RDFising PubMed Central

Alexander Garcia[1] Leyla Jael Garcia[2] Casey McLaughlin[1] and Stephen Flager[1]*

1 Florida State University, School of Library and Information Science, Tallahassee, Florida, USA
2 Universität der Bundeswehr, E-Business and Web Science Research Group, Munich, Germany

ABSTRACT

Motivation: The Web has succeeded as a dissemination platform for scientific and non-scientific papers, news, and communication in general. However, most of that information remains locked up in discrete documents, which are poorly interconnected to one another and to the Web itself. The connectivity tissue provided by RDF technology and the Social Web have barely made an impact on scientific communication. In this paper, we present our approach to scholarly communication, which entails understanding of the research paper as an interface to the web of data. Our RDF model makes extensive reuse of existing ontologies and semantic enrichment services. We expose the model over our SPARQL endpoint: http://virtuoso.idiginfo.org/sparql.

1 INTRODUCTION

Semantic Digital Libraries (SDL) make extensive use of meta-data in order to support information retrieval and classification tasks. Within the context of SDLs, ontologies can be used to: (i) organize bibliographic descriptions, (ii) represent and expose document contents, and (iii) share knowledge amongst users (Kruk, et al., 2006). There have been some efforts that aim to make use of ontologies and Semantic Web technology in digital libraries. For instance, JeromeDL (http://www.jeromedl.org) allows users to semantically annotate books, papers, and resources (Kruk, et al., 2007). Similarly, the Bricks project (http://www.brickscommunity.org/) aims to integrate existing digital resources into a shared digital memory; it relies on OWL-DL in order to support, organize and manage meta-data (Kruk, et al., 2006).

Efforts such as DOMEO (Ciccarese, et al., 2011) and the Living Document (Garcia, et al., 2009) illustrate how Semantic and Social Web technologies are being used in Digital Libraries within the biomedical domain. DOMEO is a web component developed using the Google Web Toolkit and JavaScript. It allows users to manually or semi-automatically create unstructured or semi-structured semantic annotation that can be kept private, shared within selected groups, or made public and therefore available to the entire web. The Living Document made use of the paper as an interface to the web of data; a self-descriptive document fully interoperable with the Web. The Living Document (LD) is a document that also acts as a document router, operating by means of structured and organized social tagging and using existing ontologies. UTOPIA (Attwood, et al., 2010) also exemplifies the same trend; by combining Semantic and Social Web principles and technologies, the authors aim to improve interoperability and user experience. Publishers are also actively improving programmatic access to their content. Nature, for instance, recently released 20 million Resource Description Framework (RDF) statements, including primary metadata for more than 450,000 articles published by NPG since 1869. In this first release, the dataset includes basic citation information (title, author, publication date, etc) as well as ontologies specific to NPG (http://www.nature.com/press_releases/linkeddata.html).

Similarly, Elsevier provides an Application Programming Interface (API) that makes it possible for developers to build specialized apps (http://www.developers.elsevier.com/).

In this paper, we present our approach to generating a knowledge model for biomedical literature with the ultimate goal of improving information retrieval from digital libraries and facilitating the discovery of hidden relationships among papers. Existing ontologies are brought together in order to facilitate the representation of sections in scientific literature as well as meaningful fragments within those previously identified sections. Our model makes it possible to localize meaningful pieces in sections across an entire digital library. In this way, it is possible to find papers that are similar to one another in a highly accurate manner.

2 RDFISING PMC, OUR MODEL

We are RDFising biomedical literature, PubMed Open Central in this case, by orchestrating ontologies such as DoCO (http://purl.org/spar/doco/), BIBO (http://purl.org/ontology/bibo/), DC (http://dublincore.org/), and FOAF (http://xmlns.com/foaf/0.1/). We use BIBO and DC to model the metadata, DoCO to explicitly identify sections, and FOAF to identify authors and institutions. Meaningful fragments within sections are automatically annotated and enriched, and these annotations are modeled with the Annotation Ontology (AO) (Ciccarese, et al., 2011). In our model, we follow the four principles proposed by Tim Berners-Lee for publishing linked data: (i) using URIs to identify things; (ii) using HTTP URIs to make it possible for things to be referenced and looked up (dereferenced) by software agents; (iii) representing things in RDF and provid-

* To whom correspondence should be addressed.

ing a SPARQL endpoint; and (iv) providing links to other external URIs in order to facilitate knowledge discovery.

2.1 RDFication process

PMC offers a dump in XML corresponding to its subset of open access articles. These files are the beginning of our process. Initially, we use BIBO, DC, DoCO and FOAF in order to model the document as RDF. As shown in Fig 1, the data originally provided by the XML can be enriched with publication identifiers such as DOI; this is achieved by using NCBI services, particularly the eFetch service at http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi. The same process can be used for references. Once the identifiers have been resolved, the XML is parsed into an RDF document: the article itself is modeled as a *bibo:AcademicArticle* with PMC, PMID, and DOI as identifiers. The publishing data is identified with BIBO, including data such as publisher name and ISSN, volume, issue, and page range, all modeled with BIBO. Authors are modeled as a *bibo:authorList* where each member is a *foaf:Person*. The abstract and sections are modeled as a *doco:Section* with a *rdfs:comment* containing the text. Formatting such as bolding and italics are omitted. The references are modeled as *bibo:Article*, and the relation used is *bibo:cites*. References are available for both the document and the section level. We produce one file for each publication. For example, an article retrieved from the URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2971111 would generate a file named *PMC271111.rdf*.

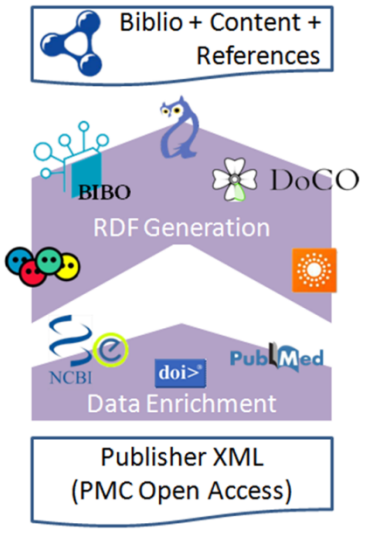

Fig 1. RDFification process

Fig 2 shows the corresponding graph for the bibliographic data, and Fig 3 shows the corresponding graph for sections and content. Fig 2 shows that it is possible to identify both DOI and PMID as identifiers for the paper. Title and keywords are represented by DC terms, and the abstract is a BIBO element, which is also be represented as a *doco:Section*. Published data is shown at the bottom of the figure while authors are on the right side, represented as a list of *foaf:Person* objects where the actual members have been omitted for space and readability reasons.

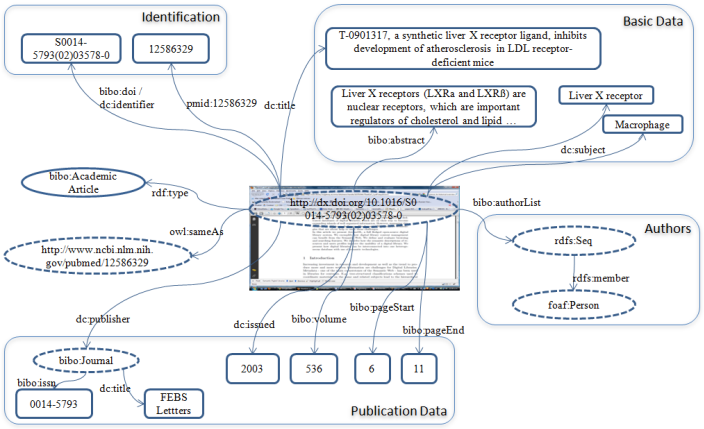

Fig 2. Bibliographic data (metadata)

Fig 3. shows the sections on the top and the references on the bottom. References are linked to the section where they are used so that further analysis on argumentative and rhetorical elements is simpler. Sections consist of a title and a set of paragraphs, and the actual content is a *rdf:comment*. References include similar data to that shown in Fig. 2. Since the content is enriched with specialized vocabularies, whenever the publisher decides not to expose the content in RDF, it is still possible to offer enough information to users willing to use the RDF format and its links to other resources.

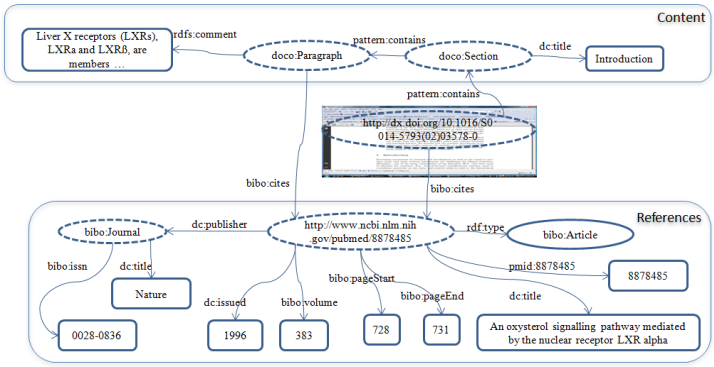

Fig 3. Sections and content

Using our RDF representation, it is possible to look for papers with terms present in a particular section. An example SPARQL query might look like:

```
SELECT *
WHERE {
  ?article a bibo:AcademicArticle .
  ?article doco:contains ?section .
  ?section dcterms:title ?title .
```

```
  FILTER (regex(str(?title), "introduction")) .
  ?section doco:contains ?paragraph .
  ?paragraph rdfs:comment ?text .
  FILTER (regex(str(?text), "cholesterol"))
}
```

This query would yield all academic articles with "*Introduction*" in the title and the world "*cholesterol*" in the text.

2.2 Annotation process

Once the RDF has been generated, the abstract and sections are enriched with automatic annotations modeled using AO. These automatic annotations are generated by the NCBO Annotator (Jonquet, et al., 2009) and Whatizit (http://www.ebi.ac.uk/webservices/whatizit), two text-mining tools publicly available (see Fig 4). Adding the annotations to the RDF rather than to the original XML makes it easier to apply the same process to compatible RDF documents coming from different publishers. Our process takes all of the *doco:Section* elements, uses the NCBO annotator, and produces a consolidated RDF identified as *xxx_ncboAnnotator*, where *xxx* corresponds to the PMC identifier. A second RDF document is produced by Whatizit, identified as *xxx_whatizit<pipeline>*; *xxx* being the same as before, while *<pipeline>* corresponds to the pipeline used. For our dataset, we used *UkPmcAll*, since it is one of the most reliable when dealing with proteins and genes. Annotations generated by the NCBO Annotator are preferred over those generated by Whatizit, because the former process faster. Thus, we use Whatizit only for UniProt, since it is not supported by NCBO, and UniProt Taxonomy (mapped to NCBI Taxon), as NCBO does not recognize as many organisms as Whatizit.

Another possibility is to generate the consolidated Bio2RDF for a particular paper. Currently, Bio2RDF offers a download option of the RDF for a particular term, but the consolidated option would contain the RDF for all the terms annotated in a paper. This is not currently part of our dataset; we plan to support it latter as it is a computationally intensive process, particularly for papers with proteins since Whatizit can identify up to 20 or more UniProt accessions for the same term.

The NCBO Annotator is used with these ontologies: CHEBI for chemicals; GO, Pathway, and MGED for genes and proteins; MDDB, NDDF, and NDFRT for drugs; medline, SNOMED, symptom, MedDRA, MESH, OMIM, FMA, ICD9, OBI, and UMLS for medical and general terms; plant ontology for plants, and NCIThesaurus for general terms. On the other hand, Whatizit is used for UniProt and UniProt Taxonomy; the latter is also mapped to NCBI Taxon. For those terms which exist in identifiers.org, this link is also provided.

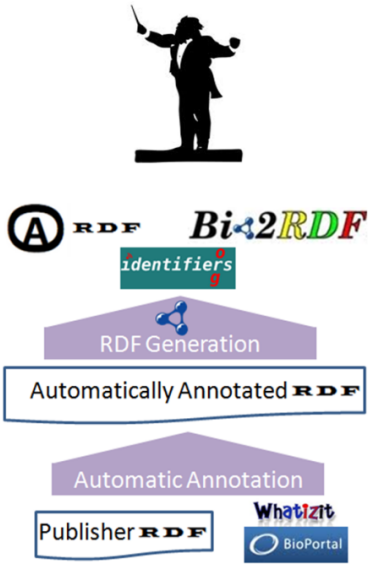

Fig 4. Orchestrating ontologies and services

In order to specify the place in a document where an annotated term is located, we have extended AO so we can select portions of text within *rdfs:comment*. The first selector is used with Whatizit as this tool does not provide information about the start and end positions of the annotated text. The second selector is used with the NCBO annotator.

- aold:ElementSelector extends aos:TextSelector → identifies an exact text within a unique *rdfs:comment* in an RDF element
- aold:StartEndElementSelector extends aos:StartEndSelector→ identifies the start and end positions of a text within a unique *rdfs:comment* in an RDF element

Fig. 5 shows how annotations are represented. This example shows a text snippet in the Introduction, "*cholesterol*", that is annotated with the ontological term *chebi:16113*. The provenance corresponds to the annotation pipeline used, either NCBO annotator or Whatizit, which is modeled as a *foaf:Agent*. The selector specifies the section and, whenever possible, the start and end positions of the annotated text.

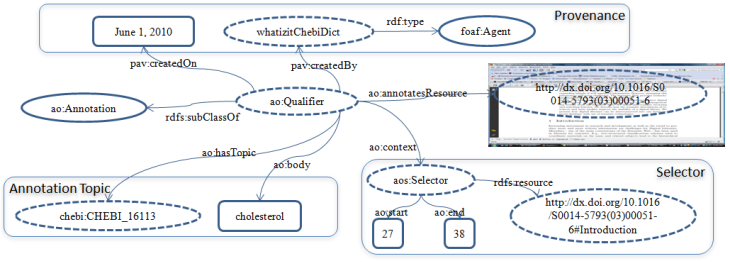

Fig 5. Annotations

A portion of the annotated RDF is provided in Fig. 6; the object of the annotation is the paper identified with the URL

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2971111. The topic of the annotation is the CHEBI term 60004, which corresponds to the text "*mixture*". Here, the annotation comes from the NCBO Annotator; thus start and end positions are identified. Provenance shows the automatic annotator used as well as the date and time when the annotation was generated. Additional information is provided by using *rdfs:seeAlso*; in this case, it is a link to http://identifiers.org/obo.chebi/CHEBI:60004

```
<aot:ExactQualifier rdf:about="http://www.biotea.ws/pubmedOpenAccess/rdf_ao/ExactQualifier_d8c0840dd5c1de1d78145b5e23944e37_mixture">
  <pav:createdBy rdf:resource="http://bioportal.bioontology.org/annotator/NCBOAnnotator"/>
  <ao:annotatesResource rdf:resource="http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2971111"/>
  <rdfs:seeAlso rdf:resource="http://identifiers.org/obo.chebi/CHEBI:60004"/>
  <ao:hasTopic rdf:resource="http://purl.obolibrary.org/obo/CHEBI_60004"/>
 - <ao:context>
   - <aold:StartEndElementSelector rdf:about="http://www.biotea.ws/pubmedOpenAccess/rdf_ao
     /Selector_bfe1d9b4c1afb4c2ef374f87d943d00f_mixture">
       <rdfs:resource rdf:resource="http://www.biotea.ws/pubmedOpenAccess/rdf/paper_PMC2971111/Refinement_paragraph_1"/>
       <aos:end xml:lang="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">304</aos:end>
       <aos:init xml:lang="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">298</aos:init>
       <ao:onResource rdf:resource="http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2971111"/>
     </aold:StartEndElementSelector>
  </ao:context>
  <sioc:num_items xml:lang="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1</sioc:num_items>
  <pav:createdOn rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2012-04-07T00:14:05.424Z</pav:createdOn>
  <ao:body rdf:datatype="http://www.w3.org/2001/XMLSchema#string">mixture</ao:body>
</aot:ExactQualifier>
```

Fig 6. Annotation RDF with AO

With this model, many SPARQL queries become available. For example, a query that returns documents containing a particular CHEBI term becomes possible; see the SPARQL code depicted below. Furthermore, it is also possible to retrieve articles including one term but excluding another one, as well as specifying the section where the terms should or should not appear.

```
SELECT ?pmid
WHERE {
  ?article a bibo:AcademicArticle .
  ?article bibo:pmid ?pmid .
  ?annotation a aot:ExactQualifier .
  ?annotation ao:annotatesResource ?article .
  ?annotations                ao:hasTopic                <
http://purl.obolibrary.org/obo/CHEBI_60004> .
}
```

3 FINAL REMARKS

We have scaffolded annotations by using the AO, reused domain ontologies, and we have structured the document by means of DOCO, BIBO, DC and others. Our model is very flexible. Modifying the annotators is a simple task, and the software makes it possible to redefine the model. Working with XML in formats other than what PMC provides is also possible. For example, we are currently experimenting with documents from BioMedCentral.

Models such as that of Nature do not link to existing vocabularies, *e.g.* MESH, in a semantic way. They instead include plain literals, which makes it difficult to use this information for knowledge discovery. Our model does link to well-known vocabularies, relevant in the biomedical domain. Similar to Nature, we also rely on ontologies such as BIBO in order to model metadata. Since we are targeting only open access documents within PubMed Central, we also include the content. Some of the difficulties we have had are: (i) at least four different formats are used to model references in PubMed, and it is not possible to parse some of them, so they end up incomplete in our RDF, (ii) authors are represented as first initial and last name, making them difficult to disambiguate, (iii) FOAF for authors and institutions are not provides, so we have had to create created dummy nodes and hope that they will later be provided by the community, and (iv) annotation services were sometimes unavailable during processing causing some annotations to be missing

To ensure the reproducibility of science, we envision that papers will provide access to raw data and to computer understandable descriptions of methodologies in order to support the re-creation of the experiments being described. To aid in resolving inconsistencies, we expect in the future to relate and compare information across multiple documents. Semantic Web technology should help deliver a self-descriptive document that makes it possible to improve the user experience and change our understanding of scholarly communication. Similar to DOMEO, we also plan to provide a community-based platform that facilitates the completion of data such as references and FOAFs for authors and institutions. We actually believe that, for the sake of disambiguation, these data should be provided at the time of submission, in the same way that emails indicate the sending party as part of the underlying protocol.

We are currently testing our RDF model and SPARQL endpoint with members of the research community, so modifications and improvements can be expected before the official release.

REFERENCES

Attwood, T.K., Kell, D.B., Mcdermott, P., Marsh, J., Pettifer, S.R. and Thorne, D. (2010) Utopia Documents and The Semantic Biochemical Journal experiment, EMBNet News, 15.

Ciccarese, P., Ocana, M. and Clark, T. (2011) DOMEO: a web-based tool for semantic annotation of online documents. BioOntologies. Vienna, Austria.

Ciccarese, P., Ocana, M., Garcia Castro, L., Das, S. and Clark, T. (2011) An open annotation ontology for science on web 3.0, Journal of Biomedical Semantics, 2, S4.

Garcia, A., Garcia, L.-J., Labarga, A., Giraldo, O., Montana, C. and Bateman, J. (2009) The Semantic Web and the Social Web heading towards a Living Document in life sciences, Journal of the Semantic Web.

Jonquet, C., Shah, N.H., Youn, C.H., Callendar, C., Storey, M.-A. and Musen, M.A. (2009) NCBO Annotator: Semantic Annotation of Biomedical Data. International Semantic Web Conference, Poster and Demo session.

Kruk, S., Haslhofer, B., Piotr, P., Westerski, A. and Woroniecki, T. (2006) The Role of Ontologies in Semantic Digital Libraries. European Networked Knowledge Organization Systems (NKOS) Workshop. Alicante, Spain.

Kruk, S.R., Woroniecki, T., Gzella, A. and Dabrowski, M. (2007) JeromeDL - a Semantic Digital Library. International Semantic Web Conference - Semantic Web Challenge. Busan, Korea.