Global Philology - Digital Infrastructure for Named Entities Data

Leipzig, January 11-13, 2017

Final Report

Introduction

On January 11-13, 2017, the Humboldt Chair for Digital Humanities in Leipzig hosted a series of talks, focusing on current issues in spatial and social research applied to historical languages. The workshop was supported by the two Special Interest Groups in Greek and Roman antiquity of Pelagios Commons, and by the Global Philology Project of Leipzig.
In order to address our main concerns, we used the notion of "Named Entities"[1] as a broader concept, not just in the meaning of "proper names" but as an expression of cultural/cognitive patterns representing information on space, time and people in premodern sources. Named Entities offer a gateway towards a completely new concept of "edition". Named Entities are the primary vehicle for additional information, which allows a more comprehensive and complex understanding of the world: social information, community networks, spatial cognition, imaginary geographies, history of language and culture, political and economical turnovers. All these fields involve Named Entities as lead actors, and require highly developed and refined methods to be fully explored.
The aim of the workshop was not, therefore, to display showcases, but to address current issues of representation and classification typically connected to names in Humanities data. We especially focused on historical and premodern languages, where a limited set of resources is currently available in digital research, compared to modern studies.
The strong premise of the discussion was to go beyond the traditional Western-centred notion of Classical languages, by including premodern societies as a whole. This has been the strongpoint of the Global Philology Project so far: every civilization has a cultural heritage which is considered authoritative for the community, and is therefore considered "Classical". We, therefore, considered as Classical languages not only Greek and Latin, but also Chinese, Arabic, Farsi, Hebrew, Egyptian, Coptic, Syriac, and so on.
This amplification in scope is the consequence of our firm belief that scholarship should aim at a "global" understanding of premodern civilizations, as an interlinked world which must be investigated as broadly and deeply as possible through the exploration of primary sources. Hence, the idea that there should be a shared set of tools and standards across language domains, *i.e.* a "shared infrastructure" of the Premodern world. Pelagios provides a successful instance of generalized and decentralized infrastructure for Linked Open Geodata across disciplines and language domains, by focusing on the common concept of "place" and supporting the community with shared services: it seemed, therefore, the ideal partner for such an effort.

Report

The discussion started, in fact, from issues of definition. We faced an immense variety of concepts in terms of "place", "name", "event", "period", let alone their multiple connections with both concrete reality and imaginary models (Carlson, *Named concepts between reality and imagination*; Franz, *Obsessed with names?*; Razanajao, *Egyptian places and place names in a digital world*; DePauw, *Trismegistos and the complexities of Named Entities of the Ancient World*; Rusinek, *Kima: places in a language*; Jovanovic and Simrell, *Digital commenting on place names in early modern Latin texts*; Bucciantini, *FGrHist V: Editorial and conceptual problems of a geographical project*).

[1] D. Nadeau & S. Sekine, *A survey of named entity recognition and classification*, "Linguisticae Investigationes" 30:1, 2007.

Such differentiated and sometimes difficult to grasp concepts are the first problem affecting modern attempts of classification of names from primary sources: the essential task of disambiguation is still substantially limited by the surprisingly scarce amount of authority lists, such as lexica, gazetteers and prosopographies. There is an entire world of authoritative scholarship, which could be used at least to provide an initial framework to the building of authority lists, but this has been done only in minimal part. This scarcity of comprehensive resources is close to nothing in some language domains, such as Hebrew (Rusinek), but also totally insufficient in the bigger fields of Greek, Latin and Arabic. Current and successful methods of classification and authority lists[2] are fundamental, but need expansion, not only to other sources, but especially to the needs of other languages and scripts (DePauw; Vitale, *Named Entities for cross cultural places: languages, boundaries, identities*).

Such classifications should also aim at including more semantically complex issues for a comprehensive understanding of the information conveyed by Named Entities (Schäfer, *Local gazetteers and named entities recognition*): for instance, the analysis of non-mapped spatial narratives (Podossinov, *Sprachliche Repräsentation des geographischen Raums in der Antike*; Brillante, *Reading a Greek Periplous*); or the interaction with additional data, such as biographies and time periods (Seydi, *Geospatial analysis of premodern Arabic sources*; Horne, *People, Places, Time: representing Entities in the Big Ancient Mediterranean Project*).

Current efforts of semantically refined classification follow different theoretical premises by necessity: some of them are derived from ontologies (Lana, *The narrow and the wide gate*; Müller, *Prosopography and its problems in the Digital Edition of the Inscriptions of Metropolis in Ionia*; Schäfer, *Local Gazetteers and named entities recognition*); others from cognitive frameworks (Goerz and Thiering, *Spatial cognition in historical geographic texts*); others were directly derived from the study of the sources, with no previous semantic scheme to comply with (Jovanovic and Simrell).

Tools for Named Entity Recognition showed impressive variety and high standard, often combined with complex semantic annotation environments: Graeco-Roman studies and Classical Chinese figured prominently in this field, with tools such as Recogito 2 and MARKUS[3] (Barker, *Investigating place*; De Weerdt, *Named entity recognition for Classical Chinese*; Schäfer, *Local gazetteers and named entities recognition*). The integration of such tools with existing NLP- and corpus linguistics-based toolkits, such as the CLTK[4], is also auspicable for the future. Semitic languages still suffer especially problematic scripts with different kinds of transliteration (Franz; Vitale), and the lack of expansive OCR services, including an OCR correction platform[5] on the model of Transkribus[6] or the Leipzig/Mount Allison OCR pipeline adaptation[7]. Such different ways of dealing with NER are also problematic in the perspective of a LOD-friendly framework: a uniform system of canonical citation is essential for this kind of harmonization, by means of a standard in the creation and use of URIs (Jovanovic and Simrell; DePauw).

Named Entities also challenges our way of visualizing edited texts. There is no doubt that we are facing the birth of an epochal transformation in the field of publishing, where the digital medium offers the ideal opportunity to investigate, explore and visualize the inner world of space, time and society as they emerge from the sources.

It was especially interesting to see how efficient and appealing environments are being developed from the exploration of Named Entities. Named Entities were used to display a global view of the

---

[2] Famous examples of gazetteers and canonical references are Trismegistos (http://www.trismegistos.org/), SNAP (https://snapdrgn.net/), Pleiades (https://pleiades.stoa.org/), PeriodO (http://perio.do/), Chronontology (http://chronontology.dainst.org/).

[3] http://recogito.pelagios.org/; http://dh.chinese-empires.eu/beta/.

[4] http://docs.cltk.org/en/latest/about.html.

[5] This is hopefully going to change soon for Arabic: M. Romanov, M. Thomas Miller, S. B. Savant and B. Kiessling, *Important new developments in Arabographic Optical Character Recognition (OCR)*, https://www.academia.edu/28923960/Important_New_Developments_in_Arabographic_Optical_Character_Recognition_OCR_.

[6] https://transkribus.eu/Transkribus/.

[7] http://heml.mta.ca/lace/.

scope of a text across several information panes, such as lexical references, maps, images, in a multilingual environment (Mambrini, *Persons and places in the iDAI publications*), but were also used to conceptually expand the research framework of classification by means of the concept of network (Broux, *TM networks: visualizing relations in Trismegistos*; De Weerdt), therefore showing the potential in the adaptation of existing standards and tools for the visualization of named entity information[8]. The use of such editions as embedded tools of a wider "user experience" was emphasized by initiatives such as ToposText[9], which has also highlighted the importance of side development of mobile apps. Mobile applications and Citizen Science platforms, combined with appropriate displays of the results in an appealing visualization environment, are important to ensure the involvement of a more general public of non specialized users, who are, however, the ultimate recipients of scholarly research, and are the essential part of any collaborative effort (Kiesling, *ToposText: toward an ecosystem of free-range Big Data in the Classics*).

One of the crucial problems emerging from the discussion, and especially relevant for the understanding of Named Entities, was the language barrier. This issue is relevant from multiple points of view: most of the Digital Humanities resources are notoriously in English, with poor resources in other languages[10], hence the need for collective translation efforts to bridge the gap between different communities, and also different traditions of scholarship (Bucciantini). But in the case of Named Entities, the interaction of languages across primary sources is crucial as well: historical changes and the multiple turnovers of empires throughout Antiquity and Late Antiquity have brought up immense linguistic changes, which have primarily affected names of people, places and periods. It is definitely not rare to have multiple instances of a place-name in several languages (Rusinek; Vitale).

Consequently, the need of multilingual text editors is crucial. Previous efforts in translation alignment have been attempted in environments of manual annotation of ancient texts, such as Alpheios[11], which provides now a model for alignment interfaces but needs to be adapted to the needs of different scripts and forms of text[12]. Cross-language automatic alignment in the shape of an interlingual translator is also an important effort in the field of Natural Language Processing, and represents the ideal goal of this research field[13].

Such multi-cultural integration should also be evaluated in mapping. The spatial information conveyed by Named Entities is obviously also connected to highly visual ways of representation. Recent developments in this field show that it is possible to develop a model for the integration of existing maps and lists of place-names coming from different sets of natural and human geography, different time span, different linguistic contexts, which often imply also different mapping systems (Åhlfeldt, *The Digital Atlas of the Roman Empire*; Franz). The integration with existing services of geospatial visualization, such as GoogleEarth, MapBox, Carto and QGIS, is also important in the perspective of collaborative projects and of the involvement of non specialized users.

Conclusions

[8] One example being Palladio, which allows the visualization of different types of named entity-related information (http://hdlab.stanford.edu/palladio/); a more generalized and text oriented tool is Voyant (https://voyant-tools.org/).

[9] http://topostext.org/.

[10] Relevant exceptions being, for example, the series of tutorials on QGIS (http://docs.qgis.org/2.14/en/docs/user_manual/).

[11] http://alpheios.net/.

[12] Poetry and text-bearing objects have, for example, a specific focus on the layout of the text, which cannot be properly visualized on Alpheios; multi-lingual alignment is also very difficult to perform, and needs a more refined display. Cf., by contrast, the visualization of the Persian *Diwan Hafez* aligned with two different translations (http://dynamiclexicon.com/hafez/).

[13] In this case, the ideal model is obviously provided by Google Translator, which is, in fact, based on an immense amount of collective data input: https://translate.google.com/.

When dealing with “Named Entities Data” with regard to historical languages, one of the most immediate problems is the challenge of dealing with a different mindset in the conceptualization of the world, which is affected both by a completely different mental model and by the scarcity of comprehensive evidence. Such discrepancy in the cognitive structures is also the most visible limit to generalized approaches of classification: we should probably accept that some level of “variety and complexity of the world” cannot be put into the order we globally envisage, because absolute ordering happens at the price of leaving something outside the classification. In fact, this is the nature of Humanities data.
This, of course, should not stop global efforts of systematization, because they are essential to the progress of research, especially in the digital field, where the standards of Linked Open Data seems nowadays to open to a completely new world of opportunities and perspectives: we should rather always be aware of this theoretical restraint also in view of its potential richness, as a measure of control and relativity to our research methods.

References

Programme:
http://www.dh.uni-leipzig.de/wo/events/global-philology-digital-infrastructure-for-named-entities-data/

Global Philology Project proposal:
https://docs.google.com/document/d/1La3PFvEMCJAAIjZ-L85kr2eNH_7CMHPcJxbBXEe8rg8/edit?usp=sharing

Conference announcement on Pelagios Commons:
http://commons.pelagios.org/groups/ancient-greek-sig/forum/topic/conference-announcement-digital-infrastructure-for-named-entities-data-2/

Pelagios Commons: http://commons.pelagios.org/

Contact: chiara.palladino@dh.uni-leipzig.de