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: Closing remarks.
- 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:** A discussion of the problem and an exploration of the steps completed to arrive at final model and tool.

- 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

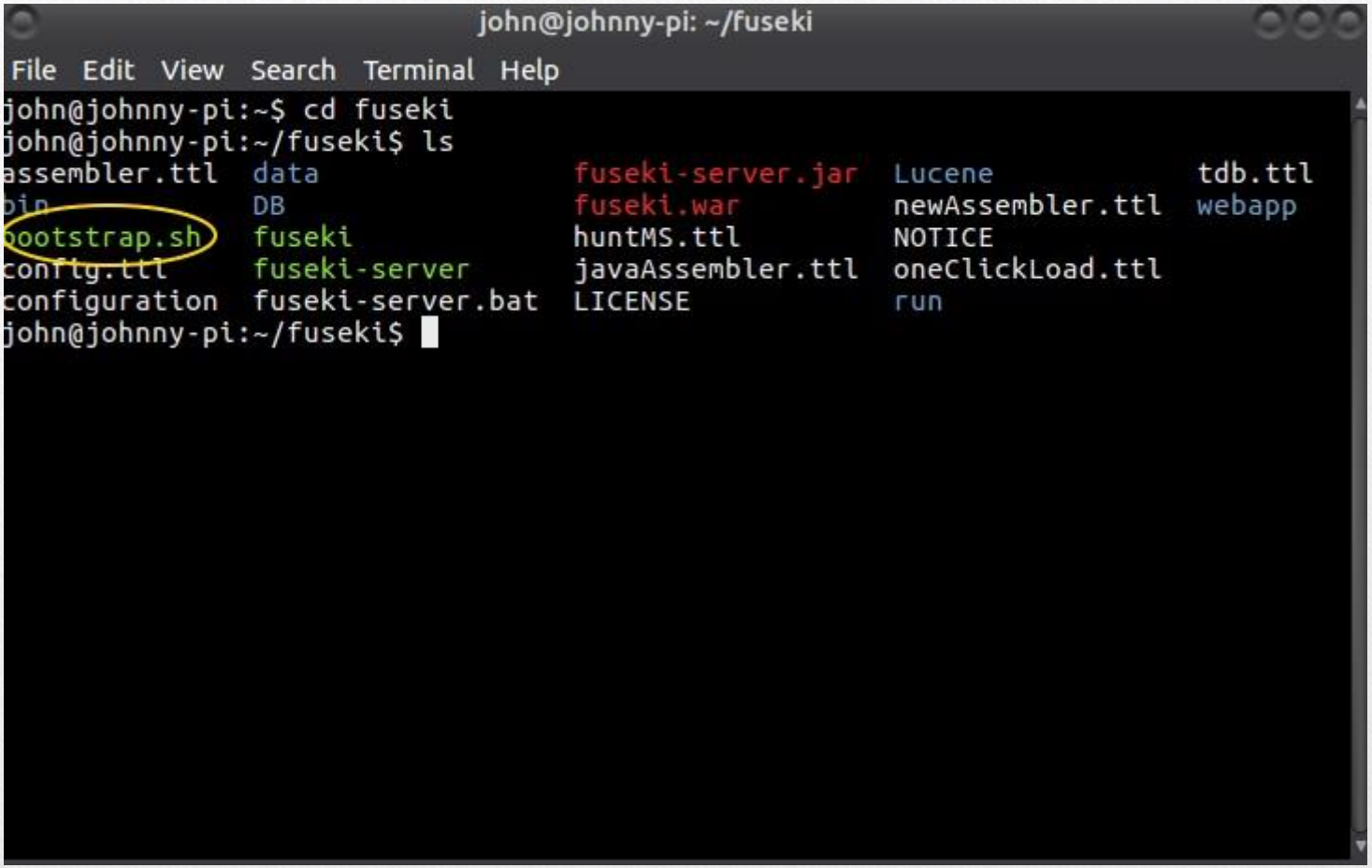
- Finally, the two .ttl files were combined and stored in a Jena TDB database, and the combined model output as a single .ttl file.

- The TDB database and the output file were then loaded into a Fuseki Server.

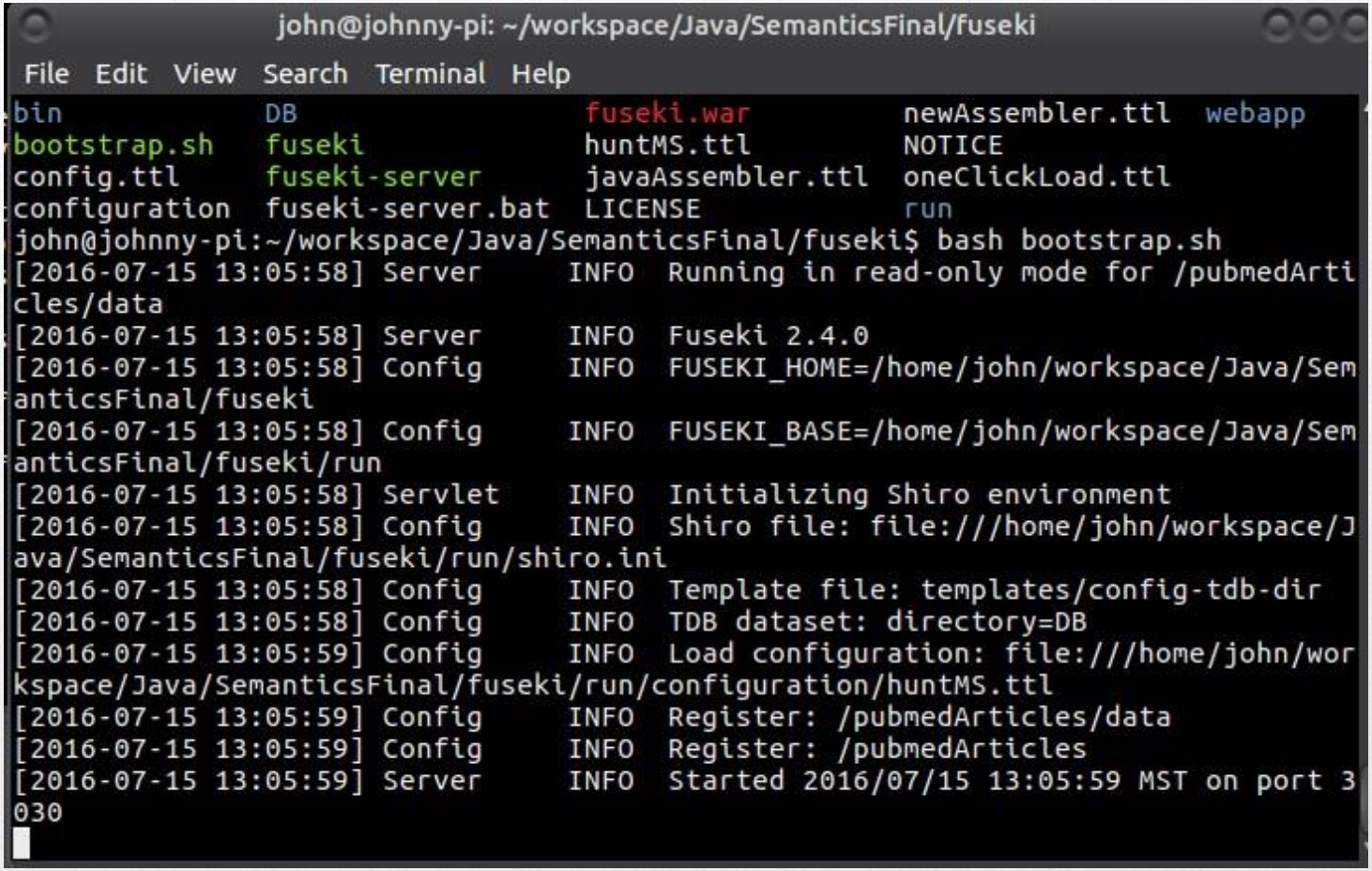
Part II: A demonstration of the Rare Disease Ontology and Search Tool.

- 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.

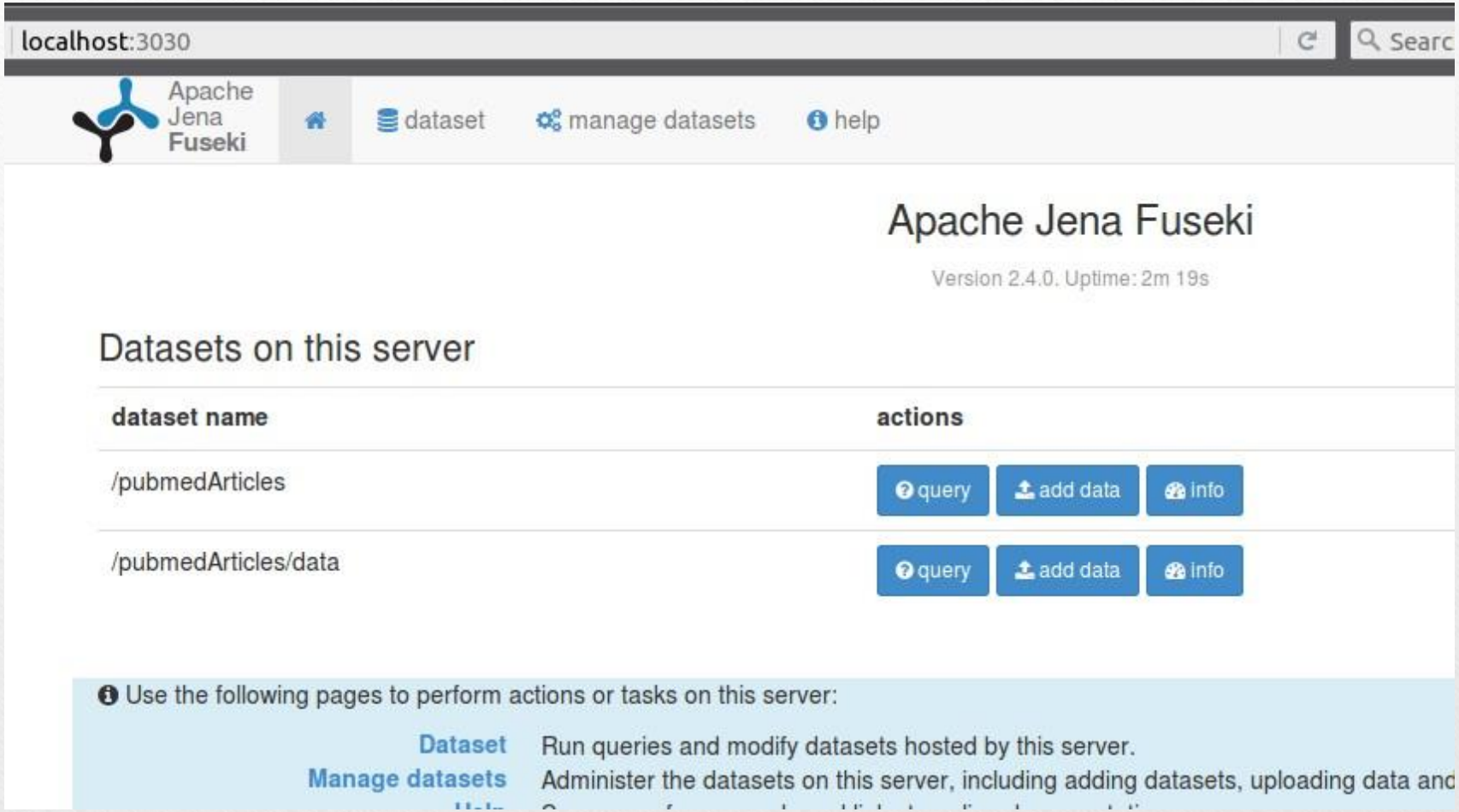
Navigate to fuseki folder:

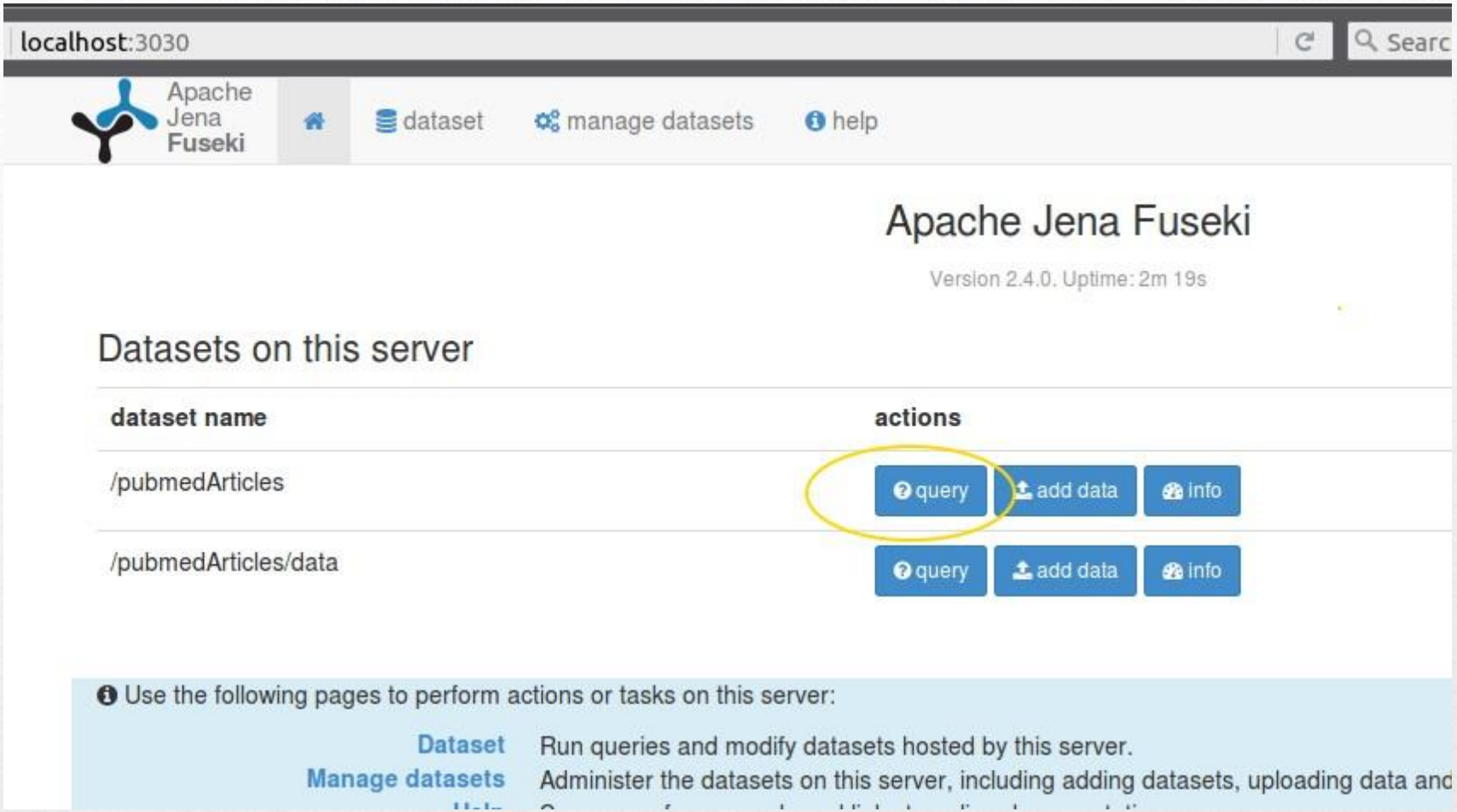


Launching Fuseki



Navigate to localhost:3030




localhost:3030
Searc
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
Run queries and modify datasets hosted by this server.
Manage datasets
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: [redacted]

	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?

```
PREFIX sbma: <http://www.ncbi.nlm.nih.gov/pubmed/>
PREFIX hunter: <http://www.ncbi.nlm.nih.gov/pubmed/#>


SELECT DISTINCT $sbma $title

WHERE{{
      $w sbma:KeyWords $sbma .
      $x sbma:Title $title .


      FILTER($sbma ="androgen receptor")}}
LIMIT 700
```

QUERY RESULTS

Table | Raw Response

Showing 1 to 50 of 546 entries

Search: [redacted] Show [redacted]

	sbma	title
1	"androgen receptor"	"[A novel primer extension method to detect the number of CAG repea 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?

```
PREFIX sbma: <http://www.ncbi.nlm.nih.gov/pubmed/>
PREFIX hunter: <http://www.ncbi.nlm.nih.gov/pubmed/#>


SELECT DISTINCT $hunter $title

WHERE{{
      $w hunter:KeyWords $hunter .
      $x hunter:Title $title .


      FILTER($hunter ="hunter's")}}
LIMIT 700
```

QUERY RESULTS

Table | Raw Response

Showing 1 to 50 of 518 entries

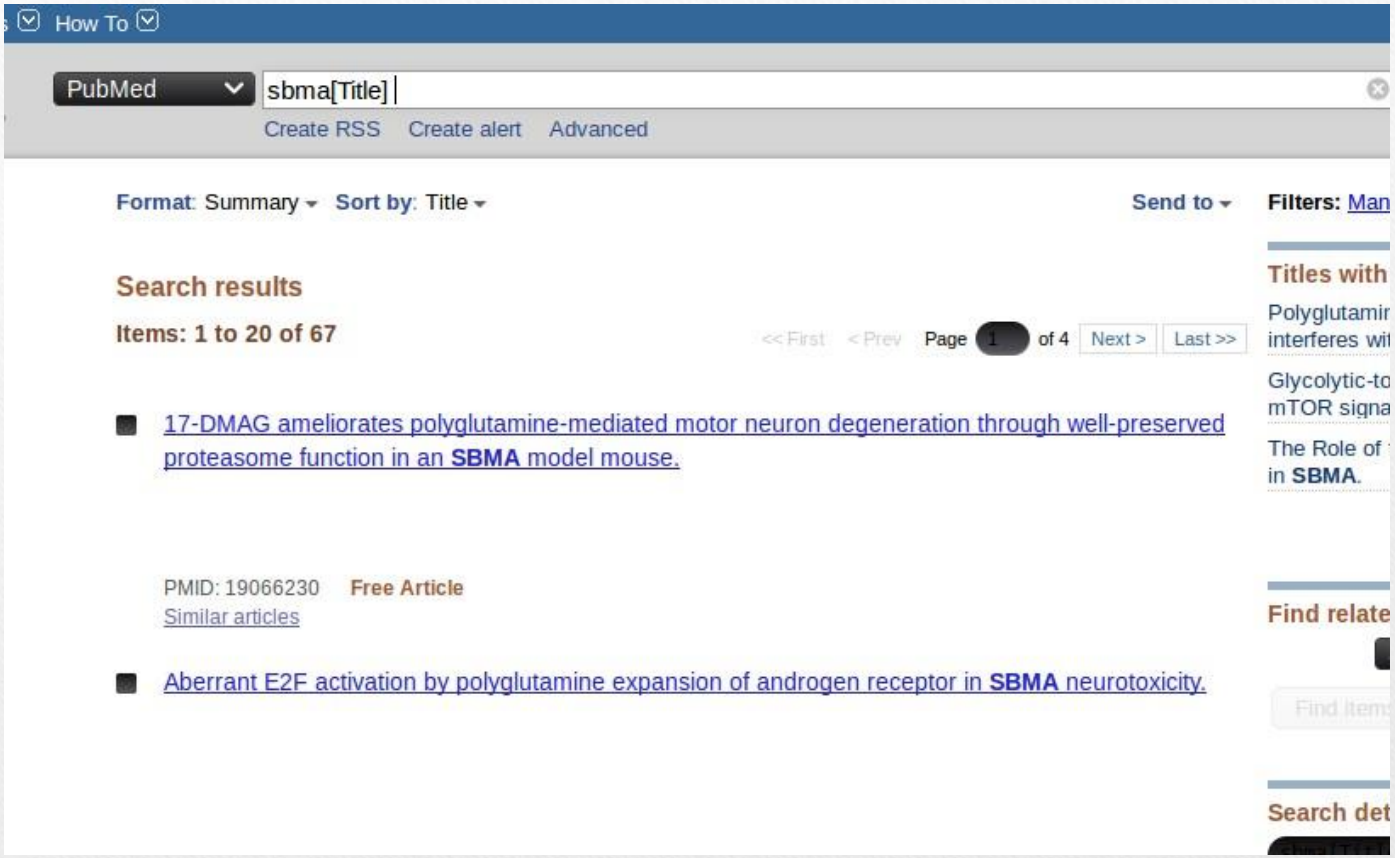
Search: [redacted] Show [redacted] entr

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

Part III: Closing Remarks

- 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. Awesome…
- 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


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
Polyglutamir
interferes wi
Glycolytic-to
mTOR signa
The Role of
in SBMA.
Find relate
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: [redacted]

	sbma	title
1	"sbma"	"[A novel primer extension method to detect the numb rogen receptor gene in families with X-linked spinal ar
2	"sbma"	"[X-linked recessive bulbospinal 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:samesAs. Through reasoning, three of them disappeared; the variations on "androgen receptor."
- For Hunter's "lyosomal 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.