Gregory, Ian. "Recogito: Semantic Annotation without the Pointy Brackets."

Citation details

Early Modern Digital Review, vol. 4, no. 3, 2021, https://doi.org/10.33137/rr.v44i3.38005.

Renaissance and Reformation / Renaissance et Réforme, vol. 44, no. 3, 2021, pp. 243-249, https://doi.org/10.33137/rr.v44i3.38005.

Peer review

This is a peer-reviewed article in *Early Modern Digital Review*, distributed in print by *Renaissance and Reformation / Renaissance et Réforme*.

Copyright

Early Modern Digital Review materials are published under a CreativeCommons 4.0 license (CC BY 4.0) that permits the right to share (copy and redistribute the material in any medium or format) and adapt (remix, transform, and build upon the material for any purpose, even commercially) the material, provided that the author and source are credited. The full description of CC BY 4.0 can be found on creativecommons.org/licenses/by/4.0/.

Early Modern Digital Review

Early Modern Digital Review is an online, open-access, and refereed journal publishing high-quality reviews of digital projects related to early modern society and culture. It is committed to productive evaluation of both established digital resources and recent tools and projects. Its publications are distributed online by the journal and its partners, and in print by *Renaissance and Reformation / Renaissance et Réforme.*

Barker, Elton, Leif Isaksen, Rebecca Kahn, Rainer Simon, Valeria Vitale, and the Pelagios Network, project creators.
Recogito: Semantic Annotation without the Pointy Brackets. Other.
Vienna: AIT Austrian Institute of Technology GmbH, 2016. Accessed 15 June 2021.
recogito.pelagios.org.

Annotating digital sources to identify features such as named entities (for example, place names, people, and events) is one of the more time-consuming and technically challenging parts of many digital humanities projects. Researching what the annotations reveal presents further challenges. In the spatial humanities, an additional layer of complexity is added by the need to convert the source into a "mappable" format, a process known as "georeferencing," and then explore it spatially. Recogito provides a free, easy-to-use tool to achieve all of these tasks in an environment designed for the needs of humanities scholars. Recogito also allows for collaborative working such that a group of people can access and work on your documents if wanted with appropriate licenses being added.

Annotation is usually done by adding tags to a textual source in formats such as eXtensible Markup Language (XML) or Text Encoding Initiative (TEI) to provide information that might include what type of feature the word refers to, a standardized version of its spelling, the provenance of any additional information, and so on. Georeferencing involves giving features a coordinate-based location, often using latitude and longitude or a map projection. The features that are georeferenced are typically either a point that might represent anything from a building to a city depending on scale; a line that represents, for example, a river or a road; or a polygon that represents zones including administrative units such as parishes or countries, or physical features such as lakes. Georeferencing using geographical information systems requires a significant degree of technical knowledge. It frequently involves the use of a place-name gazetteer, effectively a database table that provides a set of coordinates for every place name it contains along with some additional information.[1] GeoNames (geonames.org) is one of the most widely used

1. Merrick Lex Berman, Ruth Mostern, and Humphrey Southall, eds. *Placing Names: Enriching and Integrating Gazetteers* (Bloomington: Indiana University Press, 2016), doi.org/10.2307/j.ctt2005zq7.

gazetteers but mainly consists of modern place names. Georeferencing humanities sources, frequently texts, presents an additional complexity as place names need to be first identified and then georeferenced, a process known as "geoparsing."[2] Automated geoparsing tools such as the Edinburgh Geoparser (see ltg.ed.ac.uk/software/geoparser) are available, but these also require technical knowledge and can be inaccurate where there are, for example, non-standard spellings or spelling variations as commonly occurs in early modern documents.

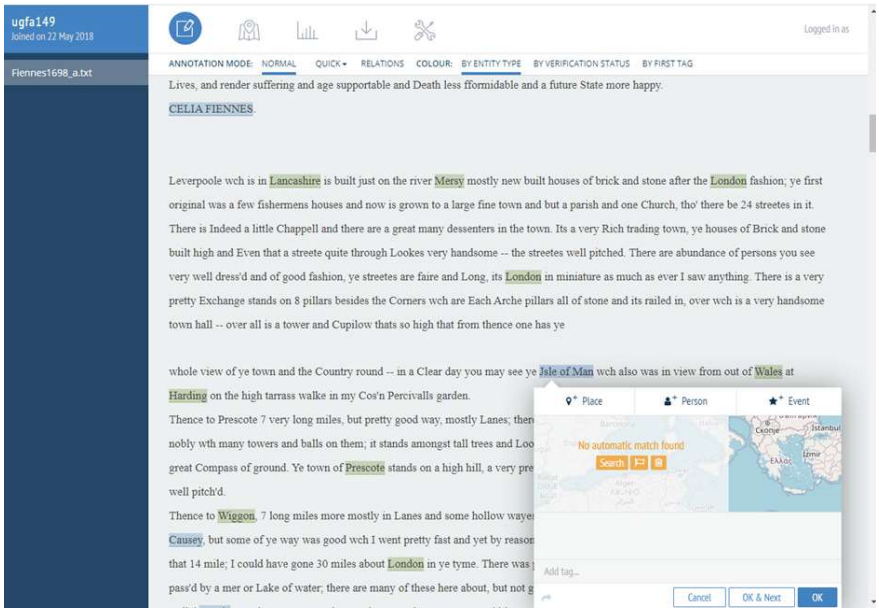

Figure 1. Marking up a place. The researcher has identified "Jsle of Man" as a place, but it cannot be automatically matched to a gazetteer.

Recogito provides a tool that enables humanities researchers to georeference textual and image-based sources (such as scanned maps) in an easy-to-use manner, while also retaining close control of the process which

2. Ian Gregory, Christopher Donaldson, Patricia Murrieta-Flores, and Paul Rayson, "Geoparsing, GIS and Textual Analysis: Current Developments in Spatial Humanities Research," *International Journal of Humanities and Arts Computing* 9.1 (March 2015): 1–14, dx.doi.org/10.3366/ijhac.2015.0135.

is primarily manual or semi-automated. In its simplest form, Recogito allows users to upload a document such as a text file or XML document. The user then works through the document, double clicking or highlighting words to start a dialogue to annotate them as a place, person, or event. Where a place is identified, Recogito will attempt to match the place name with gazetteer records from a range of publicly accessible gazetteers including GeoNames, the Map of Early Modern London (reviewed by David Wrisley in this issue), and the HGIS de las Indias which covers eighteenth-century Spanish America. Where more than one possible match is available, the user can select between them. Where no direct match is found, a common problem with early modern documents due to variant spellings, the user can search the gazetteers for the most appropriate match. Once a word has been tagged, the tag can be applied to all other occurrences of the word in the document.

Recogito also allows features to be annotated on image files. As with text annotation, this is a relatively simple process where files can be uploaded and the appropriate places identified as points or boxes (which can be tilted to represent annotation on the map); these can then be identified as places (potentially linked to a gazetteer), people, or events. These annotations can also be exported to a similar range of formats.

As an example, in Figure 1, a transcription of part of Celia Fiennes's *Through England on a Side Saddle in the Time of William and Mary* (1888) has been uploaded to Recogito. The words "Jsle of Man" have been highlighted as a place name, and the dialogue that allows users to annotate place, person, and events has been started. Perhaps due to the digitizing error ("Jsle" for "Isle"), no automatic gazetteer match has been found. The user then starts the search function as shown in Figure 2. "Man" has been searched for as shown in the top-right corner, returning fifty possible matches. The user has then selected the option that they feel is correct, in this case "Isle of Man, Isle of Man" taken from GeoNames, and this is shown on the map to confirm its location. This record will be updated; optionally, all other matches will be as well.

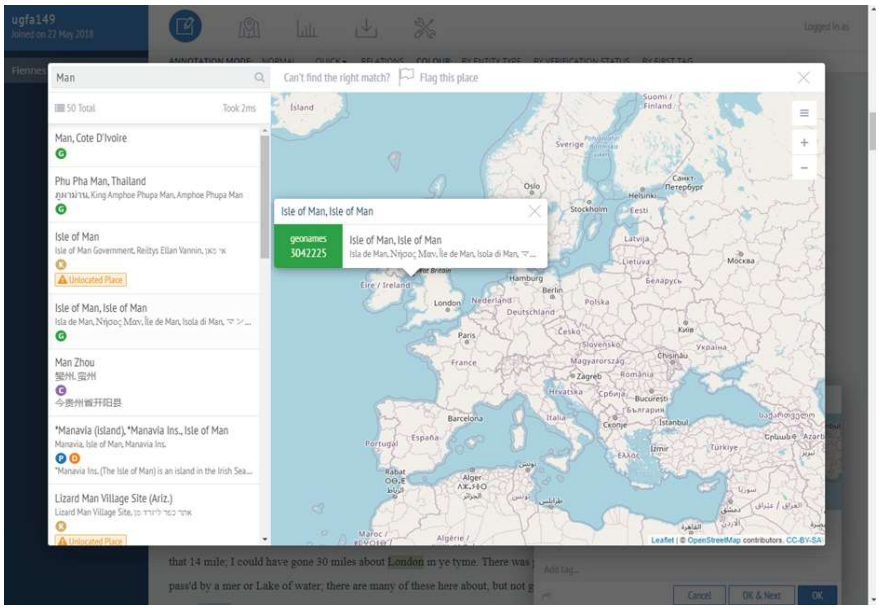

Figure 2. A search for "Man" provides a range of options. The researcher has selected "Isle of Man, Isle of Man" on the left, whose location has been highlighted on the map.

Once a document has been annotated, or during the annotation process, the user can switch to a map view which shows all the georeferenced locations. A partial annotation of the Fiennes text is shown in Figure 3 where most places have been georeferenced as points, but the Isle of Man has been georeferenced as a polygon. Clicking on a point or polygon will show the text fragment in which the place name was found, with the option of then jumping to the place name instance in the text view shown in Figure 4. During the annotation process, this provides a tool for checking. Once the document has been annotated, it provides an analytic platform for exploring the geographies within the text to ask questions such as which places is the text describing, and what is it saying about them?

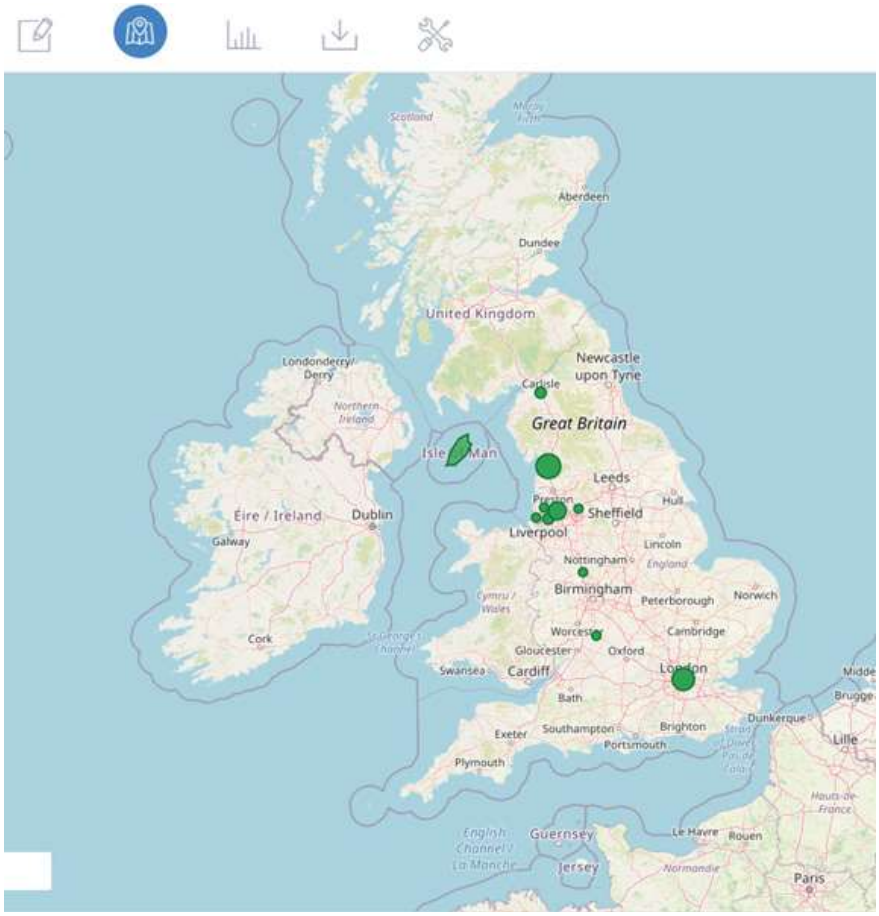

Figure 3. The map view of a partial annotation of the Fiennes text.

As well as being able to analyze the text within Recogito itself, a wide range of export formats are provided, from comma-separated values (CSV) for reading data into spreadsheets and a geographic information system (GIS) to Keyhole Markup Language (KML), which is used by Google Earth, and TEI, which provides an XML-based version of the text. Manually annotating texts has a range of advantages in that it allows the researcher to think carefully about the design decisions made. Working through a text and annotating it is a very detailed form of close reading that forces the researcher to think closely about

decisions over, for example, what constitutes a place. In Figure 4, the first line of the final paragraph talks about "upon ye high Causey." Is this a place? If so, is it called "Causey" or "high Causey," can we locate this place using a gazetteer, and, if we think we can, have we got the right place as there are many instances of "Causey" in place names in the north of England which may not appear in the gazetteers provided? Thus, annotation becomes an important part of the research process.

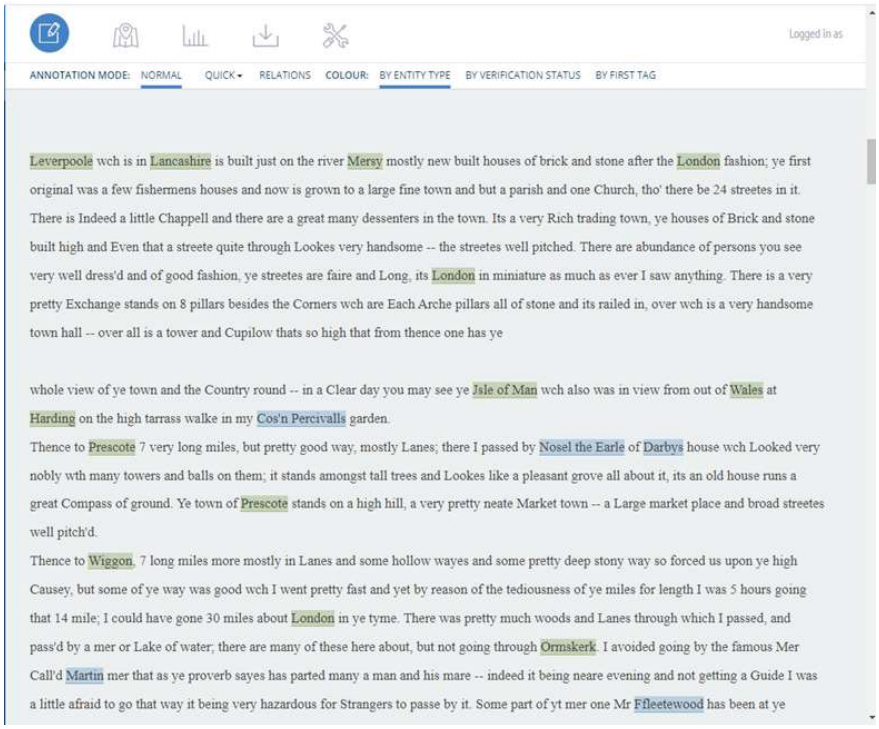

Figure 4. The text view of a partial annotation of the Fiennes text. Places are highlighted in green, people are in blue.

The major drawback to manual annotation is, of course, the time it takes. Recogito also provides functionality to allow the markup to be applied automatically with named entity recognition (NER), using the Stanford CoreNLP (natural language processing) tool for English, French, German,

and Spanish, and an additional tool for Latin. This is far quicker than manual annotation but may be error-prone for early modern sources where spellings and punctuation often do not follow modern conventions, and optical character recognition (OCR) or transcription errors may also be a problem. Manual updates can subsequently be applied to correct these.

Overall, Recogito provides an excellent platform to annotate texts and images either as a lone scholar or as a collaborative group. It is intuitive and easy to use with a good range of help and tutorial options. It also provides an integrated platform for the geographical analysis of texts, something that is not available in GIS software such as Google Earth or textual analysis packages such as AntConc (see the review in *EMDR* 4.1 by Emily Louisa Smith). For other types of analyses, a good range of export formats is provided.

Its major limitation is that while what is there is excellent, more functionality could make it even better. At a basic level, although there is a good range of export formats such as CSV, KML, or TEI, the way that the data is exported to these formats is fixed and may not provide enough information for some types of analysis. More automated tools to allow, for example, part-of-speech tagging which, at a basic level, identifies the types of words such as nouns, verbs, and adjectives, would also be a useful staging post for many subsequent forms of annotation. The gazetteer functionality could also usefully be extended. Currently, if a place name is not found in a gazetteer there is no way of adding it. If there was functionality to create new gazetteers or to update existing ones, this would greatly increase the utility. Gazetteers are currently a major obstacle to early modern spatial humanities research, particularly of multi-lingual sources. A range of researchers working on their own documents could allow such a gazetteer to be effectively crowd-sourced almost as a by-product of this work. More functionality could also be added to allow more sophisticated forms of analysis on the texts and the maps.

IAN GREGORY
Lancaster University
https://doi.org/10.33137/rr.v44i3.38005