The Next Generation of Ontology Development — Usage-based Ontology Engineering

Preface

Markus Luczak-Rösch, STI Summit 2011
My Semantics of “Train”

<owl:Thing rdf:about="#LevisTrain">
  <rdf:type rdf:resource="#Train"/>
  <rdfs:label>Levi’s Train</rdfs:label>
  <madeOf rdf:resource="#Wood"/>
</owl:Thing>

fail everytime!
A Usage-dependent Life Cycle

- Enter the room
- Request to put away the “train”
- Negotiate understanding

- toy train
- made of plastic

- Select * WHERE ?t a:madeOf a:Plastic
- Select * WHERE ?t b:madeOf b:Wood

- toy train
- made of wood

Requirements
Yet another...

OTK

NeOn

METHONTOLOGY

DILIGENT

The Maintenance Black Box

Make it less a methodology but support the people to get their “Things” done!
Who is hurt by that?

- rather small/simple ontologies
  - min. effort for OE
  - “under-engineered”
- unknown user requirements
Hey “LOD people”, do you think that ontology engineering matters?

Usage-based ontology engineering
Publishers of 99% of the dataset do not feel responsible for their data?

Survey ran in October 2010
Concrete Example of Usage-based Approach

digging in log files
Usage?

• SELECT * WHERE ?t a:madeOf a:Plastic
• SELECT * WHERE ?t b:madeOf b:Wood

Request to put away the “train”

Yes*! But beyond?

• What about the future of SPARQL endpoints on the WoD?

* W.r.t. an architecture proposed by a famous “Web-Extremist”
You should have a query endpoint!

- You get something valuable out of it which helps you to play your role on the WoD!

Effort Distribution between Publisher and Consumer

- Consumer generates/data mines links
- Publisher provides links
- Third Party effort
- Consumer's effort

Effort Distribution

Data Publisher
- Publishes data as RDF
- Reuses terms from common vocabularies
- Sets links and publishes mappings

Third Parties
- Set links pointing at your data
- Publish mappings to the Web

Data Consumer
- Has to do the rest using data mining techniques for identity resolution and schema matching

Usage Analysis

- queries
- patterns
- triples
- primitives

visualize heat maps
zoom in and see details
Some Results (DBpedia Analysis)

- ns:Band ns:knownFor ?x
- ns:Band ns:nationality ?y
- inconsistent data
  - ns:Band ns:instrument ?x
  - ns:Band ns:genre ?y
  - ns:Band ns:associatedBand ?z
- missing facts
  - ns:Band ns:knownFor ?x
  - ns:Band ns:nationality ?y

Complete analysis can be found at http://page.mi.fu-berlin.de/mluczak/pub/visual-analysis-of-web-of-data-usage-dbpedia33/
Some Thoughts about Benefit

• usage analysis helps to acquire new knowledge
  – links between data
    → helps to increase the quality of data on the Web
  – external schema

• lightweight approach helps to bootstrap linked data

It is not necessary to automate everything if the result has enough (business) value in a problem domain anyway.
• LOD vocabularies are specific ontologies
• provide detailed usage statistics for your data
• usage analysis can help to maintain them
• it is not implicitly necessary to automate
• things that benefit the dataset publisher and the Web of data as a whole

Hey “LOD people”, do you think that dataset maintenance matters?

Markus Luczak-Rösch (luczak@inf.fu-berlin.de)
Freie Universität Berlin, Networked Information Systems (www.ag-nbi.de)
Actual Addition

• “15.500.000 people in Germany are not willing to use the internet”

– emphasis on the ESWC discussion: bridging the gap (directly or indirectly) between these people and the internet/Web has a high potential to influence societal transformation (they are not going to use a browser or an iPhone and they do not care for semantics)

Source: ARD-Morgenmagazin, 08-07-2011