

# Rare Disease Ontology & Search Tool: Demonstration and Discussion

- John Bellamy
- Indiana University

# Agenda:

- Part I: A discussion of the problem and an exploration of the steps completed to arrive at final model and tool.
- Part II: A demonstration of the Rare Disease Ontology and Search Tool.
- Part III: Discussion.
- Conclusion.



# Introduction:

- As highlighted in the report *Improving Rare Disease Information Retrieval with Semantic Web Technology* this tool is built on keywords that appear frequently in two datasets.
- An ontology was created to improve information retrieval, through semantic reasoning.
- An ontology and model were then persisted and stored for serving in a Fuseki Sparql endpoint.

# Part I



# Problem:

- PubMed does a great job of indexing articles, in order to improve article retrieval, but doesn't index all or even most articles about rare diseases. Furthermore, many medical research databases do not index rare diseases at all, thus one is left to search only by disease name or title.
- Can we beat PubMed's results using *just* the "title" field?

# About the data:

- Accessing <http://www.ncbi.nlm.nih.gov/pubmed/> (PubMed) I searched for common names for two rare diseases: Hunter's Syndrome and Spinal Bulbar Muscular Atrophy. The results were then combined into two datasets, one for each disease.
- Cluster analysis was then performed on the two datasets in Python. The most popular terms (disease names) were taken out to discover the other terms (pathophysiological) that make up the rest of the results.
- Finally, once a set of significant keywords were obtained for both datasets, *one* keyword/phrase was assigned as a tuple and the output further analyzed in Jena. A keyword was found for 72% of SBMA articles (545) and 71% of Hunter's Syndrome articles.



# Protege/Jena

- Using Protege, I developed an ontology for both datasets. The ontologies were mostly identical, except the URI's and named instances (keywords) were different so that when the two datasets were combined, two prefixes could be used to query. Disease names and pathophysiological terms that referred to the same thing were assigned owl:sameAs.
- Once I had the datasets and ontologies, I used the great library CSVReader to parse the files into .rdf format, in cooperation with Jena. The two datasets were then reasoned separately and output as .ttl files.

# Model Persistence

- Finally, the two .ttl files were combined and stored in a Jena TDB database, and the combined model output as a single .ttl file.
- The TDB database and the output file were then loaded into a Fuseki Server.



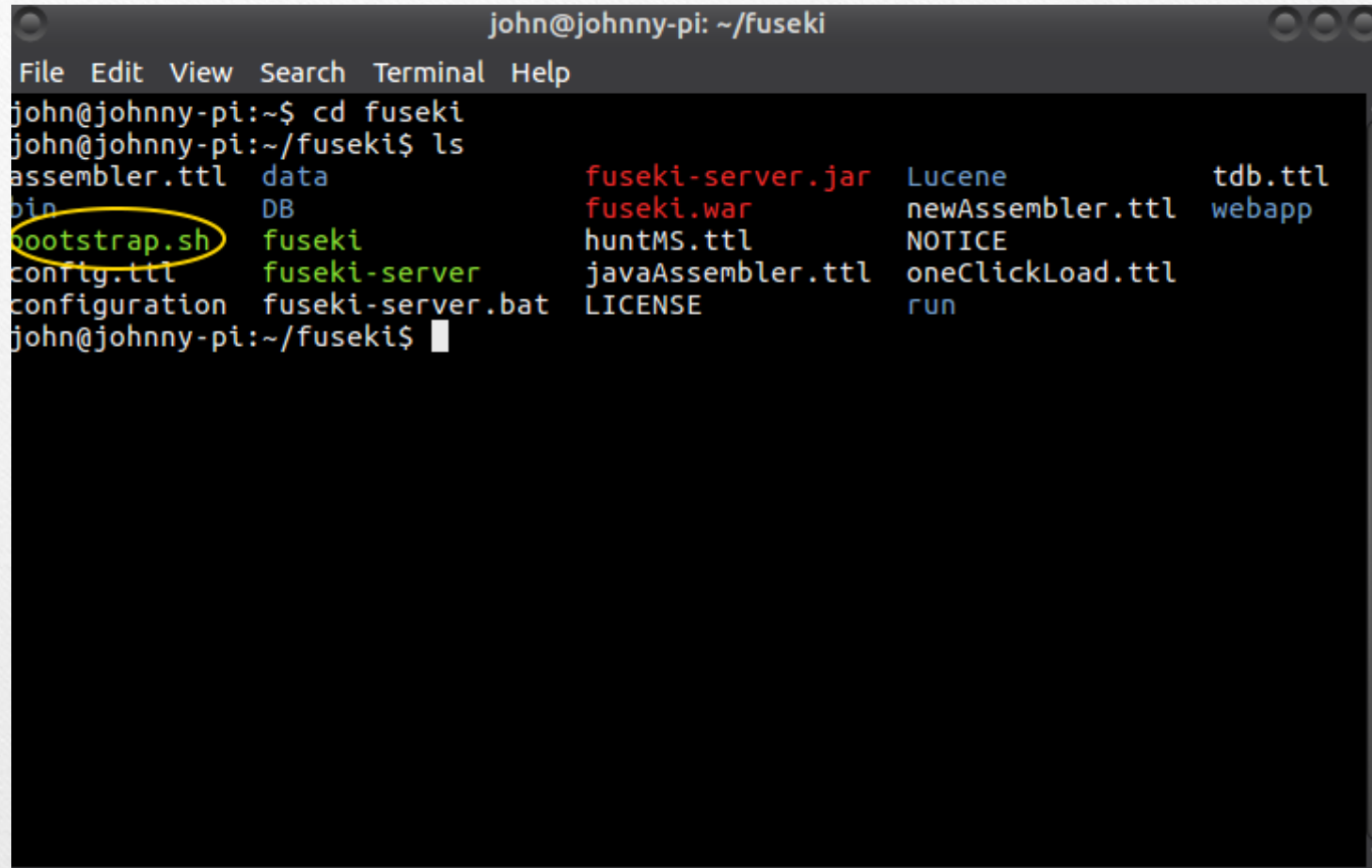
# Part II

# IDE

- The demo follows after the model has been created. Because the database alone is 128MB, and because you must include Fuseki, I recommend you use an IDE. I used the latest version of Fuseki and Jena 3.0.1. If you load the .jar file as it stands now, you will need to follow directions in the package's "readme."
- The demo will follow from the creation of the model, which generates everything needed in the fuseki folder.



# Navigate to fuseki folder:



A terminal window titled "john@johnny-pi: ~/fuseki" with a menu bar (File, Edit, View, Search, Terminal, Help). The user has executed "cd fuseki" and "ls". The output of "ls" is a multi-column listing of files and directories. The file "bootstrap.sh" is circled in yellow.

```
john@johnny-pi: ~/fuseki
File Edit View Search Terminal Help
john@johnny-pi:~$ cd fuseki
john@johnny-pi:~/fuseki$ ls
assembler.ttl  data          fuseki-server.jar  Lucene          tdb.ttl
bin            DB            fuseki.war         newAssembler.ttl webapp
bootstrap.sh   fuseki        huntMS.ttl         NOTICE
config.ttl     fuseki-server javaAssembler.ttl  oneClickLoad.ttl
configuration  fuseki-server.bat LICENSE            run
john@johnny-pi:~/fuseki$
```

# Launching Fuseki

```
john@johnny-pi: ~/workspace/Java/SemanticsFinal/fuseki
File Edit View Search Terminal Help
bin DB fuseki.war newAssembler.ttl webapp
bootstrap.sh fuseki huntMS.ttl NOTICE
config.ttl fuseki-server javaAssembler.ttl oneClickLoad.ttl
configuration fuseki-server.bat LICENSE run
john@johnny-pi:~/workspace/Java/SemanticsFinal/fuseki$ bash bootstrap.sh
[2016-07-15 13:05:58] Server INFO Running in read-only mode for /pubmedArticles/data
[2016-07-15 13:05:58] Server INFO Fuseki 2.4.0
[2016-07-15 13:05:58] Config INFO FUSEKI_HOME=/home/john/workspace/Java/SemanticsFinal/fuseki
[2016-07-15 13:05:58] Config INFO FUSEKI_BASE=/home/john/workspace/Java/SemanticsFinal/fuseki/run
[2016-07-15 13:05:58] Servlet INFO Initializing Shiro environment
[2016-07-15 13:05:58] Config INFO Shiro file: file:///home/john/workspace/Java/SemanticsFinal/fuseki/run/shiro.ini
[2016-07-15 13:05:58] Config INFO Template file: templates/config-tdb-dir
[2016-07-15 13:05:58] Config INFO TDB dataset: directory=DB
[2016-07-15 13:05:59] Config INFO Load configuration: file:///home/john/workspace/Java/SemanticsFinal/fuseki/run/configuration/huntMS.ttl
[2016-07-15 13:05:59] Config INFO Register: /pubmedArticles/data
[2016-07-15 13:05:59] Config INFO Register: /pubmedArticles
[2016-07-15 13:05:59] Server INFO Started 2016/07/15 13:05:59 MST on port 3030
```



# Navigate to localhost:3030

The screenshot shows the Apache Jena Fuseki web interface. The browser address bar displays 'localhost:3030'. The page header includes the Apache Jena Fuseki logo and navigation links: 'dataset', 'manage datasets', and 'help'. The main content area displays the title 'Apache Jena Fuseki' and the version 'Version 2.4.0. Uptime: 2m 19s'. Below this, a section titled 'Datasets on this server' contains a table with two columns: 'dataset name' and 'actions'. The table lists two datasets: '/pubmedArticles' and '/pubmedArticles/data'. Each dataset has three action buttons: 'query', 'add data', and 'info'. At the bottom, a light blue box contains a message: 'Use the following pages to perform actions or tasks on this server:'. Below this message, there are three links: 'Dataset', 'Manage datasets', and 'Help'. The 'Dataset' link is highlighted in blue. The descriptions for these links are: 'Dataset: Run queries and modify datasets hosted by this server.', 'Manage datasets: Administer the datasets on this server, including adding datasets, uploading data and', and 'Help: Get help with the Fuseki web interface.'

localhost:3030

Apache Jena Fuseki

dataset manage datasets help

## Apache Jena Fuseki

Version 2.4.0. Uptime: 2m 19s

### Datasets on this server

dataset name	actions
/pubmedArticles	<a href="#">query</a> <a href="#">add data</a> <a href="#">info</a>
/pubmedArticles/data	<a href="#">query</a> <a href="#">add data</a> <a href="#">info</a>

**i** Use the following pages to perform actions or tasks on this server:

<a href="#">Dataset</a>	Run queries and modify datasets hosted by this server.
<a href="#">Manage datasets</a>	Administer the datasets on this server, including adding datasets, uploading data and
<a href="#">Help</a>	Get help with the Fuseki web interface.

localhost:3030

Search



Apache  
Jena  
Fuseki



dataset



manage datasets



help

# Apache Jena Fuseki

Version 2.4.0. Uptime: 2m 19s

## Datasets on this server

dataset name

actions

/pubmedArticles

query

add data

info

/pubmedArticles/data

query

add data

info

Use the following pages to perform actions or tasks on this server:

**Dataset**  
**Manage datasets**

Run queries and modify datasets hosted by this server.

Administer the datasets on this server, including adding datasets, uploading data and



# What publishers do the two rare diseases have in common?

```
1 PREFIX sbma: <http://www.ncbi.nlm.nih.gov/pubmed/>
2 PREFIX hunter: <http://www.ncbi.nlm.nih.gov/pubmed/#>
3
4
5 SELECT DISTINCT $sbma $hunter
6
7 WHERE {{
8     $w sbma:ShortDetails $sbma.
9     $x hunter:ShortDetails $hunter .
10
11     FILTER($sbma =$hunter)}}
12 LIMIT 25
```

QUERY RESULTS



Table

Raw Response



Showing 1 to 14 of 14 entries

Search:

	sbma	hunter
1	"Clin Genet. 1998"	"Clin Genet. 1998"
2	"PLoS One. 2015"	"PLoS One. 2015"
3	"Am J Med Genet A. 2003"	"Am J Med Genet A. 2003"
4	"Hum Mol Genet. 2006"	"Hum Mol Genet. 2006"
5	"Hum Mol Genet. 1995"	"Hum Mol Genet. 1995"

# What articles mention both diseases?

- None :(
- Surprising, because these are both x-linked (affecting mostly the genes in males). These diseases also affect a similar number of males.



# How many articles can we return with just the keywords appearing in the title?

```
1 PREFIX sbma: <http://www.ncbi.nlm.nih.gov/pubmed/>
2 PREFIX hunter: <http://www.ncbi.nlm.nih.gov/pubmed/#>
3
4
5 SELECT DISTINCT $sbma $title
6
7 WHERE {{
8     $w sbma:KeyWords $sbma .
9     $x sbma:Title $title .
10
11     FILTER($sbma = "androgen receptor") }}
12
13 LIMIT 700
```

QUERY RESULTS



Table

Raw Response



Showing 1 to 50 of 546 entries

Search:

Show

	sbma	title
1	"androgen receptor"	"[A novel primer extension method to detect the number of CAG repeat rogen receptor gene in families with X-linked spinal and bulbar muscul
2	"androgen receptor"	"[X-linked recessive bulbospinal neuronopathy--Kennedy's syndrome].

# How many articles can we return using just the disease name?

```
1 PREFIX sbma: <http://www.ncbi.nlm.nih.gov/pubmed/>
2 PREFIX hunter: <http://www.ncbi.nlm.nih.gov/pubmed/#>
3
4
5 SELECT DISTINCT $hunter $title
6
7 WHERE{{
8     $w hunter:Keywords $hunter .
9     $x hunter:Title $title .
10
11
12     FILTER($hunter = "hunter's")}}
13 LIMIT 700
```

QUERY RESULTS



Table

Raw Response



Showing 1 to 50 of 518 entries

Search:

Show

entries

	hunter	title
1	"hunter's"	"Clinical response to long term enzyme replacement treatment in children, adolescent and adult patients with Hunter syndrome."
2	"hunter's"	"The role of enzyme replacement therapy in severe Hunter syndrome-an expert panel consensus."
3	"hunter's"	"Feasibility of first trimester prenatal diagnosis of Hunter syndrome."



## Part III

- While I expected the results with the disease name keyword, I was really surprised that the reasoner figured out the non-disease keywords were also related to the disease names. The result: for SBMA, “androgen receptor gene,” “androgen receptor cag” and other similar terms were replaced by “androgen receptor.” And this single term can return ALL of the articles loaded relating to SBMA. Astounding...
- What about searching disease names? As we see searching for “hunter’s” (I chose Hunter’s rather than Hunter’s disease or Hunter’s syndrome, because we see both.), searching by one of the popular names for the disease returns all of the results. Not 40% or so, as expected. How does this compare to PubMed?

# PubMed HubSchmed

The screenshot shows a PubMed search interface. At the top, there's a navigation bar with a 'How To' dropdown. Below it, a search bar contains 'sbma[Title]' with a dropdown menu set to 'PubMed'. To the right of the search bar are links for 'Create RSS', 'Create alert', and 'Advanced'. Below the search bar, there are options for 'Format: Summary' and 'Sort by: Title'. On the right side, there's a 'Send to' dropdown and a 'Filters: Man' link. The main section is titled 'Search results' and shows 'Items: 1 to 20 of 67'. Navigation links include '<< First', '< Prev', 'Page 1 of 4', 'Next >', and 'Last >>'. The first search result is a blue link: '17-DMAG ameliorates polyglutamine-mediated motor neuron degeneration through well-preserved proteasome function in an SBMA model mouse.' Below this link, it says 'PMID: 19066230' and 'Free Article', with a link to 'Similar articles'. The second search result is another blue link: 'Aberrant E2F activation by polyglutamine expansion of androgen receptor in SBMA neurotoxicity.' On the right side, there are sections for 'Titles with' (listing 'Polyglutamin...', 'interferes wi...', 'Glycolytic-to...', 'mTOR signa...', 'The Role of...', 'in SBMA.') and 'Find relate' (with a 'Find items' button). At the bottom right, there's a 'Search det' section.

How To

PubMed sbma[Title]

Create RSS Create alert Advanced

Format: Summary Sort by: Title

Send to Filters: Man

**Search results**

Items: 1 to 20 of 67

<< First < Prev Page 1 of 4 Next > Last >>

■ [17-DMAG ameliorates polyglutamine-mediated motor neuron degeneration through well-preserved proteasome function in an SBMA model mouse.](#)

PMID: 19066230 Free Article  
[Similar articles](#)

■ [Aberrant E2F activation by polyglutamine expansion of androgen receptor in SBMA neurotoxicity.](#)

**Titles with**

Polyglutamin  
interferes wi  
Glycolytic-to  
mTOR signa  
The Role of  
in SBMA.

**Find relate**

Find items

**Search det**



# Rare Disease Ontology & Search Tool:

```
1 PREFIX sbma: <http://www.ncbi.nlm.nih.gov/pubmed/>
2 PREFIX hunter: <http://www.ncbi.nlm.nih.gov/pubmed/#>
3
4
5 SELECT DISTINCT $sbma $title
6
7 WHERE{{
8     $w sbma:KeyWords $sbma .
9     $x sbma:Title $title .
10
11
12     FILTER($sbma ="sbma")}}
13 LIMIT 700
```

QUERY RESULTS



Table

Raw Response



Showing 1 to 50 of 546 entries

Search:

	sbma	title
1	"sbma"	"[A novel primer extension method to detect the numb rogen receptor gene in families with X-linked spinal an
2	"sbma"	"[X-linked recessive bulbosplinal neuronopathy--Kenne
3	"sbma"	"Testosterone treatment fails to accelerate disease in spinal and bulbar muscular atrophy."

# Remarks:

- The SBMA dataset had many more pathophysiological terms to describe it than the Hunter dataset. Therefore, there was more processing at the reasoner level; i.e. many pathophysiological declared owl:sameAs. Through reasoning, three of them disappeared; the variations on “androgen receptor.”
- For Hunter’s “lysosomal storage” disappeared, leaving 10/11 terms. It is interesting to note that the only “owl:sameAs” in this model, were the various disease names.



# Conclusion

- While the model still needs some fine-tuning, it is abundantly clear that semantic web technology and Jena reasoning delivers a very powerful database that outperforms traditional information retrieval methods based on disease names and indexing.