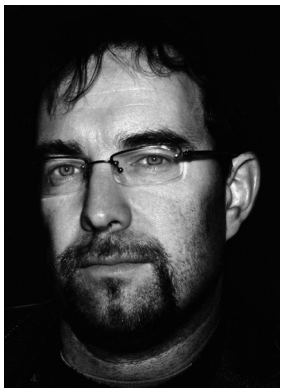
Freeing up digital content with text mining: new research means new licences



ALASTAIR DUNNING
Digitization Programme Manager
JISC

IAN GREGORY
Senior Lecturer in Digital Humanities
Lancaster University

ANDREW HARDIE
Lecturer in Corpus Linguistics
Lancaster University

The method by which users have traditionally exploited digital resources such as Early English Books Online (EEBO) has been via keyword search. However, researchers are increasingly finding new ways to exploit entire corpora of digitized resources, treating the resource as a single entity to be analysed, rather than searching or sifting through the resource for individual parts.

This article looks at the work of one research team at the University of Lancaster, exploring how they are using a corpus of seventeenth-century newsbooks to leverage open new areas of research. Using tools borrowed from linguistics and geography, the researchers can analyse the place names mentioned in the newsbooks and see which linguistic concepts (e.g. war, money) were associated with which geographical areas.

Such work has implications not only for future research but also for the resource managers to negotiate and manage the licences related to such resources.

The past ten years have seen significant amounts of spending on the digitization of cultural and scholarly material, both in the UK and abroad. Historic pamphlets, fine art images, medieval manuscripts, news-film footage and parliamentary papers are amongst some of the many digital collections developed with both public and private sector funding and made available for educational usage[1].

The impulse behind such digitization is quite straightforward. Previously, interested parties had limited access to such material, and it was a time-consuming process, perhaps involving poor quality microfilm, long-distance travel, or, as was usually the case with undergraduates, no access to the material at all. Now that digitization has been eagerly undertaken by many cultural institutions, a whole range of crucial primary evidence can be accessed whether it be from halls of residence, the seminar room, the university library or the private home.

Such mass digitization projects have created colossal amounts of data. Take the Research Library UK (RLUK)'s Nineteenth-Century Pamphlets project: over 26,000 pamphlets, comprising just over a million pages.[2] Those building the web interfaces for such digitized data have their work cut out to

develop something that can allow the user to forge through this huge jungle of data and arrive at the exact information they need.

Understandably, digitization projects have relied on a tried and tested model for allowing users to search through such material. The paradigm of the library catalogue, where users can search and browse through the index and whittle down the answers to a handful of suitable books, has been the dominant mode of search.

So, for example, the Nineteenth-Century Pamphlets project allows users to enter a search term according to their specific interests or browse through the particular collections of pamphlets digitized under the project.

The use of optical character recognition (OCR) within such projects gives added leverage to query such resources. Not only is the pamphlet title indexed, but every single word within the pamphlets is open for the user to search over, offering an exceptionally clear window to view individual aspects of the digitized resource.

Despite OCR, the philosophy remains the same; users undertake searches on particular words or phrases, or browse pre-defined categories to analyse such resources.

However, as the development of digital resources matures and, more broadly, as technology continues its inexorable process, creators and users of digitization projects are looking for other ways to analyse and exploit this huge jungle of words. While traditional searching and browsing allows for rich exploration of the content, it is restrictive in other senses. The catalogue paradigm only allows the user to look at one digital object, such as a pamphlet or newspaper page, at a time. The collection itself may be a vast paradise of digital evidence, but searching and browsing only offers a very partial sight of the entire collection. It is like looking at the collection through a keyhole.

So scholars exploiting digital resources have been looking at ways in which new technologies and software allow them to exploit the content in a holistic way, rather than piece by piece. Text mining and the exploitation of geographical information via geographical information systems (GIS) are two popular methodologies. Text mining is a family of methodologies that allow for the computer-assisted analysis of text for varying purposes, while scholars involved in GIS exploit geographical data (e.g. names of places, regions, countries, and their related co-ordinates) to visualize and analyse data.

In order to make use of such methodologies, search and browse access is not enough; scholars need access to the entire resource. Ideally, a data set can be downloaded to a scholar's own computer, from which they will use a range of software and related tools to analyse and re-visualize the data.

As we shall see, this has implications not just for scholarly enquiry but also for the information providers and gatekeepers. The typical catalogue paradigm for accessing digital content is no longer useful, for the access required is to the content as a whole rather then through a search interface. The data needs to be al fresco rather than being seen through a keyhole.

To achieve this, those involved in creating and licensing digital content, if they wish to respond to these scholarly advances, need to start to think in a radically different way about the provision of digitized content.

Rather than look at the many different ways that text mining and GIS can interrogate digital content, the rest of this article presents a case-study of the work of one particular interdisciplinary team.

Ian Gregory and Andrew Hardie are at the same institution (the University of Lancaster), but have different subject backgrounds; Gregory in using GIS in historical research and Hardie in linguistics. Nevertheless they have a mutual interest in seeing how new methodologies can leverage open new areas of research.

Their particular research interests demand the exploitation of the entirety of a digitized collection. The resource in this case-study is not a particularly large one, but it has the obvious merit of being downloadable and searchable, and without any restrictive licensing conditions. Therefore, the collection is fully amenable to the kind of novel methodologies in discussion.

The resource in question is the Lancaster Corpus of Newsbooks, two collections of seventeenth-century English news pamphlets, held at the British Library. The digitized corpus, which consists of 312 files, one for each document, includes every surviving newsbook from mid-December 1653 to the end of May 1654, and is a total of 800,000 words in size.[3]

Despite differences in physical appearance, such seventeenth-century newsbooks were not unlike contemporary newspapers – ephemeral articles and snippets of news reporting on national and international points of interest. These include births of princesses, deaths of queens, results of wars and

rebellions, news from abroad and various curiosities of interest to the seventeenth-century reader.

Using a traditional catalogue paradigm, a scholar would be able to search through the newsbooks and uncover titbits or incidental details, or indeed reports on more significant events. One of the newsbooks in the corpus offers details of how Queen Christina of Sweden abdicated; another that the English imports arriving via Dunkirk have been blocked.

But Gregory and Hardie wanted to look at some larger issues, treating the corpus in a holistic way, and seeing if it could open up new areas for scholarly inquiry for those working on the seventeenth century. Employing new technology and their particular skill sets allowed them to do this.

The Newsbooks Corpus incorporates, of course, only a very small sample of the entirety of printed material from the 1600s. However, as a *comprehensive* collection of news text (albeit for a fairly short window of time), it is an excellent test-bed for procedures that may be applied to data sets that are comprehensive across much greater periods.

The initial stage in a large-scale analysis of the content of the entire corpus is the application of various forms of automatic textual analysis to the text of the newsbooks. Linguists have been employing text parsing techniques as a staple part of their research for many years. Increased computer power just serves to further facilitate this approach on ever-larger bodies of data.

Using tools such as CLAWS[4], which incorporates embedded dictionary resources, it is possible to parse an entire corpus of text, and make sophisticated interpretations of the grammatical purpose of each word, identifying which words are proper nouns, which are adjectives, which are prepositions, and so on.

An example from the newsbooks demonstrates how such parsed text is created (Figure 1).

In the top half of the image is a snippet of original text from a news pamphlet. In the lower half, each word has been tagged and categorized as part of a larger grammatical family. So for example, usages of proper nouns, such as cities or people, are tagged with NP1 (*Patrick*, *Liverpool*), while common nouns in the plural are marked with NN2 (*Ships*, *Seas*). The analysis is not always entirely correct (for example, one example of *Brest* in the example above has been analysed correctly as NP1, while the other has been incorrectly analysed as JJT (an adjective in superlative form) – but software such as CLAWS typically succeeds in tagging around 95–97% of the words in a text accurately.

When the researcher is happy with the quality of interpretation being made by the software tool, the grammatical categories for each word can be recorded within a copy of the original text file, thus enriching the entire document.

Once this process is complete, the enriched data set can be handed on for geographical analysis (although in practice, the data set will be passed

```
In the Patrick of Liverpoole - which we lately
recovered from the Brest men of War - was one Walter
Roche - who was to carry her to Brest - and he
informed us - that there are these Ships following
belonging to Brest - who do so vex us in these Seas -
viz.
```

```
<p> In_II the_AT <em> Patrick_NP1 </em> of_IO <em>
<reg orig="Liverpoole"> Liverpool_NP1 </reg> </em> ,_,
which_DDQ we_PPIS2 lately_RR recovered_VVN from_II
the_AT <em> Brest_JJT </em> men_NN2 of_IO War_NN1 ,_,
was_VBDZ one_MC1 <em> Walter_NP1 Roche_NP1 </em> ,_,
who_PNQS was_VBDZ to_TO carry_VVI her_PPHO1 to_II
<em> Brest_NP1 </em> ,_, and_CC he_PPHS1 informed_VVD
us_PPIO2 ,_, that_CST there_EX are_VBR these_DD2
Ships_NN2 following_II belonging_VVG to_II <em>
Brest_NP1 </em> ,_, who_PNQS do_VD0 so_RR vex_VVI
us_PPIO2 in_II these_DD2 Seas_NN2 ,_, <em> viz._REX
</em> </p>
```

Figure 1. Creation of parsed text from the original text of a news pamphlet

back and forward between the two, as the interpretations are improved) and, in this example, Gregory could now begin to experiment with how they could exploit the geographical aspect of the tagged corpus. He began by creating a related database that contained every instance of a place name, drawn from the family NP1 for proper nouns, within the corpus, matching up place names to the glossary of places names made freely available on the website (http://www.world-gazetteer.com/). Following the identification of the place names, they were able to use further information from the same website to give each place name geographical co-ordinates. Finally they turned to another piece of software (called ArcGIS) which allowed them to visualize all the instances of place names in map-based form.

Once all this was in place, Gregory and Hardie were able to produce a map showing every instance of a place name within their selected corpus. The figure (Figure 2) below shows a screenshot from ArcGIS where the places named in the corpus have been mapped using a technique called *density smoothing*. This is a technique that is highly effective at identifying clusters of observations at or near each location. Effectively, the more place names that occur in an area, the darker the shading becomes. Using this, one can see which towns and cities crop up most frequently in the corpus of newsbooks.

As one might expect for material printed in Britain, British towns feature heavily, but there are other interesting discoveries. Paris, towns in The Netherlands, and Hamburg occur often, as do Rome, Venice, Naples and Stockholm – the latter related, no doubt, to the abdication of Queen Christina. Romania gets scattered references whilst other areas of Eastern Europe are untouched. A degree of 'noise' is present in the map due to errors made in different stages of the software processing. For example, place-name ambiguity can add misleading points to the map: frequent mentions of *Newcastle*, referring to the town in north-east England, produce a cluster of points in south-west Ireland, where there is a town of the same name. However, importantly, such 'noise' does not drown out the overall pattern which emerges from the large mass of data points extracted from the newsbooks. Such problems can, furthermore, be dealt with by making additional refinements to the automatic analysis procedure.

Such analysis only took the researchers so far; what really interested them was being able to produce such maps in the context of particular thematic areas, e.g. when the newsbooks spoke about military issues (featuring words such as *war*,

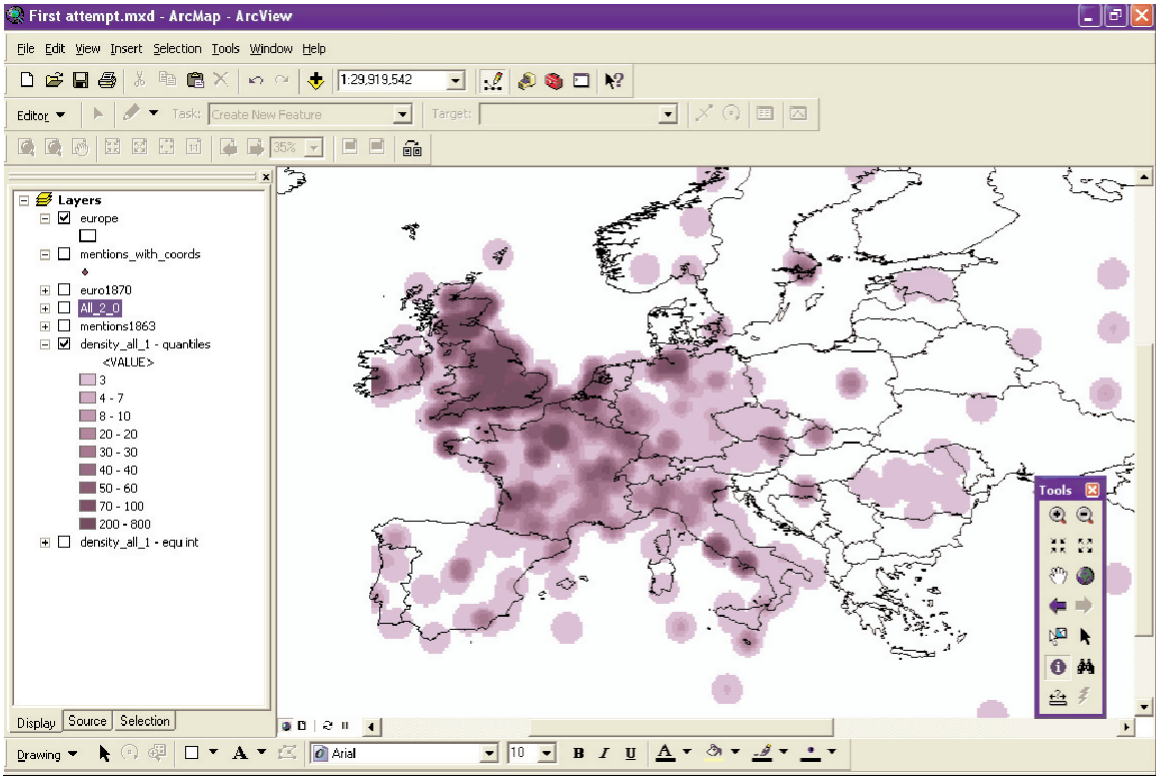


Figure 2. Screenshot from ArcGIS showing use of density smoothing to identify place-name clusters

soldier or *rifle*), what were the cities that were being cited in context?

Asking such questions meant returning to the corpus and undertaking a second round of tagging, marking up the corpus not only according to grammatical structure, but semantic structure as well.

Using a system called USAS[5], which is embedded with specially prepared thesauri, an entire corpus of text can be parsed semantically. This system makes sophisticated interpretations of the meanings of individual words, thus grouping particular words into larger families of concepts. The advantage of this is that it allows the scholar to search not just for instances of precise words within a corpus, but for particular concepts. So, for example, a scholar searching for *war* will, with a semantically parsed text, not only be able to locate instances of the string 'war' but also conceptually related words, like *military*, *battle*, *violence*, etc.[6] A short example of (part of) a sentence in which the place name *Dunkirk* is mentioned follows:

... two_N1 ships_M4 from_Z5 Dunkirk_Z2
have_Z5 brought_M2
Men_S2.2m Arms_B1 ,_PUNC
and_Z5 Ammunition_G3 to_Z5
Middleton_Z1mf ,_PUNC
but_Z5 of_Z5 all_N5.1+ that_Z8 join_A2.2
with_Z5 them_Z8mfn we_Z8 hear_X3.2
of_Z5 few_N5- that_Z8 have_A9+
estates_M7 to_Z5 subsist_A3+ on_M6 ...

In this example, there are a number of words which are tagged as indicating no particular concept. For instance, we see that grammatical words such as *of* or *to* are under category Z5; proper nouns are under Z1 or Z2. However, the content words have tags indicating particular conceptual fields. For instance, *ships* is classified as M4 (shipping, swimming, etc.), *men* is classified as S2.2m (People:- Male) and *ammunition* as G3 (warfare, defence and the army; weapons). A search for the *warfare* concept would therefore retrieve this sentence (and the associated mention of *Dunkirk*), even though the word *war* itself is not mentioned.

Tagging the semantic category of each word within the corpus thus provides a much richer set of evidence for the researcher to exploit. Gregory took this enriched data set and asked some more focused questions about the cities featured in the newsbooks.

Gregory could undertake this by searching through the corpus for all words related to the concept of *war* (which will have been tagged with category G3 by the USAS system) and then see which cities are cited within a particular distance (for example within five words' distance of the cited word).

The following map (Figure 3) is the result, when the data was exported into ArcGIS and smoothed.

This map shows clearly that the rebellion that was under way in Scotland at the time received considerable attention in London and that these mentions were concentrated on the eastern side of Scotland around the Forth and the Tay. Some of the Channel ports such as Dunkirk and Ostend are mentioned heavily in response to the ongoing Anglo-Dutch war, as is the nearby town of Clermont. Mentions from Brest and Hamburg are mainly a consequence of stories reported *via* these towns, rather than stories that are about these places (i.e. via newsbooks published there).

Similarly, the following two images (Figures 4 and 5) are the result of analysing the corpus for cities mentioned in the context of the concept *money*, labelled I1 in the USAS framework, and the

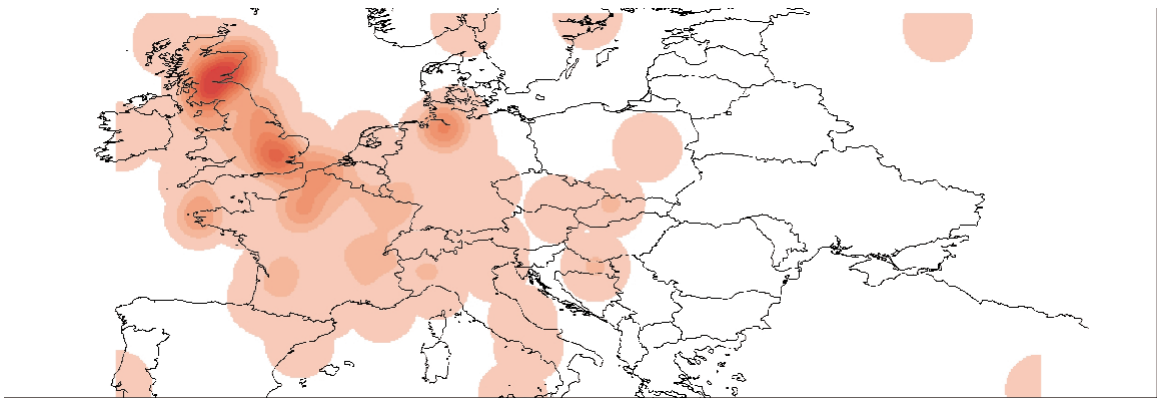
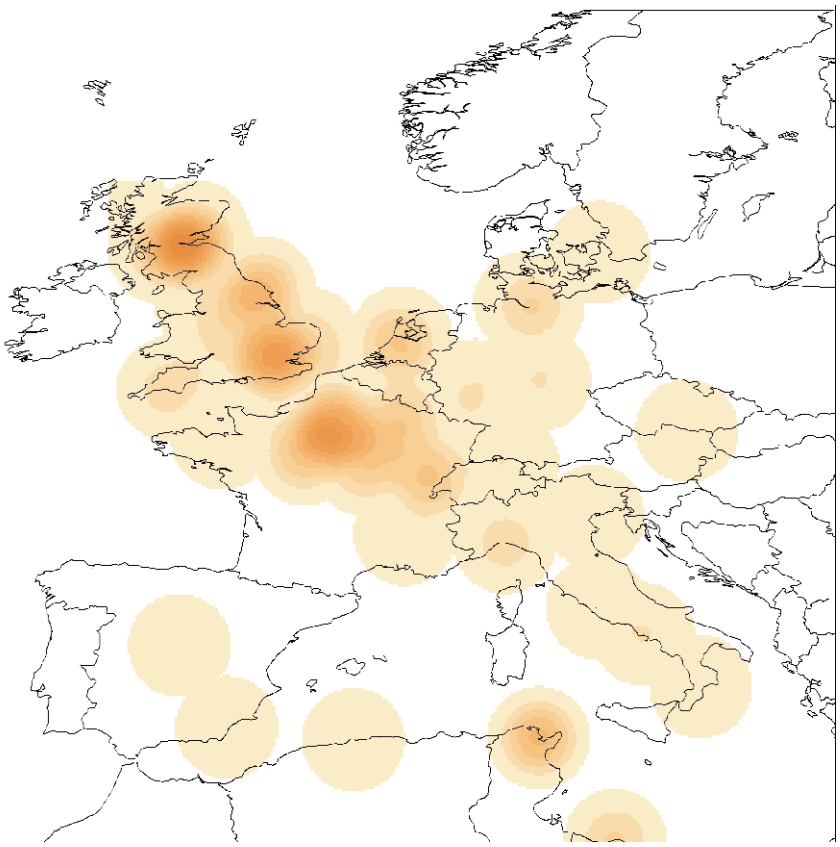
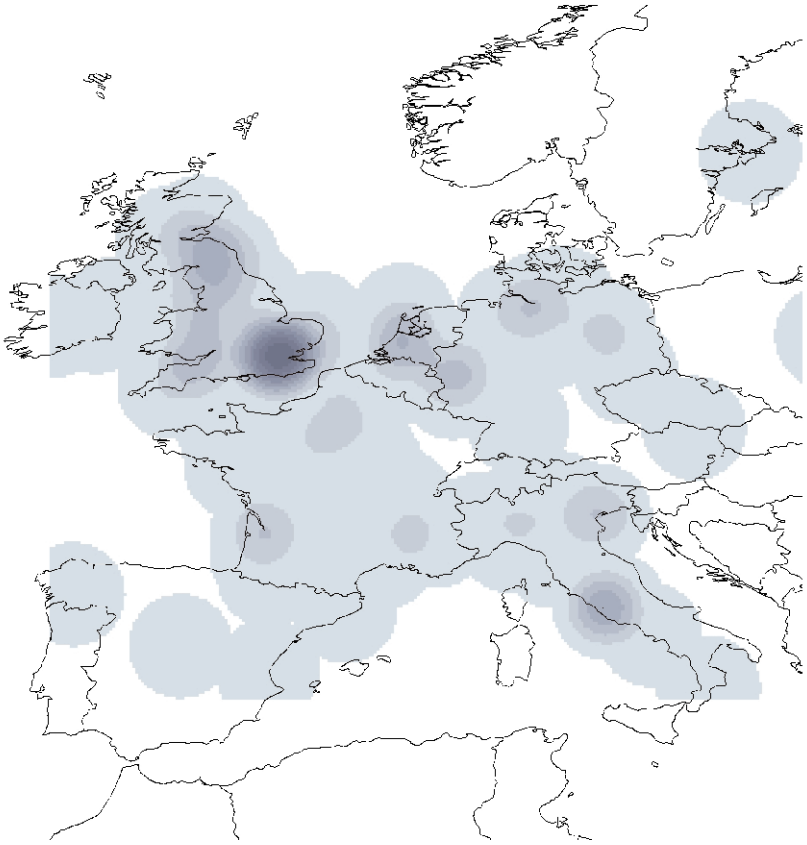


Figure 3. Place-name map generated in response to a search for words related to the concept of 'war'



Place-name map generated in response to a search for words related to the concept of 'money'



Place-name map generated in response to a search for words related to the concept of 'government'

concept *government*, category G1 under the USAS system.

These maps illustrate that for the historians of the seventeenth-century there is much to explore. This brings us to the next important point – that such computer-based visualizations are not providing new historical answers to earlier historical questions. Rather, they open up new alleys of exploration, uncovering untapped areas of scholarly enquiry. The maps above do not provide any kind of precise answer to the nature of war, money or government in seventeenth-century Europe, but instead pose new questions: *'why do the spots on the maps appear as they are?'*, *'why are some places mentioned and others not?'* The approach can also refocus old questions: *'why was the notion of government so centred on London?'* As with most tools, the power of this methodology will only be realized when historians use it as a springboard to other types of analysis; the maps are tools that allow exploration of the data in new ways, rather than being finished products.

To consider this in more detail, the large numbers of references to money and finance in London, Edinburgh and Paris are perhaps not surprising. However, why are other places that might be expected here, such as Amsterdam, not mentioned? Is this to do with the Anglo-Dutch war? Conversely, why is there a cluster on the east coast of England centred on Scarborough? The answer to this latter question is easily found by querying the underlying corpus.[7] From this we find that there are several mentions of captured ships being landed at Scarborough and these being 'rich prize[s].' Further investigation of this map reveals another issue, namely that automated tools such as these will inevitably produce some ambiguities. An example of this is the apparent cluster of money- and finance-related references to Tunis. This is in fact an error in the semantic tagging. The corpus contains several demands to 'call the Turks to an account at Tunis… for the injuries they have done unto the Christians' and such like. The word 'account' has been automatically tagged as a literal reference to *money*, whereas to the human interpreter it clearly has a different (metaphorical) meaning in this context.

Nevertheless, while it needs to be used with care, this approach does open up new ways of exploring large bodies of text, asking the fundamental question of *'what is being said about which places?'* Potentially, *'how has this changed over time?'* could be added to this. No other aproach is able to ask this of such large databases.

So we have seen what the impact of such technology is on researchers. But the ramifications of this can impact on a much broader range of related parties. If the broader educational and publishing community is committed to facilitating researchers to undertake work of this nature, it needs new approaches to handling the digital content that underpins this work. Those developing digital resources have to start providing mechanisms whereby the entire data of a resource can be made available and queried by users, i.e. getting away from the keyhole approach to data which the traditional library catalogue underpins. This will involve significant technical work; not just in getting the data in a form in which it can be used, but also in the challenge of sending the data from content provider to the user. The newsbook corpus is a relatively small corpus, but giving access to an entire data set which could comprise billions of words (such as Google's entire corpus of digitized books) seems, even in 2009, an eye-wateringly difficult task.

However, it is likely that the inevitable step of technological progress will diminish such difficulties. A more intractable problem is likely to involve the licensing of such data. Even those digital resources which are freely available on the internet rarely have explicit licences which permit the usage of the collection in such a context; in some cases owners and creators of free resources are happy to free them up for this type of work, but in others there is still a reluctance to allow everything to be opened up.

When it comes to licences developed by commercial publishers, it is rare for the publisher to have even considered the exploitation of the data for the type of research executed by Gregory and Hardie. Many of the licensing arrangements between commercial bodies and the educational sector presuppose the content being made available via the traditional search and browse paradigm. But if we want our scholars to continue pushing forward the intellectual and methodological boundaries, both the licensing and technological frameworks within which they operate must continue to keep pace.

References and notes

1. Some of the most notable examples include the British Library's Historic Newspapers project: http://www.bl.uk/britishnewspapers (Accessed

8 May 2009), the University of Oxford's First World War Poetry Archive: http://www.oucs.ox.ac.uk/ww1lit/ (Accessed 8 May 2009), or

the National Archives' Cabinet Papers 1915–1978: http://www.nationalarchives.gov.uk/cabinet papers/ (Accessed 8 May 2009).

A growing list of digitization projects funded by public sector bodies is at: http://web.me.com/xcia0069/uk-digitisation.html (Accessed 14 May 2009)

2. Information is available at: http://www.britishpamphlets.org.uk/ (Accessed 8 May 2009)

   while the resource itself is available via JSTOR at: http://www.jstor.org/page/info/participate/other/britishPamphlets.jsp (Accessed 14 May 2009)

3. The collection is available from: http://ahds.ac.uk/catalogue/collection.htm?uri=lll-2531-1 (Accessed 8 May 2009)

   and its creation is described at: http://www.ling.lancs.ac.uk/newsbooks (Accessed 8 May 2009),

4. CLAWS is a grammatical annotation system developed by University Centre for Computer Corpus Research on Language (UCREL), University of Lancaster. See: http://ucrel.lancs.ac.uk/claws/ (Accessed 8 May 2009)

5. USAS, the UCREL Semantic Analysis System, was devised by researchers at the University of Lancaster, and, based on the previous labour of lexicographers, USAS breaks down words in 26 categories, e.g. 'government and public' or 'arts and crafts'. More details are at: http://ucrel.lancs.ac.uk/usas/ (Accessed 8 May 2009)

   The USAS framework now also forms part of the WMatrix software, which UCREL has made available as a web service: http://ucrel.lancs.ac.uk/wmatrix (Accessed 8 May 2009)

6. The process by which such dictionaries are created and the rules under which such parsing takes place are obviously complex, subject to all kinds of scholarly argument. For greater detail on the issue see Wilson, A and Thomas, J A (1997) Semantic annotation, in Garside, R, Leech, G, and McEnery A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 53–65.

7. The web-based interface used to search the Lancaster Newsbooks Corpus is publicly available: http://juilland.comp.lancs.ac.uk/hardiea/newsbooks/ (Accessed 8 May 2009)

■ **Alastair Dunning**
Digitisation Programme Manager, JISC
JISC Office (1st Floor)
Brettenham House (South Entrance)
5 Lancaster Place
London WC2E 7EN, UK
Tel: +44 (0)203 006 6065
E-mail: a.dunning@jisc.ac.uk

■ **Ian Gregory**
Digital Humanities
c/o Department of History
Lancaster University
Lancaster, LA1 4YG, UK
T: +44 (0)1524 594967
E-mail: I.Gregory@lancaster.ac.uk

■ **Andrew Hardie**
Department of Linguistics and English Language
Lancaster University
Lancaster LA1 4YT, UK
Tel: +44 (0)1524 593024
Email: a.hardie@lancaster.ac.uk