The need for Lexicalization of Linked Data

John McCrae

Cognitive Interaction Technology Excellence Center – Universität Bielefeld
Linked Data

- Linked data is growing rapidly...
- … but mostly it looks like this:
Linked Data

- We need:
  - Natural Language Generation/Interface
    - Description in text
  - Question Answering
    - Mapping natural language description to (SPARQL) queries
  - Machine Translation
    - Adapting linked data vocabularies to new languages
Welcome

If you come to the Netherlands to live, work or study, you are likely to have some questions about the arrangements you need to make. This site will guide you to the government organizations you may have to deal with.

What can you do?

If you fill in your profile, you will see a list of subjects that are relevant for you, with links to information from the organizations concerned. You can also choose a subject directly.

Your profile

<table>
<thead>
<tr>
<th>What is your age?</th>
<th>Make your choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is your nationality?</td>
<td>Make your choice</td>
</tr>
<tr>
<td>Length of your stay?</td>
<td>Make your choice</td>
</tr>
</tbody>
</table>

The purpose of your stay?
(You may give more than one answer)

- work in Holland
- study in Holland
- live in Holland

Create my checklist

Subjects

- Education
- Employment
- Permits and visa
- Social security
- Taxes
- Vehicles

See also
- Employers (dutch)
- The forum (dutch)
how many rivers run through texas?

I found 1 answer.
ANSWER: 44

Are you satisfied with this answer? ☐ Yes ☐ No

WHAT?
PYTHIA is an ontology-based question answering system. This demo is made available for research purposes - feel free to play around with it.

HOW?
PYTHIA translates natural language into SPARQL by means of a deep linguistic analysis. Learn more »

WHO?
PYTHIA is developed by the Semantic Computing Group @ CITEC, Bielefeld University.
Labels

- Linguistic description of linked data terms by `rdfs:label`

- Usage statistics:

Simple labels are very ambiguous, e.g.,
- “addresses” (from openEHR Demographic)
  - The “addresses” of an organization?
  - Someone “addresses” an audience?
  - A set of web “addresses”??

Use URIs for labels not/as well as strings!
Linguistic Linked Data

Draft of the Linguistic Linked Open Data (LLOD) cloud

As of February, 2012

- lexical-semantic resources (LSRs), schemes for LSRs
- metadata and terminology repositories, linguistic KBs
- annotated corpora, schemes for annotated corpora

Open Linguistics Working Group
http://linguistics.okfn.org/lloid
Lexicon model for ontologies

• Common format for describing lexical information relative to 'ontologies' (OWL, RDF(S))

• Built on existing models
  - Lexical Markup Framework (ISO 24613)
  - SKOS

• Design:
  - Modular
  - Concise
  - RDF-native
  - Not prescriptive
• Allows full linguistic description
• Further development under W3C OntoLex community group
• Described in cookbook
People will not create a *lemon* model for each vocabularies
Instead refer to repositories on *lemon* data
- Such as *lemon source*

Before *lemon*
- `openehr:addresses rdfs:label “Addresses”@en`

With *lemon*
- `openehr:addresses lemon:lexicalization lemonsourcsource:address__noun__sense1__en`

Full linguistic description available by dereferencing URI
Thank you!

- Ontolex Community group
  - http://www.w3.org/community/ontolex
- Lemon cookbook
- Monnet project
  - http://www.monnet-project.eu/