GAP: A NeoGeo Approach to Classical Resources

Leif Isaksen, Archaeology, University of Southampton, SO17 1BJ, UK

Elton Barker, Classical Studies, The Open University, MK7 6AA, UK

Eric C. Kansa, iSchool, UC Berkeley, Berkeley, CA, 94720-4600, USA

Kate Byrne, Informatics, University of Edinburgh, EH8 9AB, UK

E-mail: <l.isaksen@soton.ac.uk>.

Submitted: <leave for Editor to date>

Abstract

Google Ancient Places (GAP) is a Google Digital Humanities Award recipient that will mine the Google Books corpus for classical material that has a strong geographic and historical basis. GAP will allow scholars, students, and enthusiasts world-wide to query the Google Books corpus to ask for books related to a geographic location or to ask for the locations referred to in a classical text. We will overcome the traditional difficulty of identifying place names by using a combination of URI-based gazetteers and an identification algorithm that associates the linear clustering of *places* within narrative texts with the geographic clustering of *locations* in the real world.

Background

The GAP project has its foundations in two projects: The Herodotus Encoded Space-Time Imaging Archive (HESTIA) [1] and Open Context [2]. HESTIA was a two-year collaboration (2008-2010) between The Open University and the Universities of Oxford and Birmingham, funded by the UK Arts and Humanities Research Council. Its aim was to explore new methods for visualizing geographical concepts and their relationships to each other in Herodotus' *Histories*. HESTIA applied multiple approaches, including: (1) mapping the frequency of references to specific locations (in both spatial and narrative terms); and, (2) manually and automatically generating maps of the network connections between places.

. A particularly powerful approach developed by the project was a Narrative Timeline that enables readers to see the locations appear and then fade away as they move through the pages of the text (Fig. 1). The project made use of Greek and English versions of the text from the Perseus Digital Library [3] which are marked up with the Text Encoding Initiative (TEI) XML schema, including geographical locations based on automated string-matching with the Perseus internal gazetteer and the Getty Thesaurus of Geographic Names (TGN). Closer analysis revealed that many of the locations were misidentifications however, and a relatively labor-intensive process was required to correct them. The utility of visualizing locations within a narrative was demonstrated but a question remained: could the approach be automated so as to scale beyond manually processing individual texts?

Open Context is an Open Repository of archaeological excavation data developed and maintained by the Alexandria Archive and UC Berkeley. As an archaeological archive with wide-ranging collections and multiple contributors one of its key functions is to provide users with information about similar sites. To this end, Open Context provides a map and timeline on its splash page to enable both providers and users to quickly identify related research. The ability to identify relevant works outside the repository would be a ground-breaking extension to this service.

Although based in the separate disciplines of Classics and Archaeology, these two projects face a common problem. Much of the relevant literature is rare or out of print, limiting access to those with both the legal and geographical access to major libraries such as the British Library or Library of Congress. References to ancient places are often brief or fragmentary. This makes inter-library loans a slow and inefficient method for research sharing, especially problematic in today's climate of increasing financial and time pressures in the Humanities.

Thus, access to even short extracts is of great value and Digital Libraries will play an increasingly vital role in improving efficiency for researchers.

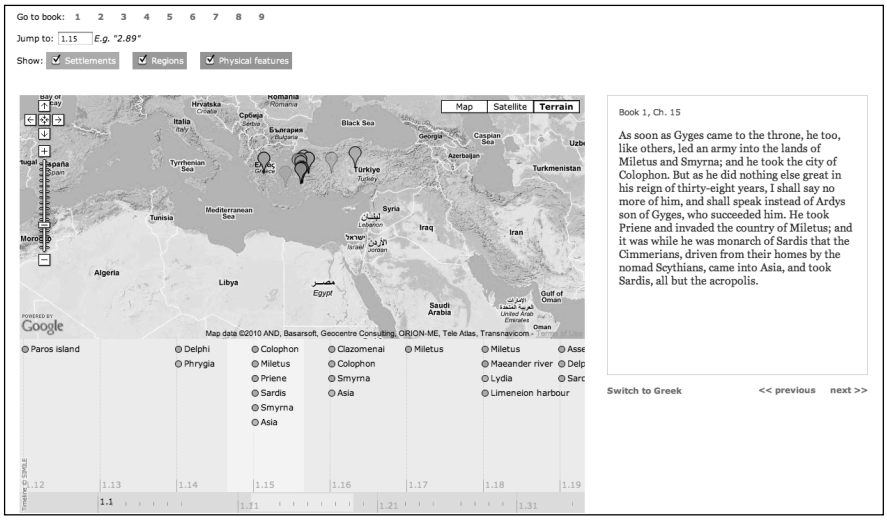
Current search services tend to be either metadata or content-based but both present difficulties for entity-based search. Metadata searches typically require manual input of the metadata itself, thereby limiting it to no more than a handful of keywords when cataloguing large numbers of books. In contrast, full text search provides access to the very words we are looking for but the problems of toponymic homonyms and synonyms (different places that share names, and single places with multiple names) lead to high numbers of false positives and false negatives.

Conceived in 2002, the Google Books project is a major initiative by Google to scan, OCR, and disseminate online all the world's books [4]. In 2010 they announced a Digital Humanities Award program that would enable researchers to "apply quantitative research techniques for answering questions that require examining thousands or millions of books." Investigators on both Open Context and HESTIA saw this as an ideal opportunity to realize their common dream of tying classical resources together by referencing places and submitted a successful joint proposal. The project commenced in October 2010 and is now underway.

Methodology

The central principle behind the GAP project, identified while developing the HESTIA Narrative Timeline, is that places referenced in narrative texts generally cluster together to maintain narrative coherency. In other words, given a set of toponyms with multiple possible identifications, the set of identifications with the shortest overall path between them is likely to be correct. We can further weight the influence of

Fig. 1. The HESTIA Narrative Timeline ***(© Nicholas Rabinowitz and HESTIA)*** **<http://www.open.ac.uk/Arts/hestia/herodotus/basic.html>**



each toponym on our decision by the number of possible locations it could refer to. Somewhat counter-intuitively, this means that small, obscure places with unusual names are much better guides to location than well-known places with many namesakes.

Naturally, however, the story is complicated by a number of additional factors: 1. The approach does not work well for fragments or with arbitrary higher-level structures such as the alphabetic organization of an encyclopedia. 2. The author may assume that the anticipated audience will be able to contextualize by other narrative elements (such as well-known individuals) and thus mention only a single location (or even none at all). 3. The author may contextualize by giving a territory in which the place is located. These can confuse point-based algorithms as there is no single 'best' point that represents them. 4. The author may have confused the place they are discussing with another, especially if they are commenting on another work or reporting independent sources. 5. Occasionally the location clustering assumption simply does not hold. This is especially the case for places that do not perform an active function in the text such as personal names derived from places of origin (e.g. Herodotus of Halicarnassus).

Fortunately a number of additional features that Digital Libraries make possible can assist us in improving both precision and recall. The most important of these is a new generation of Semantic Gazetteers such as GeoNames and Pleiades. These provide a unique HTTP URI for each place to which multiple names (toponyms), locations (such as spatial coordinates) and categories (like 'settlement') can be assigned. These gazetteers make it much easier to handle the problem of synonymy. They also mean that once an identification is made it can permanently fixed with a non-ambiguous identifier.

Because the algorithm produces probability estimates rather than binary results we can easily identify 'hard cases': those in which there is either insufficient or conflicting evidence. These can then be handled by more sophisticated but computationally expensive procedures. First, there are multiple levels at which we can look for clustering, including the chapter, book, and corpus (of the author or even genre). Looking at higher levels may provide us with broader contextual clues. A further advantage of working with massive digital corpora is that they frequently provide multiple translations and editions. In such cases we can use the linear chain of places in one edition to inform the processing of another and vice versa. Finally, as we process more books the system itself can record additional metadata about the places as well as the books. In particular it may detect that in cases of homonymy, one location is much more frequently mentioned than all the others (such as the Egyptian Alexandria, as opposed to the many other cities of that name). This can help in cases where we have no other contextual clues to draw on.

It is also important to remember that there are some hard limits imposed on the process and some pragmatic aspects to our goals. First, we are only able to identify those places for which we have an entry in a gazetteer. No amount of Natural Language Processing will be able to find those places which were previously unknown to us. Secondly, we are not looking for a 'perfect' set of results for the simple reason that natural language is ultimately indeterminate. The best we can hope for is to get as close as possible to the precision and recall rate of a human with a reasonable set of reference books. In this way our results will provide probabilities to assist the work of Ancient World researchers.

We are by no means the first to experiment with such approaches (see, for example, [5]) but we are not aware of similar methods being applied to ancient geography or massive corpora. The results of this processing will be RDF annotations for each text that provide an ordered directory of places. Such annotations are extremely useful to search engines but less helpful for humanities researchers who require a human interface. To help them we will provide Web mapping tools, like those on the HESTIA and Open context websites, that enable searches in both directions – from text to places, and from a place to the texts which reference it. To lower adoption barriers, we will use RESTful Web services, enabling other developers to incorporate our results into the next generation of Humanities Virtual Research Environments.

Current and Future Work

GAP is still in its infancy but moving apace. We are now working in conjunction with Pleiades [6] to associate the local identifiers used by the HESTIA project with Pleiades URIs. This will provide a benchmark against which the precision and recall of the automated system can be measured. Our next step will be to test this system on the raw text of the *Histories* used by HESTIA. Once a satisfactory retrieval rate has been achieved (we aim for 80% recall and 90% precision) we will test and develop the algorithm further on an 1828 translation provided by Google [10], using the previous work to inform our results. Finally we will apply the algorithm for general use on the Google Books corpus, focusing on Out of Copyright texts with Library of Congress Headings DE-DG (Greco-Roman World; Greece; Italy).

GAP is intentionally focused on the restricted domain of the Ancient World, but we do not see this approach as in any way limited to dealing with historic places. We are currently engaging with a number of other initiatives worldwide to establish generic processes for referencing places in digital documents (including tables, maps and images as well as texts) and would be delighted to hear from others working in this area.

Acknowledgements

We are grateful to Jon Orwant, Leslie Yeh Johnson, and the Google Digital Humanities Award scheme for making this research possible.

References and Notes

* This paper was presented as an invited talk at the High Throughput Humanities satellite symposium at ECCS2010.
See <http://hth.eccs2010.eu>

1. Elton Barker, Stefan Bouzarovski, Christopher Pelling and Leif Isaksen, "Mapping an Ancient Historian in a Digital Age: the Herodotus Encoded Space-Text-Image Archive (HESTIA)," *Leeds International Classical Journal* **9** (2010) pp. 1-24.

2. Eric Kansa and Sarah Whitcher Kansa ,"Open Context: Collaborative Data Publication to Bridge Field Research and Museum Collections," in Jennifer Trant and David Bearman (eds). *International Cultural Heritage Informatics Meeting (ICHIM07): Proceedings* (Toronto: Archives & Museum Informatics, 2007) <http://www.archimuse.com/ichim07/papers/kansa/kansa.html>.

3. Perseus (2010), <http://www.perseus.tufts.edu/>, accessed 31 October 2010.

4. Google (2010), <http://books.google.com/intl/en/googlebooks/history.html>, accessed 31 October 2010.

5. Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn and Julian Ball, "Use of the Edinburgh Geoparser for Georeferencing Digitized Historical Collections" *Phil. Trans. R. Soc. A* **368** (2010) <http://rsta.royalsocietypublishing.org/content/368/1925/3875.full>

6. Pleiades (2010) <http://pleaides.stoa.org/>, accessed 31 October 2010.

7. Herodotus, *The Histories,* W Beloe trans. (New York: Berresford, 1828)