Social Semantic Annotation with Recogito 2

Leif Isaksen
l.isaksen@lancaster.ac.uk
Lancaster University, United Kingdom

Rainer Simon
rainer.simon@ait.ac.at
Austrian Institute of Technology, Austria

Elton T. E. Barker
elton.barker@open.ac.uk
The Open University, United Kingdom

Pelagios Commons (Pelagios Commons 2017) is a community of practice dedicated to supporting Linked Open Data-related activity within the humanities. Comprising a series of interconnected Special Interest Groups, it operates through the collective establishment of digital conventions for semantic data and produces software to facilitate its production and use. Such data can ultimately be used to interconnect independently maintained and heterogeneous online resources about the past. While Pelagios Commons is engaged in a variety of activities to achieve these ends, this paper focuses on the development of a second generation implementation of its most popular tool, Recogito (Simon et al., 2015). Recogito is a semantic annotation system (Andrews et al., 2012) with similarities in some regards to platforms such as Pundit (Grassi et al., 2012) and Hypothesis (Hypothesis, 2017), but with particular focus on geographic content, and accessibility to humanists without high levels of technical literacy.

Recogito's origins lie in the earlier *Pelagios 3* project cycle dedicated to the semantic annotation of early geographic documents so as to provide a 'critical mass' of content to which other historical materials could be related. The intended goal was to identify place references, and their referents, in as many pre-1500 geographic documents as possible. Such documents, or more specifically their digital surrogates, could take a wide variety of different formats, but are typically represented as texts, images and tables. While the first phases of Pelagios saw annotation carried out by a community of resource curators, much - though not all - of the annotation for Pelagios 3 was conducted within the Investigative Team. As most of the documents to be annotated had no formal structure, it was necessary to develop a tool which allowed us to rapidly produce annotation records (recording a reference, a global gazetteer entry and an annotator). The result was *Recogito*, a web-based platform which allowed its users to upload digital surrogates and produce annotations pointing back to the original digital version of an online document.

The first version of Recogito offered multiple interfaces for the discrete tasks of: identifying place references (whether in text, image or table); transcribing them where necessary; relating them to an entry in one of multiple possible gazetteers (depending on the period of the document in question). Semi-automated processes based on Natural Language Processing and Named Entity Extraction techniques allowed the software to accelerate these tasks, while ultimately requiring all annotations to be verified by a human individual. Semantic interpretation is always under-determined by a text or symbol, and thus human intervention remains a fundamental principle of the Pelagios methodology. While document surrogates were kept behind a log-in interface to prevent the possibility of copyright infringement, the annotations themselves were made publicly available in real time under a CC0 (public domain license).

Pelagios 3 and Recogito successfully met their aims and have drawn interest from potential stakeholders working throughout the humanities. These range not only across different periods and geographic regions, but between disciplinary fields (such as archaeology, classics and history), in diverse forms of text, and for the production and alignment of gazetteers to boot. Thanks to an Open Knowledge Foundation/DM2E Open Humanities Award, Recogito was also tested with several undergraduate student classes. This not only allowed us to refine its user interface, but also to see the various ways it changed students' approaches to, and understanding of, the material they were annotating. The very process of annotation, as many digital humanists can attest, forces a level of systematic reflection upon the annotator which might otherwise be elided.

Despite - indeed because of - Recogito's success, a number of limitations came to light during the course of Pelagios 3. First and foremost was its centralised architecture which provided a single workspace for all users. While well suited to a small team in regular

communication with one another, the increasing number of users and documents meant that managing document metadata, preventing errors caused by concurrent edits, and keeping users abreast of changes to documents of interest became increasingly difficult. Furthermore, administrative tasks like uploading documents could only be carried out centrally, creating unnecessary barriers to use. Above all, while document copyright was protected through a restriction on public access, it was clear that if the system user base continued to grow over time, then the risk of copyright abuse would grow with it.

In addition to this central issue were a considerable number of feature requests which were not in the project's original scope of works and which we were unable to introduce within the time available. These include support for various input and output formats (including TEI XML, KML, and GeoJSON); the ability to add non-semantically defined commentary; overlapping annotations; simple points for identifying symbols on images (rather than textboxes for toponyms); competing interpretations of place references; and the ability to make annotated documents publicly available where copyright permits. With renewed financial support from the Andrew W. Mellon Foundation, the Pelagios initiative has been able to redesign and implement Recogito from the ground up in order to address these deficiencies and introduce additional features as well.

Recogito 2 presents an entirely new interface for the semantic annotation of place references. Users can self-register and are now provided with their own workspace (with an initial storage allowance of 200MB) in which they can upload and annotate documents. Each user's cataloguing page has its own URL and is publicly visible, although any uploaded documents are not. Documents can be uploaded singly or as a batch of related files (such as images of pages within a manuscript). Documents are automatically pre-parsed for possible place references at the upload stage unless the user declines to do so. Currently, Recogito 2 makes use of the Stanford NLP Toolkit (Manning et al., 2014), but we intend to make it extensible so as to support alternative parsing engines, such as the Classical Language Toolkit (Johnson et al. 2014-17) which may be better suited to specific languages or use cases. Whereas Recogito 1 only supported plain text documents and image files, Recogito 2 also supports TEI XML and images held in IIIF-compliant repositories, as well as JPEG, TIFF and PNG.

Once uploaded, an 'annotation view', allows the user to confirm automatically identified place references or create new ones (Figure 1). This takes place by means of a pop-up dialog box which determines the type of annotation (currently only places references, with free-text commentary and tags, but future development will include additional support for person references and events). Assuming the reference is to a place, the system proposes a probable gazetteer candidate which can either be confirmed or corrected as appropriate. Where the same place definition has been aligned across multiple gazetteers, these will be merged into a single entity for consideration, but the user is able to select which gazetteer they wish to formally associate the reference to.

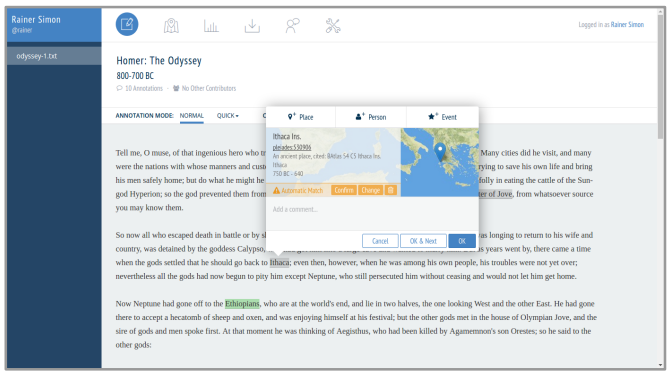

Fig. 1. Recogito 2 text annotation view.

A major development in Recogito 2 is the introduction of 'social tools' for collaborative annotation. Annotators can share their documents with other registered users, allowing them to view, edit and create annotations dependent on the permission settings. Edits effectively act as discussion threads so that the full history of changes and commentary can be seen for each annotation (and where necessary, rolled back by the owner). We are also aware that for many users, producing Linked Open Data is not their primary objective. There are many other benefits to be derived from annotating place references, not least of which is the ability to map content. Recogito 2 offers a simple mapping interface that shows the distribution of place references - where coordinates can be derived from a gazetteer - against a range of possible backing maps. Symbol size reflects the number of references to the place, and selecting one provides the user with the specific references within the text itself (Figure 2).

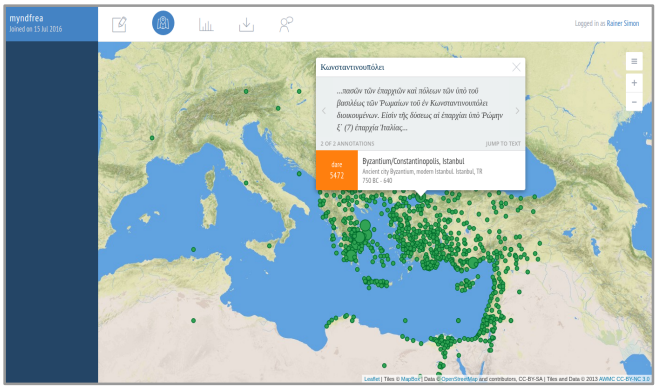

Fig. 2. Recogito 2 map view.

A series of additional features replicate popular functions within Recogito 1. This includes a statistics dashboard, which provides data about both documents (such as place reference frequency, or the proportion of references lacking identification) and places (such as the toponyms by which they are referred to, and tags associated with them). It is our intention to add 'social statistics' that may allow people to see which of their documents are proving most popular or perhaps to help identify other users with similar interests. We have also extended the annotation download formats from solely CSV and RDF, to include KML, TEI and GeoJSON. This allows for their incorporation within a much wider range of software commonly used by humanists.

Already in public Beta-testing, Recogito 2 offers a next generation approach to semantic annotation of humanities documents with specific emphasis on place references. Nevertheless, we believe strongly that the value of this contribution lies not simply in its technical innovation but in facilitating community contributions to the Web of Linked Open Data.

Bibliography

Andrews, P., Zaihrayeu, I., Pane, J. (2012) "A Classification of Semantic Annotation Systems." In *Semantic Web*, 3(3): 223-248.

Grassi, M., Morbidoni, C., Nucci, M., Fonda, Ledda, G. (2012) "Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries."In *Proceedings of the 2nd International Workshop on Semantic Digital Archives* (SDA 2012).

Hypothesis. 2017. http://hypothes.is/

Johnson, K. P. et al. (2014-2017). CLTK: The Classical Language Toolkit. DOI 10.5281/zenodo.60021

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky, D. (2014) "The Stanford CoreNLP Natural Language Processing Toolkit." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Pelagios Commons. 2017. http://commons.pelagios.org

Simon, R., Barker, E., Isaksen, L., and de Soto Cañamares, P. (2015) "Linking Early Geospatial Documents, One Place at a Time: Annotation of Geographic Documents with Recogito." In *e-Perimetron.* 10(2): 49-59. ISSN.