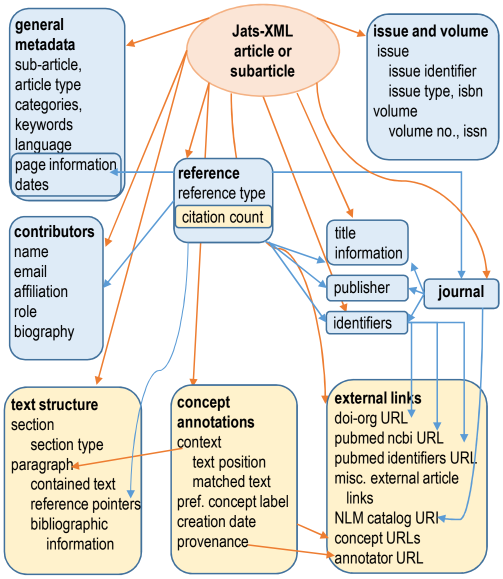

Figure 2: SE4OJS RDF model. ©ZPID

To create the 'Metadata RDF' (cf. Figure 1) we adapt the JATS2RDF mapping by Peroni et al [7] (triples created by this mapping are marked blue in Figure 2, additional information created by SE4OJS and external links are marked yellow). We reuse the model developed by Peroni because it accurately represents most of the data elements from the original JATS-XML files, capturing article meta-information such as title information, article language, keywords and subject-terms. It distinguishes 30 different article types; it also models contributor roles such as author, corresponding author, editor-in-chief, their affiliations and biographies. This level of detail, which sets the model apart from previous approaches, is facilitated by the extensive use of the SPAR family of ontologies. These are used in combination with ontologies such as Dublin Core and FOAF[8]. The generated RDF relates the article to the journal, volume and issue in which it has been published. The references for the article

[6] http://psychopen.eu

[7] http://bioportal.bioontology.org/annotate

[8] http://www.foaf-project.org

and their types are also linked to the article. We extended the model by adding several external links, e.g. the NLM catalog URL for the journal. The semantic representation for the structure and content annotation of the document is not addressed by the model from Peroni.

3.1.1 *Structure for the Text*

The 'text structure' box in Figure 2 lists what information is RDF-encoded by the respective SE4OJS module (e.g. sections, section types, paragraphs and paragraph texts for the text structure description). Following the work by Garcia et al [2] we used SPAR-DoCO[9] for sections and paragraphs. As a refinement to this previous approach we additionally employed SPAR-DEO[10] and SRO[11] concepts to classify section types. We also rdfized in-text citations with SPAR-C4O concepts, annotating their location at paragraph level and relating them to the targeted bibliographic reference; this is enriched by the in-text citation count for the reference. We make extensive use of the po:contains relation from the Pattern Ontology to relate the structural parts and extend DoCO by adding a transitive 'contains' relation; this allows us to retrieve deeply nested structures. Article and sections are linked to 'List-instances' from the Collections Ontology. Thus, child elements can be counted and their order of occurrence is preserved, which enables context-sensitive SPARQL queries. For example, we can query for annotated content from the 'Methods' section. The RDF for a section looks like:

```
<http://example.org/resource1/Conclusions>
a     doco:Section ,
      sro:Conclusion;
dcterms:title  'Conclusions';
pattern:contains
<http://example.org/resource1/
      ConclusionsChildElementList>
```

3.1.2 *Annotating Content*

For annotating the content we are using the AO [3]; it allows us to express the relations between the text and those ontology concepts that have been automatically matched by means of the MetaMap tool or the NCBO annotator. We are closely following the approach presented in the Biotea project. As illustrated in Figure 2, and in more detail in Figure 3, an annotation has information about:

- the concept. This has an identifier that is stored as an external link (corresponding to the 'annotation topic' in our RDF), we also work with the preferred label for the concept (the 'annotation body');
- the 'annotation context'. This specifies the location of the matched text (through the ao:exact relation), relating it to the article (through ao:resource);
- the metadata for the annotation. This includes, author, creation date, number of occurrences for the concept within the article.

The orange boxes in Figure 3 represent the described entities and blue boxes represent the relations between two entities (corresponding to the predicate of the triplet).

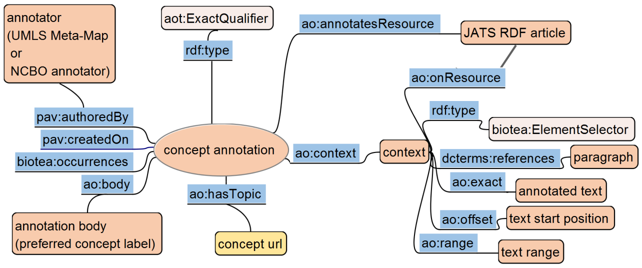

Figure 3: RDF for the content. ©ZPID

The code below illustrates an annotation

```
<http://example.org/resource1/Introduction_Para1/
     aoContext/cogat.owl#CAO_00602_10'>
ao:exact  'EMPATHY';
ao:range  '6'^^xsd:nonNegativeInteger;
ao:offset  '334'^^xsd:nonNegativeInteger;
ao:onResource <http://example.org/resource1 >;
dc:references <http://example.org/resource1/
     Introduction_Para1 >;
a <http://www.biotea.ws/ontology/ao_biotea.owl#/
     ElementSelector>
```

3.2 Querying KOBPSY

Our dataset currently comprises more than 300 articles from PsychOpen, distributed across 6 journals; a sample dataset can be obtained from [12]. The RDF output allows for bibliometric queries not possible with index-based search engines; for instance, to 'identify the sections containing the highest number of citations and, if available, their section types' or, as the following code illustrates, to 'extract the most cited authors of an article having a given DOI':

```
SELECT (SUM(?inTextCount) AS ?count)
       ?authorName ?firstName
WHERE {
  ?art prism:doi '10.5964/ejcop.v3i1.23'.
  ?art cito:cites ?reference.
  ?inTextCitation biro:references ?reference.
  ?inTextCitation c4o:hasInTextCitationFrequency
                  ?inTextCount.
  ?reference dcterms:title ?refTitle.
  ?role pro:relatesToDocument ?reference.
  ?role pro:withRole pro:author.
  ?group pro:holdsRoleInTime ?role.
  ?group foaf:member ?member.
  ?member foaf:familyName ?authorName.
  ?member foaf:givenName ?firstName.
}
GROUP BY ?authorName ?firstName
ORDER BY DESC (?count)
```

Queries addressing domain knowledge by relating annotations are also possible. For instance, 'retrieve papers about effects of low income in immigrant populations'; in this example, 'low income' and 'immigrant' are the pivots for the query. The following example retrieves all child concepts of Mental Disorders and selects the context by returning the title and the text where the concept is located. A small set of sample queries is available at[13]

```
SELECT ?title ?textMatch ?text ?label ?concept
       ?startPosition
WHERE{
  ?annotation ao:body ?body.
  ?annotation ao:context ?context.
```

[9] SPAR Document Components ontology (DoCO); http://goo.gl/chyfID

[10] SPAR Discourse Elements Ontology (DEO); http://purl.org/spar/deo

[11] SALT Rhetorical Ontology; http://goo.gl/XXHr7Q

[12] https://goo.gl/FRP8DC

[13] https://goo.gl/3uxLfD

```
?context ao:exact ?textMatch.
?context dcterms:references  ?para.
?context ao:offset ?startPosition.
?para c4o:hasContent ?text.
?annotation ao:annotatesResource ?doc.
?doc dcterms:title ?title.
?annotation ao:hasTopic ?concept.
?concept  skos:prefLabel ?label.
?concept rdfs:subClassOf ?parentConcept.
?parentConcept skos:prefLabel ?parentLabel.
FILTER(STR(?parentLabel) = 'Mental disorders')
FILTER(STR(?label) = ?body)
```

4. CONCLUSIONS AND FINAL REMARKS

We have developed a reusable tool, SE4OJS. The tool makes it possible to semantically annotate articles encoded in JATS-XML. By using SE4OJS, we have generated a knowledgebase for psychology, KOBPSY. The output of SE4OJS is an interlinked set of documents rooted in existing ontologies; this makes it possible to establish conceptual relations across documents within the corpus as well as to external resources in the larger web –e.g. DBPEDIA as well as the Cognitive Atlas. The quality of the knowledgebase depends on the quality of the ontologies and the accuracy of NLP tools used to generate KOBPSY. Controlled vocabularies and ontologies in psychology are not as precise and readily available as those in the biomedical domain. Very often terminology for psychology is part of larger biomedical ontologies. Furthermore, resources such as the APA thesaurus are structured as flat hierarchies; little inference is supported. Concept names are often artificial labels, unlikely to occur in written texts –the automatic mapping process would benefit from additional alternative labels for such concepts. The analysis of academic texts in psychology requires the use of ontologies; however, the lack of psychology-specific controlled vocabularies make it difficult to be precise. For instance, specific concepts such as 'Methylenedioxymethamphetamine Measurement' are often related to abuse of substances; in order for recognizing the entity some context is also needed –e.g. making clear the difference between 'substance abuse' and 'clinical methods for measuring substances'. Inheriting vocabularies from the biomedical domain is an advantage; however, it becomes a burden if there is not explicit contextualization for the terminology. Psychology requires the formalization of its own hierarchies, structures of object properties and classification schemata; The quality of the annotation could benefit from such effort, idem. an ontology for psychology networked with biomedical vocabularies as well as methodologies encompassed with those from the biomedical domain.

More sophisticated NLP methods could prevent errors and thus improve the quality of the annotation in KOBPSY. For instance, avoiding common english terms, relating and identifying acronyms and ruling out the use of biomedical terminology that is contextually inaccurate. the use of techniques such as Named Entity Recognition, anaphora resolution and relation extraction, could help verify automatically mapped concepts and discover entities and relations previously hidden to the mapping tools.

We are reusing existing infrastructure and vocabularies. KOBPSY makes it possible to define concept-based queries that are not easily executed with indexes. For instance, retrieving papers about 'mental disorders' with no explicit mention to a particular disorder requires coding in the database; our approach makes use of annotations based on ontologies, such queries may be easily executed because the reasoning system that is part of inference engines traverses the graph inferring the disorders associated to mental disorders according to the ontology. More complex scenarios are also possible; these are, however, always dependent on the knowledge encoded in the ontologies.

Our annotation dataset complements the metadata usually available for scientific articles. If the publisher does not want to expose the content, then publishing the metadata with the annotations is possible. Such enrichment to the metadata provides a much more informative landscape that can be transmitted to the end user via more elaborated user interfaces. A fully annotated dataset also makes it possible to determine the degree of semantic similarity across scientific publications. We relied on the NCBO recommendation REST-service for selecting the most suitable ontologies for annotating our content. Annotations may easily be linked to external resources; the graph is thus enriched. For instance, 'generalized anxiety disorder' may be resolved by the Cognitive Atlas [8] as well as by DBPEDIA.

5. ACKNOWLEDGMENTS

AGC acknowledges the KOPAR, H2020-MSCA-IF-2014, Grant Agreement nr: 655009

6. REFERENCES

[1] A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[2] L. G. Castro, C. McLaughlin, and A. Garcia. Biotea: Rdfizing pubmed central in support for the paper as an interface to the web of data. *Biomedical semantics*, 4(Suppl 1):S5, 2013.

[3] P. Ciccarese, M. Ocana, L. J. Garcia-Castro, S. Das, and T. Clark. An open annotation ontology for science on web 3.0. *J. Biomedical Semantics*, 2(S-2):S4, 2011.

[4] P. Ciccarese and S. Peroni. The collections ontology: creating and handling collections in owl 2 dl frameworks. *Semant Web J*, 2013.

[5] L. Gallagher Tuleya. Thesaurus of psychological index terms. *Washington: American Psychological Association*, 2007.

[6] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, page gkp440, 2009.

[7] S. Peroni, D. Lapeyre, and D. Shotton. From markup to linked data: Mapping niso jats v1. 0 to rdf using the spar (semantic publishing and referencing) ontologies. In *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012 [Internet]*. Bethesda (MD): National Center for Biotechnology Information (US), 2012.

[8] R. A. Poldrack, A. Kittur, D. Kalar, E. Miller, C. Seppa, Y. Gil, D. S. Parker, F. W. Sabb, and R. M. Bilder. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Frontiers in neuroinformatics*, 5, 2011.

[9] D. Shotton. Introduction the semantic publishing and referencing (SPAR) ontologies. *URL: http://goo.gl/J6Btdt*, October 2010.