RDF4PMC, RDFizing PubMed Central

Alexander Garcia\textsuperscript{1}, Leyla Jael García Castro\textsuperscript{2}, Casey McLaughlin\textsuperscript{1}
\textsuperscript{1}Florida State University
\textsuperscript{2}Universitat Jaume I
Outline

• The Biotea project
• Why Semantic Web Technologies?
• RDF4PMC in a nutshell
• Architecture
• RDFization process
  • PMC RDFization
  • Content enrichment
  • Some numbers for RDF4PMC
  • Architecture
• Using the data
  • SPARQL
  • Bio2RDF integration
  • Web services
  • A first prototype
• Challenges and Lessons
• Currently working on...
• Future Work
• Conclusions
• Acknowledgments
Biotea

Scholarly data and documents are of most value when they are interconnected rather than independent

Christine L. Borgman

- Methodologies, methods and techniques supporting semantic enrichment of scholarly communication
- Once enriched, then how is this changing our user experience?
Biotea

Scholarly data and documents are of most value when they are interconnected rather than independent

Christine L. Borgman

• How are publications connected to each other?
• Putting together explicit assertions from different papers to form new implicit assertions
• Semantic Web Technology supporting scholarly communication, Literature Based Discovery and the Search-Retrieval-and-Interacting-with-the-Document (SRID) processes
Why SWT for research documents

• Generates an adaptable open approach, the data becomes the platform
• The SW delivers an integrative platform
• Makes it easier for the community to build over the platform
• Simplifies programmatic access to information
  • Retrieve all papers that have a component X (CHEBI) and the cellular location in GO terms
    • As simple as relating terminologies
• Delivers Social Network ready content
RDF4PMC in a nutshell

- Delivers an interoperable, interlinked, and self-describing document model in the biomedical domain.
- A network of interconnected documents
- Semantic infrastructure for PMC
- An interface to the Web of Data
- A knowledge model for biomedical literature – easily extendible
RDF4PMC in a nutshell

• RDFizing biomedical literature by orchestrating ontologies such as
  - DoCO, BIBO, DC, FOAF, W3CPROV, and others
• Datasets are available
  - RDF for metadata and content
  - RDF for annotations from text-mining
• RDFizator will be available
  - Adding other ontologies and annotators is possible
  - Working with XML from other sources is possible
PMC RDFization

Metadata+ Content + References

References Enrichment

RDF Generation

RDF Reactor

PMC NXML
Annotations: Content Enrichment

Enriched RDF

RDF Generation

Automatic Annotation

Metadata + Content + References

Web service

Web service
RDF4PMC, some numbers

Articles per journal

Coverage of biological entities.
RDF4PMC Server Architecture

Master Server

RDF DB Master

Replicate

RDF DB Slave

Web & SPARQL Server (development)

Replicate

RDF DB Slave

Web & SPARQL Server (production)

Import scripts + RDF files

PMCRDFization
## Consuming the data: SPARQL

<table>
<thead>
<tr>
<th>SPARQL query</th>
<th>Query expressed in natural language</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>WHERE {</code></td>
<td></td>
</tr>
<tr>
<td><code> ?article a bibo:Document ;</code></td>
<td></td>
</tr>
<tr>
<td><code>bibo:pmid ?pmid ;</code></td>
<td></td>
</tr>
<tr>
<td><code>?section a doco:Section ;</code></td>
<td></td>
</tr>
<tr>
<td><code>dcterms:isPartOf ?article ;</code></td>
<td></td>
</tr>
<tr>
<td><code>dcterms:title ?secTitle .</code></td>
<td></td>
</tr>
<tr>
<td><code>FILTER (regex(str(?secTitle), &quot;introduction&quot;, &quot;i&quot;)) .</code></td>
<td></td>
</tr>
<tr>
<td><code>?para a doco:Paragraph ;</code></td>
<td></td>
</tr>
<tr>
<td><code>dcterms:isPartOf ?section ;</code></td>
<td></td>
</tr>
<tr>
<td><code>cnt:chars ?text .</code></td>
<td></td>
</tr>
<tr>
<td><code>FILTER (regex(str(?text), &quot;cancer&quot;, &quot;i&quot;)) .</code></td>
<td></td>
</tr>
<tr>
<td><code>} LIMIT 50</code></td>
<td></td>
</tr>
</tbody>
</table>
**Consuming the data: SPARQL**

<table>
<thead>
<tr>
<th>SPARQL query</th>
<th>Query expressed in natural language</th>
</tr>
</thead>
</table>
| 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> .}  | Retrieving PubMed identifier for those articles that have been semantically annotated with the biological entity CHEBI:60004. The semantic annotation comes from the occurrence of the term “mixture” in any paragraph of the retrieved articles. |

**CHEBI:60004**

*A mixture is a chemical substance composed of multiple molecules, at least two of which are of a different kind*
Bio2RDF Integration

Metadata & References

BIBO

- pmc_vocabulary:Document
- pmc_vocabulary:Article
- pmc_vocabulary:AcademicArticle
- pmc_vocabulary:Proceedings
- pmc_vocabulary:Book
- pmc_vocabulary:CollectedDocument
- pmc_vocabulary:EditedBook
- pmc_vocabulary:Issue
- pmc_vocabulary:DocumentPart
- pmc_vocabulary:BookSection
- pmc_vocabulary:Chapter
- pmc_vocabulary:Thesis
- pmc_vocabulary:WebPage
- pmc_vocabulary:Journal
- pmc_vocabulary:authorList
- pmc_vocabulary:citedBy
- pmc_vocabulary:cites
- pmc_vocabulary:editorList

Content

DoCO

- pmc_vocabulary:Paragraph
- pmc_vocabulary:Section
- pmc_vocabulary:Figure
- pmc_vocabulary:Table

W3C CNT

- pmc_vocabulary:Content
- pmc_vocabulary:ContentAsText

Annotations

RDF

- pmc_vocabulary:Annotation
- pmc_vocabulary:ExactQualifier
- pmc_vocabulary:Selector
- pmc_vocabulary:ElementSelector
- pmc_vocabulary:StartEndElementSelector
- pmc_vocabulary:annotatesResource
- pmc_vocabulary:context
- pmc_vocabulary:createdBy
- pmc_vocabulary:hasTopic
- pmc_vocabulary:onResource

W3C PROV

- pmc_vocabulary:wasAttributedTo
- pmc_vocabulary:wasDerivedFrom

- object
- organization
- person
- user account
- has attribute
- has attribute

- has creator
- has attribute

- annotation
- symbol
- text span
- computational entity
- computational entity
- is annotation of
- refers to
- has creator
- refers to
- is about
# Consuming the data: Web services

<table>
<thead>
<tr>
<th>Retrieval</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>A list of terms and their related topics</td>
<td><a href="http://biotea.idiginfo.org/api/terms">http://biotea.idiginfo.org/api/terms</a></td>
</tr>
<tr>
<td>A list of topics and their related vocabularies</td>
<td><a href="http://biotea.idiginfo.org/api/topics">http://biotea.idiginfo.org/api/topics</a></td>
</tr>
<tr>
<td>All topics related to a term</td>
<td>e.g., <a href="http://biotea.idiginfo.org/api/topics?term=cancer">http://biotea.idiginfo.org/api/topics?term=cancer</a></td>
</tr>
<tr>
<td>All vocabularies related to a term</td>
<td>e.g., <a href="http://biotea.idiginfo.org/api/vocabularies?term=cancer">http://biotea.idiginfo.org/api/vocabularies?term=cancer</a></td>
</tr>
<tr>
<td>All terms that start with a specific string (for autocompletion)</td>
<td>e.g., <a href="http://biotea.idiginfo.org/api/terms?prefix=canc">http://biotea.idiginfo.org/api/terms?prefix=canc</a></td>
</tr>
<tr>
<td>All topics related to a vocabulary</td>
<td>e.g., <a href="http://biotea.idiginfo.org/api/topics?vocabulary=po">http://biotea.idiginfo.org/api/topics?vocabulary=po</a></td>
</tr>
<tr>
<td>RDF of articles that include a term</td>
<td>e.g., <a href="http://biotea.idiginfo.org/api/articles?term=cancer">http://biotea.idiginfo.org/api/articles?term=cancer</a></td>
</tr>
<tr>
<td>Count of RDF of articles that include a term</td>
<td>e.g., <a href="http://biotea.idiginfo.org/api/articles?term=cancer&amp;count=true">http://biotea.idiginfo.org/api/articles?term=cancer&amp;count=true</a></td>
</tr>
<tr>
<td>A list of vocabularies and their prefixes</td>
<td><a href="http://biotea.idiginfo.org/vocabularies">http://biotea.idiginfo.org/vocabularies</a></td>
</tr>
<tr>
<td>RDF of articles that include a vocabulary</td>
<td>e.g., <a href="http://biotea.idiginfo.org/api/articles?vocabulary=po">http://biotea.idiginfo.org/api/articles?vocabulary=po</a></td>
</tr>
</tbody>
</table>
Consuming the data: a dashboard for semantic bio-publications

Semantically enriched publication

Metadata + Content + References

biomedical annotation

SPARQL

Automatically Annotated RDF

Catalase
Consuming the data: first prototype

a) Retrieval: Metadata + Cloud of annotations

Cloud of Bio-annotations (term + # of bio-entities)

Title & authors

Links

Abstract

b) Enriched content → facts-based reading

Graphical tools

Paragraphs containing the annotation selected by the user

Interactive zone
Consuming the data: A first prototype
Challenges and Lessons

• Content
  • Tables and images → Links
  • Inline tables → Format is lost
  • Supplementary material
  • Most of them follow one DTD but ...

• References
  • At least 4 different styles
  • Some times are just plain text

• Annotators
  • Not always available
  • Stop words are tricky
Challenges and Lessons

- Where are the facts? How to validate the facts?

- Delivering the expressivity of the data set to the end user is a complex issue

- Annotation is context dependent

- Maintaining the triplet store has a learning curve of its own
  - Building SW infrastructure is H A R D
Currently working on:
Literature Discovery Process

• Search
  • Usually string-based search mechanisms
  • Little cognitive support

• Retrieval
  • Simple list of DB entries
  • Little cognitive support

• Interacting with the document
  • Straight into the PDF
  • Zero cognitive support
  • Data availability
451 documents
52 journals
756 terms
48 authors

Related terms:

Catalase
Protease
Insulin
Insulin-like
Catalase
Protease
catalase
protease

Summary results
Currently working on: Literature Discovery Process

- **Search**
  - Usually string-based search mechanisms
  - Little cognitive support

- **Retrieval**
  - *Simple list of DB entries*
  - *Little cognitive support*
  - How, why and where are a set of documents similar?

- **Interacting with the document**
  - Straight into the PDF
  - Zero cognitive support
Paper list Item View

Title
Author1, Author2...
- Same as...
- See also...
Abstract

Annotation Cloud click ↓ shows the text of term context

navigate between texts
Currently working on: Literature Discovery Process

- **Search**
  - Usually string-based search mechanisms
  - Little cognitive support

- **Retrieval**
  - Simple list of DB entries
  - Little cognitive support

- **Interacting with the document**
  - Straight into the PDF
  - Zero cognitive support
Future Work

• RDF
  • URI standardization following similar patterns to identifiers.org and Bio2RDF
  • Integration into Bio2RDF
  • Dataset identification and summary (void)
  • Improve data for references

• User Experience
  • Web services for data analysis
  • RDF browser
  • More visualization tools
  • Supporting and taking advantage of the structure of the document
  • Collaborative element
Future Work

- Application in Clinical Psychology, the MSRC case
  - From PDF to XML to RDF to Enriched Metadata for the PDF
  - The PDF is gently introduced in the WoD
  - Once the metadata has been enriched then
    - Rich interaction supporting: SEARCH-RETRIEVAL-INTERACTION WITH THE DOCUMENT (PDF)
Conclusions

• We provide
  • the transformation into RDF from the original PMC files
  • the annotation of the RDF
  • an API which makes that data available.
• New vocabularies as well as annotators can easily be plugged in
• Our approach is useful for both open and non-open access datasets
  • Publishers may decide what to expose via RDF and what content to make available
• Our approach is also applicable for PDF-only environments
Acknowledgments

• The MSRC consortium
• Greg Riccardi, FSU
• Oscar Corcho, UPM
• Olga Giraldo, UPM
• Bob Morris, Harvard University
• Michel Dumontier, Carleton University
• Dietrich Rebholz-Schuhmann, University of Zurich
• Diane Leiva, FSU
• US DoD Grant MOMRP Grant w81xwh-10-2-0181
• All of those who gave us feedback about the RDFization and the quality of our RDF datasets
Thanks for your attention

Contacts

- Alexander García: agarciac@gmail.com
- L. Jael García Castro: leylajael@gmail.com