SNAP:DRGN Advisory Board

1st meeting Skype (voice only) 2014-05-09

Present: Øyvind Eide (chair), Fabian Koerner, Robert Parker, Laurie Pearce, Charlotte Roueché, Rainer Simon, Gabriel Bodard (principal investigator)

Not present: Sonia Ranade.

The meeting lasted around one hour.

Minutes written by Øyvind Eide based on notes from Laurie Pearce and Rainer Simon.

1. Welcome from Advisory Board chair

Øyvind Eide welcomed all present member and expressed a hope for critical, but not evil input to the project. It was agreed that the chair would write minutes based on notes from Laurie Pearce and Rainer Simon.

2. Welcome from principal investigator

Gabriel Bodard welcomed and thanked the members of the board and welcomes critical feedback; however, this being an advisory board also reserves the project's right to ignore some of the feedback. The SNAP team is keen to hear what comes out of this meeting.

3. Introductions

All members of the board present presented themselves and their respective projects.

4. Project presentation

In his presentation of the project Gabriel Bodard covered the points below, also referring to the following documents on the project webpage:

a) the background, aims and objectives: http://snapdrgn.net/about

b) the outcomes of the workshop: http://snapdrgn.net/archives/110

c) progress and decisions so far: http://snapdrgn.net/ontology and http://bit.ly/SNAPcookbook

The projects has no intention to build a new prosopography, or an ontology for modelling prosopographies. Rather, the aim is to create a minimalist way of expressing relationships, although not as minimalist as Pelagios, with mechanism for attributing assertions of identities. This will be used to create a 'virtual' global prosopography, with annotations on top, capturing assertive statements and allow people to annotate texts, databases, etc. with "SNAP Ids", based (potentially) directly on the Pelagios approach & annotation model. The dataset created across projects can become huge. The team is currently working on an ontology and recommendations for exposing minimal information and annotations.

In the second half of the project it will mirror Pelagios, allowing people to annotate texts, and databases: this person named is this person with SNAP identifier and this link will take you to further information.

The London workshop was briefly summarised: discussions on technical approaches to named entity recognition; how to identify co-references between different datasets; how to encode dates

and places. See Website for more information. The draft for a cookbook was presented and the feedback was good, including robust criticism. A data model for exposing basics (id, names, place, and date) in the SNAP graph is more or less complete; covered in the emerging cookbook.

The project will continue looking for additional data sets while getting sample data from already participating providers and add it to the graph. On top of that there will be NER approaches and tools, as well as presentation and services on top of the triple store.

5. Questions and comments

Based on the report, the members of the Advisory Board asked questions.

Øyvind Eide: is the core of the project to build a co-reference service with some extensions?

Gabriel Bodard: The project will provide unique and stable ID to each person reference. Co-references across data sets will be strong component, but project also goes beyond that.

Charlotte Roueché asked to be reminded of the overall time table.

Gabriel Bodard explained that the project is a year pilot project almost half way through. Getting sample data into the graph is due by late June, the remainder of work will happen in two strands: experiment with new methods for new data and build services to show utility. The last three months will be used for writing up.

Øyvind Eide asked for a definition of "pilot": what is the hope for continuation.

Gabriel Bodard: no funding is lined up beyond this project. Big Data scheme for the pilot is the current funding. However, after this project he thinks they will be well placed for further, larger funding. He would hope to expand on multiple axes: services/technology, robustness of data, expand data set chronologically, geographically, linguistically.

Rainer Simon had some question about the data model: First, is it the same data model for prosopographical data and for annotating docs?

Gabriel Bodard confirmed.

Rainer Simon: are both data models complete?

Gabriel Bodard: Of the four parts, 1 (expose data) and 2 (adding co-references) are developed, 3 (adding annotations) is under discussion, whereas for 4 (use data to further annotation) they expect to borrow heavily from Pelagios.

Rainer Simon said he would like to see an alignment with Pelagios.

Øyvind Eide: in the work for conceptual development of prosopography, the workshop at the DH conference in Lausanne is an important step. Is CIDOC-CRM enough to express prosopographical relationships? He sees fundamental problems with the concept of persons, as CRM has a realist view of persons which prosopographies does not. A prosopography person can be mythical or a god, a CRM person has to be treated as a historical person. He suggested taking part in the next CRM SIG meeting to present the SNAP graph for feedback.

Gabriel Bodard, claiming not to be an RDF expert agreed that CRM cannot be used out of box. They would need additions and/or emendations to CRM, but want to map wherever they can – identical classes, superclasses, subclasses. They would be happy to present mapping the the next CRM SIG meeting.

Fabian Koerner addressed the time frame and wanted more details on service software to be finished in March-June.

Gabriel Bodard: this timeframe is about the services being built on top of the triple store, that is, queries via URI and other basics to free users from typing SPARQL queries. These mostly invisible

services is getting into place. Further to search function the development of visualisations and SNA will be done in parallel with further development and will continue beyond June.

Fabian Koerner brought up the dating question, suggesting to have searchable dates converted toa single standard, in response to Laurie Pearce's question of means and scope of date conversion.

Øyvind Eide suggested monitoring use of dates. Simple data models can end up giving messy data sets, whereas more complex data models could lead to a simpler and more elegant result. Even if the model is simple it should be very clear what dates mean, such as inner or outer bounds – it could be dangerous to mix inner and outer bounds in same field.

Gabriel Bodard: we will try to encode any complexity, but the kind of dates are so varied, and it is not necessarily marked in the data sets what dates mean; they do not know what dates they get from the various data providers. They will ask providers to map dates that are meaningful to the persons' life spans. Not clear how to get closer to Øyvind Eide's suggestion.

Øyvind Eide: dates may add more noise than they solve problems. There is a difference between automatic and manual co-referencing. In manual co-referencing, dates extracted to SNAP may not be the ones used in the process anyway.

Gabriel Bodard: SNA would be useful to develop patterns, but he does not expect software to be able to say person A = person B now; maybe in the next decade. Co-reference creation is one of many tools SNA visualisation can use.

Øyvind Eide in reference to the cookbook: consider, especially after the workshop at DH in Lausanne, to give stronger indication of what format the project would want people to submit data in.

Gabriel Bodard: this is not about SNAP, it is about a general publication outside SNAP. This will be clarified in future versions of the cookbook.

Rainer Simon points out that the SNAP annotation format as documented in the cookbook is based on the intial (pre-Pelagios 3) model, which had several issues and has been revised. He suggested a discussion at some point to solve that.

Gabriel Bodard suggested that everybody would join the Ancient People Google Group: https://groups.google.com/forum/#!forum/ancient-people.

6. Targeted feed-back

a) advise on whether the project is meeting its aims

Charlotte Roueché: the project is meeting its aims. It requires restraint to keep from bulging, as we saw in the discussion of dates.

b) suggest data sets which SNAP should include in its outreach

A public list of data sets would be desirable.

Gabriel Bodard: we have a list of sets whose providers have agreed.

Charlotte Roueché: could you set up a location where one can formalise commitments to data sets?

Gabriel Bodard shares the link: http://wiki.digitalclassicist.org/Greco-Roman_Prosopographies

c) how can the SNAP work benefit other research

No comments at this point.

d) ideas about representation of the SNAP project in other for a

Gabriel Bodard: the project will be presented at Digital London. At the DH conference in Lausanne there will be a poster and a workshop. Other possibilities include EAGLE.

Fabian Koerner suggested a presentation of the project at one of the Digital Classicist Seminars in Berlin in the upcoming winter term. The Call for Proposals is out.

Laurie Pearce announced the DH-CASE II Workshop at DocEng2014. Call up at: http://research-it.berkeley.edu/dhcase2014

7. Any other business

Charlotte Roueché praised the effective workshop in London.

The next two meetings will be after the DH conference in Lausanne, that is, mid to late July, and in early October.

The Advisory Board and the Principal Investigator were mutually thanked for a good and effective meeting.