# Geographical Text Analysis:
# GIS approaches to analysing large volumes of texts

Ian Gregory, [1] Alistair Baron,[2] Patricia Murrieta-Flores,[1]
Andrew Hardie,[3] Paul Rayson[2] and C.J. Rupp[1]

[1]Department of History, Lancaster University, Lancaster LA1 4YT
Tel. +44 (0)1524 594967
Email, I.Gregory@lancaster.ac.uk Web, http://www.lancs.ac.uk/spatialhum
[2]School of Computing & Communications, Lancaster University, Lancaster, LA1 4WA
[3]Department of Linguistics & English Language, Lancaster University, Lancaster, LA1 4YL

KEYWORDS: Natural Language Processing; Spatial Analysis; Corpora; Texts

## 1. Introduction

Traditionally GIS, and indeed IT more generally, has been associated with quantitative sources which have been analysed in human geography using social science approaches. Even Qualitative GIS has not really changed this as many of the techniques it uses are based around attempting to structure the unquantifiable, for example fear of crime, in ways that then allow social science-style approaches to be applied (Cope and Elwood 2009; Kwan 2008). This has meant that researchers with suitable data have adopted GIS, however to many other geographers and people beyond the discipline it has simply been seen as irrelevant. While the bulk of the content produced for IT was quantitative this was undesirable but perhaps inevitable. In recent years, however, there has been a huge, and largely under-reported, shift in IT that means that the bulk of digital content that is now being created is text including digital libraries and archives up to the size of Google Books, and email archives, the world wide web, social networks and so on.

This provides a major opportunity and a major challenge. The opportunity is to make use of a vast array of new content and to bring geographic analyses to a wide new audience. The challenge is that while analysing numeric data is well established through the use of statistics, analysing large *corpora* – as collections of digital text are called – is more difficult. The traditional way of analysing these – close reading of the texts – is too slow for corpora that contain millions or billions of words, however the more mechanistic approaches used by many quantitative techniques cannot hope to deal with the subtlety and nuance that is essential if a text is to be properly understood.

This paper presents some initial research that begins to resolve this dilemma. It brings together GIS-based spatial analysis approaches with Natural Language Processing (NLP) techniques that enable the automated analysis of text. By using techniques from both fields we are able to both summarise the broad geographies within the text and also identify which parts of the text need to be read in more detail.

## 2. The data

Geo-referencing a text has received a attention from a number of authors and it is not the intention to describe the process here beyond saying that candidate place-names are identified using NLP techniques, they are extracted, linked to a gazetteer to provide coordinates, and subsequent processing is done to resolve disambiguation issues and resolve errors (see Gregory and Hardie, 2011; Grover et al, 2010; Yuan, 2010). The more interesting question, and the one explored here, is what does one do to analyse a geo-referenced corpus in ways that can both provide broad summaries and also allow us to explore in detail what is forming these patterns.

To explore this, this paper uses the Histpop collection (http://www.histpop.org). These are the printed volumes that accompany the census and Registrar General's reports from 1801-1937. Of particular interest are the Registrar General's reports from 1851-1911 for England and Wales. The Registrar General collected and published statistics on births, marriages and deaths and paid particular attention to mortality. This period was of vital importance as it marked the start of the mortality decline that was to characterise the 20th century, the causes of which remain controversial but which had pronounced and poorly understood geographies (see Gregory 2008; Szreter 1991; Woods et al 1988 and 1989; Woods and Shelton 1997). This corpus contains around 2.5 million words: too much to read from start to finish or to get a clear impression of geography from. The corpus was georeferenced by Claire Grover and her colleagues at the University of Edinburgh (Grover et al, 2010).

## 3. Identifying *where* a corpus is talking about

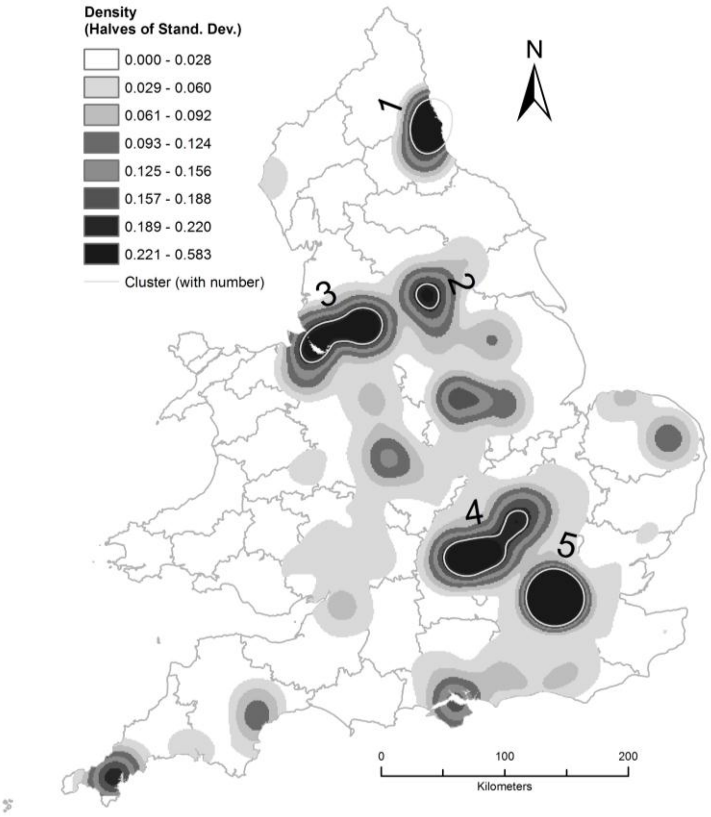


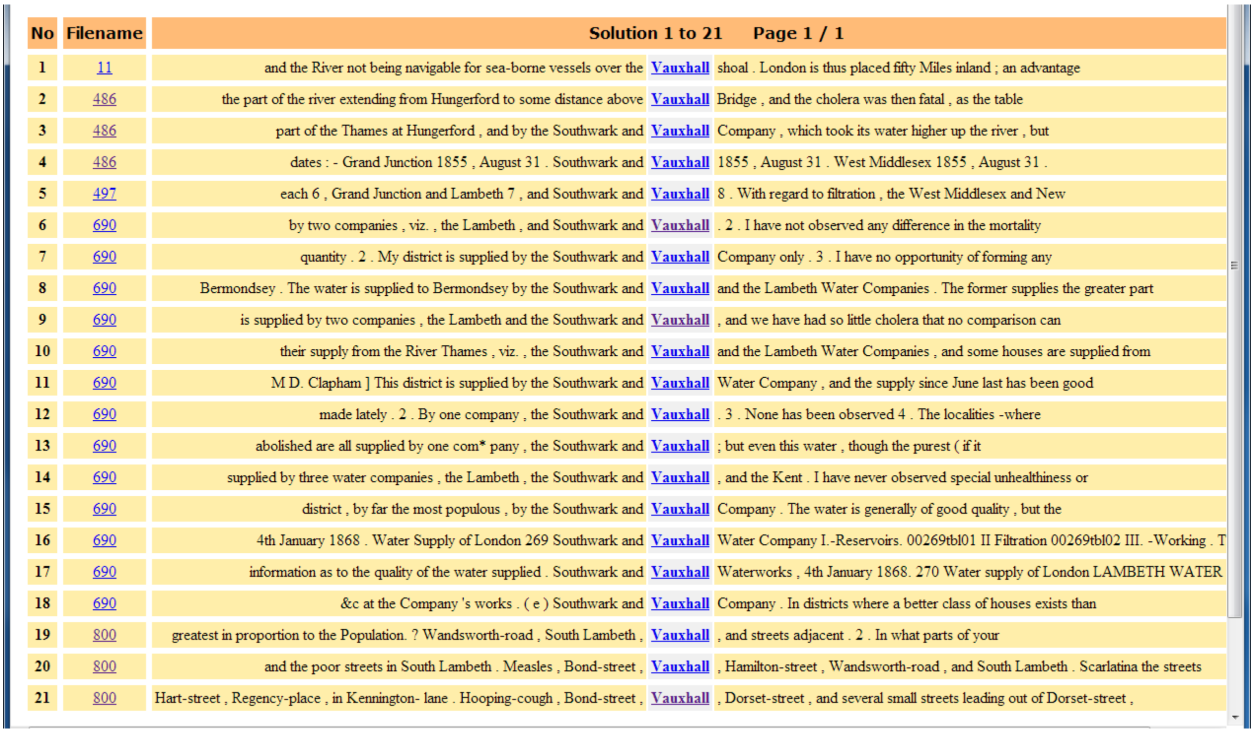
**Figure 1.** Clusters of place-name instances from the Registrar General's reports for the 1850s.

**Table 1.** Alternative ways of ranking the importance of place-name instances

| Word Frequency Count | Kernel Density |
|---|---|
| London | London |
| Manchester | Holborn |
| Liverpool | Shoreditch |
| Nottingham | Vauxhall |
| Leicester | Kennington |

A geo-referenced corpus is, in effect, a point pattern in which each *place-name instance* – or mention of a place-name – is represented as a point. Spatial analysis provides an obvious starting point to exploring such patterns. Figure 1 shows an example of this where kernel density analysis has been used to smooth the pattern and clusters have then been based on standard deviations above the mean. This provides us of a map-based summary of where the corpus is talking about. We are likely to also want to know what particular place-name instances make up these clusters or, indeed, other parts of the pattern. One approach to this would be to use Electronic Text Analysis techniques (Adolphs 2006) such as word frequency counts to identify which place-names occur most often. These are shown in the left hand column of table 1 which shows the five most common place-names – the spellings have been standardised in a gazetteer – that occur in the corpus. The right-hand column of table 1 shows an alternative way of measuring the importance of places within the gazetteer. These are ranked by the kernel density scores derived by overlaying the point locations of the place-name instances with the density smoothed surface from figure 1. Using this approach, all of the top-five place-name instances are within London. Alternative measures of clustering such as Gi* statistics could also have been used. While both approaches have limitations such as choice of bandwidth and the fact that ‘London’ is represented by a single point as are places within London, they provide alternative ways of identifying which places the corpus regards as most important. The spatial analysis-based approaches have the advantage of identifying the areas that are most concentrated on rather than the individual place-names.

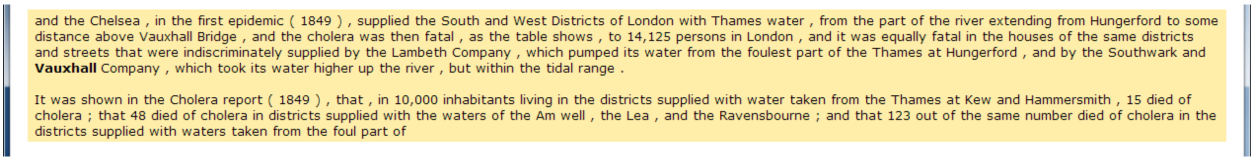
## 4. Identifying *what* a corpus is saying about places



| No | Filename | Solution 1 to 21 Page 1 / 1 | | |
|---|---|---|---|---|
| 1 | 11 | and the River not being navigable for sea-borne vessels over the | Vauxhall | shoal . London is thus placed fifty Miles inland ; an advantage |
| 2 | 486 | the part of the river extending from Hungerford to some distance above | Vauxhall | Bridge , and the cholera was then fatal , as the table |
| 3 | 486 | part of the Thames at Hungerford , and by the Southwark and | Vauxhall | Company , which took its water higher up the river , but |
| 4 | 486 | dates : - Grand Junction 1855 , August 31 . Southwark and | Vauxhall | 1855 , August 31 . West Middlesex 1855 , August 31 . |
| 5 | 497 | each 6 , Grand Junction and Lambeth 7 , and Southwark and | Vauxhall | 8 . With regard to filtration , the West Middlesex and New |
| 6 | 690 | by two companies , viz. , the Lambeth , and Southwark and | Vauxhall | . 2 . I have not observed any difference in the mortality |
| 7 | 690 | quantity . 2 . My district is supplied by the Southwark and | Vauxhall | Company only . 3 . I have no opportunity of forming any |
| 8 | 690 | Bermondsey . The water is supplied to Bermondsey by the Southwark and | Vauxhall | and the Lambeth Water Companies . The former supplies the greater part |
| 9 | 690 | is supplied by two companies , the Lambeth and the Southwark and | Vauxhall | , and we have had so little cholera that no comparison can |
| 10 | 690 | their supply from the River Thames , viz. , the Southwark and | Vauxhall | and the Lambeth Water Companies , and some houses are supplied from |
| 11 | 690 | M D. Clapham ] This district is supplied by the Southwark and | Vauxhall | Water Company , and the supply since June last has been good |
| 12 | 690 | made lately . 2 . By one company , the Southwark and | Vauxhall | . 3 . None has been observed 4 . The localities -where |
| 13 | 690 | abolished are all supplied by one com* pany , the Southwark and | Vauxhall | ; but even this water , though the purest ( if it |
| 14 | 690 | supplied by three water companies , the Lambeth , the Southwark and | Vauxhall | , and the Kent . I have never observed special unhealthiness or |
| 15 | 690 | district , by far the most populous , by the Southwark and | Vauxhall | Company . The water is generally of good quality , but the |
| 16 | 690 | 4th January 1868 . Water Supply of London 269 Southwark and | Vauxhall | Water Company I.-Reservoirs. 00269tbl01 II Filtration 00269tbl02 III. -Working . T |
| 17 | 690 | information as to the quality of the water supplied . Southwark and | Vauxhall | Waterworks , 4th January 1868. 270 Water supply of London LAMBETH WATER |
| 18 | 690 | &c at the Company 's works . ( e ) Southwark and | Vauxhall | Company . In districts where a better class of houses exists than |
| 19 | 800 | greatest in proportion to the Population. ? Wandsworth-road , South Lambeth , | Vauxhall | , and streets adjacent . 2 . In what parts of your |
| 20 | 800 | and the poor streets in South Lambeth . Measles , Bond-street , | Vauxhall | , Hamilton-street , Wandsworth-road , and South Lambeth . Scarlatina the streets |
| 21 | 800 | Hart-street , Regency-place , in Kennington- lane . Hooping-cough , Bond-street , | Vauxhall | , Dorset-street , and several small streets leading out of Dorset-street , |

**Figure 2.** Concordances on the word 'Vauxhall'

This enables us to identify *where* a corpus is talking about both in terms of the map patterns and the place-name instances that make up these patterns, however we want to go beyond this to ask *what* the corpus is saying about these places. The simplest way of doing this involves using a *concordance* which presents a list of the words occurring around a particular search term. This allows a quick assessment to be made about what is being said about this place. Figure 2 presents a concordance for 'Vauxhall', one of the place-names that has among the highest densities of place-name instances surrounding it but a low words frequency count. The concordance reveals that most of the 21 instances of 'Vauxhall' occur in relation to the Southwark and Vauxhall Water Company. This is being investigated in relation to the quality of its water supply with suggestions that this is in turn prompted by its relationship with disease. These suggestions can be examined further by following hyperlinks to the full text where the instance occurs, as shown in figure 3 which is a broader concordance of the text from row 3 in figure 2.



and the Chelsea , in the first epidemic ( 1849 ) , supplied the South and West Districts of London with Thames water , from the part of the river extending from Hungerford to some distance above Vauxhall Bridge , and the cholera was then fatal , as the table shows , to 14,125 persons in London , and it was equally fatal in the houses of the same districts and streets that were indiscriminately supplied by the Lambeth Company , which pumped its water from the foulest part of the Thames at Hungerford , and by the Southwark and **Vauxhall** Company , which took its water higher up the river , but within the tidal range .

It was shown in the Cholera report ( 1849 ) , that , in 10,000 inhabitants living in the districts supplied with water taken from the Thames at Kew and Hammersmith , 15 died of cholera ; that 48 died of cholera in districts supplied with the waters of the Am well , the Lea , and the Ravensbourne ; and that 123 out of the same number died of cholera in the districts supplied with waters taken from the foul part of

**Figure 3.** Expanded text from row 3 of figure 2.

This simple approach can be expanded further to create much more sophisticated queries. For example, we might want to create a concordance of all of the place-name instances from the clusters in figure 1 and explore what the key themes that are being discussed in relation to these clusters are and whether the texts is referring to similar themes for each cluster or whether there are differences between them. We might also want to compare the clusters (individually or as a group) with the background pattern. As well as the simple concordances shown here, more sophisticated statistical approaches to explore instances of different words can also be used, such as for example, log-likelihood statistics.

By implication, what we are doing here introduces another NLP concept, that of *collocation*. Collocation simply asks the question 'what words occur near this search term?' The importance of this from a geographical perspective is that we can ask either what words are collocated with a place-name or, alternatively, what place-names collocate with search terms of interest. One technical issue within this is how can we define 'near', however for our purposes here we are simply defining 'near' as within the same sentence. As sentences are a major linguistic building block this has some justification but further research will be conducted to test the implications of this.

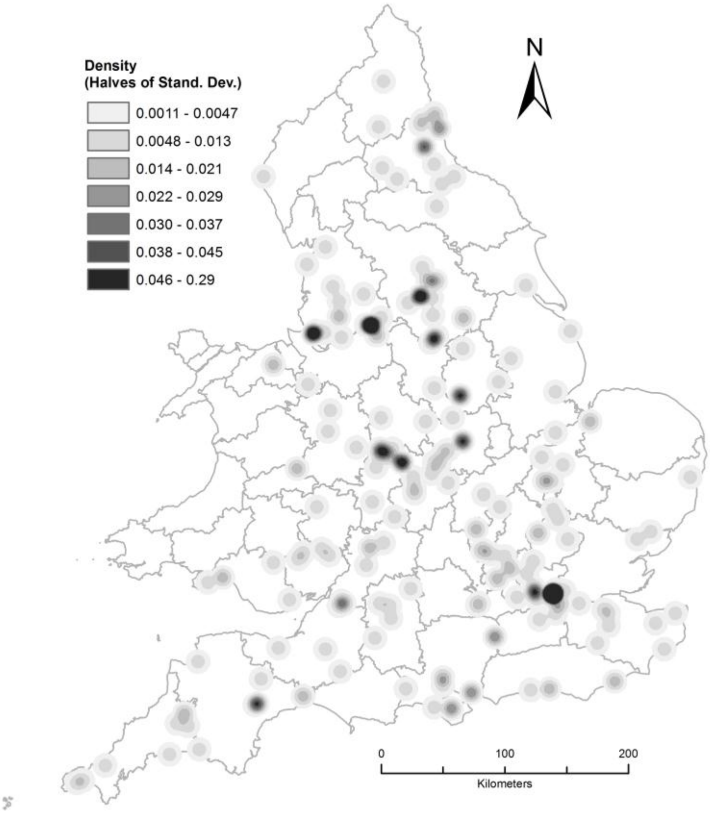


**Figure 4.** The distribution of places that collocate with 'measles'

Collocation can be used to explore what themes are associated with a particular place or cluster of places either qualitatively in the manner shown above or in more quantitative ways using statistics such as log-likelihood to explore how unusual these particular collocates are compared to the corpus as a whole. It can also be used to explore what places are associated with a particular theme. The literature tells us that infectious diseases were among the major killers of infants and children in this period (see, for example, Woods and Shelton, 1997). This is supported by NLP analysis that shows that 'measles' and 'diarrhoea' were among the most common disease terms found in the corpus for the 1850s. Figure 4 is thus a density smoothed map of place-name instances that collocate with the word 'measles.'

This is thus a simple map of where the Registrar General was most interested in in relation to this particular disease. From here a number of further analyses can be done. The Registrar General also published a wide range of statistical information on mortality including numbers of deaths from diseases sub-divided into districts. Simple correlation analysis allows us to compare whether there is a relationship between place-name instances that collocate with a disease and actual deaths from the disease. In general we have found that they correlate well suggesting that the places that the Registrar General focussed on in his reports were, in fact, the places with high death rates from the disease.

One possible criticism of the map above is that if the word measles was randomly located through the text it might be expected to produce the pattern shown simply because this is a random subset of place-name instances through the corpus. Again, spatial analysis can help us with this. Techniques such as the Kulldorf spatial scan statistic (Kulldorff, 1997) can be used to identify where there are clusters of place-name instances that collocate with 'measles' compared with the background distribution of either place-name instances in the corpus or perhaps the background population distribution as measured by census data. As in section 3, once these clusters have been identified, we can explore them in more detail using concordances and the underlying text to close read what the text is actually saying about these places.

## 5. Conclusions

By bringing together the field of spatial analysis and NLP we are able to explore texts in ways that ask explicitly geographical questions. These can be used to summarise the broad geographies within the corpus, the geographies associated with particular words or themes, the themes associated with particular clusters of instances, and so on. It can also be used to compare patterns of perception, as recorded in texts, with empirical patterns recorded statistically. Drawing on examples beyond the current one, this could be used to explore whether, for example, patterns of writing about crime match actual crime statistics or whether government documents are bias towards certain types of areas, such as urban areas or affluent areas.

We are in the early stages of this work but we believe that it has major implications. If we can analyse large bodies of text in a GIS environment then the potential uses of the technology start to spread way beyond its traditional boundaries and many of the criticisms of GIS being overly empirical and descriptive can start to be overcome.

## 6. 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). We are very grateful to Claire Grover (University of Edinburgh) for providing the georeferenced Histpop corpus to us, and the Richard Deswarte for his advice on this collection.

## References

Adolphs S (2006) *Introducing Electronic Text Analysis* Routledge: London

Cope M and Elwood S (2009) *Qualitative GIS: A mixed methods approach* Sage: London

Gregory IN (2008) Different places, different stories: Infant mortality decline in England & Wales, 1851-1911 *Annals of the Association of American Geographers* **98** pp 773-794

Gregory IN and Hardie A (2011) Visual GISting: Bringing together corpus linguistics and Geographical Information Systems *Literary and Linguistic Computing* **26** pp 297-314

Grover C, Tobin R, Woollard M, Reid J, Dunn S and Ball J (2010) Use of the Edinburgh geoparser for georeferencing digitized historical collections *Philosophical Transactions of the Royal Society A* **368** pp 3875-3889

Kulldorff M (1997) A spatial scan statistic *Communications in Statistics: Theory and Methods* **26** pp 1481-1496

Kwan M-P (2008) From oral histories to spatial narratives: Re-presenting the post-September 11 experiences of the Muslim women in the USA *Social & Cultural Geography* **9** pp 653-669

Szreter S (1991) The GRO and the Public Health Movement in Britain, 1837-1914 *Social History of Medicine* **4** pp 435-463

Woods RI, Watterson PA and Woodward JH (1988) The causes of rapid infant mortality decline in England and Wales, 1861-1921. Part I *Population Studies* **42** pp 343-366

Woods RI, Watterson PA and Woodward JH (1989) The causes of rapid infant mortality decline in England and Wales, 1861-1921. Part II *Population Studies* **43** pp 113-132

Woods RI and Shelton N (1997) *Atlas of Victorian Mortality* Liverpool University Press: Liverpool

Yuan M (2010) Mapping text. In Bodenhamer D, Corrigan J and Harris T (eds.) *The Spatial Humanities: GIS and the future of humanities scholarship*. Indiana University Press: Bloomington pp 109-123

## Biography

Alistair Baron is a research fellow in the School of Computing and Communications at Lancaster University. He previously worked as a research associate on the *Spatial Humanities: Text, GIS, Places* project. His research interests include Natural Language Processing with noisy data, Corpus Linguistics, Named Entity Recognition and Machine Learning.

Ian Gregory is professor of Digital Humanities in the Department of History at Lancaster University. He is PI on the ERC-funded *Spatial Humanities: Text, GIS, Places* project. His main research interests concern applying GIS to the humanities.

Andrew Hardie is a lecturer in Corpus Linguistics in the Department of Linguistics and English Language at Lancaster University. He is Deputy Director of the ESRC *Centre for Corpus Approaches to Social Science*. His main research interests are corpus-based methodologies, especially quantitative methods and corpus annotation, especially grammatical. He is a lead developer of the Open Corpus Workbench (CWB) and CQPweb software.

Patricia Murrieta-Flores holds a PhD in Archaeology and currently works as a research associate for the *Spatial Humanities: Texts, GIS, Places* project. Her research is focused in the development of theories and methodologies of spatial analysis for Archaeology, History and Literature. She is also interested in theories of perception of place, space and time.

Paul Rayson is a senior lecturer in Computer Science at Lancaster University. He is director of the UCREL interdisciplinary research centre which carries out research in corpus linguistics and natural language processing. His applied research is in the areas of online child protection, learner dictionaries, and text mining of historical corpora and annual financial reports.

C.J. Rupp is a Computational Linguist, currently working as a research associate for the *Spatial Humanities: Texts, GIS, Places* project. His interest is in the application of Natural Language Processing and Language Technology to large corpora in research domains, such as Chemistry, Biomedicine and, now, Literary Studies, combined with the development of semantic search applications appropriate for researchers, and other users, in that domain.