Strange experience

- How I feel today.
- Johan: long tail is so long he cannot even see the end
- Maarten: writes programs targetting the long tail
- And I?

The Semantic Web has a long tail!
And we have to deal with it

So, what is the Semantic Web (aka Web of Data)
**Semantic Web in 5 principles**

1. Give all things a name
2. Make a graph of relations between the things at this point we have (only) a Giant Graph
3. Make sure all names are URIs at this point we have (only) a Giant Global Graph
4. Add semantics (= predictable inference)

---

**Examples of “semantics”**

- married-to
- lowerbound
- upperbound

- Φρονίκ is male
- married-to relates males to females

- married-to relates
- 1 male to 1 female
- Λυγνία = Ηαζέλ.

**Semantics = predictable inference**

---

**A success story**

---

**Who cares about head or tail?**

- Formally there is no problems with the long tail.
- That’s not a feature, but a bug.
How big is the Semantic Web now?

Jupiter

Denny Vrandečić – AIFB, Universität Karlsruhe (TH)

Building a WoD observatory?
An observatory for the biggest Knowledge Base ever

- Add LOTUS: from words to resources
- A centralised infrastructure to work with and analyse decentralised data

LOD Laundromat:
clean other peoples dirty data
crawl from registries + user driven
clean syntax errors
compute meta-data information
publish triples: gzip, hdt, ldf
Publish meta-data: SPARQL
harvest 1B triples/day

LOD Laundromat

Models, Heads and Tails
What is this picture telling us?

Does the meaning of a node depend on the cluster it appears in?
- Does path-length correlate with semantic distance?
- Are highly connected nodes more certain?
- Mutual influence of low-level and high-level structure?

Web of Data: Meet the long tail

BTC: Degree distribution

Tails versus heads: Social Semantics?

Comparing WoD 2009 & 2010:
- increasing powerlaw behaviour.
- top 5 by degree centrality in sameAs-aggregated

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SameAs Degree centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverb.com</td>
<td>0.039</td>
</tr>
<tr>
<td>Semantwebo.org</td>
<td>0.027</td>
</tr>
<tr>
<td>Wikipedia.org</td>
<td>0.027</td>
</tr>
<tr>
<td>Data.semanticweb.org</td>
<td>0.019</td>
</tr>
<tr>
<td><a href="http://www.defile">www.defile</a></td>
<td>0.017</td>
</tr>
</tbody>
</table>
Head and tail matter!

\[ 2 + 2 = 5 \]

Warning: Which head, which tail?
Hotspots in Knowledge Graphs:
Observations

- Realistic queries only hit a small part of the data (< 2%)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>#queries</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBPedia 3.9</td>
<td>459M</td>
<td>1640</td>
<td>0.003%</td>
</tr>
<tr>
<td>Linked Geo Data</td>
<td>289M</td>
<td>81</td>
<td>1.917%</td>
</tr>
<tr>
<td>MetaLex</td>
<td>204M</td>
<td>4933</td>
<td>0.016%</td>
</tr>
<tr>
<td>Open-BioMed</td>
<td>79M</td>
<td>911</td>
<td>3.100%</td>
</tr>
<tr>
<td>Rec2PROF/RECGG</td>
<td>50M</td>
<td>1297</td>
<td>2.013%</td>
</tr>
<tr>
<td>SW Dog Food</td>
<td>240K</td>
<td>193</td>
<td>39.438%</td>
</tr>
</tbody>
</table>

What does that imply?

Understanding heads and tails and how to use these insights
Hotspots in Knowledge Graphs: Benefits?

1) Ignore the long tail (for efficiency)!
2) Look at the long tail (for completeness)!

Closing

The Web of Data requires Semantics that “understand” heads and tails.