Towards a methodology for the identification, extraction and analysis of place in the Republic of Letters

Minutes

WG1 and WG2 Meeting

Chester, UK 16th and 17th of February 2017

The meeting took place at the Digital Humanities Research Centre at the University of Chester, UK. The purpose of the meeting was threefold. Firstly, we discussed the work that WG1 & WG2 have been carrying out in the context of the project, related to spatial information in datasets relevant to scholars working with the RofL. Secondly, we aimed to define a clear strategy to follow in the next stages in order to accomplish the effective identification, disambiguation, extraction and analysis of place-names, and possibly other spatial information focusing on these historical datasets. This informed not only the general research agenda established within the project, but also our forthcoming funding applications. Finally, we aimed to establish a combined front with WG2, looking to address common methodological issues both groups are encountering, and design joint solutions.

Participants in this meeting

Name	Email address	Role in the Action (if known)/ External speaker	Country of residence
Patricia Murrieta-Flores	p.murrietaflores@chester.ac.uk	Deputy Leader WG1	UK
Ian Gregory	i.gregory@lancaster.ac.uk	Leader WG1	UK
Bruno Martins	bgmartins@gmail.com	WG1 Collaborator	Portugal
Miguel Won	miguelwon@gmail.com	WG1 Collaborator	Portugal
Howard Hotson	howard.hotson@st-annes.ox.ac.uk	Action Chair	UK
Eero Hyvonen	eero.hyvonen@aalto.fi	WG2 Leader	Finland
Ikkala Esko	esko.ikkala@aalto.fi	WG2 Collaborator	Finland
Ruth Ahnert	r.r.ahnert@qmul.ac.uk	WG2 Subgroup Leader	UK
Sebastian Ahnert	sea31@cam.ac.uk	WG2 Subgroup Leader	UK
Valeria Vitale	valeria.vitale@sas.ac.uk	Invited speaker	UK
Leif Isaksen	l.isaksen@lancaster.ac.uk	WG1 collaborator	UK
Arno Bosse	arno.bosse@history.ox.ac.uk	WG5 collaborator	UK

The meeting consisted on two working days. The first comprised short presentations about particular subjects and the second day consisted mainly on guided discussions regarding the challenges we encounter in both groups, possible solutions to these and how to move forward.

Thursday 16th

- Patricia Murrieta-Flores (PMF) gave general welcome, introduced the purpose of the meeting and the different participants.

- Ian Gregory (IG) gave a presentation on the work that WG1 has been carrying out in the context of the RofL, particularly the research done in collaboration with Ruth Whelan.
- Bruno Martins (BM) gave a presentation of the state of the art on NLP and GIR for textual documents.
- Miguel Won (MW) presented the results of the STSM carried out at Chester.
- Valeria Vitale (VV) gave a presentation on the Pelagios Project, introducing the tool 'Recogito' and the work they have been doing with communities.
- Ruth (RA) and Sebastian Ahnert (SA) gave a presentation on their work related to their tool for cleaning datasets and network analysis.
- Eero Hyvonen (EH) and Ikkala Esko (IE) presented the tools they have developed on ontology services and maps.
- At the end of the day PMF lead a general discussion on problems related to the identification, disambiguation, extraction and annotation of named identities.

Friday 17th

Discussion

Multiple problems were identified. Some of them are related to (1) place-name identification; and others are related to (2) geographic disambiguation. Both are related to NER, but also to Gazetteers because NER pulls from lists of place-names for identification and geographic disambiguation relies on them for putting coordinates to them.

- Issues related to all datasets:
    - Proper transliteration or translation
    - Cleaning the datasets//automatization techniques// establishment of pipelines
    - Access to metadata and digitised collections
    - Access to gold standard collections annotated so we can do comparisons
- Issues related to place-name identification include:
    - Spelling variations/ Early vs. Modern English
    - Use of different languages in one source
    - Abbreviations
    - Changes in names
    - Disappearance of place-name
    - Distinction between places that can be identified through coordinates and places that might be more ambiguous. For instance, Great Britain vs. London or an actual address.
- Issues related to geographic disambiguation include:
    - The use of an adequate gazetteer
    - Scope and scale of the gazetteer
    - Disappearance of place-name
- Issues related to standardisation and working together
    - An historical-gazetteer e.g. Early modern
    - Annotation Models? Should we aiming to use one data model?
    - How can we go from the data we are producing in terms of NER and integrate them in Linked DATA?
    - Could we connect EMLOD data annotator with Recogito info?

- Can we take the Finish Ontology Service and adapt it to other regions?

We discussed the solutions we are already implementing:

- Spelling variations: Rui Santos, BM and PMF are working on string and topopnym matching to solve some of the problems of spelling variation and changes over time.
- EH and IE are working on the use of gazetteers in different languages
- IG, VV and Leif Isaksen (LI) are working on the creation of gazetteers
- RA and SA have implemented solutions in terms of cleaning datasets

We also discussed the next steps to take after the work done during the STSM (Hartlib & Hamilton Papers):

- We need to assess the accuracy of place-names identified
- We need to assess the level of geographical disambiguation we have at the moment
- Experiment with other gazetteers and tools

Actions

Article

MW, BM and PMF will finish writing the article related to the STSM called provisionally 'Democratic NER: Evaluating Name Entity Recognition tools for the identification and geocoding of place-names in historical epistolary'. We are planning to send this for publication in the Journal of Geohumanities.

H2020 Proposal

We agreed to move forward and write an H2020 proposal together addressing the issues discussed during the meeting. The drafting of this application will be lead by PMF and will be written during summer 2017. The preliminary idea behind this would be to create a methodological pipeline and associated tools for the use of researchers and the public that enable a reiterative process of annotation, enhancement, extraction, and analysis of named entities in historical corpora using as example datasets from the RofL, particularly the Samuel Hartlib collection.