Semantic Data Analytics – The key to challenging Big Data

Pascal Hitzler
Kno.e.sis Center
Wright State University, Dayton, OH
http://www.pascal-hitzler.de/
Semantic Web journal

- **EiCs:** Pascal Hitzler
  Krzysztof Janowicz

- New journal with significant initial uptake.

- We very much welcome contributions at the “rim” of traditional Semantic Web research – e.g., work which is strongly inspired by a different field.

- Non-standard (open & transparent) review process.

- [http://www.semantic-web-journal.net/](http://www.semantic-web-journal.net/)
Textbook

Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph

Foundations of Semantic Web Technologies

Chapman & Hall/CRC, 2010

Choice Magazine Outstanding Academic Title 2010 (one out of seven in Information & Computer Science)

http://www.semantic-web-book.org
Contents

• Big Data, Linked Data, Semantic Web
• An Example: Linked Data Querying
• The Big Data Added Value Pipeline
Gartner: Big Data Will Generate 6 Million U.S. Jobs by 2015

Susan Hall | CHARTING YOUR IT CAREER | 23 OCT, 2012

Gartner predicts that 4.4 million IT jobs will be created to support Big Data by 2015, with 1.9 million of them to be in the United States.

In addition, every Big Data-related role in the United States will create employment for three people outside of IT, pushing the total to 6 million U.S. jobs, Peter Sondergaard, senior vice president at Gartner and global head of research, told those attending the Gartner Symposium/ITxpo. He said:

But there is a challenge. There is not enough talent in the industry. Our public and private education systems are failing us. Therefore, only one-third of the IT jobs will be filled. Data experts will be a scarce, valuable commodity,” he said. “IT leaders will need immediate focus on how their organization develops and attracts the skills required. These jobs will be needed to grow your business. These jobs are the future of the new information economy.

Though I don’t follow Sondergaard’s math, we know there’s a shortage of analytics talent for Big Data and for engineering talent as well.
Big Data

Big Data is characterized not only by the enormous volume or the velocity of its generation but also by the heterogeneity, diversity and complexity of the data.

Suzi Iacono, source: http://community.topcoder.com/coeci/nitrd/

- **volume**: the sheer size of the data
- **velocity**: new data is added at breathtaking speed
- **variety**: different formats and different perspectives
- *(value: how useful is the data?)*
- *(veracity: how good/reliable is the data?)*
Contents

• Big Data, Linked Data, Semantic Web

• An Example: Linked Data Querying

• The Big Data Added Value Pipeline
### Linked Data: Volume

<table>
<thead>
<tr>
<th>Number of Datasets</th>
<th>Number of triples (Sept 2011)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011-09-19  295</td>
<td>31,634,213,770 with 503,998,829 out-links</td>
</tr>
<tr>
<td>2010-09-22  203</td>
<td></td>
</tr>
<tr>
<td>2009-07-14  95</td>
<td></td>
</tr>
<tr>
<td>2008-09-18  45</td>
<td></td>
</tr>
<tr>
<td>2007-10-08  25</td>
<td></td>
</tr>
<tr>
<td>2007-05-01  12</td>
<td></td>
</tr>
</tbody>
</table>

From http://www4.wiwiss.fu-berlin.de/lodcloud/state/
Information as RDF triples / graph

- **LOTR** hasAuthor **Tolkien**.
- **Hobbit** hasAuthor **Tolkien**.
- **LOTR** hasCharacter **Bilbo**.
- **Hobbit** hasCharacter **Bilbo**.

![RDF Diagram]

- **Tolkien** hasAuthor **LOTTR**.
- **Bilbo** hasCharacter **Hobbit**.
- **Tolkien** hasCharacter **LOTR**.
- **Hobbit** hasAuthor **Tolkien**.
- **Bilbo** hasCharacter **Hobbit**.
### DBpedia: LOTR page

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia-owl:thumbnail</td>
<td><img src="http://upload.wikimedia.org/wikipedia/commons/thumb/6/62/Jrlotr_cover_design.jpg/200px-Jrlotr_cover_design.jpg" alt="LotR cover" /></td>
</tr>
<tr>
<td>dbpedia-owl:wikiPageExternalLink</td>
<td>![LotR links](<a href="http://lotr.wikia.com/">http://lotr.wikia.com/</a></td>
</tr>
<tr>
<td>dbpprop:author</td>
<td>dbpedia:J_R_R_Tolkien</td>
</tr>
<tr>
<td>dbpprop:books</td>
<td>dbpedia:The_Two_Towers</td>
</tr>
<tr>
<td></td>
<td>&quot;Volumes:&quot;</td>
</tr>
<tr>
<td>dbpprop:country</td>
<td>England</td>
</tr>
<tr>
<td>dbpprop:expiry</td>
<td>20 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:genre</td>
<td>dbpedia:Adventure_novel</td>
</tr>
<tr>
<td>dbpprop:hasPhotoCollection</td>
<td><img src="http://www.w1.wiwiss.fu-berlin.de/flickr/wrappr/photos/The_Lord_of_the_Rings" alt="LotR photos" /></td>
</tr>
<tr>
<td>dbpprop:imageCaption</td>
<td>Tolkien's own cover designs for the three volumes</td>
</tr>
<tr>
<td>dbpprop:language</td>
<td>English</td>
</tr>
<tr>
<td>dbpprop:mediaType</td>
<td>Print</td>
</tr>
<tr>
<td>dbpprop:name</td>
<td>The Lord of the Rings</td>
</tr>
<tr>
<td>dbpprop:pages</td>
<td>1216 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:precededBy</td>
<td>dbpedia:The_Hobbit</td>
</tr>
<tr>
<td>dbpprop:pubDate</td>
<td>21 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:publisher</td>
<td>dbpedia:Allen &amp; Unwin</td>
</tr>
<tr>
<td>dbpprop:small</td>
<td>yes</td>
</tr>
</tbody>
</table>
Linked Data: Volume

Geoindexed Linked Data – courtesy of Krzysztof Janowicz
http://stko.geog.ucsb.edu/location_linked_data
Data Velocity

- Weather sensors
- Tweets
- Satellite images
- …
Linked Data: Variety

Copernicus lunar crater located on earth – courtesy of Krzysztof Janowicz http://stko.geog.ucsb.edu/location_linked_data (missing reference coordinate system)

Copernicus is a lunar impact crater named after the astronomer Nicolaus Copernicus, located in eastern Oceanus Procellarum. It is estimated to be about 800 million years old, and typifies craters that formed during the Copernican period in that it has a prominent ray system.

Characteristics
Copernicus is visible using binoculars, and is located slightly northwest of the center of the Moon's Earth-facing hemisphere. South of the crater is the Mare Insularum, and to the south-southwest is the crater Reinhold. North of Copernicus are the Montes Carpatus, which lie at the south edge of Mare Imbrium. West of Copernicus is a group of dispersed lunar hills. Due to its relative youth, the crater has remained in a relatively pristine shape since it formed.

The circular rim has a discernible hexagonal form, with a terraced inner wall and a 30 km wide, sloping rampart that descends nearly a kilometer to the surrounding mare. There are three distinct terraces visible, and arc-shaped landslides due to slumping of the inner wall as the crater debris subsided.

Most likely due to its recent formation, the crater floor has not been flooded.
“Nancy Pelosi voted in favor of the Health Care Bill.”

Bills:h3962

Vote: 2009-887

Votes:2009-887/+ vote:hasOption

Aye rdfs:label

people/P000197 vote:votedBy

Nancy Pelosi name

H.R. 3962: Affordable Health Care for America Act dc:title

On Passage: H R 3962 Affordable Health Care for America Act dc:title

vote:vote vote:hasAction
Geoindexed Linked Data – courtesy of Krzysztof Janowicz
http://stko.geog.ucsb.edu/location_linked_data
Linked Data: Veracity

Courtesy of Krzysztof Janowicz
http://stko.geog.ucsb.edu/location_linked_data

**SPARQL Query**

<table>
<thead>
<tr>
<th>place</th>
<th>populationCount</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia.Keta</td>
<td>18077</td>
</tr>
<tr>
<td><a href="http://sws.geonames.org/2304548/">http://sws.geonames.org/2304548/</a></td>
<td>29748</td>
</tr>
<tr>
<td>w-flick:Aneho</td>
<td>47579</td>
</tr>
<tr>
<td><a href="http://sws.geonames.org/8295630/">http://sws.geonames.org/8295630/</a></td>
<td>6814400000</td>
</tr>
<tr>
<td>dbpedia.Lome</td>
<td>749700</td>
</tr>
<tr>
<td><a href="http://sws.geonames.org/2393947/">http://sws.geonames.org/2393947/</a></td>
<td>9847</td>
</tr>
</tbody>
</table>
Linked Data: Veracity

Courtesy of Krzysztof Janowicz

http://stko.geog.ucsb.edu/location_linked_data
EarthCube requires

- information integration
- interoperability
- conceptual modeling
- intelligent search
- data-model intercomparison
- data publishing support

Semantic Web studies

- information integration
- interoperability
- conceptual modeling
- intelligent search
- data-model intercomparison
- data publishing support

Pascal Hitzler, WSU; Krzysztof Janowicz, UCSB
Linked Data and Big Data

• Linked Data is a kind of structured Big Data

• Linked Data is Big Data in a nutshell

Many of the same problems

Testbed for Big Data solutions

Intermediate stage for getting semantics into Big Data
Contents

• Big Data, Linked Data, Semantic Web
• An Example: Linked Data Querying
• The Big Data Added Value Pipeline
Basic Idea of the Semantic Web

- **Person 1** (left): Exchange of symbols (e.g., "Duck")
  - HA1: Concept
  - HA2: Concept

- **Person 2** (middle): Ontology description
  - Main concept: Ontology
  - Agreement

- **Agent 1** (right): Exchange of symbols
  - MA1: Specific Domain, e.g., Animals
  - Agreement

- **Agent 2** (far right): Specific Domain, e.g., Animals

**Concept**: Basic Idea of the Semantic Web

**Symbol**: MA1, MA2

**Thing**: Specific Domain, e.g., Animals

**Ontology**: Agreement
Ontology Example

subClass

instantiation

Declaration of properties

Declaration of classes

schema knowledge
PhDStudent v 9advisedBy.Professor

rules
responsible_for(y,x) Æ Professor(y)
! Employee(x)
Basic Idea of the Semantic Web

Ontology represents general domain knowledge

e.g. every publication has an author

DL Rules
Krötzsch, Rudolph, Hitzler
ECAI 2008

Data e.g. on Websites

February 2013 – Siemens Princeton NJ – Pascal Hitzler
Basic Idea of the Semantic Web

e.g. every publication has an author
Basic Idea of the Semantic Web

Ontology represents general domain knowledge

e.g. every publication has an author

Data e.g. on Websites

DL Rules
Krötzsch, Rudolph, Hitzler
ECAI 2008
The Science Behind an Answer
Watson performs so fast that it can rival the greatest human contestants in understanding a Jeopardy! clue and arriving at a single, precise answer. The significance of this accomplishment can be difficult to comprehend. Watch the video to see how the computing system designed to play Jeopardy! works.

The DeepQA hypothesis is that by complementing classic knowledge-based approaches with recent advances in NLP, Information Retrieval, and Machine Learning to interpret and reason over huge volumes of widely accessible naturally encoded knowledge (or “unstructured knowledge”) we can build effective and adaptable open-domain QA systems. While they may not be able to formally prove an answer is correct in purely logical terms, they can build confidence based on a combination of reasoning methods that operate directly on a combination of the raw natural language, automatically extracted entities, relations and available structured and semi-structured knowledge available from for example the Semantic Web.

What is Watson?
Implications for analytics, system design and industry transformation

Watson for a Smarter Planet™
Apple Buys Siri: Once Again The Back Story Is About Semantic Web

According to Robert Scoble who got it from tracking FTC, Apple is buying Siri. (This has yet to be confirmed by Siri or Apple). The front story is mobile, specifically a bruising battle between Apple and Google. But once again the back story is semantic technology. Siri is not some cute iPhone app hanged together in a garage over a Red Bull fueled long weekend. Siri has hard core semantic tech that originated from Darpa (just like that little system called the Internet).

Like the Facebook OpenGraph story, this is another example of semantic web going mainstream. The Open Graph front story was all about social media, but the back story was their adoption of RDFa. That has been a big boost to the semantic web community.

Siri looks like a good exit for investors and will give them confidence to invest more in companies
schema.org for enhancing web search
joint effort including Bing, Google, Yahoo, Yandex
Google and the future of search: Amit Singhal and the Knowledge Graph

Google has revolutionised the way we holiday, shop, work and play. Now, with Knowledge Graph, it plans to radically transform the way we search the internet… again. But some voice qualms about the company’s ambitions.

senses, that attribute is in the process of changing. This year, Google will roll out what it calls its Knowledge Graph, the closest any system has yet come to creating what Tim Berners-Lee, originator of the web itself, called "the semantic web", the version that had understanding as well as data, that could itself provide answers, not links to answers.

The Knowledge Graph is a database of the 500 million most searched for people, places and things in the Google world. For
Open Graph Protocol

We announced the next version of the Open Graph at f8 2011. It is currently released to the public. You can view the Open Graph documentation here. The below documentation refers to the first version of the Open Graph that shipped with the Like Button at f8 2010.

We have also updated how the Like Button will function with respect to the next version of Open Graph. Please read the following developer doc about the Like Button Migration.

The Open Graph Protocol enables you to integrate your Web pages into the social graph. It is currently designed for Web pages representing profiles of real-world things — things like movies, sports teams, celebrities, and restaurants. Including Open Graph tags on your Web page, makes your page equivalent to a Facebook Page. This means when a user clicks a Like button on your page, a connection is made between your page and the user. Your page will appear in the "Likes and Interests" section of the user’s profile, and you have the ability to publish updates to the user. Your page will show up in the same places that Facebook pages show up around the site (e.g. search), and you can target ads to people who like your content. The structured data you provide via the Open Graph Protocol defines how your page will be represented on Facebook.
Contents

• Big Data, Linked Data, Semantic Web

• An Example: Linked Data Querying

• The Big Data Added Value Pipeline
Example problem

“Identify films, the nations where they were shot and the population of these countries”

Issues:

• Where is the data? (what is in which linked dataset)
• How to query each dataset? (internal structure)

Here: Need to merge knowledge from IMDB and DBPedia datasets
Linked Data federated querying

Query → Upper level ontology → Answer

- LOD IMDB Dataset
- LOD Wikipedia Dataset (DBPedia)

Joshi, Jain, Hitzler et al. ODBASE 2012
Linked Data federated querying

Where do the mappings come from?

Query → Upper level ontology → Answer

LOD IMDB Dataset

LOD Wikipedia Dataset (DBPedia)

Joshi, Jain, Hitzler et al. ODBASE 2012
Table 4. Results of various systems for LOD Schema Alignment. Legends: Prec=Precision, Rec=Recall, M=Music Ontology, B=BBC Program Ontology, F=FOAF Ontology, D=DBpedia Ontology, G=Geonames Ontology, S=SIOC Ontology, W=Semantic Web Conference Ontology, A=AKT Portal Ontology, err=System Error, NA=Not Available

<table>
<thead>
<tr>
<th>Test</th>
<th>Prec</th>
<th>Rec</th>
<th>Prec</th>
<th>Rec</th>
<th>Prec</th>
<th>Rec</th>
<th>Prec</th>
<th>Rec</th>
<th>Prec</th>
<th>Rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>M,B</td>
<td>0.4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>err</td>
<td>err</td>
<td>0.04</td>
<td>0.28</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M,D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>err</td>
<td>err</td>
<td>0.08</td>
<td>0.30</td>
<td>0.45</td>
<td>0.01</td>
</tr>
<tr>
<td>F,D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>err</td>
<td>err</td>
<td>0.11</td>
<td>0.40</td>
<td>0.33</td>
<td>0.04</td>
</tr>
<tr>
<td>G,D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>err</td>
<td>err</td>
<td>0.23</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S,F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>0.2</td>
<td>0.52</td>
<td>0.11</td>
<td>0.30</td>
<td>0.20</td>
</tr>
<tr>
<td>W,A</td>
<td>0.12</td>
<td>0.05</td>
<td>0.16</td>
<td>0.03</td>
<td>err</td>
<td>err</td>
<td>0.06</td>
<td>0.4</td>
<td>0.38</td>
<td>0.03</td>
</tr>
<tr>
<td>W,D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>err</td>
<td>err</td>
<td>0.15</td>
<td>0.50</td>
<td>0.27</td>
<td>0.01</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.07</td>
<td>0.01</td>
<td>0.17</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>0.17</td>
<td>0.43</td>
<td>0.25</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Jain, Hitzler et al, ISWC2010
Table 1. Results on the oriented matching track. Results for RiMOM and AROMA have been taken from the OAEI 2009 website. Legends: Prec=Precision, A-API=Alignment API, OMV=OMViaUO, NaN=division by zero, likely due to empty alignment.

<table>
<thead>
<tr>
<th>Test</th>
<th>A-API</th>
<th>OMV</th>
<th>S-Match</th>
<th>AROMA</th>
<th>RiMoM</th>
<th>BLOOMS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>Prec</td>
<td>Rec</td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>1XX</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.06</td>
<td>0.01</td>
<td>0.71</td>
</tr>
<tr>
<td>2XX</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.03</td>
<td>0.05</td>
<td>0.30</td>
</tr>
<tr>
<td>3XX</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
<td>0.047</td>
<td>0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.04</td>
<td>0.03</td>
<td>0.38</td>
</tr>
</tbody>
</table>
1. **Pre-processing of the input ontologies** in order to (i) remove property restrictions, individuals, and properties, and to (ii) tokenize composite class names to obtain a list of all simple words contained within them, with stop words removed.

2. **Construction of the BLOOMS forest** $T_C$ for each class name $C$, using information from Wikipedia.

3. **Comparison of constructed BLOOMS forests**, which yields decisions which class names are to be aligned.

4. **Post-processing** of the results with the help of the Alignment API and a reasoner.
Fig. 1. BLOOMS trees for Jazz Festival with sense Jazz Festival and for Event with sense Event. To save space, some categories are not expanded to level 4.
1. **Pre-processing of the input ontologies** in order to (i) remove property restrictions, individuals, and properties, and to (ii) tokenize composite class names to obtain a list of all simple words contained within them, with stop words removed.

2. **Construction of the BLOOMS forest** $T_C$ for each class name $C$, using information from Wikipedia.

3. **Comparison of constructed BLOOMS forests**, which yields decisions which class names are to be aligned.

4. **Post-processing** of the results with the help of the Alignment API and a reasoner.
Big Data Bootstrapping for Big Data

We

• use big data
• for aligning big data
• in order to query big data

Idea

Upper level ontology

Query

LOD IMDB Dataset

LOD Wikipedia Dataset (DBPedia)

Answer
“Identify films, the nations where they were shot and the population of these countries”

SELECT ?film ?nation ?pop

WHERE {
}
protonu:ofCountry maps to lmdb:country
protonu:Movie maps to lmdb:film
protont:populationCount maps to dbprop:populationCount

Alignment confidence > 0.9
(a) SELECT ?film ?nation ?pop
WHERE {
}

(b) SELECT ?nation ?pop
WHERE {
Querying Illustration – Results

(a)

```
lmdb-film:11446 protonu:ofCountry lmdb-country:IN.
lmdb-film:11446 rdfs:label "Run".

lmdb-film:17091 protonu:ofCountry lmdb-country:LK.
lmdb-film:17091 rdfs:label "Getawarayo".

lmdb-film:16973 protonu:ofCountry lmdb-country:IN.
lmdb-film:16973 rdfs:label "Kabeela".
```

(b)

```
dbpedia:India protont:PopulationCount 1210193422.
```
Querying Illustration – Result I

- With proxy identifiers

<table>
<thead>
<tr>
<th>film</th>
<th>name</th>
<th>nation</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>lmdb-film:17091</td>
<td>“Getawarayo”</td>
<td>aloqus:2908ba82</td>
<td>21324791</td>
</tr>
<tr>
<td>lmdb-film:16973</td>
<td>“Kabeela”</td>
<td>aloqus:9bc35ca1</td>
<td>1210193422</td>
</tr>
<tr>
<td>lmdb-film:11446</td>
<td>“Run”</td>
<td>aloqus:9bc35ca1</td>
<td>1210193422</td>
</tr>
</tbody>
</table>
**Querying Illustration – Result II**

- Without proxy identifiers

<table>
<thead>
<tr>
<th>film</th>
<th>name</th>
<th>nation</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>lmdb-film:17091</td>
<td>“Getawarayo”</td>
<td>lmdb-country:LK</td>
<td>21324791</td>
</tr>
<tr>
<td>lmdb-film:16973</td>
<td>“Kabeela”</td>
<td>lmdb-country:IN</td>
<td>1210193422</td>
</tr>
<tr>
<td>lmdb-film:11446</td>
<td>“Run”</td>
<td>lmdb-country:IN</td>
<td>1210193422</td>
</tr>
<tr>
<td>lmdb-film:11446</td>
<td>“Run”</td>
<td>nytimes:india_geo</td>
<td>1210193422</td>
</tr>
<tr>
<td>No.</td>
<td>Query</td>
<td>Datasets</td>
<td>Primary Ontology</td>
</tr>
<tr>
<td>-----</td>
<td>----------------------------------------------------------------------</td>
<td>---------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Q1</td>
<td>Identify movies, countries where they were shot and the latest population for these countries.</td>
<td>LinkedMDB, DBpedia</td>
<td>PROTON</td>
</tr>
<tr>
<td>Q2</td>
<td>List the semantic web people and their affiliation.</td>
<td>Semantic Web Dog Food</td>
<td>N/A</td>
</tr>
<tr>
<td>Q3</td>
<td>Find all Jamendo artists along with their image, home page, and the population of city they are near.</td>
<td>Jamendo, Geonames</td>
<td>N/A</td>
</tr>
<tr>
<td>Q4</td>
<td>Software companies founded in the US</td>
<td>DBpedia</td>
<td>PROTON</td>
</tr>
<tr>
<td>Q5</td>
<td>Find list of movies, director and actors and the population of their birth cities.</td>
<td>DBpedia, LinkedMDB, Factbook</td>
<td>PROTON</td>
</tr>
<tr>
<td>Q6</td>
<td>List the countries, birth rates and sex ratios.</td>
<td>DBPedia, Factbook</td>
<td>PROTON</td>
</tr>
<tr>
<td>Q7</td>
<td>Is Mayotte a country?</td>
<td>DBPedia</td>
<td>PROTON</td>
</tr>
<tr>
<td>Q8</td>
<td>Get the birthdates of folks who acted in Star Trek</td>
<td>DBPedia, LinkedMDB</td>
<td>PROTON</td>
</tr>
<tr>
<td>Q9</td>
<td>List Music artists and birth dates.</td>
<td>DBPedia, BBC Music, Jamendo</td>
<td>DBpedia</td>
</tr>
<tr>
<td>Q10</td>
<td>Find list of movies made in countries with population greater that 1 Billion.</td>
<td>DBpedia,LinkedMDB</td>
<td>DBPedia</td>
</tr>
</tbody>
</table>
## Comparison with other systems

<table>
<thead>
<tr>
<th>Features</th>
<th>ALOQUS</th>
<th>DARQ</th>
<th>SQUIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach</td>
<td>Uses upper level ontology (PROTON) or any other ontology as primary ontology for query serialization and execution.</td>
<td>Requires formal description of datasets in the form of Service Description.</td>
<td>Requires an initial URI to execute queries.</td>
</tr>
<tr>
<td>Query Creation</td>
<td>Creates query corresponding to every mapping for a concept.</td>
<td>Creates queries only corresponding to the concepts mentioned in the query.</td>
<td>Creates queries only corresponding to the concepts mentioned in the query.</td>
</tr>
<tr>
<td>Failsafe</td>
<td>Executes all sub-queries for multiple mappings. Hence retrieves at least partial answers if a specific endpoint doesn’t work.</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Detect Entity co-references</td>
<td>Crawls and also consumes sameAs.org webservices.</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Result Processing</td>
<td>Query answers, retrieved from different datasets are merged and presented to user.</td>
<td>Retrieves answers from multiple dataset based on service description.</td>
<td>Retrieves answers from multiple dataset through link traversal.</td>
</tr>
<tr>
<td>Write queries using ontology not present in LOD</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Support for open-ended queries like ?s ?p ?o</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Result Storage for later Retrieval</td>
<td>Yes</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DESCRIBE Query Form</td>
<td>Yes</td>
<td>N/A</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Linked Data federated querying

Query ➔ Upper level ontology ➔ Answer

LOD IMDB Dataset

LOD Wikipedia Dataset (DBPedia)

Joshi, Jain, Hitzler et al. ODBASE 2012
Contents

• Big Data, Linked Data, Semantic Web
• An Example: Linked Data Querying
• The Big Data Added Value Pipeline
The Big Data Added Value Pipeline

- big noisy data
- Data Analytics
  - meaning/information via patterns
- Formal Semantics
  - intelligent systems applications
- ontologies / metadata
- data
The Big Data Added Value Pipeline

- **big noisy data**
- **Learning, Mining**
- **meaning/information via patterns**
- **Reasoning, Deduction**
- **intelligent systems applications**

- **ontologies / metadata**
- **data**
Other stuff I’m doing

• Pushing the limits of ontology modeling languages (OWL, RDF).

• Ontology reasoning algorithms.

• Ontology modeling.

• EarthCube.
Thanks!

Happy to talk about collaboration opportunities.
References


References


References
