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; it 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 instantiated model over our SPARQL endpoint.

Availability: http://biotea.idiginfo.org/

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 aiming 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 annotations that can be kept private, shared within selected groups, or made public and therefore available to the entire web. The Living Document (LD) made use of the paper as an interface to the Web of Data. The LD is a document that also acts as a router, operating by means of structured and organized social tagging and using existing ontologies; it is a self-descriptive document fully interoperable with the Web. 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 Publishing Group (NPG), 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 datasets include basic citation information (title, author, publication date, etc) as well as ontologies specifics 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 knowledge model for biomedical literature. In order to facilitate the representation of sections and meaningful fragments we are using existing ontologies. Our model makes it possible to localize meaningful pieces, *e.g.* concepts, in sections and paragraphs across the entire digital library. Such infrastructure facilitates information retrieval, makes it easier to discover hidden relationships across papers, as well as to accurately find similar papers based on the semantics of the content.

2 RDFISING PMC, OUR MODEL

We are RDFising the open access subset of PubMed Central (PMC) 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/); these namespaces have been added to our SPARQL endpoint so that users do not need to define them as prefixes. We use BIBO and DC to model the bibliographic metadata, DoCO to explicitly identify sections, and FOAF to identify authors and organizations. Meaningful fragments within sections are automatically marked and enriched; such 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 Link Data: (i) using URIs to identify things, (ii) using HTTP URIs so we make it

[*] To whom correspondence should be addressed.

possible for things to be referenced and looked up by software agents, (iii) representing things in RDF and providing a SPARQL endpoint, and (iv) providing links to external URIs in order to facilitate knowledge discovery.

2.1 RDFication process

PMC offers a dump in XML corresponding to the subset of open access articles; these files are the raw-input for our process. Initially we use BIBO, DC, DoCO and FOAF in order to model the document as RDF. We are using identifiers for scientific literature, namely PMC IDs, PubMed IDs and DOIs; these are part of the generated RDF, the same principle is also applied to the references. For incomplete references, *e.g.* “Allen, F. H. (2002). Acta Cryst. B58, 380-388” in PMC2971765, services such as Mendeley (http://www.mendeley.com/), CrossRef (http://www.crossref.org), or eFetch (http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi) could be used in order to complete the information.

In Fig 1 the ontologies used to convert an article to RDF are illustrated. The article is modeled as *bibo:Document*; whenever possible a more accurate class is also added, *e.g. bibo:AcademicArticle* for research articles. The data of the publication is identified with BIBO; it includes the name of the publisher, ISSN, volume, issue, and starting and ending page. Authors are modeled as a *bibo:authorList* where each member is a *foaf:Person*. Abstract and sections are modeled as a doco:Section with an *rdf:value* containing the actual text. The references are modeled as *bibo:Document*, the relations we are using are *bibo:cites* and *bibo:citedBy*. References are available for both the document and the section level. We are producing one RDF file per publication; for example, for http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2971111 the http://biotea.idiginfo.org/pubmedOpenAccess/rdf/PMC2971111.rdf is generated.

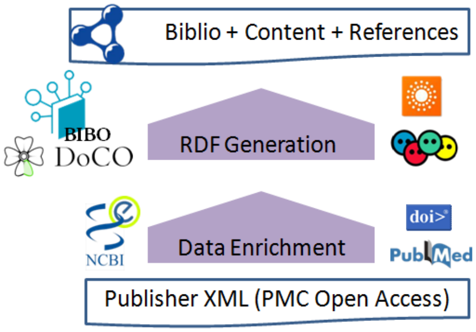
Biblio + Content + References
BIBO
DoCO
RDF Generation
NCBI
Data Enrichment
doi>
PubMed
Publisher XML (PMC Open Access)

Fig 1. RDFication process

In Fig 2 the graph representing bibliographic data as well as the identifiers is presented. Title and keywords are represented by means of DC terms; the abstract is a BIBO element, this is also represented as a doco:Section. Published data is shown at the bottom of the figure. Authors, on the right side, are represented as a list of foaf:Person objects.

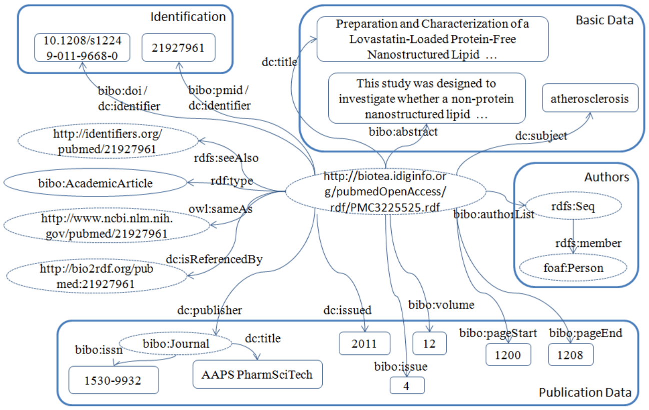
Identification
10.1208/s12249-011-9668-0
21927961
bibo:doi / dc:identifier
bibo:pmid / dc:identifier
Basic Data
Preparation and Characterization of a Lovastatin-Loaded Protein-Free Nanostructured Lipid …
This study was designed to investigate whether a non-protein nanostructured lipid …
atherosclerosis
dc:title
bibo:abstract
dc:subject
http://identifiers.org/pubmed/21927961
rdfs:seeAlso
bibo:AcademicArticle
rdf:type
http://www.ncbi.nlm.nih.gov/pubmed/21927961
owl:sameAs
http://bio2rdf.org/pubmed:21927961
dc:isReferencedBy
http://biotea.idiginfo.org/pubmedOpenAccess/rdf/PMC3225525.rdf
Authors
rdfs:Seq
bibo:authorList
rdfs:member
foaf:Person
dc:publisher
dc:issued
bibo:volume
bibo:pageStart
bibo:pageEnd
bibo:issn
bibo:Journal
dc:title
1530-9932
AAPS PharmSciTech
2011
12
bibo:issue
4
1200
1208
Publication Data

Fig 2. Bibliographic data

The sections and the references are illustrated in Fig 3. Sections entail a title and a set of paragraphs; the content is an *rdfs:value*. Since the content is clearly identified and enriched with specialized vocabularies, publishers may decide not to expose the entire content in RDF but just the concepts that are available in the content. Even for such a case it is still possible to deliver useful information *i.e.* annotations fully identified within the localized sections of the document.

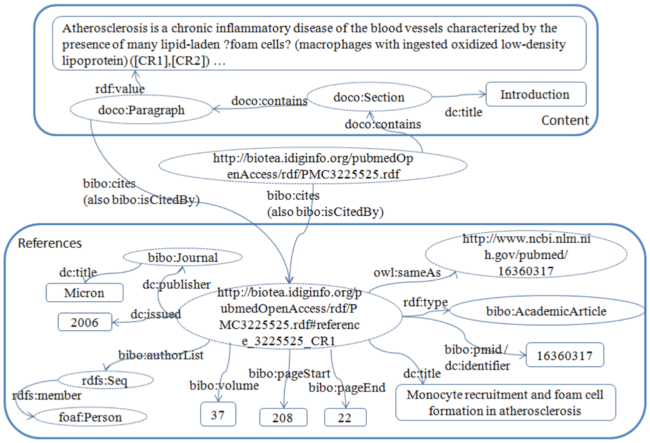
Atherosclerosis is a chronic inflammatory disease of the blood vessels characterized by the presence of many lipid-laden ?foam cells? (macrophages with ingested oxidized low-density lipoprotein) ([CR1],[CR2]) …
rdf:value
doco:Paragraph
doco:contains
doco:Section
Introduction
dc:title
doco:contains
Content
http://biotea.idiginfo.org/pubmedOpenAccess/rdf/PMC3225525.rdf
bibo:cites (also bibo:isCitedBy)
bibo:cites (also bibo:isCitedBy)
References
bibo:Journal
dc:title
Micron
dc:publisher
2006
dc:issued
http://biotea.idiginfo.org/pubmedOpenAccess/rdf/PMC3225525.rdf#reference_3225525_CR1
owl:sameAs
http://www.ncbi.nlm.nih.gov/pubmed/16360317
rdf:type
bibo:AcademicArticle
bibo:pmid / dc:identifier
16360317
bibo:authorList
rdfs:Seq
rdfs:member
foaf:Person
bibo:volume
37
bibo:pageStart
208
bibo:pageEnd
22
dc:title
Monocyte recruitment and foam cell formation in atherosclerosis

Fig 3. Sections and content

Our RDF representation makes it possible to search for articles with terms present in specific sections. For instance:

```
SELECT distinct ?article ?title ?text
WHERE {
  ?article a bibo:AcademicArticle .
  ?article doco:contains ?section .
  ?section dc:title ?title .
  FILTER (regex(str(?title), "introduction")) .
  ?section doco:contains ?paragraph .
  ?paragraph rdfs:comment ?text .
  FILTER (regex(str(?text), "cancer"))
}
```

2.2 Annotation process

Once the RDF for the document has been generated, the abstract and sections are enriched with automatic annotations modeled with the AO. These automatic annotations are generated by using the NCBO Annotator (Jonquet, et al., 2009) and Whatizit (http://www.ebi.ac.uk/webservices/whatizit), an illustration of the process is presented in Fig 4. Adding the annotations to the RDF rather than to the original XML makes it easier to apply the same process to compatible RDFs coming from a different publisher. Our process takes all the *doco:Section* elements, uses the NCBO annotator and produces a consolidated RDF identified as PMC*identifier*_ncboAnnotator. A second RDF is produced by Whatizit; it is identified as PMC*identifier* _whatizit<*pipeline*>, where <*pipeline*> corresponds to the used pipeline. For our dataset, we have used *UkPmcAll*, since it is very reliable when dealing with proteins and genes.

We are using the NCBO Annotator with the following 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, and OBI for medical, Plant Ontology for plants, and MESH, SNOMED, and NCIThesaurus for general terms. Whatizit is used with UniProt, UniProt Taxonomy, and diseases mapped to UMLS; UniProt Taxonomy is also mapped to NCBI Taxon. For vocabularies supported by both annotation tools, we chose NCBO over Whatizit as the former is faster; accuracy is similar in both cases. For the identification of organisms, however, we have chosen Whatizit because it recognizes more organisms than the NCBO tool; for instance "human" or "mouse" were not recognized by NCBO in any of our tests.

Bio2RDF is a project that provides interlinked life science data supported by Semantic Web technologies such as RDF and SPARQL (Nolin, et al., 2006). Bio2RDF brings together information from diverse public databases such as Kegg, PDB, Uniprot, NCIThesaurus, PubMed, amongst others (Belleau, et al., 2008). Generating a consolidated Bio2RDF view for a specific article is possible with our dataset. Currently, Bio2RDF provides a "download" option for the RDF corresponding to a particular term; for instance uniprot:P38398 –a breast cancer susceptibility protein for humans, can be viewed and downloaded in RDF/XML, N3 and JSON. Since we have annotated articles with UniProt terms it is possible to download and consolidate all of these related RDFs provided by Bio2RDF into a single one. Although this use case is not currently supported by our dataset, we are planning to offer this service on a demand basis because it is a computationally intensive process.

We have extended the AO so that specifying the location of an annotated term in the document is possible. In this way, we are able to select portions of text represented as *Literal* objects in RDF triplets; the property whose object is the literal must be used only once in the annotated element. For instance, for a *doco:Paragraph*, we can annotate the text modeled as object of the property *rdf:value*; there should be only one triplet <*doco:Paragraph*> *rdf:value* <*literal*> for this paragraph.

- aold:ElementSelector: identifies an exact text in a literal, *e.g.* *rdf:value*, in an RDF element (extends aos:TextSelector)
- aold:StartEndElementSelector : similar to the previous one but also including the start and end positions of the snippet in the text (extends aos:StartEndSelector)

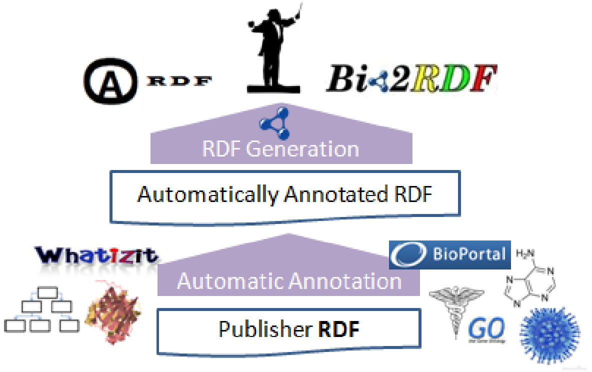

Fig 4. Orchestrating ontologies and annotation services

In Fig 5 the representation of annotations is presented. The object of the annotation is the RDF that corresponds to *pmc:XXX*; the annotation refers to a text snippet in the Introduction section; "cholesterol", has been annotated with the ontological term *chebi:16113*. CHEBI terms are identified by the NCBO annotator, this is modeled as a *foaf:Agent* in the provenance section; NCBO also provides the start and end positions within the text: from X to Y in the selector section. Additional information on the topic is provided by *rdfs:seeAlso* relation as well as *dc:isReferencedBy*.

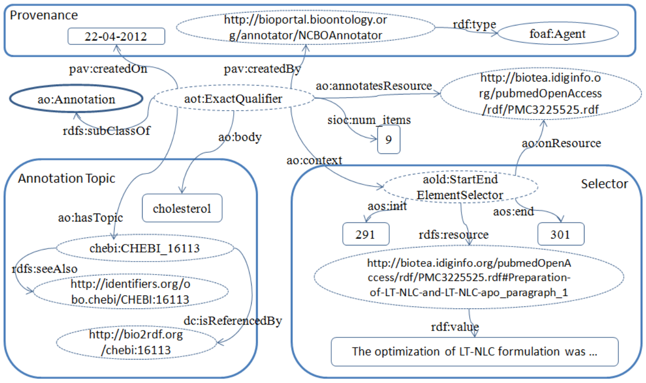
Provenance
22-04-2012
http://bioportal.bioontology.org/annotator/NCBOAnnotator
rdf:type
foaf:Agent
pav:createdOn
pav:createdBy
ao:Annotation
aot:ExactQualifier
ao:annotatesResource
http://biotea.idiginfo.org/pubmedOpenAccess/rdf/PMC3225525.rdf
rdfs:subClassOf
ao:body
sioc:num_items
9
ao:onResource
ao:context
Annotation Topic
cholesterol
ao:hasTopic
chebi:CHEBI_16113
rdfs:seeAlso
http://identifiers.org/obo.chebi/CHEBI:16113
dc:isReferencedBy
http://bio2rdf.org/chebi:16113
aold:StartEnd ElementSelector
Selector
aos:init
aos:end
291
rdfs:resource
301
http://biotea.idiginfo.org/pubmedOpenAccess/rdf/PMC3225525.rdf#Preparation-of-LT-NLC-and-LT-NLC-apo_paragraph_1
rdf:value
The optimization of LT-NLC formulation was ...

Fig 5. NCBO annotation for a chemical term

The model we are presenting here allows the execution of several SPARQL queries. For example, it is easy to retrieve documents containing a particular CHEBI term –see code 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 distinct ?pmid
WHERE {
  ?article a bibo:AcademicArticle ;
    bibo:pmid ?pmid .
  ?annotation a aot:ExactQualifier ;
    ao:annotatesResource ?article ;
    ao:hasTopic
<http://purl.obolibrary.org/obo/CHEBI_60004> .
}
```

3 FINAL REMARKS

We have scaffolded annotations by using the AO; we reused domain ontologies as well structured the document by means of DOCO, BIBO, DC and others. Our model is very flexible and the software can be easily customized; modifying the annotators is a simple task. Working with XML different from that provided by PubMed 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 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 full text open access documents within PMC, annotating the content is important. Some of the difficulties we have had are for instance: (i) at least four different formats are used to model references in PMC XMLs, (ii) authors names are represented with initials an last name, making it difficult for disambiguation purposes, (iii) FOAF for authors and institutions are not provided, and, and (iv) annotations services were sometimes unavailable during processing –services are not always reliable.

To ensure the reproducibility of science, we envision that papers will provide access to raw data as well as to machine-processable descriptions of methodologies, experimental protocols, in order to support the recreation of the experiments being described –towards a self-descriptive document. In order to resolve inconsistencies, we expect in the future to relate and compare information across multiple documents. Semantic Web technologies should help in delivering a self-descriptive document that makes it possible to improve the user experience and change our experience of scholarly communication. There should be a community-based platform that provides FOAF for authors and institutions; such platform could easily be part of submission systems. In this way disambiguating authors may be much simpler.

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.

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. and Morissette, J. (2008) Bio2RDF: Towards a mashup to build bioinformatics knowledge systems, Journal of Biomedical Informatics, 41, 706-716.

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

Nolin, M.-A., Belleau, F., Ansell, P. and Dumontier, M. (2006) Bio2RDF: Linked Data for the Life Sciences (available at http://bio2rdf.blogspot.co.uk/).