Making sense out of the Wikipedia categories

A Google Summer of Code 2013 project

Mentor: Marco Fossati @hjfocs <fossati@spaziodati.eu>

Student: Kasun Perera <kkasunperera@gmail.com>

Intro

The latest version of the DBpedia ontology has 529 classes. It is not well balanced and shows a lack of coverage in terms of encyclopedic knowledge representation.
Furthermore, the current typing approach involves a costly manual mapping effort and heavily depends on the presence of infoboxes in Wikipedia articles.
Hence, a large number of DBpedia instances is either un-typed, due to a missing mapping or a missing infobox, or has a too generic or too specialized type, due to the nature of the ontology.
The goal of this project is to identify a set of senseful Wikipedia categories that can be used to extend the coverage of DBpedia instances.

How we used the Wikipedia category system

Wikipedia categories are organized in some kind of really messy hierarchy, which is of little use from an ontological point of view.
We investigated how to process this chaotic world.

Here's what we have done

We have identified a set of meaningful categories by combining the following approaches:

1. **Algorithmic**, programmatically traversing the whole Wikipedia category system.
   Wow! This was really the hardest part. Kasun made a great job! Special thanks to the category guru Christian Consonni for shedding light in the darkness of such a weird world.
2. **Linguistic**, identifying conceptual categories with NLP techniques.
   We got inspired by the YAGO guys.
3. **Multilingual**, leveraging interlanguage links.
   Kudos to Aleksander Pohl for the idea.
4. **Post-mortem**, cleaning out stuff that was still not relevant
   No resurrection without Freebase!

Outcomes

We found out a total amount of 3751 candidates that can be used to type the instances.
We produced a dataset in the following format:

```
<Wikipedia_article_page> rdf:type <article_category>
```

You can access the full dump here. This has not been validated by humans yet.

If you feel like having a look at it, please tell us what do you think about.