Juliette Hueber and Antonio Mendes da Silva (ed.)

Keys for architectural history research in the digital era
Handbook

Publications de l'Institut national d'histoire de l'art

Crossing Boundaries: Using GIS in Literary Studies, History and Beyond

Ian Gregory, Alistair Baron, David Cooper, Andrew Hardie, Patricia Murrieta-Flores and Paul Rayson

DOI: 10.4000/books.inha.4931
Publisher: Publications de l'Institut national d'histoire de l'art
Place of publication: Publications de l'Institut national d'histoire de l'art
Year of publication: 2014
Published on OpenEdition Books: 5 December 2017
Serie: Actes de colloques
Electronic ISBN: 9782917902592

http://books.openedition.org

Electronic reference
GREGORY, Ian ; et al. *Crossing Boundaries: Using GIS in Literary Studies, History and Beyond* In: *Keys for architectural history research in the digital era: Handbook* [online]. Paris: Publications de l'Institut national d'histoire de l'art, 2014 (generated 18 décembre 2020). Available on the Internet: <http://books.openedition.org/inha/4931>. ISBN: 9782917902592. DOI: https://doi.org/10.4000/books.inha.4931.

Crossing Boundaries: Using GIS in Literary Studies, History and Beyond

Ian Gregory, Alistair Baron, David Cooper, Andrew Hardie, Patricia Murrieta-Flores and Paul Rayson

AUTHOR'S NOTE

The work on the Literary Mapping of the Lakes project was made possible by an award from the British Academy (SG46004). The work on the Spatial Humanities project was funded by the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant "Spatial Humanities: Texts, GIS, places" (agreement number 283850).

1 Geographical Information Systems (GIS) have become widely accepted in historical research and there are increasing calls for them to be used more widely in humanities disciplines. The difficulty is, however, that GIS comes from a quantitative, social science paradigm that is frequently not well suited to the kinds of sources that are widely used in the humanities. The challenge for GIS, if it is to become a widely used tool within the humanities, is thus two-fold. First, approaches need to be developed that allow humanities sources to be exploited within a data model that is usable by GIS. Second, and more importantly, researchers need to demonstrate that by adopting GIS they can make significant new and substantive contributions to knowledge across humanities disciplines. This paper explores both of these questions focussing primarily on examples from literary studies, in the form of representations of the English Lake District and history, looking at nineteenth century public health reports.

2 A GIS is effectively a form of database. It differs from a conventional database in that every item of data within it is linked to a location on the map, thus a typical GIS will consist of a table of quantitative data where each row within the table is linked to a

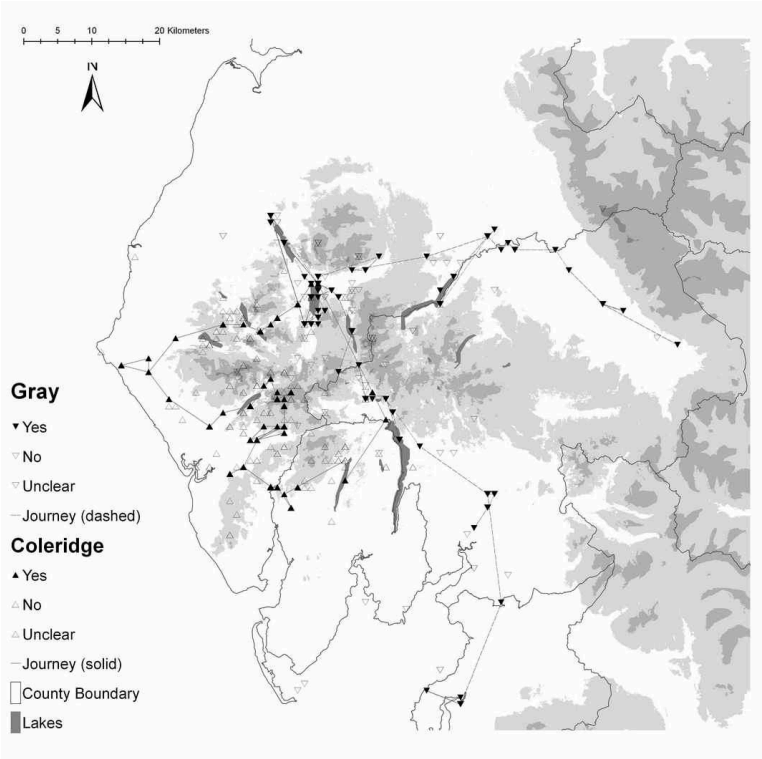
point, line or polygon (representing an area) that maps the location to which the row of data refers. The key advantage of this structure is that it allows the user to explore not only what is occurring but also where it is occurring and, by extension, how things occur differently in different places. This structure has been very successful in quantitative history[1] but its use within the humanities is limited by its reliance on quantitative sources. To be an effective tool within the humanities, GIS must be able to manage non-quantitative sources and, since the major source used by humanities scholars is text, it must by definition be able to handle textual sources. This paper reports on two different examples of how this can be done using different types of digital texts, a small study using writings from literary studies and a much larger scale approach using sources from nineteenth century history.

The Mapping the Lakes project

3 Our initial work on using texts within GIS was called the "Mapping the Lakes" project.[2] This was deliberately small-scale and focussed on two early descriptions of tours of the English Lake District: Thomas Gray's proto-Picturesque tour of 1769 and Samuel Taylor Coleridge's 1802 "circumcursion". These tours were selected for two reasons. First, Gray's tour became well known as a precursor of the classic Picturesque tour, while Coleridge is closely associated with the Romantic movement. This distinction is important. The Picturesque movement is closely connected with the early development of landscape tourism. It is associated with an observer travelling around a landscape and observing it from defined beauty spots in a stylised manner. The Romantic movement, of which Wordsworth is the leading figure, both developed this and reacted against it. While continuing to stress the aesthetic quality of the landscape, the Romantic writer became part of the landscape rather than being a detached observer. From an intellectual viewpoint, therefore, the differences we can find between these two accounts is clearly important. From a more practical point of view, both of these accounts are relatively short, at around 10,000 words each, making them relatively easy to handle within the limitations of the project.

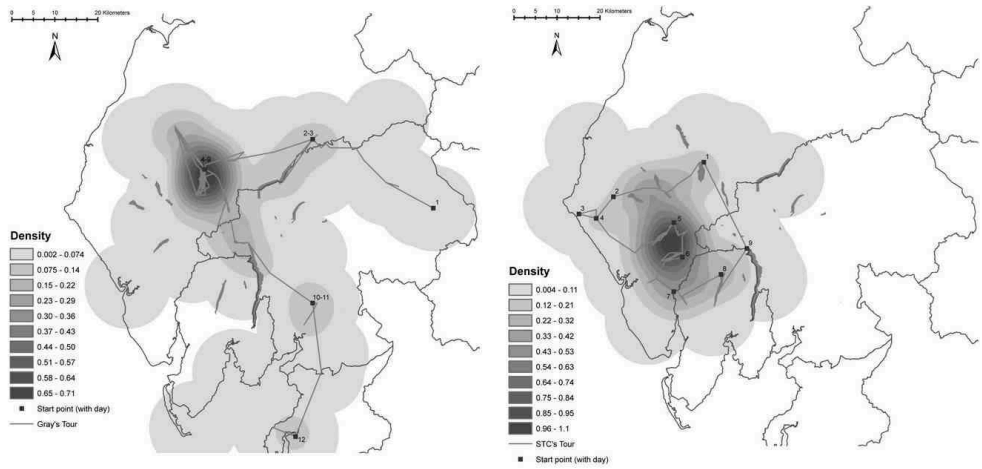
4 The texts were typed up by hand and, during this process, place-names were identified and tagged manually using XML (eXtensible Mark-up Language). Tagging the place-names in this way meant that subsequently extracting them from the text is relatively simple. To convert this into a GIS the essential next stage is to give a co-ordinate to every place-name. This can be done by using a relational join to link the raw place-names to a place-name gazetteer, effectively a database table that gives a coordinate for every name. In this project the Ordnance Survey's 1:50,000 gazetteer was used to provide a British National Grid reference for every place-name. One issue in doing this is the need to resolve spelling variations, such as the differences between "Bow-fell" and "Bow Fell". Names also need to be disambiguated where the same name can refer to more than one location. Given the relatively small size of the texts and the study area, neither of these presented a major challenge. There were also issues to do with the accuracy of the grid references, which are at best only to the nearest kilometre but for linear features, such as rivers, or vague features, such as valleys, may be somewhat misleading.

Figure 1: Simple dot mapping the tours' of Gray and Coleridge.



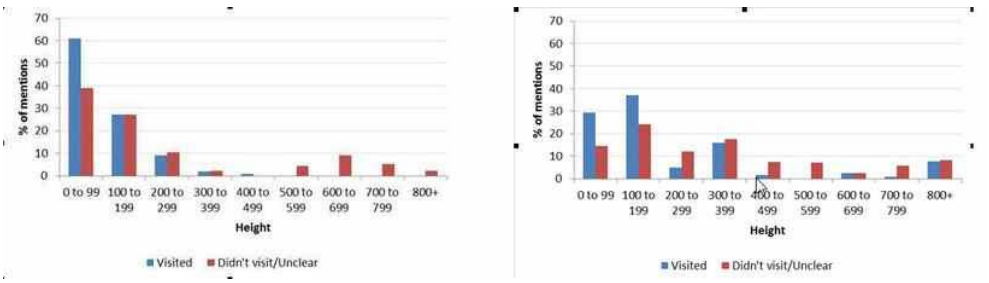
5 Once the place-names have been allocated to co-ordinates, converting these to point locations in a GIS is simple. Figure 1 shows both tours on a single map with straight lines being used to join the points mentioned together to help illustrate the route taken. Gray started at Brough to the east of the Lake District, moved on to Penrith where he spent two nights, going down to Ullswater for the day in between. He then journeyed on to Keswick where he spent six nights travelling out on day trips to the surrounding countryside. Leaving Keswick, he went south, over Dunmail Raise, the main route through the central Lake District, to spend two nights in Kendal, and finally on to Lancaster where the Lake District part of his tour finishes. By contrast, Coleridge started in Keswick where he lived and journeyed south-west through the Newlands Valley to spend three nights in and around St Bees on the coastal plain, west of what is now the National Park. He then went back into the Lake District up Wasdale valley and climbed Sca Fell, his account of descending this mountain is particularly famous. Once down he travelled on through the south-western Lake District and over to Coniston before going north over Dunmail Raise to return home.

Figure 2 : Density smoothed maps of (left) Gray and (right) Coleridge.



6 It is well known cartographically that maps such as those in figure 1 are difficult to interpret. For this reason spatial analysis techniques have been developed that attempt to simplify them and make them more readily comprehensible. One example of this, pioneered in disciplines such as epidemiology and crime mapping, is kernel density smoothing in which the density of events around each location is mapped with denser locations being shaded in darker colours. The density is calculated using a distance decay model in which near events have more impact than those that are further away. In this case an "event" is a place being named in a text. As well as simplifying the pattern, this has the second advantage of reducing the accuracy implied by the point map. Figure 2 shows density smoothed versions of the two tours. Figure 2 (left) shows the central importance of the area around Keswick to Gray's account although other clusters such as Penrith and Ullswater, Kendal, and Lancaster are all apparent. It is clear from this that urban centres and valleys are the most talked about areas within Gray's text. Coleridge, by contrast shows a very different pattern with the account being particularly clustered on the area around Sca Fell.

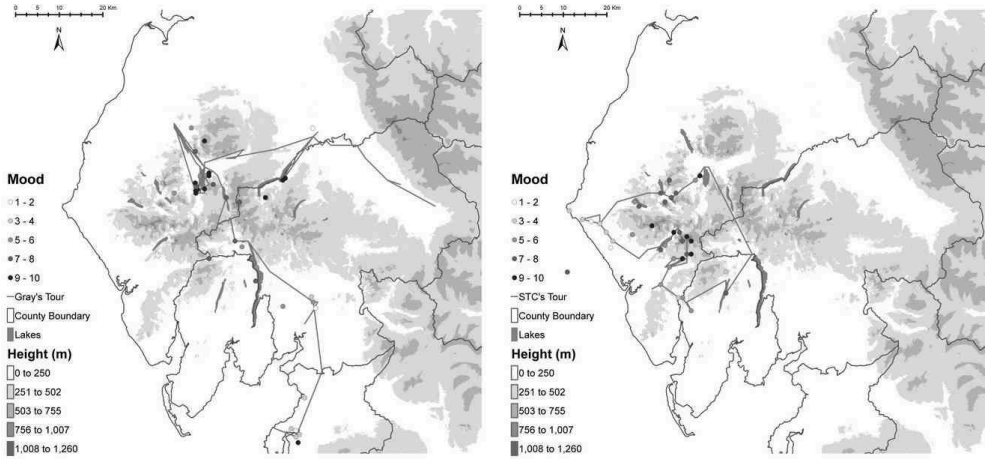
Figures 3: Heights of places mentioned by the two authors (left Gray and right Coleridge).



7 One of the big advantages of GIS is its ability to integrate data from apparently disparate sources. The previous maps imply that Gray concentrated on the more urban areas and valleys, while Coleridge consciously sought out the more remote and upland parts of the Lake District. Using location to integrate data from other sources can help us explore this further. A useful GIS-based source of information about height is a *Digital Terrain Model* (DEM), a representation of the Earth's surface that gives heights for every location. Integrating a DEM with the point data on place-name references allows us to allocate a height to every mention. Rather than mapping them, these can then be

graphed. The graph in figure 3 (left) shows heights of places cited by Gray distinguishing those places that he visits from those that he talks about from a distance. A clear pattern is apparent. He spends all of his time at low altitudes, with over 60% of visited places being under 100m and all being under 1000ft. Most of the places he mentions but does not visit are similarly low although some are at altitude, particularly over 600m which represent the higher Lake District peaks. He almost completely ignores places in mid-altitudes. This pattern seems to fit well with the concept of Gray as a Picturesque tourist: he spends his time in the valleys and passes, describing the areas around him and looking up to the high peaks. The similarities and differences between this and Coleridge's pattern, shown in figure 3 (right), are interesting. Like Gray, Coleridge spent much of his time at lower altitudes but not to quite the same extent. Coleridge also visits places across the height range including a cluster of references in the very highest intervals, over 800m when he climbs Sca Fell. It is interesting though that, while his account is famous for this ascent, it only occupies a relatively small proportion of the heights of the places that he visits. It is also noticeable that Coleridge does not ignore mid-height places.

Figure 4: The emotional response to places by (left) Gray and (right) Coleridge.



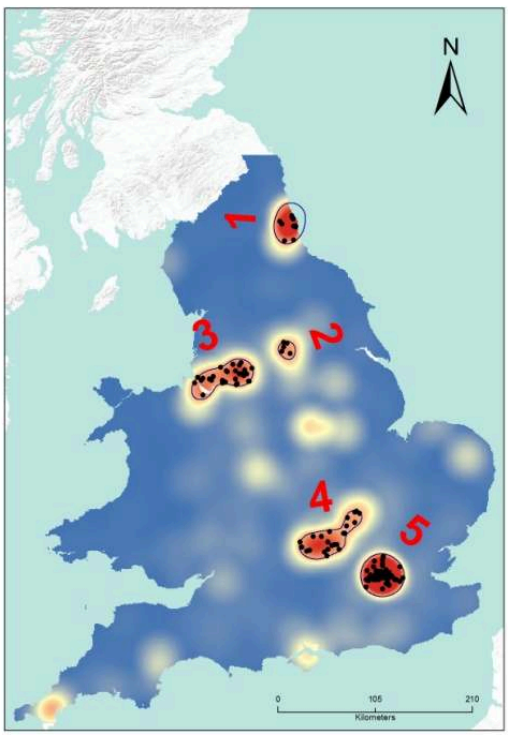
8 As well as mapping where the writers were talking about, we were also interested in what they were saying about the landscape. To do this a ten point scale was devised that associated the emotional response that the writers had to the places that they were talking about. At the bottom of the scale were words such as "dull" and "tedious" while at the opposite end, words such as "sublime" and "terrifying" were given a score of 10. As shown in figure 4, mapping these for the two authors gives a somewhat different pattern than the simple maps of where they were talking about. For Gray, rather than Keswick, the emotional centre is Borrowdale, the valley south of Keswick. Ullswater is also prominent. For Coleridge, perhaps more predictably, the area around Sca Fell is clearly the emotional centre, the area around the Newlands Valley also attracts him, but he seems indifferent to the coastal areas to the west where he spent much of the early part of his tour.

The Spatial Humanities project

9 The above project showed two things: first that we could create a GIS from texts, and secondly that this would allow us to explore the geographies within these texts in new ways and glean new knowledge from them. Its major limitation was that the two texts involved were only 20,000 words long in total and the place-names had to be identified by hand. To be truly effective in the emerging world of digital libraries and archives as well as born-digital material, these techniques have to be scaled-up such that they can be applied to corpora – large volumes of digital text – that consist of millions, if not billions of words.

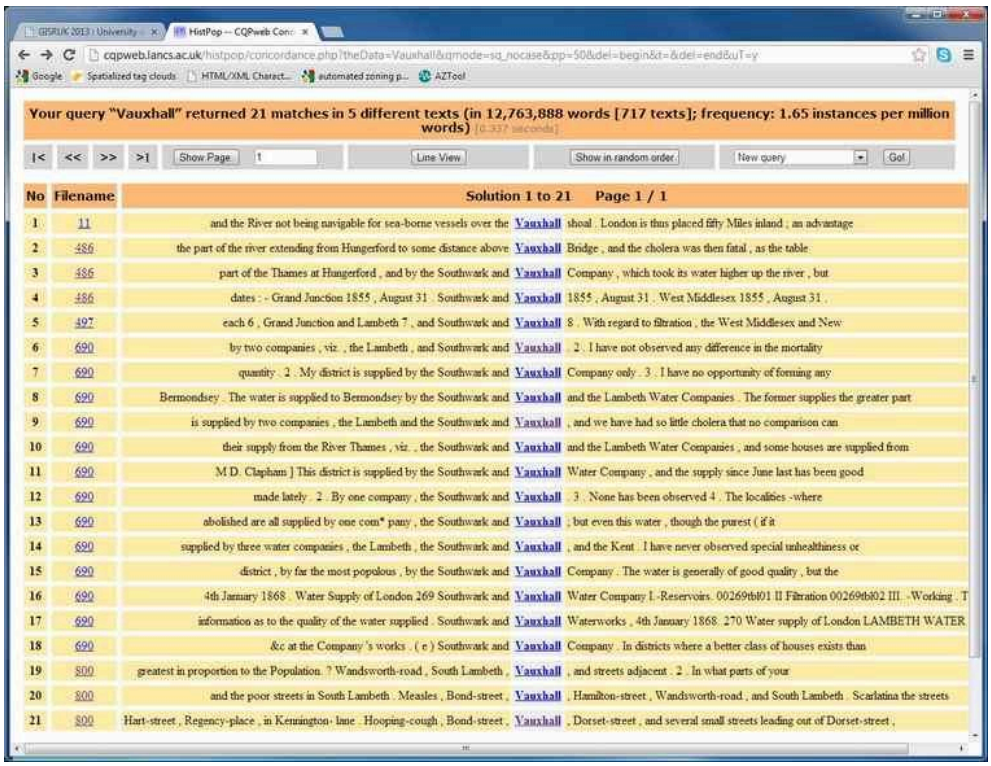
10 The first challenge in doing this lies in geo-referencing the text: identifying the place-names and linking them to a co-ordinate from a gazetteer has received attention from a number of authors. It is not the intention to describe this process here beyond saying that candidate place-names are identified using natural language processing (NLP) techniques. They are then extracted, linked to a gazetteer to provide coordinates, and disambiguated automatically.[3] Here we explore the second challenge: once we have a large georeferenced corpus how can it be analysed? The work is based on the Registrar General's reports from 1851-1911 for England and Wales, taken from the Histpop collection.[4] This source is particularly interesting as the Registrar General was commenting on, and influential in, the start of the period of mortality decline that was to characterise the 20th century. This corpus contains around 2.5 million words and was georeferenced by Claire Grover and colleagues at the University of Edinburgh (Grover et al, 2010).

Figure 5: Clusters of place-name instances from the Registrar General's reports for the 1850s.



11 Having geo-referenced the corpus, the challenge, as with Mapping the Lakes, was then to use appropriate techniques to explore both what places are being mentioned and what is being said about these places. As the corpus is 2.5 million words rather than 20,000, automated techniques need to be used to a greater extent than they were in the Mapping the Lakes project. Figure 5 shows an example of one of the ways this has been done. Kernel density analysis has again been used, this time to smooth the pattern of place-names from the 1850s. This example has gone further than this, the resulting densities have been used to identify clusters which are defined as those areas with a density more than one standard deviation above the mean. Place-name instances lying within these clusters are marked in figure 5.

Figure 6: Concordances on the word "Vauxhall".

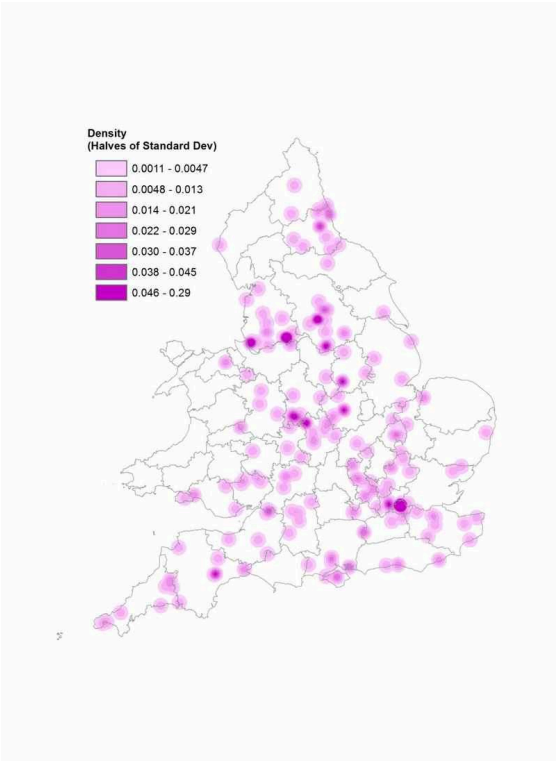


12 This enables us to identify *where* a corpus is talking about both in terms of the general map patterns and the specific place-names that make up these patterns. The next stage is to ask *what* the corpus is saying about these places. The simple approach of "mood mapping" used in Mapping the Lakes is not appropriate here as it only applied to a specific sense of place theme that was encoded by hand. Instead, techniques from corpus linguistics are used.[5] The most basic corpus linguistics technique for exploring what a text is saying about a particular theme or place involves using a *concordance.* This presents the text surrounding each instance of a particular search term which allows a quick assessment to be made about what is being said about a particular place-name. Figure 6 presents a concordance for "Vauxhall", one of the place-names that has among the highest densities of place-name instances surrounding it. The concordance reveals that most of the 21 instances of "Vauxhall" occur in relation to the Southwark and Vauxhall Water Company which in turn points to the Registrar General's interest in water quality and its link to health in London. The software that allows this, CQPweb[6]

allows the concordance lines to be investigated further by following hyperlinks to the full text.

13 This simple approach can be expanded further to create much more sophisticated queries. For example, we might want to create a concordance of all of the place-name instances from the clusters in figure 5 and explore what the key themes that are being discussed in relation to these clusters are and whether the texts are referring to similar themes for each cluster or whether there are differences between them. We might also want to compare the clusters, individually or as a group, with the background pattern.

Figure 7: The distribution of places that collocate with “measles”.



14 This idea introduces another concept from corpus linguistics, that of *collocation* which asks the question “what words occur near this search term?” Collocation can be used to explore what themes are associated with a particular place or cluster of places using statistics that explore how significant the collocates are based on word frequencies in the corpus as a whole. It can also be used to explore what places are associated with a particular theme. The literature tells us that infectious diseases were among the major killers of infants and children in this period.[7] This is supported by a corpus linguistics analysis that showed that “measles” was among the most common disease terms found in the corpus for the 1850s. Figure 7 is thus a density smoothed map of place-names that collocate with the search-term “measles” This is a simple map of the places in which the Registrar General was most interested, in relation to this particular disease. It shows that there was a particular emphasis on the major urban centres of London, Birmingham, Liverpool and Manchester.

Conclusions

15 This work is in its early stages but it clearly has much potential. Firstly, we have illustrated that at a technical level it is possible to create GIS databases from large volumes of text. Secondly, we are developing techniques that draw on the geographical traditions of spatial analysis and the textual traditions of corpus linguistics to allow us to understand both where a corpus is talking about and what it is saying about these places. Thirdly, and most importantly, we have illustrated that this provides a useful scholarly tool in helping to understand texts from both literary studies and from history. The main conclusion is thus that GIS has much to offer to scholarship within the Digital Humanities.

NOTES

1. Ian N. GREGORY and Paul S. ELL, *Historical GIS: Techniques, methodologies and scholarship*, Cambridge; New York, NY: Cambridge University Press, 2007; Anne Kelly KNOWLES (ed.), *Placing History: How GIS is changing historical scholarship*, Redlands, CA: ESRI Press, 2008. URL: http://www.lancaster.ac.uk/staff/hardiea/cqpweb-paper.pdf. Accessed September 5, 2014.

2. Ian N. GREGORY and David COOPER, "Thomas Gray, Samuel Taylor Coleridge and Geographical Information Systems: A Literary GIS of Two Lake District Tours" *International Journal of Humanities and Arts Computing*, vol. 3, no. 1-2, October 2009, p. 61-84; David COOPER and Ian N. GREGORY, "Mapping the English Lake District: A literary GIS", *Transactions of the Institute of British Geographers*, vol. 36, no. 1, 2011, p. 89-108.

3. Ian N. GREGORY and Andrew HARDIE, "Visual GISting: Bringing together corpus linguistics and Geographical Information Systems" *Literary and Linguistic Computing,* vol. 26, no. 3, 2011, p. 297-314; 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", *Philosophical Transactions of the Royal Society A,* vol. 368, no. 1925, 2010, p. 3875-3889.

4. URL: http://www.histpop.org. Accessed September 4, 2014.

5. Svenja ADOLPHS, *Introducing Electronic Text Analysis: A Practical Guide for language and literary studies*, London; New York, NY: Routledge, 2006;Tony MCENERY and Andrew HARDIE, *Corpus Linguistics: Method, theory and practice*, Cambridge; New York, NY: Cambridge University Press, 2012 (Cambridge textbooks in linguistics).

6. Andrew HARDIE, "CQPweb – combining power, flexibility and usability in a corpus analysis tool", *International Journal of Corpus Linguistics*, vol. 17, no. 3, 2012, p. 380-409.

7. Robert I. WOODS and Nicola SHELTON, *Atlas of Victorian mortality,* Liverpool: Liverpool University Press, 1997.

AUTHORS

IAN GREGORY

Department of History, Lancaster University, Lancaster

ALISTAIR BARON

School of Computing and Communications, Lancaster University

DAVID COOPER

Department of Interdisciplinary Studies, Manchester Metropolitan University, Crewe

ANDREW HARDIE

Department of Linguistics and English Language, Lancaster University

PATRICIA MURRIETA-FLORES

Department of History and Archaeology, University of Chester, Chester

PAUL RAYSON

School of Computing and Communications, Lancaster University