Open Archaeology 2015; 1: 166–175 DE GRUYTER OPEN

Original Study **Open Access**

Patricia Murrieta-Flores*, Ian Gregory

Further Frontiers in GIS: Extending Spatial Analysis to Textual Sources in Archaeology

Abstract: Although the use of Geographic Information Systems (GIS) has a long history in archaeology, spatial technologies have been rarely used to analyse the content of textual collections. A newly developed approach termed Geographic Text Analysis (GTA) is now allowing the semi-automated exploration of large corpora incorporating a combination of Natural Language Processing techniques, Corpus Linguistics, and GIS. In this article we explain the development of GTA, propose possible uses of this methodology in the field of archaeology, and give a summary of the challenges that emerge from this type of analysis.

Keywords: GIS, Digital Humanities, Spatial Humanities, Geographic Text Analysis, NLP, Corpus Linguistics, corpora, digital archaeology

DOI 10.1515/opar-2015-0010
Received December 19, 2014; accepted March 31, 2015

1 Archaeology, space, GIS and texts

What (happened), when, where and why, are some of the central questions that a researcher investigates when confronted with the archaeological record. Although the enquiries might seem simple in themselves, the answers to these questions constitute part of a complex puzzle that archaeologists put together in the form of rationalisations from the evidence available, becoming eventually interpretations. In this challenge, we get only fractions of the riddle, which correspond to partial aspects of the 'what', 'when', 'where' or 'why'. Epistemologically, it can be argued that in archaeology, dealing regularly with tangible evidence and the archaeological record, the spatial context –or 'where'-, usually turns out to be the point of departure for any further enquiry. Moreover, the relationships that such evidence sustain often become key to explain the reasons behind the observed phenomenon. In this sense, a particular distribution of settlements in the landscape might be explained through the observation of their spatial association to routes of communication, resources and connections to other dwellings, among many other factors. At a different scale, the location of particular material in an excavation in relation to other spaces might give hints on the actual use of such space. From these examples, is easy to realise that the spatial aspect of the archaeological record has a fundamental role in our discipline. As such, archaeology has considered and discussed space and place as central in almost all of its theoretical and methodological developments. The early adoption of Geographic Information Systems (GIS) and other spatial technologies in this field, is the result of this strong tradition in spatial thinking, the development at universities of graduated degrees in archaeology specialising in such technologies, testify for the matureness and advances in their use.

Article note: This article is a part of Topical Issue on Challenging Digital Archaeology.

***Corresponding author: Patricia Murrieta-Flores**: History & Archaeology Department, University of Chester, E-mail: p.murrietaflores@chester.ac.uk
Ian Gregory: History Department, University of Lancaster. i.gregory@lancaster.ac.uk

Employed constantly in Cultural Heritage Management and increasingly to address archaeological research questions, GIS and other spatial approaches in archaeology have reached an interesting level of sophistication over the past decade. It only takes a brief look at the proceedings from the international conference 'Computer Applications in Archaeology and Quantitative Methods', held every year and from which this collection of essays is derived, to realise this. Hundreds of papers addressing topics from the usefulness of descriptive spatial statistics to a combination of methodologies and theories such as Time Geography, 3D GIS, and Geographic Network Analysis have been written. These appear in the proceedings of this and other specialist conferences and also in many general scientific journals.

Although GIS has been used in archaeology for more than 30 years now, the refinement of the spatial methodologies used and the developments in the spatial thinking behind them, has not however, reached one particularly important but almost completely overlooked resource in this field: textual sources. What we mean by this is that despite great part of the field is informed by, deals with, and works at its core with textual sources, from medieval charts to chronicles and antiquarian reports, GIS in archaeology have been primarily used to analyse the spatial relationships of features, landscapes, and objects among others, that are recorded from the 'real' world. In this sense, although textual sources of archaeological interest contain geographic information that is invaluable to researchers, GIS has been rarely used to analyse the places portrayed in the textual realm. This is of great relevance because particular geographies play a substantial role in the narratives depicted in the large textual collections that we use to understand from economic and social developments in the Anglo-Saxon world, to the constitution and material expression of a national identity, and that are a constant source of information in the archaeological enquiry. This is not to say that there have not been historiographical or technological attempts in archaeology to understand geographies in text (see the works by [1] and [2]; and [3] on Natural Language Processing for archaeology). However, it is easy to wonder why the analysis of the spatial information *contained in* textual sources has not been more developed beyond traditional means using spatial technologies.

One reason for this could be that the mention of places in texts has been usually overlooked regarding them as implicit to the arguments or narratives on them. Another is that so far, GIS has been conventionally used for quantitative analysis and is thus not seen as relevant to research on texts. Borrowing from spatial literary criticism, researchers are increasingly realising of the significance of places in text and the role and importance they play in either historical or fictional narratives. In the past years, archaeologists have moved in this direction, making substantial progress in the visual exploration of place-names recorded on historical texts, where projects such as HESTIA and GAP have developed innovative Web-based resources for the examination of geographies in ancient literature [4-5]. However, in order to understand the meaning of the geographies depicted in a text and the realities underlying them, a combination of quantitative and qualitative methods is certainly necessary. The benefits of even basic mapping of the places mentioned in an historical source are easy to realise, and this might constitute a powerful visualisation of the geographical scope of a text, creating a general overview of its context. However, this can hardly be regarded as a full analytic approach. Recently, with the advance of Natural Language Processing (NLP) techniques it has become possible to incorporate corpora – as large bodies of text are called – within a GIS. This provides the first stage in allowing us to spatially analyse the corpus. Conducting this analysis effectively requires the development of an interdisciplinary approach enabling the understanding of space and place in textual sources through the integration of qualitative and quantitative methods from Geoscience, Linguistics, Computer Science, and other Humanities fields. In this paper we argue that this combination will prove beneficial to archaeological research. We demonstrate this by drawing on the experiences of other fields such as History and Literature that, like Archaeology, deal extensively with textual sources.

2 From exploration to analysis: extending GIS to textual sources

In recent years, particularly with the advent of Digital Humanities, archaeologists have started to realise the potential of the use of spatial technologies for the examination of texts. This has only been possible due to the creation of, and increasing online access to, digitised collections of historical and literary

documents and other textual sources. In addition, research leading to the development of important textual tools like the Edinburgh Geoparser [6], has made it possible to automatically annotate such texts, creating geo-referenced corpora – where the place names have been automatically identified and allocated to a coordinate-based location – that can be integrated with GIS. Pioneering projects in archaeology like Hestia [4] and Google Ancient Places [5] as said before, have developed Web resources, facilitating the study of geographies portrayed in ancient literature. Providing tools for powerful geographic visualisations integrated with the textual sources, these projects have extended the way in which geographies referenced in corpora can be explored.

The advances that these projects have brought in terms of visual geographic exploration of texts cannot be denied, however, it can also be argued that Archaeology has not yet moved into the realm of spatial analysis of corpora. In this sense, the development of Spatial Humanities [7], a field that has developed from Historical GIS, has achieved already some results. Moving away from simple mapping and traditional quantitative exploration of economic and social data, examples of spatial analysis of texts in the fields of History and Literature, have recently opened new ways for understanding patterns and overarching geographies that can be identified in corpora. For instance, to move beyond a cartographic approach, the project 'Spatial Humanities: Texts, GIS, Places' based at Lancaster University, is working to bring textual sources into GIS, and analyse them using quantitative methods combined with qualitative research that uses nuanced textual spatial analyses. In order to achieve this, it had to consider the traditional approach in Humanities-based fields to textual sources. In History and Literature, as in Archaeology, the regular method is to study a small number of sources in great detail, an approach termed *close reading*. In History this might mean to study a limited set of sources, in Literature a number of texts, and in Archaeology detailed documentation of a particular case study. There are several problems with this approach, two of the main ones are (1) that while close reading might allow a detailed understanding of a phenomenon, this is based in a very limited amount of information leaving much un-researched; and (2) it is difficult to know to what extent what we learnt from this restricted evidence is actually applicable to the broader pattern. A contrary approach, termed *distant reading*, can also be considered in which, rather than read individual texts, broad summaries derived from large numbers of texts are used [8]. However, taking large volumes of data and looking to summarise large amounts of textual material can also have limitations as the broader the material studied is, the shallower the research is in danger of becoming. The challenge then resides in using GIS to bring together the strengths of both approaches. While analysing large textual collections, the distant reading approach allows us to summarise overall patterns and identify particular points in the texts that call for subsequent close reading. The results from the close reading of these parts are, in turn, contextualised by the broader patterns. This allows us to identify, describe and potentially explain the geographical patterns within the texts.

By developing pilot studies [9], experimenting with combining approaches from corpus linguistics and GIS [10] and eventually developing tools, the Spatial Humanities project has created integrated methodologies through an interdisciplinary approach that combines methods and theories from Corpus Linguistics, Natural Language Processing (NLP), Geoscience, and the Humanities. This has resulted in some of the first examples of textual spatial analysis that combines distant and close reading.

3 Spatial Humanities: interdisciplinarity and experiences from History and Literature

In a recent study we used the digital collection of the Registrar General's Reports for England and Wales (http://www.histpop.org), which contain more than 200,000 pages of descriptions, census data and vital statistics to conduct a study of the Registrar General's interest in cholera. To do this we developed a methodology called Geographical Text Analysis (GTA), a combination of techniques from NLP and corpus linguistics, and GIS and spatial analysis. The goal was to test whether this approach could automatically identify textual and geographic patterns in the corpus, and also identify particular parts of the text associated

to cholera and related diseases that require close reading within the millions of running words. In this case the corpus had already been geoparsed using the Edinburgh Geoparser [6]. Once geoparsed, in order to identify those places related to cholera, we used a linguistic analysis called 'collocation'. This technique allows to establish whether a sequence of terms co-occur often. In this case we wanted to identify whether a key term was located in close proximity or 'collocated' to another in the text. Therefore, we developed a tool called the Geographical Collocates Tool (GCT) which focuses specifically on geography and detects where a particular word or search-term appears in the text near to one or more place-names (Figure 1).

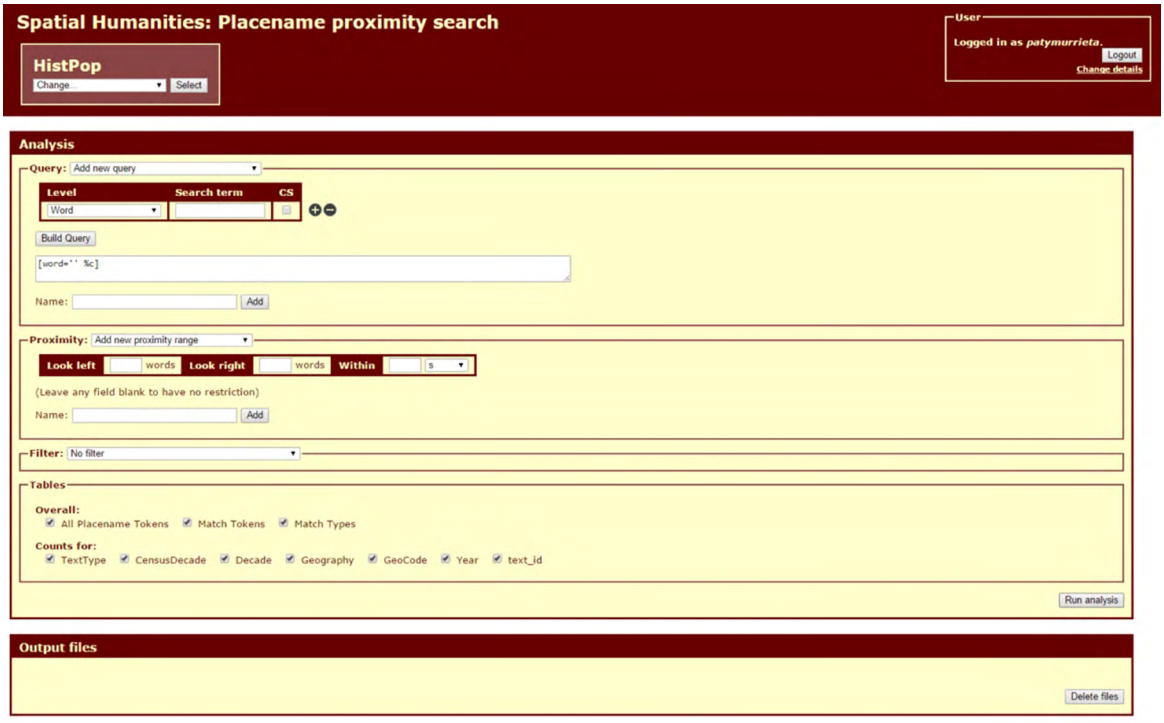

Figure 1: Interface screenshot of the Geographical Collocates Tool.

The exact process has been described elsewhere [11], but the assumption is that where place-names and the search-term occur near to each other, the search-term is being used in relation to the place. The exact definition of 'near' needs some experimentation, but typically will be within five or ten words or within the same sentence. As said, in the terminology of corpus linguistics this means that the place-name *collocates* with the search-term [12-13]. The tool then extracts: all of the collocates identified; their co-text – the text that surrounds the search term and place-name and thus provides context on what is being said; and any relevant metadata such as which volume the mention occurs in (Figure 2). It creates a table that is ready to transfer into GIS for further spatial analysis (Figure 3). From the table, a GIS point layer is created that shows every place that collocates with the particular search term. Patterns can then be analysed using a variety of spatial analytical approaches such as density smoothing, cluster detection and Kuldorff's spatial scan statistic [14].

Using 'cholera' as a search term, this approach allowed us to identify not only what the Registrar General described as the major (and minor) events and places related to this major cause of nineteenth century mortality, but also peculiar events in the history of the Registrar General Reports which have passed to some extent unnoticed, such as the creation by William Farr of a dedicated volume to investigate the disease (Figure 4). What this highlights is that GTA allowed us to effectively carry out geographic enquiries that comprehensively explored all mentions of this disease that occur in a very large volume of textual material. This would have taken years to achieve using close reading. At the same time, it enabled the recognition of particular portions of this corpus, such as the specialised study on cholera by William Farr

3014018: <text_id 27>: wn to the two last decen- naries ; when the public health has suffered from epidemics of influenza , --	cholera	-- , and other diseases 5 while emigration from the United Kingdom has proceeded at an accelerated ra
3015944: <text_id 27>: y , and life of the nation . The pestilences of the middle ages-the famine , the influenza , and the --	cholera	-- of modern times- are examples of one class of these agencies ; the security , and freedom which Eng
3032049: <text_id 27>: f Scotland " that the number of adults in the island was at one time reduced to 4 by small-pox ; and --	cholera	-- in the firs ; epidemic was fatal in this remote region . The dwellings of the poor people , who bre
3731618: <text_id 30>: , who have been left , -as well as their companions that have been taken , -by fever , consumption , --	cholera	-- , and the cloud of diseases that at present surround mankind , -stand like sad monuments of our mor
3753835: <text_id 30>: has been arithmetical . The assertion falls to the ground , that the disappearance of small-pox , of --	cholera	-- , or of other epidemics , must be followed immediately by famine , or by an increase of other disea
4239644: <text_id 356>: hypothetical , will not be thought wholly irrelevant . We must also remark , that the ravages of the --	Cholera	-- in 1832 and 1833 might perhaps be supposed to have operated as a considerable check to the increa
4270870: <text_id 356>: seases and causes of death to 93 . Owing to the length of time that had elapsed from the outbreak of --	Cholera	-- in 1832 and 1833-from the circumstance of so many poor and destitute persons , as well as whole fa
4270922: <text_id 356>: and surgical hospitals having been closed against the admittance of persons labouring under Asiatic --	Cholera	-- , the return of deaths from this disease either in the A. Forms or in the hospital registries was ,
4270963: <text_id 356>: might have been anticipated , very defective . To remedy this omission , a return of the deaths from --	Cholera	-- , that occurred in 1832 and 1833 , amounting to 25,378 , was procured from the Office of the Board
4278865: <text_id 356>: m that upwards the males are again in excess . It is remarkable , that with the exception of Asiatic --	Cholera	-- , two or more epidemics have generally prevailed in Ireland at the same time ; and that from the ea
4279942: <text_id 356>: 86 persons , the sexes being in the proportion of 100 males to 95.97 females . With the exception of --	Cholera	-- , this disease has proved more fatal in towns , and among large and closely united masses of the po
4287518: <text_id 356>: disease which then appeared was so malignant in character contrasted with common European or Bilious --	Cholera	-- , that all other records of this affection fall into comparative insignificance . For the reasons s
4287551: <text_id 356>: ification . For the reasons set forth at page ii . of this Report , the return of deaths from Asiatic --	Cholera	-- was naturally defective in those documents from which the notices of other diseases were obtained
4287678: <text_id 356>: iled most . xx CENSUS OF IRELAND FOR THE YEAR 1841 . TABLE showing the Number of CASES and DEATHS of	CHOLERA	-- , with their relative proportions , in the Rural and Civic Districts of the Counties of Ireland . 0
4287870: <text_id 356>: he Dublin hospitals ; subsequently a return was received of 272 deaths which happened in the Belfast --	Cholera	-- Hospital , the only good and extensive country return of the kind afforded by the Census inquiry ;
4288026: <text_id 356>: e a large portion of the deaths returned subsequent to that period , must have been from the Bilious --	Cholera	-- incidental to this country . In the records afforded by competent medical authorities from the Dubl
4288043: <text_id 356>: ncidental to this country . In the records afforded by competent medical authorities from the Dublin --	Cholera	-- Hospitals , the sexes were 100 males to 137 " 34 females . In the Belfast Hospital , however , the
4288179: <text_id 356>: Hart , Professor of Anatomy , Royal College of Surgeons , Dublin , and Medical Superintendent of the --	Cholera	-- Hospital , Townsend-street , in 1832 and 1833 , and subsequently to that in St. Peter 's parish , i
4288246: <text_id 356>: . REPORT UPON THE TABLES OF DEATHS . xxi Health , we learn , that upon its first outbreak in 1832 , --	Cholera	-- prevailed most in the towns of the civic districts ; and that in 1833 it had spread throughout the
4288429: <text_id 356>: cause fell rapidly till the years 1837-38 , when it again rose to 968 and 1,222 . In the returns of --	Cholera	-- received in the A. Forms , where the sexes have been specified , they are in the proportion of 100
4288518: <text_id 356>: e appearance of this epidemic , that although there may be a deficiency of the specified deaths from --	Cholera	-- , yet that several of the deaths recorded from that disease may have arisen from other causes ; thi
4288565: <text_id 356>: rity of the sexes afforded by the A. Forms , and those of the Dublin hospitals . During the years of --	Cholera	-- , Fever and other epidemic diseases , fell below the usual standard ; but the years 1832 and 1833 d
4290455: <text_id 356>: of Fever in Ireland for ten years , both in and out of hospital , not being much above 112,000 ; and --	Cholera	-- in its three years progress , carried off little more than 45,000 . Rutty notices an epidemic Fever
4290891: <text_id 356>: t of any pestilential Fever or other formidable epidemic occurring in Ireland , until the arrival of --	Cholera	-- in 1832-3 . The total deaths from Fever in Ireland , during the ten years included between June , 1
4291433: <text_id 356>: eared in 1708 ; 1718-21 ; 1728-31 ; 1740-43 ; 1763-64 ; 1771-73 ; and 1817-21 . In the years 1832-33 --	Cholera	-- took its place , but in 1837 it again appeared ; and the year 1842 has been marked by a most fatal
4301105: <text_id 356>: buted to , or indeed mistaken for , other diseases of the abdomen , as Inflammation of the Bowels or --	Cholera	-- Morbus . It is often mentioned in Irish MSS. as Maidhm seicne , -a bursting of the covering of the
4323025: <text_id 356>: Fever 46 , Insanity 39 , Diseases of the cavity of the abdomen 24 , Childbed 17 , Mortification 16 , --	Cholera	-- 9 , Concussion of the Brain 6 , and Cancer 1 . Of these Inquests 663 were held in Gaols and Prisons
4324807: <text_id 356>: ent appearance of other affections ; of the remaining 191 deaths from epidemic diseases , 73 died of --	Cholera	-- , and 98 of Fever . The diseases of the circulating and digestive organs , compared with those from
4331118: <text_id 356>: occurred from Erysipelas , 92 males and 46 females ; 73 cases of Small Pox ; 50 of Influenza ; 33 of --	Cholera	-- ; 37 of Croup ; 11 from Hydrophobia , 8 males and 3 females ; 5 from Glanders , 4 males and 1 femal
4333686: <text_id 356>: uring the period The late House of Industry at Clonmel g No satisfactory record previous to 1836. of --	Cholera	-- in.1833. merged into the Poor-house on its formation h Opened in 1840. s Closed in 1841 ; no record

Figure 2: Sample of the co-text output created from the GCT.

A	B	C	D	E	F	K	L	M	N
PN_Token	PN_StartIndex	PN_EndIndex	PN_LeftContext	PN_CorpusText	PN_RightContext	PN_enamex_gazref	PN_enamex_lat	PN_enamex_long	PN_enamex_name
70312	6372434	6372434	glish poor law , the suffering and	England	. Persons of advanced ages among all classes have been cut o	geonames:6269131	52.1604546	-0.703125	England
70313	6372835	6372835	; nor is it evident that the northe	London	exceeded by 3947 , -or 1-fifth part , the deaths in the winter q	unlock:9654368	51.51659393	-0.104514991	London
70314	6372859	6372859	ed by 3947 , -or 1-fifth part , the	England	were 19,452 , or nearly , in the same degree in excess of that s	geonames:6269131	52.1604546	-0.703125	England
70315	6372891	6372891	ess of that season . In the South-	Epsom	district suffered from scarlatina ; Guild ford from small-pox an	unlock:9700432	51.33030891	-0.27019611	Epsom
70316	6372904	6372904	lace of 8400 . The Epsom district	Farnham	from fever , measles , hooping cough , and diarrhoea . The dea	unlock:9698908	51.21092987	-0.790143132	Farnham
70317	6372930	6372930	hooping cough , and diarrhoea .	Bexley	sub-district in Kent there were as many as forty cases of small	unlock:9756679	51.45900345	0.108984184	Bexley
70318	6372933	6372933	arrhoea . The deaths for the first	Kent	there were as many as forty cases of small-pox at one time : c	unlock:9438867	51.52469254	-0.318665028	The Kent
70319	6372957	6372957	cases of small-pox at one time :	Elham	, Portsmouth , the Isle of Wight , Kintbury , Faringdon , and Fyf	unlock:9701146	51.14884186	1.110216856	Elham
70320	6372959	6372959	f small-pox at one time : only the	Portsmouth	, the Isle of Wight , Kintbury , Faringdon , and Fyfield . In the b	unlock:9492167	50.8087616	-1.070197552	Portsmouth
70321	6372962	6372964	e time : only those unvaccinated	Isle of Wight	, Kintbury , Faringdon , and Fyfield . In the barracks at Win che	unlock:9668650	50.67547798	-1.299371004	Isle of Wight
70322	6372966	6372966	ose unvaccinated died . Scarlatin	Kintbury	, Faringdon , and Fyfield . In the barracks at Win chester , occu	unlock:9662343	51.39573288	-1.446552753	Kintbury
70323	6372971	6372971	rlatina prevailed in Folkstone , El	Fyfield	. In the barracks at Win chester , occupied by about 2000 men	unlock:9693975	51.73487473	0.266653925	Fyfield
70324	6373067	6373067	ry cottage . The South Midland C	Oxford	25 deaths occurred from small-pox , and the deaths exceeded	unlock:9180537	51.75434494	-1.253751814	Oxford
70325	6373084	6373084	ts . In Oxford 25 deaths occurred	Cambridge	the mortality was high . In Leighton Buzzard the deaths were n	unlock:9733776	52.20480728	0.144198786	Cambridge
70326	6373091	6373092	small-pox , and the deaths exce	Leighton Buzzard	the deaths were nearly double the average . In some -districts	unlock:9660597	51.91989708	-0.654833645	Leighton Buzzard
70327	6373122	6373122	the Eastern Counties measles , s	Norwich	and several other places . The mortality in the South-western	unlock:9194466	52.62779045	1.3028844	Norwich
70328	6373148	6373148	mortality in the South-western D	Salisbury	the deaths " in the winter quarters of 18,53 and 1855 were 77	unlock:9462898	51.0730896	-1.79294318	Salisbury
70329	6373270	6373270	istrict may be lower in a cold tha	Truro	; typhus in Lerrin , Liskeard ; Plymouth and the surrounding dist	unlock:9356772	50.2597332	-5.051959038	Truro
70330	6373276	6373276	old than it is in a mild winter . Inf	Liskeard	; Plymouth and the surrounding districts are still in an unsatisfa	unlock:9659271	50.45374298	-4.458462954	Liskeard
70331	6373278	6373278	is in a mild winter . Influenza wa	Plymouth	and the surrounding districts are still in an unsatisfactory sanit	unlock:9498131	50.38795471	-4.145573378	Plymouth
70332	6373294	6373294	d ; Plymouth and the surrounding	Clifton	, and Cheltenham the mortality was above the average . The V	unlock:9726842	52.11619186	-2.226390362	Clifton
70333	6373297	6373297	nd the surrounding districts are s	Cheltenham	the mortality was above the average . The West Midland Cour	unlock:9729052	51.90057945	-2.079949498	Cheltenham
70334	6373339	6373339	he mortality was high in , Herefo	Gloucester	, Shrewsbury , Stafford , Worcester , and Warwick . 2094 deatl	unlock:9689020	51.86437225	-2.239719987	Gloucester
70335	6373341	6373341	was high in , Hereford , where m	Shrewsbury	, Stafford , Worcester , and Warwick . 2094 deaths were re gis	unlock:9440742	52.70746422	-2.747530341	Shrewsbury
70336	6373343	6373343	Hereford , where measles was e	Stafford	, Worcester , and Warwick . 2094 deaths were re gistered in Bi	unlock:9416281	52.80866623	-2.111278772	Stafford
70337	6373345	6373345	where measles was epidemic ; a	Worcester	, and Warwick . 2094 deaths were re gistered in Birmingham a	unlock:9284353	52.19711876	-2.212242126	Worcester
70338	6373348	6373348	as epidemic ; and somewhat abo	Warwick	. 2094 deaths were re gistered in Birmingham and Aston ; 112	unlock:9329885	52.28650665	-1.582137167	Warwick
70339	6373356	6373356	ge in Gloucester , Shrewsbury , St	Birmingham	and Aston ; 112 less than the deaths in the winter , quarter of	unlock:9756016	52.4849472	-1.860012591	Birmingham
70340	6373358	6373358	r , Shrewsbury , Stafford , Worce	Aston	; 112 less than the deaths in the winter , quarter of 1854 , but	unlock:9762686	52.50296783	-1.87476337	Aston

Figure 3: Sample of GIS-ready table created from the GCT.

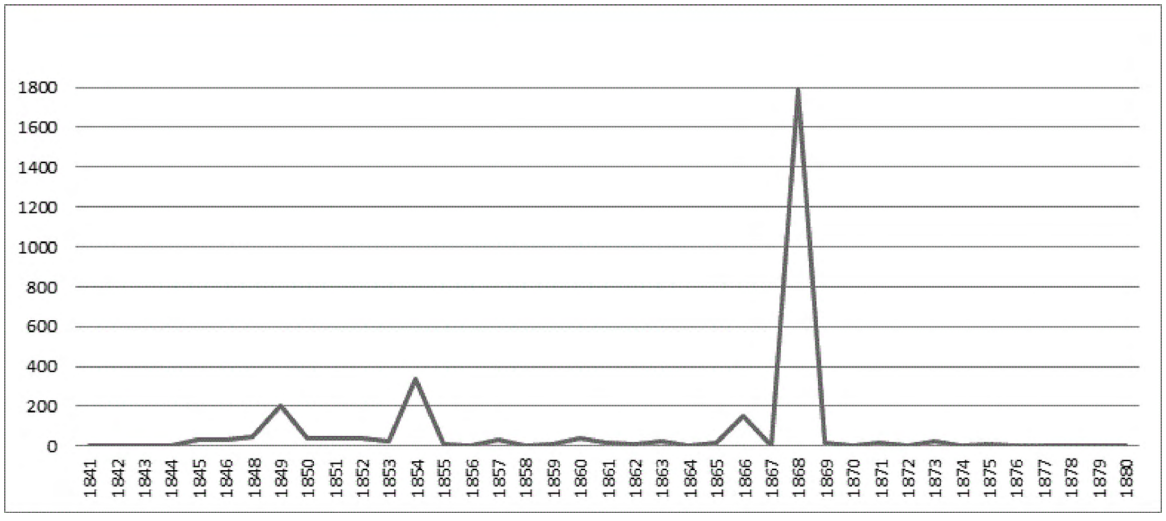

Figure 4: Place-name collocations of the word 'cholera' by year. Minor and major cholera events were identified (epidemics of 1849, 1854 and 1866) as well as the creation of Farr's dedicated volume to the disease published in 1868.

and specific sections within it that called for closer inspection. Further, it made it possible to identify the geographic patterns related to the diseases, highlighting places to which more emphasis was made within the texts as well as places that were not discussed as much as perhaps the incidence of the disease might lead us to expect. This gives us a new insight into the social context of cholera in the nineteenth century and particularly the Registrar General's perception of where and when it was important.

From a literary perspective we have created a corpus of nearly 1,500,000 words of writings about the English Lake District. This includes eighty texts covering the period from early writing in the seventeenth century to 1900. In this case we were interested in the ways in which different places were represented. To do this a list of adjectives and other words thought to be commonly used to describe Lakeland places was created. The places that these words collocated with were then identified. To ensure that we had all of the appropriate words, a second stage of collocation was then performed in which place-names became the search-terms and we explored their collocates for other descriptive words. In this way we could be confident that all descriptive words frequently related to places had been identified. The search-terms were then mapped and some comparisons done between the different patterns.

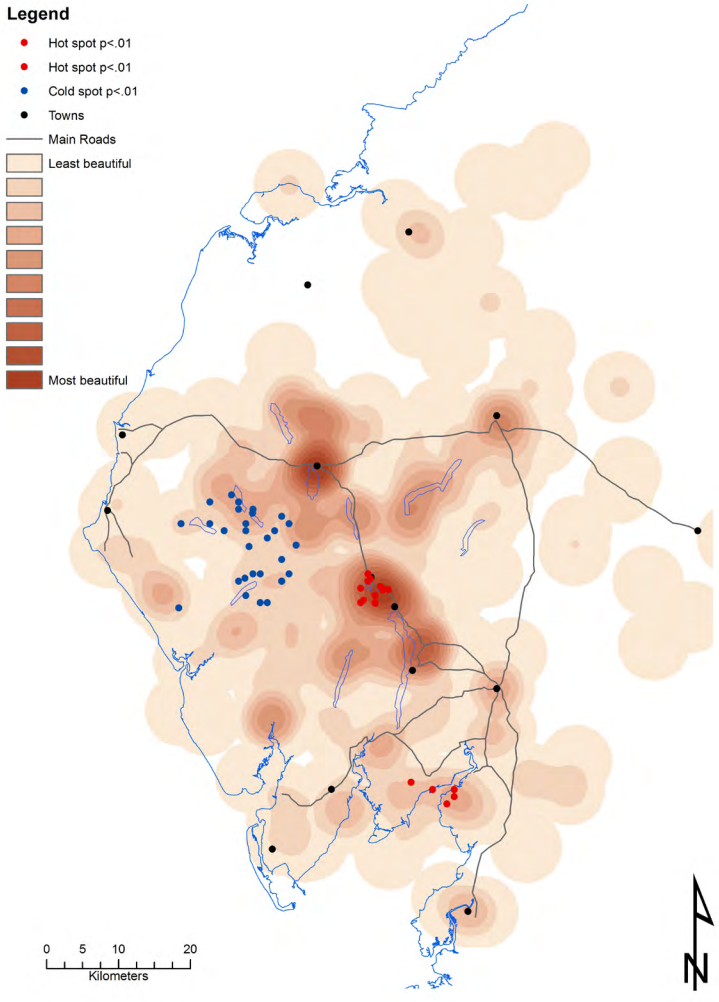

Figure 5: Places associated with the search-term 'beautiful' in the Lake District corpus. The overall distribution of place-names that occur in the same sentence as 'beautiful' are shown as the smoothed brown surface. Places the occur near to 'beautiful' more or less frequently than would be expected from the overall distribution of place-names within the corpus are shown as red and blue dots respectively.

One interesting result from doing this was found in comparing the patterns of three of the more commonly used adjectives: beautiful, sublime and dreary. As shown with the smoothed brown surface in figure 5, the pattern of places described as beautiful very closely follows the overall pattern of place-name mentions for the Lake District as a whole: concentrating on the main tourist hotspots such as Keswick, Grasmere, Ambleside and Windermere. It might be thought that this pattern could be caused simply because names from these places occur frequently in the corpus. To compensate for this Kulldorf's scan statistic was used to identify places that occur more or less often than would be expected from this background geography. These areas are shown in red and blue respectively in figure 2. It shows that the Grasmere area is called beautiful more often than would be expected while the area around the Kent Estuary, a relatively infrequently discussed part of the Lake District also appears as a hotspot. Interestingly the western fells, covering an area from Buttermere down to Wasdale are identified as being described as beautiful less often than would be expected. Comparing this pattern with the geography of place-names associated with sublime reveals an interesting contrast as the western fells are the very area that stands out as being characterised as sublime more often that would be expected. To further complicate matters, this area is also a hotspot of descriptions of dreary places.

Explaining these patterns requires referring in detail to the relevant parts of the relevant texts. Doing this reveals that beautiful is mostly associated with the most visited touristic areas. Only one writer, Edwin Waugh, frequently describes places such as Arnside, Silverdale, Hampsfell and Humphrey Head as beautiful. As these places are rarely mentioned by other writers his text makes the area around the Kent Estuary stand out. The western fells were (and still are) relatively inaccessible and require more effort to get to. Writers that did go there tended to describe them as sublime rather than beautiful suggesting a differing language being used by the more adventurous travellers. The western fells are described as dreary in four texts written by three authors between 1819 and 1833 suggesting perhaps that this type of landscape was not fashionable at this time, or that these writers drew heavily on each other's work.

Both examples show how GTA allows us first to identify patterns of words associated with places drawing on corpus linguistics to ask what words are associated with place-names, and on spatial analysis to describe the patterns. It then explains the patterns by referring back to the appropriate places within the texts for close reading. The approach has already proved fruitful and useful in the disciplines of History and Literature. In Archaeology, it is yet to be implemented, however, there are some ideas that might be developed in the future through which we expect to exemplify the potential value of this approach in forthcoming research as explained in the next section.

4 Possible uses of Geographical Text Analysis in Archaeology

As large digitised corpora such as the '19th century newspaper collection' (British Library) or the 'Colonial Archives of New Spain' (AGN-Mexico) become available, the possibilities for new geographic textual research in Archaeology grows exponentially. Invaluable information of archaeological and historical interest is contained in resources such as these and geographic textual approaches like the explained above, might open new venues of research and lines of collaboration.

For instance, in Britain, since the 17th century and particularly towards and during the 19th century, a precipitated rise of interest on prehistoric sites was experienced [15]. Anthropological and ethnographic accounts from abroad and areas perceived as remote within the country, fuelled the curiosity and imagination of the Victorian society which, in response, initiated research and archaeological excavations on an unprecedented scale. While many of the so called 'Royal Societies' conducted some of this research, much information survives in articles within the periodical press of the time. Using the 19th century newspaper collection of the British Library that includes much of the information where these kinds of news and articles were published, a Geographical Text Analysis like the ones described above, could provide for the very first time not only a full geographical picture of the archaeological activities carried out during Victorian times, but would also allow the semi-automatic extraction of important information regarding the excavations and research that took place during that time at a national scale. In addition,

using a combination of linguistic and spatial analyses it could shed light on the changing interpretations and evolution in the perceptions of the Victorian society regarding topics such as prehistory in general and particular prehistoric sites. As with this technique is possible to review in a semi-automated way thousands of documents identifying those places associated to key themes of our interested, one could trace for instance, all what was said on the news during the 19th century about not only an archaeological site, but also its association to particular ideas. Using this digital resource and in collaboration with the 'Past in its Place Project' based at Chester and Exeter universities, we are aiming to explore in the near future the 'emergence' in the Victorian imagination and public perception of prehistoric places such as Wayland's Smithy among others with this approach.

Another possible use of this kind of analysis could be the identification of archaeological sites and places of historical interest. The General Archive of the Nation in Mexico (AGN-Mexico) and the General Archive of Indies in Seville for instance, contain most of the administrative documents and social correspondence related to the establishment and management of the Spanish colonies in America, and particularly New Spain. The role that the development and importance of trades such as the sugar industry had in the economic and political spheres from the 16th to the 20th century, has been studied from the historical perspective, particularly in the modern territory of Mexico. In archaeological terms, however, only a few examples of sugar plantations have been excavated and there is little actual information in the national monuments record regarding the many sugar plantations of middle and small size that were founded and developed during colonial times. Archaeological recordings and explorations have rarely been carried out due, among other reasons, to the lack of information regarding the location of these plantations. This information exists in these archives. Nevertheless, just the collection of 'colonial institutions' of the AGN-Mexico contains millions of words in the thousands of documents available. The application of Geographic Text Analysis in these collections to look for the places once related to the colonial sugar industry, would be a significant way forward in the identification of sugar plantations, and therefore, this could lead to a vast improvement in the national monuments record related to historical archaeology. In addition, from spatial analyses looking at changes over time in the geographies of these plantations, it would be possible to investigate the emergent economic relationships and social changes within and due to this important industry during colonial times.

These are just two examples, but Geographical Text Analysis could have many other uses within archaeological research. With access to adequate resources such as articles and archaeological reports now added as linked data in projects like Europeana (http://www.europeana.eu/), in the future, using GTA might make possible to identify and analyse for instance, general trends in the interpretations given about particular material culture worldwide, or prehistoric periods at a national, regional or continental scale.

5 Challenges and Future

There is, therefore, clear potential for Archaeologists to use GIS for the exploration of texts. This type of research is, however, fairly new – there are still many obstacles to overcome and issues to address and improve. There are several areas of research and particular topics that need to be tackled as we advance in this field and we will mention here only few. The immediate challenges we recognise can be divided in three broad categories and are related to: (1) the accessibility of collections and digitisation of texts, (2) geoparsing textual sources, and (3) the analysis and interpretation of texts once they have been geoparsed. Starting with the accessibility of sources, clearly for this type of work to be feasible the researcher must have access to the appropriate text as digital texts. This means that they will either have been scanned and then optical character recognition (OCR) technology used to convert the images to text, or they will have been typed. Either way, the process is slow, expensive and error-prone. The volume of digital texts has, however, snowballed in the past few years with massive public and private investment. In the UK it is estimated that between the JISC, the AHRC, the Wellcome Trust and the New Opportunities Fund £137.5 million has been spent on the creation of digital resources, much of it textual. Although accurate figures are not available, this amount must be dwarfed by private investment from companies such as Google, Cengage, ProQuest

and BrightSolid [16]. Using these resources has proved more difficult than had perhaps been anticipated. There are some practical reasons for this: OCR quality is often poor [17], and many of the resources are only available through search interfaces that are limited and potentially give unpredictable results. Beyond this, however, there is the more basic issue that as yet, beyond keyword searching, we are still in the very early stages of knowing how to analyse very large volumes of text. We hope that GTA provides one example of how this might be achieved.

Geoparsing texts effectively is clearly vital if we are to place any trust in the results of a geographical text analysis. Evaluation of the Edinburgh Geoparser shows that while it is quite effective, the results are far from error free [18]. This is hardly surprising as there are a number of challenges to an automated geoparser including: determining what is and is not a place-name ("city of Lancaster", "Lancaster bomber", "Stuart Lancaster", "Duke of Lancaster"); disambiguating names that can refer to more than one place (there are variants of "Grizedale" in the east, south and north-west of the Lake District), and dealing with spelling variations which may be genuine or may result from data capture errors. While geoparsers and the gazetteers that they require to allocate coordinates to place-names can and will be improved, a significant amount of user intervention is still likely to be required to have full confidence in this process. One approach that has shown promise is that rather than geoparse the whole corpus, to only geoparse the text surrounding a particular search-term. The results can then be explored and corrected with the updates being saved to file and used to correct the results of subsequent searches [19].

Many of the analytical techniques required to analyse a geoparsed text exist however they are split between Corpus Linguistics, which focuses on the texts, and GIS and spatial analysis, which analyse their geographies. This is clumsy from a software perspective, and few people have skills in both subjects so more education and training is required. Training is increasingly becoming available through, for example, the Digital Humanities Summer Institute (DHSI) and the Lancaster Summer Schools in Interdisciplinary Digital Methods.[1]

There are also, perhaps, more fundamental issues around the analysis of texts in this way. As yet we are still in the very early stages of understanding how to pose research questions based on large corpora and how to fully understand and interpret the answers. There is clearly much potential as the above case studies from the disciplines of History and Literature show. In the case of Archaeology, the approach is yet to be adopted, however, we believe that this will happen in the not too distant future. Although there is still much to be done, the meetings organised by the Spatial Humanities Project in the summer of 2013 and the round table from which this edited collection emerges, are testimony of the interest in the possible applications that the GTA approach might offer to archaeology. As said, we are still in the early days of these developments and although relatively uncertain, the path ahead is certainly exciting.

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, places" (agreement number 283850).

References

[1] L. Isaksen, 'The application of network analysis to ancient transport geography: A case study of Roman Baetica', *Digital Medievalist*, vol. 4, 2008.

[2] J. Baker and S. Brookes, 'Outside the gate: sub-urban legal practices in early medieval England', *World Archaeology*, vol. 45, no. 5, pp. 747–761, Dec. 2013, doi: 10.1080/00438243.2013.865330.

[3] S. Jeffrey, J. Richards, F. Ciravegna, S. Waller, S. Chapman, and Z. Zhang, 'The Archaeotools project: faceted classification and natural language processing in an archaeological context', *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1897, pp. 2507–2519, Jun. 2009.

[4] E., Barker, S., Bouzarovski, C., Pelling, L., Isaksen, 'Mapping an ancient historian in a digital age: the Herodotus Encoded Space-Text-Image Archive (HESTIA)', *Leeds International Classical Studies*, vol. 9, no. 1, 2010.

1 See: http://www.dhsi.org and http://ucrel.lancs.ac.uk/summerschool respectively [viewed 25/9/14].

[5] E., Barker, K., Byrne, L., Isanksen, E., Kansa, N., Rabinowitz, Google Ancient Places. (2012). at http://googleancientplaces.wordpress.com/2012/02/25/ the-story-continues/
[6] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, and J. Ball, 'Use of the Edinburgh geoparser for georeferencing digitized historical collections', *Phil. Trans. R. Soc. A*, vol. 368, no. 1925, pp. 3875–3889, Aug. 2010.
[7] T. Harris, J. Corrigan, and D. Bodenhamer, 'Challenges for the Spatial Humanities: Toward a Research Agenda', in *The Spatial Humanities: GIS and the Future of Humanities Scholarship*, Bloomington: Indiana University Press, 2010, pp. 167–176.
[8] F., Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, 2005.
[9] D., Cooper and I., Gregory, 'Mapping the English Lake District: A literary GIS', *Transactions of the Institute of British Geographers*, vol. 36, no. 1, pp. 89–108, 2011.
[10] I. N. Gregory and A. Hardie, 'Visual GISting: bringing together corpus linguistics and Geographical Information Systems', *Lit Linguist Computing*, vol. 26, no. 3, pp. 297–314, Jan. 2011.
[11] P. Murrieta-Flores, A. Baron, I. Gregory, A. Hardie, and P. Rayson, 'Automatically Analyzing Large Texts in a GIS Environment: The Registrar General's Reports and Cholera in the 19th Century', *Transactions in GIS*, 2015, doi: 10.1111/tgis.12106.
[12] S. Adolphs, *Introducing electronic text analysis*. New York: Routledge, 2006.
[13] T. McEnery and A. Hardie, *Corpus linguistics: method, theory and practice*. Cambridge ; New York: Cambridge University Press, 2012.
[14] M. Kulldorff, 'A spatial scan statistic', *Communications in Statistics - Theory and Methods*, vol. 26, no. 6, pp. 1481–1496, Jan. 1997.
[15] M. D. Eddy, 'The prehistoric mind as a historical artefact', *Notes Rec. R. Soc.*, vol. 65, no. 1, pp. 1–8, Mar. 2011.
[16] T. Hitchcock, 'Confronting the Digital: Or How Academic History Writing Lost the Plot', *Cultural and Social History*, vol. 10, no. 1, pp. 9–23, Mar. 2013.
[17] S. Tanner, T. Munoz, and P. Hemy Ros, 'Measuring Mass Text Digitization Quality and Usefulness : Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive', *Dlib Magazine*, vol. 15, no. 78, 2009.
[18] R., Tobin, C., Grover, K., Byrne, J., Reid & J., Walsh, Evaluation of georeferencing. in 1 (ACM Press, 2010). doi:10.1145/1722080.1722089
[19] C. J. Rupp, P. Rayson, I. Gregory, A. Hardie, A. Joulain, and D. Hartmann, 'Dealing with heterogeneous big data when geoparsing historical corpora', 2014, pp. 80–83.