Knowledge and the Web 2015/2016: Homework 3

November 5, 2015

In the previous three exercise session we focused on using Semantic web to answer question and domain modelling supporting such question answering. Your projects will go in the same direction.

In this homework you will focus on finding interesting questions for your projects, based on the discussion we had on the last exercise session. You have seen a good and bad examples during the exercise session. The following components should be discussed in your homework:

- submit **one research question per group member, and compile this into one document** to increase coherence. Content-wise, these four/three questions can relate to one another, or be as divergent as in the first round. But please make sure you discuss everything with your group, so that also easy-to-fix problems, such as wordings that people other than yourself can't understand, are avoided. Provide a clear statement of a question, but also provide a more detailed explanation of the goal of the question.
- In addition to focusing on **a good research question**, we ask you to do three things for this homework:

  1. Combine data from different sources

  2. Be concrete about the **concepts** in the question. For example, if researching the connection between the educational background of the Parliament speakers and the topics of discussions they participate in, instead of *educational background*, specify things such as *obtained higher-education degree*, *university*, *jobs held before joining the parliament*, and so on. As you can see, *educational background* can be interpreted in different ways, and we ask you to take this into account - if any concept in your question can be interpreted in different ways, elaborate on it in the homework.

  3. Please take a closer look at the data. You need one or more **datasets** in addition to ToE. Look at each of them carefully:
     - We have compiled a first list of potentially interesting datasets on *Datasets* in the Homeworks folder, but of course please feel free to add to it. (The document is editable.) Please specify the **identity** of your chosen dataset, if applicable with version, and give an URL.
     - Describe, for the dataset, how it would support you in answering your question. The following **quality** factors could be helpful:
       * Relevance (does the information really help you to answer your question; for example, a US media source may not contain much news about Europe)

* Availability (can you actually get it? Sources must be gratis and open.)
* Size / number of data points that pertain to your question
* Correctness (i.e. look at selected data records to check whether they make sense)
* Completeness (are there too many missing values with respect to what you are interested in?)
* Timeframe to which the data refer (should match ToE)
* Use of common ontologies (e.g. FOAF, Geonames, dbpedia, ...)

For now, you can define these factors informally; later, we will ask those of you in the 6-ECTS version of the course to describe the properties of your chosen dataset formally. Please note that in **THIS** homework 3, **EVERYBODY** (also the 4-ECTS people) should think about data quality at the informal level described above.

Practicalities

The **deadline** for the homework is Sunday **November 15**, 2015, 23:59h. The homeworks should be submitted in groups, in the form of a report, using the **File Exchange** tool on Toledo. **You don't have to model the questions in Protege this time**.

As a reminder of yesterday's discussion in the exercise session, here are some important properties we are looking for in your research questions:

- Relevance of the question (avoid: *the relationship between number of daisies in Portugal with the voting behaviour of Polish parliamentarians*)
- Answer requires combining data from different sources
- Not a straightforward answer (not just a data look-up)
- Answering requires data analysis (not just retrieval) – looking for things such as correlation, importance of X for Y, etc.
- Answer requires interpretation of data, and it may depend on the operationalization that you use (e.g.: how you define *economically powerful country*)
- Interpretation is subject to self-criticism (*limitations and future work* – even if you don't do this future work)

Only one member should submit the homework for the entire team. Feel free to ask any questions about this assignment in the class, on Toledo or by e-mail to Sebastijan Dumancic (be sure to put [KaW] in the subject line).

Examples

An example of a good research question is:

Can we find any specific indicators (related to the country of representation of a speaker at the European Parliament) that influence how many speeches are given by a certain country in the European Parliament?

This question is a very good starting point for a project because it does not have a straightforward answer (the result will not hold for all the countries) as it looks for an impact of certain factors to the number of speeches over a larger timeline, which requires specification which indicators might influence the number of speeches. By choosing a different set of indicators you can change the result completely. The question also combines multiple data sources. While looking for the indicators, you can choose one the describe country, problems of a country, certain information about a speaker, or even combine all of it. The result depends on analysing the data, and is entirely subject to the interpretation. Finally, it investigates an interesting potential pattern that describes a partial *modus operandi* in the European Parliament. **Note:** only the group that suggested this question can reuse it for this homework.

A bad examples of a research question for a project is:

List the number of seats each country has in the European Parliament.

The question is a simple data look-up. It does not combine different data sources. You don't even need the Semantic web to answer this question.