WP4: Voortgangsverslag

Datum: 2015-10-25

Auteur: Richard Zijdeman(richard.zijdeman@iisg.nl)

Wat doet WP4?

WP4 houdt zich bezig met het opzetten van een Gestructureerde Data Hub (SDH), in eerste instantie ten behoeve van de sociaal-economische geschiedbeoefening, voor de curatie, opslag, het vinden, linken, selecteren, visualiseren en analyseren van gestructureerde datasets, waarbij de SDH geïntegreerd moet kunnen worden met de totale CLARIAH-infrastructuur. Het uiteindelijke criterium voor het slagen van het project is de mate waarin de SDH sociaal-economisch historici in staat stelt om data en tools te combineren voor innovatief onderzoek.

CLARIAH Structured Data Hub (CSDH)

The Clariah Structured Data Hub (CSDH) is one of 5 working packages in CLARIAH and aims to recover and connect datasets on social and economic history that now live in isolation. Recover in the sense that there are many projects for which datasets were created, but never appropriately cleaned and published. Connect in the sense that many of these datasets use similar variables such as occupation, sex or Gross Domestic Product, but there are virtual no procedures available to link these datasets (other than manual linkage through scripting). CSDH creates both tools to make the recovery and linkage easier, but also provides a series of strategic databases as linked open data.

The CSDH project started with a, now completed, pilot phase resulting in:

- A report describing the pilot and the most important outcomes
- Several events and presentations
- Preliminary products (software, linked historical datasets, and vocabularies).

The pilot report is available from Clariah's github https://github.com/CLARIAH/wp4-docs/blob/master/csdh_pilot_report.md and presents four main outcomes.

1 The first outcome is that few concepts and historical data are available for the field of economic and social history. That means we can borrow little from previous projects and need to create vocabularies for substantive historical research.

2 The second outcome shows that the strategic datasets that are to be transposed into linked data are very heterogeneous and requires field experts (historians) to properly transpose the data into RDF.

3 The third outcome concerns a technical practicality namely the increase in file sizes as a result of transposing the data into a linked data format. Increases of 20 to 40 times the original size have not been uncommon, one file going from a 2.5GB .csv file to a 121 GB .ttl file.

4 The final outcome regards the difference in levels to which one can create linked data. E.g. is it enough to link data by variable names or do we need to link at the cell level? The latter obviously is the more versatile, but is also less manageable.

In addition to these outcomes the report provides some guidelines for minimal requirements to transpose data within the project, for example related to file structure and file extension.

First half year

In it's first half year the CSDH project has also focused on outreach, mainly by organizing so called 'sounding board meetings' with researchers from demographic, economic and social history as well as those active in the field of cultural heritage. The third meeting will be held this December and during those meetings researchers provide feedback on our course of action (e.g. which datasets to transpose, what tools are missing) as well as advice on where to go next.

Outreach also has taken form in terms of deliverables, such as presentations, software, and linked vocabularies. Presentations are available from http://clariah-sdh.github.io, while software and vocabularies are available from: https://github.com/CLARIAH.

Coming months

In the coming months we will be finishing our prototype infra-structure that will provide vocabularies and a first release of QBer. After that we will be focusing on expanding the infra-structure to incorporate the strategic datasets that will be transposed to linked data in years 2 and 3 of the project. In the third year the focus will also turn towards 'ease of usability' when GUI's and visualization tools will be developed.

Sounding Board

PS: if you made it this far through the text, why not join us at our sounding board group meeting, December 16th, 15.00-18.00 hours at the International Institute of Social History?

Mail me at: richard.zijdeman@iisg.nl