Spatializing and analyzing digital texts: Corpora, GIS and places

Ian Gregory, David Cooper, Andrew Hardie and Paul Rayson

Abstract:

This paper argues for Geographical Text Analysis (GTA), a new approach based on combining techniques from corpus linguistics and Geographical Information Systems (GIS). First these allow a text to be converted into a GIS. Corpus linguistics allows place-names to be extracted from texts. These place-names can then be matched to a gazetteer to provide grid references which subsequently allow the place-names to be converted into a GIS layer. Once this has been done, GTA also allows the text to summarize and analyze the geographies within the texts. These approaches can be applied to very large bodies of text, potentially millions or billions of words. We argue that this approach does not replace close reading which is more commonly used in the study of texts. Instead it allows very large volumes of text to be summarized and the close reader's attention to be drawn to the parts of the text that are most relevant to their interest in particular places or the themes associated with these places. The implications of this approach in relation to deep mapping and literary studies are discussed.

Key words: geographical information system, corpus linguistics, place-names, geographical text analysis, literary criticism

1. Introduction

Once upon a time not so very long ago it was all very simple – Information Technology (IT) was concerned with storing and analyzing databases of numbers. The discipline of statistics – which pre-dated computing by centuries – provided suitable techniques for taking the large amount of numbers held in a database and summarizing them and the relationships between them, using a much smaller number of summary statistics and graphics. Thus the use of IT involved quantitative data and social science approaches and, conversely, if you did not use quantitative sources or were suspicious of social science approaches you would not use IT.[1]

Developments over the past decade or so have meant that this cozy dichotomy of mutually incompatible approaches is breaking down. Suddenly, and in many ways without much fanfare, IT has become primarily concerned with text. Through both 'born digital' sources such as email, the World Wide Web, and social networking, and through digitizing paper sources to create digital libraries and digital archives which have the potential to contain every book ever written, IT has undergone a fundamental shift. Today the bulk of the content that is created for IT is, in fact, text. This presents a major challenge. While statistics was well suited to analyzing large numeric databases, there is no similar discipline for text. The disciplines that place the study of texts at their center are those in the humanities. Unlike statistics, humanities disciplines have traditionally eschewed approaches that quickly summaries large amounts of content, and instead stress the importance of using reading – close reading – to understand the subtleties and nuances within the text. While this approach will rightly remain the gold standard for understanding texts, it has one fundamental flaw: it is far too slow to be the only approach to understanding large bodies of text in a world where the researcher has access to literally billions of words of content. This results in humanities researchers having to be highly selective, and this tends to be done in a way that is far more arbitrary than most humanities researchers would like to admit.

These developments might be thought to place the humanities at a crossroads, where researchers are faced with the choice of carrying straight on and continuing to read texts in a detailed but slow way, thus failing to exploit much of the content that is available, or alternatively taking a sharp turn to new methods that summaries millions or billions of words without the researcher ever having to read any of them. The reality is more complex and more subtle: reading will remain central to humanities-based approaches, but it needs to be enhanced and complemented by methodologies that exploit the digital nature of modern text. These methods will enable us to summarize large corpora which would be difficult, if not impossible, to read in their entirety. By extension, this process will allow us to determine which parts of the corpora we ought to read closely (and, conversely, which we should not), thereby helping us to justify textual selectivity. It will also help us to understand how what is read closely relates to the broader context of the remaining unread material within the corpus. Such methods, then, will necessarily involve a shuttling between distance and proximity, abstraction and particularity: processes which correspond, at least in part, with Franco Moretti's controversial model of 'distant reading'.[2] At the same time, the adoption of such methods will also facilitate self-reflexive thinking on the processes through which we, as

academic researchers, select those texts which become the objects of detailed critical analysis; thinking which, in turn, will raise further questions about the nature of canon formations. Finally, such methods will also allow texts to be placed within networks which stretch beyond the circumscribed boundaries of the digitized corpus: they will allow texts from a disparate range of sources, and in a diverse range of genres, to be brought together in a single analysis; and they will permit a range of other forms of data – including, for example, statistics, images, multimedia formats and, of course, maps – to be integrated within the same digital space.

This paper argues that Geographical Text Analysis (GTA) – the combination of Geographical Information Systems (GIS) and corpus and computational linguistics – offers one way to implement this. A GIS is a combination of a database and a computer mapping system in which every item of data is *georeferenced* to give it a real-world location.[3] This structure offers a number of advantages: it allows the researcher to explore the database by location to ask questions such as "what is here?" and "what is near here?"; it allows data to be mapped to summaries the geographies that the database contains; it allows data from different sources to be integrated because all data are underlain by real-world co-ordinates; and it provides a platform for spatial analysis, a form of statistical analysis in which the locations of the items under study are explicitly included with the analysis.[4] GIS has traditionally been a quantitative technology whose use within the humanities has been restricted mainly to the social science ends of history,[5] although there are increasing calls for GIS to be more generally applicable within the humanities.[6] Corpus linguistics is a methodology used to study language using a large naturally occurring body of text – a *corpus* – on a variety of levels including lexis, syntax, semantics, and pragmatics or discourse. It has been employed, notably, for lexicography or dictionary creation, where large corpora are used as source data for determining sense boundaries, definitions and examples for dictionary entries. Corpus techniques are increasingly being exploited across a wide range of areas within linguistics, such as the description of grammars, the analysis of literary style, or the investigation of language change. As a preliminary step in many corpus-based techniques, automatic language analysis techniques from the closely related area of computational linguistics, otherwise known as Natural Language Processing (NLP), are used to enhance the corpus data with some sort of annotation to code one or more levels of the analysis in a robust and consistent manner.[7]

Bringing these fields together provides an approach that offers the potential to develop spatial narratives that tell us how different places are represented in different ways. Extending this will also provide information on how representations of different places have changed over time, and between authors, genres and/or publications. It can also allow this information to be integrated with geo-referenced statistics, maps or images to see whether the different sources are telling the same or different stories. As yet, we are in the early stages of developing such techniques, however pilot work has been encouraging.[8] This paper will describe the ways in which place-names can be automatically extracted from a corpus using NLP techniques so that they can then be linked to a gazetteer to geo-reference them and read them into a GIS. Once this database creation phase is complete, the next stage is to analyses these data. Here two entirely different forms of analysis need to be brought together: those from GIS and spatial analysis concerned with the analysis of geographical patterns on the one hand, and those from corpus linguistics and NLP concerned with analyzing texts on the other. Finally we look at some of the broader implications that this approach could have to humanities disciplines. The paper is based on the early stages of a European Research Council funded project *Spatial Humanities: Texts, GIS, Places* which is developing suitable techniques to implement these approaches and applying them to two separate studies, one of which will create a *literary GIS* of the English Lake District, while the second will focus on nineteenth century social history by integrating quantitative and qualitative approaches.[9]

2. Creating spatial databases of digital texts

a. Creating and searching a corpus[10]

A computer corpus may be created in many ways; one central concept in corpus linguistics is that the researcher's corpus must be well-suited to the research question that they wish to study. In the early days of corpus linguistics the only way of creating machine-readable text was to manually type in all the texts to be included in the corpus. Nowadays, manual transcription can be avoided for some sorts of text by employing Optical Character Recognition (OCR) software to extract machine-readable data from images scanned from the printed page, though some types of text – such as spoken texts – still need to be typed in. Of course, an increasingly large proportion of the texts we might wish to include in corpora of the contemporary language are now 'born digital', especially those available via the World Wide Web. Historical corpus data is typically among the kinds of data where typing (or else extensive manual correction of OCR output) is required. Although the methods described in this chapter rely on full text sources being available and are more accurate when the quality

of the OCR or transcription is high, there are significant efforts and initiatives underway aimed at improving the quality of OCR in historical and literary texts and at undertaking large-scale full-text transcription.[11] However the text is produced, a standardized style of markup based on eXtensible Markup Language (XML) may be applied to the corpus to indicate the structure of each text, although it is also possible to apply corpus methods to a "raw" corpus without any markup.

[Figure 1: Concordance]

The first step in moving from a raw text corpus to a spatial database is rooted in one of the foundational techniques of corpus analysis, namely concordance searches. A *concordance* is a data output consisting of all matches in a corpus for a specified search pattern, plus a specified amount of the text surrounding each match, the *co-text*. This technique permits a researcher to begin to engage with the text by reading the co-text and thus carrying out qualitative analysis in order to draw meaningful conclusions about the occurrences of the words or phrases searched for. Concordancing software usually permits the researcher to extend each concordance line to provide more surrounding context, or to jump straight into the full text at that position in a manner similar to a hyperlink on a website. An example of a concordance of the place-name *Stirling*, from the Lancaster Newsbooks Corpus, [12] a corpus of newsbooks published in London in the seventeenth century, is shown in figure 1.

All corpus search software allows the corpus to be searched for a specified word or phrase, which can often include wildcard symbols. Using more advanced tools, we can also search a corpus for all the words that have been given a certain tag in the annotation process. When we wish to extract information about the places discussed within a corpus for the purpose of building a GIS, our starting point is a concordance. Critically, we do not just want to find all the different place-names. We want to find all the different *instances* (or "mentions") of all the place-names. If *London* is mentioned 8,000 times and *Foston on the Wolds* is mentioned once in a million words of text, then the list we create needs to contain 8,000 instances of *London* and one instance of *Foston on the Wolds*. This way the list accurately reflects the relative prominence of different place-names in the data.

The most straightforward way to get an exportable concordance of place-names is to search for a part-of-speech tag for *proper nouns*. 'Proper noun' is the grammatical category of names of people, places, organizations and other specific entities, which often behave differently to *common nouns*, names of categories of entities. In English, proper nouns are

typically capitalized whereas common nouns are not. Most grammatical corpus annotation will be able to identify and mark up proper nouns – for example, the CLAWS tagger developed at Lancaster University[13] marks proper nouns with the tag *NP1*. A concordance for NP1 will therefore extract all mentions of all place-names that have been correctly tagged as proper nouns (it is possible, of course, for the tagger to make mistakes). It will also extract all mentions of names of persons or organizations but, as is discussed below, these can be filtered out at a later stage.

An alternative approach would be to use Natural Language Processing (NLP) techniques for *named entity extraction*. These techniques allow the automatic discovery of names of people, places, organizations, times, dates, and other quantities, using criteria other than simply spotting grammatical markers of proper nouns. The accuracy of such techniques is quite variable and depends on the subject and quality of the input text. For example, if one has a list of all major cities in the UK stored as a gazetteer then it is quite straightforward to find all mentions of such place-names in a large quantity of newspaper text with simple fuzzy matching techniques. However, to find all names of people and distinguish them from place-names and other organizational names is more problematic. For example, *Lancaster* (a city) would have to be distinguished from *Lancaster Bomber* (a plane or a type of beer), *Stuart Lancaster* (the England rugby coach) and *Duke of Lancaster* (a nobleman or a pub) by using phrasal patterns and possibly further context in order to decide on the meaning. As with other NLP problems, a variety of approaches including knowledge-based (dictionary look-up), rule-based (hand crafted templates) and statistical methods (using probabilities and trained language models) exist and these may need to be tailored for a particular domain or input text. The problem of disambiguating our initially-extracted list of instances is one that we must address – either at the initial stage, using these automatic named entity extraction techniques; or manually at the geo-referencing stage, as discussed below. In either case we must draw on features of the co-text to decide which mentions become part of our dataset for georeferencing.

b. From a corpus to a GIS

Using corpus techniques thus provides a list of all of the suspected place-names within the corpus. Converting these into a GIS is conceptually simple; however there are a number of practical difficulties that may make this difficult and time consuming. The core of the process is to join the list of suspected place-names to one or more gazetteers to provide a location to georeference each place. A gazetteer is a database table that provides a coordinate for each

place-name that it contains. In its simplest form it has three columns: place-name, x-coordinate (or longitude) and y-coordinate (or latitude). Some gazetteers also hold additional information such as which higher-level administrative unit the place lies within, what type of feature it is, and different variations of the name's spelling.[14] A number of gazetteers are available including Geonames,[15] World-Gazetteer,[16] the Getty Thesaurus of Geographical Names,[17] and the Ordnance Survey's 1:50,000 Gazetteer.[18] In theory joining the list of suspected place-names to the gazetteer using a relational join will give coordinates to all of the suspected place-names and, from here, it is simple to read these coordinates into a GIS software package where they will form a point layer.[19] In practice, the process is more complicated as a result of a number of practical problems that can lead to three types of errors: names that are not place-names being wrongly allocated a coordinate; real place-names not being given a coordinate, and place-names being allocated to the wrong coordinate.

The first of these occurs when a suspected place-name, such as *Lancaster*, is not being used as a place-name. Errors of this type can be filtered out using the co-text in which the suspected place-name occurs using the same phrase-patterns that named entity extraction often exploits. Thus, if an instance of *Lancaster* is preceded by *Mr*, *Duke of* or a proper noun known to be a person's forename then it can be removed from the list. In the same way, if it is followed by words such as *bomber* these can also be removed. This is the first way in which the co-text of each instance of a suspect place-name, extracted via the concordance search, is crucial for the transference to a GIS. It could also be argued that occurrences of the suspected place-name that are preceded by 'the' or 'a' should be removed, as in "...the Lancaster flew many missions..." or "...a Lancaster was sighted..." one might also remove plurals such as "Lancasters were involved in the raid on..." However, if filters like these are used too mechanically, there is a danger of removing place-names that may be wanted. There may also be genuine questions about whether some words should be included as place-names or not. *A Lancaster graduate* clearly refers to the university and thus the town and arguably should be included as a place-name; but it is equally true that this is not a direct reference to the town. As a consequence, filtering cannot be applied across the board as a standardized procedure for all lists of suspected place-names, but is instead a crucial step in the research process that involves making judgments about what does, and does not, constitute a reference to a place-name. This judgment can only be made by exploring the context in which the suspected place-names are used.

The second type of potential error, where suspected place-names fail to match to the gazetteer, are conceptually simpler. They can occur because: the place-names' spellings differ between the text and the gazetteer; because the place-name is not in the gazetteer; or because the suspected place-name is genuinely not a place and should be omitted. Spelling variations are very common particularly in historical sources although they persist today. The town of 'Saint Helens' can be spelt with at least three possible variants of saint ('Saint', 'St.' or 'St') and 'Helens' may or may not have an apostrophe. This gives a total of six possible variations, excluding spelling mistakes and digitizing errors. When variants are found, whether they are genuine differences or spelling mistakes, they can be added to the gazetteer along with a standardized version of the spelling. Similarly, if places have been omitted from the gazetteer these can also be added to it. In this way, the process of georeferencing a text also becomes a way of improving and enhancing gazetteers.

The third problem is places being allocated to the wrong coordinate. This usually occurs because there are two or more places with the same name, and either the gazetteer has the wrong version or contains several places with the same name. In the second case, disambiguation can be done manually or automated procedures can be developed to decide which option is most likely to be correct.[20] The first option is more difficult and only careful checking is likely to spot problems of this sort.

It should be clear from this that the process of geo-referencing a list of suspected place-names requires a certain amount of manual intervention and even then is still error prone Careful checking can, of course, reduce these errors significantly, but requires a considerable investment of time. Even if this can be invested, if a large corpus is to be georeferenced the results will still contain errors, and while these can be minimized, subsequent analyses should be sensitive to this.

3. Analyzing georeferenced digital texts

a. Corpus-based approaches

[Figure 2: Frequency of proper nouns containing 'ford']

Once place-names have been identified and georeferenced, they can subsequently be analyzed using both corpus-based approaches which have traditionally been primarily used by linguists, and GIS-based approaches that have traditionally been used by geographers. Crossing this divide has much to offer to the methodologies used by both disciplines and, more importantly, to contribute new knowledge in the humanities and social sciences. As

briefly described in the introduction, the intention is to be able to summarize large corpora and to highlight texts, or sections of texts, which are considered worthy of more detailed analysis. We have already seen the role that the concordance technique plays in building a corpus-derived GIS. A further technique can be used to enhance the analysis: *frequency profiles*. Frequency profiling of counts of place-names – or other words – identified within a text can be used to highlight variation within one text or a collection of texts. An example is shown in figure 2 which counts the number of instances of proper nouns containing the string 'ford.' Additional information on where within the texts these instances are found would allow us, for example, to highlight sections of a text where high concentrations of certain place-names occur. By further processing the resulting frequency lists we may also be able to spot groups of place-names that regularly (or never) occur together. Using a corpus comparison technique called *keywords*, it is also possible to highlight which place-names occur more often than expected in one section of the corpus relative to the whole corpus or to another reference dataset.

While these techniques allow us to ask *where* a corpus is talking about, and how this changes, the question that is likely to be of more interest is 'what does the text say about these places?' To begin to answer this we need to engage with two further automated techniques from the corpus researcher's toolbox: *collocation* and *semantic analysis*. A collocation analysis addresses co-occurrence – which words or phrases regularly occur together in a text or corpus. This is a quantitative abstraction and summation of the co-text. For example, a reading of the concordance for *Stirling* shown in figure 1 quickly and clearly illustrates that the town was largely being talked about in relation to its military significance. Collocation analysis allows us to identify such patterns without the need to read every concordance line, although examination of at least some concrete examples by the researcher is usually necessary. Since collocations are a key indicator of semantics, lexicographers can use them to help differentiate and cluster word meanings. Collocation can be used comparatively by corpus linguists to discover differences between the representation of words and concepts in certain genres. For example, we may compare the concepts of *bachelor* and *spinster* in romance fiction, or research the representation of *immigrants* and *asylum seekers* in the UK press.[21] When exploring georeferenced texts, collocations of place-names can be used to discover what words writers use in the text surrounding the occurrences of those place-names. At some level of aggregation, this may permit overall patterns of description to emerge from a text or collection of texts that would otherwise not be available from a simple

reading of those texts. Such patterns may show what modifiers (adjectives) are being used to describe a particular place or what activities (verbs) occur close by. By combining the collocations extracted with a qualitative analysis using concordances, a researcher is able to form categories from groups of co-occurring words in the context of place-names and start to highlight the key descriptions that emerge from the underlying texts.

In fact, we can automate this grouping of terms to some extent. As described in section 2a, an automatic part-of-speech tagger can be used to mark up proper nouns in text. Using a *semantic* tagger, we can assign to the text another level of tags representing semantic fields of words and phrases. The semantic field represents each word or phrase's position within a general ontology. Each semantic tag groups words from a dictionary together into similar categories of meaning, e.g. Education (P1), Warfare (G3) and Farming (H4). The tagger used here is the UCREL Semantic Analysis System (USAS) which incorporates a dictionary of over 75,000 words and phrasal templates designed by hand, in combination with a small number of other rule-based and statistical techniques to assist in the selection of the correct semantic tag in context.[22] Although the accuracy is high (91 percent)[23] there are inevitably some meanings that the system does not know or that it will tag incorrectly in specific domains, and these need to be considered in any subsequent analysis. Using the USAS system in combination with a collocation analysis allows us to analyses patterns of semantic tags that regularly co-occur in the text near to each place-name. This therefore assists in the analysis of the *concepts* associated with a particular place, addressing question such as, is that place associated with themes such as education or warfare or, more generally, is it presented in a positive or negative way? With enough occurrences of a specific place-name within a text or corpus, we will start to see patterns emerging through the collocation analysis, to show statistically significant relationships between the place-name and a concept or group of concepts. These will then need to be confirmed through a qualitative analysis using close reading of concordance lines or extended extracts from texts in order to check for errors from the automated processing. As outlined below, the collocation analysis allows different kinds of mappings to be generated from the same overall list of mentions, for example, mappings associated with different concepts.

b. GIS-based approaches

[Figure 3: Dot map of Wordsworth]

GIS and spatial analysis techniques offer a completely different, and radically new, method of exploring texts. The easiest way of 'analyzing' a georeferenced text is simply to map the points that are named within the text. This can be done using a simple point map of the places mentioned or perhaps using more complex symbology to show for example: which source the places were named in; the date they were mentioned; or their collocates, for example, whether they are talked about in association with a certain theme or certain type of emotional response. An example of this, taken from a small corpus of material by William Wordsworth, is shown in figure 3. This is very straightforward within a GIS and presents a simple but effective overview. From here there are two very different routes for further analysis: *map-based querying* and *spatial analysis*.

[Figure 4: Google Earth]

In map-based querying an interactive map is used that allows the text or texts to be explored through a map-based interface using technology such as Google Maps or Google Earth.[24] Figure 4 shows an example of this based on a small corpus of Lake District literature where a user has clicked on the point on the map representing Sca Fell. In the top-right of the screen this has returned a concordance of all of the mentions of Sca Fell (including spelling variations) and this, in turn, has been used to find a particular mention in a text shown beneath the concordance. Clicking other place-names in the text would, in turn, highlight these on the map. This approach allows the reader to ask the question 'what has been said about this location?' and then read all of the responses in the corpus. It is thus well suited to close reading but makes use of the hyper-textuality within digital texts such that rather than reading in a linear manner from beginning to middle to end, the reader can switch from place to place within and between texts.[25] In this case, geographic location provides the hyper-textual structure through which the reader can approach the analysis (or reading) of the text. This, in turn, leads to the possibility of creating distinctly spatial narratives as space becomes the prime organizing motif behind the way in which the texts are organized.

[Figure 5: Density smoothing]

While map-based querying provides an approach that fits well with the humanities tradition, spatial analysis comes firmly from the social and Earth sciences. While this means that the approaches offered must be used sensitively, it does not mean that they should be rejected, as they provide a highly effective way of summarizing large and complex geographical patterns. One highly effective technique, originating from criminology and epidemiology, is *density*

smoothing.[26] This accepts that maps of point patterns are difficult for the human eye to comprehend, and also that the points tend to suggest a level of precision that place-names within a text cannot support. Instead, density smoothing creates a continuous surface with high values being found in areas with many points. This provides an effective way of summarizing either all of the places mentioned in a text or, when used in combination with techniques such as collocation and semantic tagging, place-names that are found near to words with specific meanings. Figure 5a shows an example of this based on mapping *all* of the places mentioned in the Lancaster Newsbooks Corpus, while figure 5b shows places that collocate with words tagged as being associated with war.[27]

While previous work has shown density smoothing to be a highly effective technique for summarizing texts,[28] it is only one of a number of well-established spatial analysis techniques that are used to explore point patterns to see if they cluster, are evenly distributed, or are randomly distributed.[29] Techniques such as Moran's I, Geary's G and Local Indicators of Spatial Autocorrelation (LISA)[30] would all be suitable for exploring this. Developing these ideas further, spatial analysis techniques could be used to explore whether places associated with certain words or themes cluster near other words or themes or whether they are found in totally different locations. In spatial analysis terms this requires a multivariate technique such as multivariate LISA or possibly Geographically Weighted Regression.[31] A slightly different question where spatial analysis will also be able to help concerns whether the pattern of place-names found throughout the texts follow a logical progression such as moving around a study area sequentially as might be expected to be the case in conventional travel literature. Finally, an additional component can also be added to ask whether places and their associated themes change over time, or between authors or genres. All of these questions can be asked using established spatial statistical approaches with some minor modifications to adapt them to the use of texts rather than statistics.

From a technical perspective, therefore, there are two well-developed fields that can be used to analyses texts: corpus-based techniques and spatial analysis. To date these have had little to do with each other; bringing them together will provide exciting new potential to understand the geographical patterns and meanings within texts.

4. Implications for Humanities: The evolution of the literary GIS

So far this paper has described the ways in which texts can be analyzed geographically. While these are becoming technically feasible, the key issue that remains is how useful they

are to applied scholarship within and beyond the spatial humanities. In considering this we focus on how such methodologies can facilitate the spatial analyses formulated by literary critics – scholars whose work has traditionally been predicated upon the close reading and qualitative interpretation of a relatively small number of texts. By extension, how can such exploratory digital humanities research feed off, and back into, contemporary theoretical debates regarding the literature of space, place and landscape and, in particular, the emergence of geocritical practice as codified by Bertrand Westphal? The potential and problems of a literary GIS were first explored, however, in the 'Mapping the Lakes' project.[32] This project used place-name georeferencing to map out the spatial narratives of two key Lake District topographical prose texts: Thomas Gray's account of his 1769 touristic tour of the region; and Samuel Taylor Coleridge's documentation of a characteristically singular walking excursion through the western half of the Lakes in August 1802.[33] This initial attempt to construct a deliberately delimited literary GIS generated both technical and critical findings. From a technological perspective, the project circumnavigated the presentational problems traditionally associated with codex-based reader-generated literary cartographies through the direct linking of electronic texts and digital maps. At the same time, 'Mapping the Lakes' transcended the limitations of problematically positivist applications of GIS through the mapping of the emotional geographies to be traced in the respective primary texts: a form of textual mood mapping which has clear correspondences with 'sentiment analysis' methodologies and which intersects with the wider multi-disciplinary evolution of qualitative GIS research.[34] Close textual analysis was used to identify those locations at which Gray and Coleridge articulated positive, negative and even inherently contradictory experiences of the Cumbrian landscape and environment. As a result, 'Mapping the Lakes' negotiated a rapprochement between the use of GIS – a digital tool almost invariably associated with the visualization of significant quantities of spatial data – and the nuanced close reading of textual textures which defines much literary critical practice.

The 'Spatial Humanities: Texts, GIS, Places' project has begun to build upon this initial development of literary GIS through the digitization and georeferencing of, among other sources, a large corpus of Lake District landscape writing, including guide books, regional histories, autobiographies, journals and poetry. One of the primary aims of this part of the project is to produce an electronic archive which will bring together, in a single digital space, a significant number of a heterogeneous range of topographical texts written between 1750 and 1900. This process is leading to the creation of a fully accessible scholarly resource: a

corpus which allows the user to identify the richly intertextual nature of both canonical and historically marginalized Lake District literary texts. This electronic archive will not remain absolutely fixed, though, but instead will exist as a textual environment which can be continually developed through the digitization of further place-specific texts.[35] The digital intertext, therefore, is characterized by a state of openness and will be subjected to ongoing growth and expansion. Crucially, as the texts are georeferenced, this facilitates the visualization of the large-scale geographical patterns which are embedded in texts written during a key period in the region's spatial history. The GIS highlights those sites which emerged as central to the dominant spatial narratives of this particular landscape: those fells, villages, lakes and pathways which were subjected to multiple layers of textual representation and re-representation. Alongside this, the GIS also draws attention, through its blank spaces, to those locations which were placed on the edges of the region's hegemonic cultural geography. This is a form of digital humanities practice, therefore, in which the corpus and the GIS are brought together to produce what might be described as a spatial intertext of this culturally over-determined terrain.

The development of this methodology clearly chimes with Bertrand Westphal's articulation of a 'geocritical' approach to literary texts. According to Westphal, 'geocriticism tends to favor a geocentered approach, which places *place* at the center of debate'; and, as a result, "the spatial referent [either a named location or a generic form of topography such as the desert or the archipelago] is the basis for analysis, not the author and his or her work".[36] He also contends that it is "uncommon that artistic works are categorized according to the geographical spaces that they explore" and he suggests that: "Databases organized around spatial data are rare indeed." Westphal goes on to acknowledge that the "Internet certainly helps" in the formation of "[i]ndices that associate a work with a place"; but he also warns that "a great deal of patience, and a certain amount of scholarship, will be indispensable in forming a corpus necessary for a fully geocritical analysis."[37] Georeferencing Lake District landscape writing between 1750 and 1900, therefore, represents a first step towards the realization of this geocritical ambition. As a result, the literary thread to 'Spatial Humanities: Texts, GIS, Places' moves away from what Westphal would define as the exclusively egocentric approach of 'Mapping the Lakes' – which is structured around the mapping of subjective geographical accounts offered by just two canonical writers – to a 'multifocalized' methodology which brings together dozens of interweaving, and frequently competing, examples of geo-specific landscape writing.[38]

The analytic techniques described above facilitate further thinking about how literary articulations of place have shaped the region's distinct cultural identity and have contributed to the sense of exceptionalism which has underpinned the concept of Lake District-ness since the middle of the eighteenth-century. Alongside this, it allows users to identify a named location – such as the village of Rydal, where William Wordsworth lived from 1813 until his death in 1850 – and to examine how the textual representations of that particular place evolved over a period of 150 years. By extension, the GIS-based spatial intertext opens up further geocritical thinking about the way in which writers, in representing place, respond to both the three-dimensional materiality of the spatial referent and (consciously or otherwise) earlier textual accounts of the named geographical site. The project then uses the tools described above as a platform for exploring the role played by literary texts in the 'placialization' of the Lakes: the term coined by Edward S. Casey to describe 'the formation of place, for example, in landscape paintings and maps, but also in historical narration and prose fiction'; and a term, therefore, which allows for the way in which literary texts inform a more general process of place-making.[39] Moreover, the use of layers of text – to use both the GIS and more general meaning of the term – allows analysis of the notion of the 'stratigraphic' which is integral to Westphal's geocritical understanding of the literature of space, place and landscape.[40] That is to say, by adding and removing multiple layers on the spatial intertext of the literary map, the user is able to examine imbrications of geography and temporality and to further his or her understanding of the 'diachronic depths' of the region's geo-specific literary history.[41]

The exploration of the collocation of place-names with specific semantic tags, described above, can, for instance, open up further thinking about the 'stratigraphic' nature of Lake District landscape writing. It is possible, for example, to trace the textual origins, of the habitual practice of prefacing the Cumbrian fell-name, 'Skiddaw' with the adjective 'lofty' and to identify how successive generations of landscape writers reinscribe this collocation. Can the construction of the spatial intertext highlight any significant subversions of, and deviances from, this familiar descriptive tag? Can such changes be placed within the wider spatial history of shifting cultural attitudes to vertical topographies? As Svenja Adolphs points out, corpus approaches enable researchers to think further about two inextricably enlaced forms of intertextuality: the way in which texts allude to and echo previous writings; and, at the same time, the way in which the 'semantic prosodies' of published texts relate to everyday language use.[42] When the 'spatial referent' of the material landscape is introduced

in order to form a triangulation of text, intertexts and place, then it becomes clear that there is potential for exploring the relationship between Lake District landscape writing and vernacular geographies: the colloquial and quotidian articulations of place and spatial experience which often sit outside, and cut across, officially recorded geographies.[43] So, for instance, do Lake District landscape writers incorporate, within their texts, any locally generated names to describe the fell marked 'Skiddaw' on the Ordnance Survey map? Do the topographical texts refer to an identifiably bounded geographical space when they name 'Skiddaw'; or, in using this toponym, do the texts allude to a vaguely conceived and imprecisely demarcated terrain? The use of GIS to explore the way in which vernacular geographies are embedded within literary texts can be expanded yet further by an examination of how, in linguistic terms, spatial practices are recorded and geographical directions are described. To return to the example of Skiddaw, how do writers record the physical movement towards this fell? Do particular expressions of spatial navigation become culturally entrenched within the spatial intertext? The use of the combination of corpus and GIS techniques, therefore, helps to move geocritically informed literary GIS research into the theoretical territory famously occupied by Michel de Certeau in his exploration of the roles played by spaces and places, maps and tours, in the formulation of the spatial stories which constitute the practice of everyday life.[44]

We are acutely sensitive, however, to the fact that these map-based approaches might appear to marginalize the practice of detailed textual analysis which was integral to the development of the literary GIS showcased in 'Mapping the Lakes'.[45] In other words, it may appear as if the critical depth evidenced in the geographical readings, and mood mappings, of the texts by Gray and Coleridge is being sacrificed in favor of the surface spatial overviews with which the scholarly use of GIS has been traditionally associated but for which it has also been critiqued.[46] 'Spatial Humanities: Texts, GIS, Places', however, functions on multiple cartographic scales by oscillating between the mapping of overarching spatial patterns to the type of textual micro-mapping which is more familiar to the literary critic. On the one hand, the project produces surface maps through the geovisualization of the evolution of linguistic patterns over time. As well as mapping and analyzing topographical texts by a diverse range of Lake District writers, however, the project simultaneously involves the spatialization of multiple versions of a single work of literary geography, such as William Wordsworth's prose *Guide to the Lakes* or Harriet Martineau's (critically neglected) *A Complete Guide to the English Lakes*: a process which illustrates how the 'multifocalization' – which, for

Westphal, can be understood by examining the work of a range of writers – can also be traced in the work of a single title which is revised and reconfigured over a period of time. The result is a multi-scalar literary GIS which endeavors to allow for both the new reading practices opened up by the digitization and spatialization of large corpora and traditional attention-to-textual-detail. Ultimately, then, 'Spatial Humanities: Texts, GIS, Places' moves beyond 'Mapping the Lakes' by developing a more fluid form of literary GIS which enables the user to negotiate his or her own path(s) through the spatial intertext. The creation of the intertext will be founded upon a geocentric approach to literary history: the GIS has the potential to identify large-scale spatial patterns within Lake District literary geographies; and, at the same time, it visualizes, in cartographic form, the textual layering of particular named places. Yet, crucially, the system also enables users to pursue particular lines of enquiry by allowing for egocentric – or writer- and text-specific – points of entry.

The further advancement of the literary GIS can be traced by returning to an idea articulated in the introduction to this paper: the possibility of combining the textual GIS with other multi-media representations of place in order to construct a 'deep map' of the Lake District. According to David J. Bodenhamer, 'deep mapping' offers 'a fresh conceptualization of humanities GIS'.[47] As Bodenhamer explains: 'In its methods deep mapping conflates oral testimony, anthology, memoir, biography, images, natural history and everything you might ever want to say about a place, resulting in an eclectic work akin to eighteenth and early nineteenth-century gazetteers and travel accounts'.[48] Many of the difficult-to-define Lake District landscape writings which we are digitizing and mapping neatly comply with Bodenhamer's definition of the 'deep map'; and our own layering of texts upon the geocentered literary GIS will lead to the further thickening of this sense of place. For Bodenhamer, though, authentically deep maps are predicated on several additional characteristics: 'They are meant to be visual, time-based, and structurally open. They are genuinely multimedia and multilayered. They do not seek authority or objectivity but involve negotiation between insiders and outsiders, experts and contributors, over what is represented and seen'.[49] A genuinely deep map of the Lake District, therefore, would necessarily involve the integration of a range of other geo-specific materials including landscape paintings, historical maps and oral histories. What is more, an authentically deep GIS would also need to provide opportunities for users to upload their own site-specific data and to contribute their own layers to the spatial palimpsest. Clearly, the creation of such an open and porous map presents challenges to the spatial humanities researcher. Do temporal and financial

parameters make it possible to incorporate *all* of the spatial narrative forms and genres to which Bodenhamer refers? Is it necessary to introduce a control mechanism to monitor the quality of contributions to the site? By extension, is it essential to create two types of mapping to differentiate between the scholarly and the user-generated: a process which would clearly problematize the democratization of the GIS? Yet, in spite of such cautionary notes, there remains much that is attractive about this democratic form of mapping which can go some way to illustrating how the process of place-making is founded upon a complex interpenetration of material and imaginative, official and vernacular geographies. The ongoing development of deep GIS, therefore, patently points the way towards new forms of practicing both digital humanities research and critical literary geography.[50]

5. Conclusion: Towards Geographical Text Analysis

We are approaching a situation where the georeferencing of very large quantities of text and, more importantly, their subsequent spatial analysis is becoming a possibility. These technical advances will bring with them some tensions as at least some of the approaches required to analyses very large datasets will be alien to current humanities research paradigms. Just because they are alien does not mean that they are wrong, however, and many lessons will have to be learned from the social sciences and elsewhere about their application. Human geography, which has a long tradition of both quantitative and qualitative approaches and has learned many lessons about the strengths and weaknesses of both, is particularly pertinent here.[51]

Briefly, however, there are implications at all three stages of geographical text analysis that this paper has considered. The key question at the georeferencing stage is how accurate does this have to be? The process of moving from an automatically extracted list of suspected place-names to a layer in which every place-name that occurs in the text is linked to an accurate coordinate while all other names are discarded requires a level of manual intervention that is time-consuming and expensive – resources that are consequently being removed from the process of research and interpretation. Thus, rather than attempting to track down every error, the researcher is likely to be faced with a situation where the major errors that distort the pattern have been spotted – a process that is relatively simple when data are mapped, as this is a good way of noticing unlikely patterns – and the research phase can begin. During this phase the researcher has to be aware that there will be errors in the data and that these will have to be updated as they are discovered. This may seem slightly

uncomfortable; however the social sciences are well used to dealing with situations where error in databases has to be managed rather than eliminated. Equally the humanities are well used to thinking critically about their sources and this approach merely extends this to the digital version.

In the analysis phase, corpus and GIS-based techniques can both be used. Both sets of approaches have some tools that could be characterized as 'broad but crude' and others that are 'in-depth but narrow.' The broad but crude techniques are those that attempt to quickly summaries the entire corpus, including techniques such as corpus-based frequency profiling or GIS-based density smoothing. They are broad in that they summaries the entire corpus but crude in that they are highly abstracted from the original texts. The in-depth but narrow techniques include the use of concordances and map-based querying, which re-structure the texts and present them in new ways. They are in-depth in that they still present the original texts to the researcher who has to develop his or her own understanding from them, and narrow in that using them is relatively slow and thus inevitably selective. Traditionally these approaches have been mutually antagonistic, but it should be clear that both have roles to play in understanding large quantities of georeferenced text. In particular, the broad techniques quickly summaries the corpus but are essentially descriptive. They can then point the researcher to where (both in terms of within the text and place) the in-depth techniques can be most usefully applied and these, in turn, provide more explanatory power. Referring back to the broader analyses then allows the researcher to ask 'where are these lessons also applicable and where appears to be different?' thus contextualizing findings and avoiding the risk of atomistic fallacy, where a lesson learned from a small subset of data or places is inappropriately applied to the whole dataset or study area.

Finally, the digitization and spatialization of a relatively large corpus of landscape writing characterized by 'transgeneric heterogeneity'[52] also highlights the integral role that GIS technology can play in the ongoing development of geocritical practice. The principal critical potentiality of GIS appears to be rooted in the facilitation of map-based reading of a corpus. That is to say, the large-scale mapping of significant quantities of texts can reveal abstract 'shapes, relations, structures', to apply Franco Moretti's cardinal terms, which demand further research and explanation.[53] This process can lead to a reconfiguration of the corpus by drawing attention to previously marginalized or even neglected texts which, in actual fact, make a striking contribution to the spatial narrative of a particular geographical location such as the English Lake District, the whole of Britain, or potentially the whole world. Crucially,

though, the exploratory methodologies set out in this chapter have been underpinned by the belief that the new reading practices which are encouraged by the emergence of large-scale digital corpora do not need to usurp traditional approaches to understanding texts. Instead, there is a need to think in terms of the new scales of reading which have been opened up by such digital corpora and to self-consciously reflect upon the ways in which the digital humanities researcher is able to move between freely the macro- and the micro-, the abstract and the concrete, the distant and the close.

Acknowledgements:

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant "Spatial Humanities: Texts, GIS and places" (agreement number 283850). Our thanks to Kirsten Hansen (Mt Holyoke College, USA) for her work on the Wordsworth material used in figure 3.

Text	Concordance
DutchDiur15	them. The General intends next week to march towards **Stirling**, and so for the hills. The Lord Craighall
FScout156	to be their Commander in chief, their coming near **Stirling**, and col. Lilburne's forcing them to the Hills
FScout156	(though the weather was unreasonable) to march from **Stirling** against them; but as he appeared, they quitted
FScout175	divers soldiers out of Leith, and about 50 from **Stirling** are gone to the Hills; from whence they descend down
FScout178	it is certified, That Gen. Monk is advancing towards **Stirling**, and intends to cut his passage through the Hills,
FScout180	, whereby he signified, That he was advancing beyond **Stirling** towards the Highlanders from whence he intends, after securing of
PerfAcc175	companies are come up. The General will be at **Stirling** next week. From Milford Haven, May 8. All
PerfAcc175	Scotland do advertise, that General Monk was lately in **Stirling**, and is now on his advance towards the Highlanders,
PerfAcc176	May 9. The General marcheth tomorrow from Dalkeith towards **Stirling**, and from thence to some of our frontier Garrisons and
PerfAcc176	General Monk hath yesterday removed his Headquarters from Dalkeith to **Stirling** from whence he intends, after securing some new passes
PerfAcc177	, provisions being so exceeding scarce in those parts. **Stirling** May 16. Thursday last the General came hither with part
PerfAcc177	to and from the hills, about 12 miles from **Stirling**, General Monk is marched that way to observe the several
PerfDiOc02	that General Monk is upon his march on this side **Stirling**, to join with us, his coming may prove very
PerfDiOc03	, May. General Monk is gone from hence to **Stirling**, at his coming to this city here was a great
PerfDiOc03	, that he was advanced a day's march beyond **Stirling** towards the Highlanders, and that he doubteth not (by

Figure 1: An example of a concordance based on 'Stirling' in the Lancaster Newsbooks Corpus. The "text" column on the left hand side provides information on which text the reference was taken from.

Wordform	Tag	Frequency

Oxford	NP1	38
Milford	NP1	28
Hereford	NP1	19
Bedford	NP1	18
Stafford	NP1	18
Seaford	NP1	14
Crawford	NP1	11
Hertfordshire	NP1	11

Figure 2: An example of a frequency profile for proper nouns containing the string 'ford' in the Lancaster Newsbooks Corpus. Proper nouns are tagged as NP1. Only words that occur more than ten times have been included.

[See attached file]

Figure 3: Proportional circles representing place-name instances from a small corpus of writing by William Wordsworth.

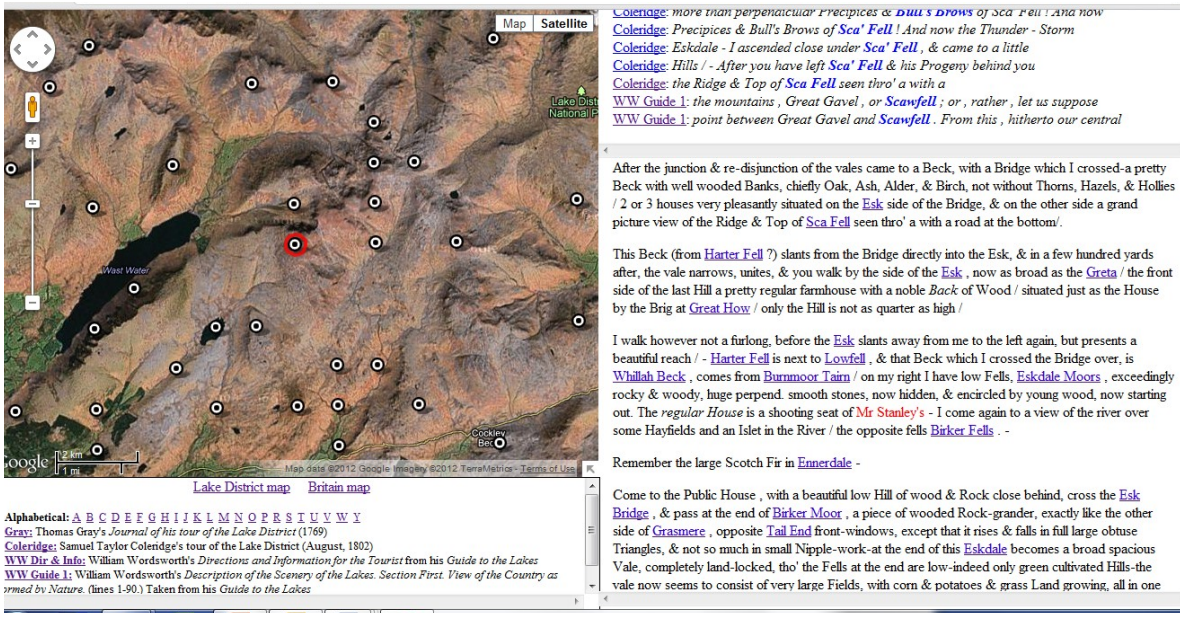

Figure 4: Map-based querying of a corpus of Lake District literature. See: http://www.lancs.ac.uk/mappingthelakes/v2 [14 Aug 2013].

[See attached files 5a and 5b]

Figure 5: Density smoothing of place-names mentioned in the Lancaster Newsbooks Corpus showing (a) all places mentioned and (b) places mentioned that collocate with words related to war (tagged as G3). A point indicates one or more mentions while the shading shows the density of mentions with darker shading indicating more mentions in an area.[54]

[1] In reality the situation is slightly more complex as approaches to the automated analysis of texts date back over thirty years; however, these are yet to be widely adopted particularly within the humanities and social sciences.

[2] Franco Moretti first articulated his model of 'distant reading' in 'Conjectures on World Literature', *New Left Review*, 1 (2000), 54-68. For a useful summary of the critical controversy generated by Moretti's methodological proposal, see R. Serlen, 'The Distant Future? Reading Franco Moretti', *Literature Compass* 7 (2010), 214-25.

[3] There are many introductions to GIS available that describe this basic structure. See for example: N.R. Chrisman, *Exploring Geographic Information Systems. 2nd edition* (Chichester: John Wiley, 2002); K.C. Clarke, *Getting Started with Geographic Information Systems* (Upper Saddle River, NJ: Prentice Hall, 1997); D. Martin, *Geographic Information Systems and their Socio-Economic Applications. 2nd edition* (Hampshire: Routledge, 1996).

[4] I.N. Gregory, K. Kemp and R. Mostern, "Geographical Information and Historical Research: Current Progress and Future Directions," *History and Computing*, 13 (2003), 7-21.

[5] See: I.N. Gregory and P.S. Ell, *Historical GIS: Technologies, Methodologies and Scholarship* (Cambridge: Cambridge University Press, 2007); or I.N. Gregory and R.G. Healey "Historical GIS: Structuring, Mapping and Analysing Geographies of the Past," *Progress in Human Geography*, 31 (2007), 638-653.

[6] D.J. Bodenhamer "The Potential of Spatial Humanities," in D.J. Bodenhamer, J. Corrigan and T.M. Harris, eds., *Spatial Humanities: GIS and the Future of Humanities Scholarship* (Bloomington: Indiana University Press, 2010), 14-30

[7] Good introductions to this field include S. Adolphs, *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies* (London: Routledge, 2006); T. McEnery and A. Hardie, *Corpus Linguistics: Method, Theory and Practice* (Cambridge: Cambridge University Press, 2012); and C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing* (Cambridge: MIT Press, 1999).

[8] See: I.N. Gregory and A. Hardie, "Visual GISting: Bringing together Corpus Linguistics and Geographical Information Systems," *Literary and Linguistic Computing*, 26 (2011), 297-314; I.N. Gregory and D. 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*, 3 (2009), 61-84; D. Cooper and I. N. Gregory "Mapping the English Lake District: A Literary GIS," *Transactions of the Institute of British Geographers*, 36 (2011), 89-108.

[9] http://www.lancs.ac.uk/spatialhum [14 Aug 2013].

[10] The account of corpus linguistics given in this section is drawn largely from: McEnery and Hardie, *Corpus Linguistics*.

[11] For example, the *IMPACT Project* (http://www.impact-project.eu) and *Early Modern OCR Project - eMOP* (http://emop.tamu.edu) as well as the *Text Creation Partnership* (http://www.textcreationpartnership.org) [all 14 Aug 2013].

[12] *The Lancaster Newsbooks Corpus*, http://www.lancs.ac.uk/fass/projects/newsbooks [14 Aug 2013].

[13] See R. Garside, G. Leech and T. McEnery, eds., *Corpus Annotation: Linguistic Information from Computer Text Corpora*. (London: Longman, 1997).

[14] L.L. Hill *Georeferencing: The Geographic Associations of Information*. (Cambridge, MA: MIT Press, 2006) provides a comprehensive overview of the use of gazetteers especially in chapter 5.

[15] http://www.geonames.org [14 Aug 2013]

[16] http://www.world-gazetteer.com [21 Feb 2012]

[17] http://www.getty.edu/research/tools/vocabularies/tgn/index.html [14 Aug 2013]

[18] http://www.edina.ac.uk/digimap [14 Aug 2013]

[19] A layer is effectively the GIS equivalent of a database table. The difference is that in a layer every row of data in the table (termed *attribute data*) is linked to a map-based location which is underlain by coordinates (termed the *spatial data*).

[20] C. Grover, R. Tobin, M. Woollard, J. Reid, S. Dunn and J. Ball, "Use of the Edinburgh Geoparser for Georeferencing Digitized Historical Collections," *Philosophical Transactions of the Royal Society A*, 368 (2010), 3875-3889 provides an example of this.

[21] See C. Gabrielatos and P. Baker, "Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press (1996-2005)," *Journal of English Linguistics*, 36 (2008), 5-38.

[22] P. Rayson, D. Archer, S.L. Piao and T. McEnery, "The UCREL Semantic Analysis System," *Proceedings of the Workshop on 'Beyond Named Entity Recognition Semantic Labelling for NLP Tasks' in Association with '4th International Conference on Language Resources and Evaluation (LREC)'*, Lisbon, 2004, 7-12.

[23] Ibid, 7.

[24] See http://www.lancs.ac.uk/mappingthelakes/v2 [14 Aug 2013] for a draft of the use of Google Maps. http://www.lancs.ac.uk/mappingthelakes/Interactive%20Maps%20Introduction.html [14 Aug 2013] provides an earlier example of how Google Earth can be used in the same way.

[25] D.J. Cohen and R. Rosenzweig, *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. (Philadelphia: University of Pennsylvania Press, 2006).

[26] T.C. Bailey and A.C. Gatrell, *Interactive Spatial Data Analysis* (Harlow: Longman, 1995).

[27] I.N. Gregory and A. Hardie, "Visual GISting: Bringing together Corpus Linguistics and Geographical Information Systems," *Literary and Linguistic Computing*, 26 (2011), 297-314.

[28] See: Gregory and Hardie, "Visual GISting," 297-314 and Cooper and Gregory, "Mapping the English Lake District," 89-108.

[29] Good texts on point pattern analysis include: Bailey and Gatrell, *Interactive Spatial Data Analysis*; A.S. Fotheringham, C. Brunsdon and M.E. Charlton, *Quantitative Geography: Perspectives on Spatial Data Analysis* (Sage: London, 2000); and C. Lloyd, *Spatial Data Analysis: An Introduction for GIS Users* (Oxford: Oxford University Press: Oxford, 2010).

[30] L. Anselin, "Local Indicators of Spatial Association – LISA" *Geographical Analysis*, 27 (1995), 93-115.

[31] A.S. Fotheringham, C. Brunsdon and M.E. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* (Chichester: Wiley, 2002).

[32] http://www.lancaster.ac.uk/mappingthelakes [14 Aug 2013].

[33] Cooper and Gregory, "Mapping the English Lake District," *Transactions of the Institute of British Geographers*, 89-108; and Gregory and Cooper, "Thomas Gray, Samuel Taylor Coleridge and Geographical Information Systems" *International Journal of Humanities and Arts Computing*, 61-84.

[34] For more on 'sentiment analysis' research see, for example, S. Piao, Y. Tsuruoka, S. Ananiadou, "Sentiment Analysis with Knowledge Resource and NLP Tools," *International Journal of Interdisciplinary Social Sciences*, 4 (2009), 17-28.

[35] The conceptualization of the archive as "a textual environment [which is] open to continuous transformation and development" owes a theoretical indebtedness to Jerome McGann's early exploratory work on the *Rossetti Archive*: J. McGann, *Radiant Textuality: Literature after the World Wide Web* (New York: Palgrave, 2001), 82.

[36] B. Westphal, *Geocriticism: Real and Fictional Space*s, trans. Robert T. Tally Jr. (New York: Palgrave Macmillan, 2011), 112-13.

[37] Westphal, *Geocriticism*, 117.

[38] Ibid., 122.

[39] E.S. Casey, *Representing Place: Landscape Painting and Maps* (Minneapolis: University of Minnesota Press, 2002), 351.

[40] Westphal, *Geocriticism*, 137.

[41] Ibid., 139.

[42] Adolphs, *Introducing Electronic Text Analysis*, 68-69.

[43] See, for example, A.J. Evans and T. Waters, "Mapping Vernacular Geography: Web-based GIS Tools for Capturing 'Fuzzy' or 'Vague' Entities," *International Journal of Technology, Policy and Management*, 7 (2007), 134-50. See also

on-going collaborative research by the Ordnance Survey: http://www.ordnancesurvey.co.uk/education-research/research/vernacular-geography.html [14 Aug 2013].

[44] M. de Certeau, *The Practice of Everyday Life*, trans. Steven Rendall (Berkeley: University of California Press, 1984), 115-30.

[45] F. Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* (London: Verso, 2005), 1.

[46] For and overview see: J. Pickles, "Arguments, Debates, and Dialogues: the GIS-Social Theory Debate and the Concern for Alternatives," in P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, eds., *Geographical Information Systems: Principals, Techniques, Management and Applications*. 2nd edition (Chichester: John Wiley, 1999), 49-60.

[47] D.J. Bodenhamer, "The Potential of Spatial Humanities," 27.

[48] Ibid., 27.

[49] Ibid, 27.

[50] A. Thacker, "The Idea of a Critical Literary Geography," *New Formations*, 3 (2005-6), 56-73.

[51] A number of good general readers on human geography are relevant here including: D.N. Livingstone, *The Geographical Tradition: Episodes in the History of a Contested Enterprise* (Oxford: Blackwell, 1992); R.J. Johnston, *Philosophy and Human Geography: An Introduction to Contemporary Approaches* (London: Edward Arnold, 1983); A. Holt-Jensen, *Geography: History and Concepts, A Student's Guide* 3rd edition (London: Sage, 1999)

[52] Ibid., 119.

[53] Moretti, *Graphs, Maps, Trees*, 1.

[54] Derived from Gregory and Hardie "Visual GISting," *Literary and Linguistic Computing*, 26, figs 3 & 5.