Digital Exposure of English Place-names (DEEP) Advisory Board Minutes

8 March 2012, Kings College London

Present

Jean Anderson, University of Glasgow
Mat Bristow, Institute for Historical Research
Robert Brook, History of Parliament Trust
Jayne Carroll, University of Nottingham
Paul Cavill, University of Nottingham
Richard Coates, University of the West of England (chair)
Stuart Dunn, King's College London
Paul Ell, Queen's University Belfast
Rob Gaizauskas, University of Sheffield
Richard Gartner, King's College London
Claire Grover, University of Edinburgh
Paola Marchionni, JISC
Andy O'Dwyer, BBC
David Parsons, University of Aberystwyth
Ed Parsons, Google
Humphrey Southall, University of Portsmouth
Jo Walsh, University of Edinburgh

Apologies

Ian Gregory, University of Lancaster

1. BACKGROUND

The Digital Exposure of English Place-names (DEEP) project's purpose is to digitise, and make available as structured machine-readable data, the Survey of English Place-names (hereafter 'the Survey'). This document summarises the Advisory Board meeting held in March 2012. It does not aim to record the views of individuals, rather it seeks to identify points of agreement on what the project is trying to achieve and should be trying to achieve, and what views the Board came to on key areas.

The project exists against a background of several deficiencies in traditional ways of handling space and time in historical and archaeological data. Historical GIS (HGIS) has not had a notable impact on digital humanities, for a number of reasons. GIS technologies, and their means of handling data, are often perceived as being overly reductionist and positivist. Of the convention vector GIS primitives of points, lines and polygons, points are easy, polygons are difficult. This being so, building connections between historical datasets using their geographic component is an elusive, but much sought-after, goal. The Survey is the primary resource for English place-name studies, but it is also a valuable source for a range of disciplines which use historical datasets containing geographic data allowing

location to be expressed in terms of place-names rather than very precise co-ordinates as required in a traditional HGIS. The purpose of the DEEP project is to digitise this information, to extract significant entities contained within the text, and to encode the structural relationships of those entities using the MADS metadata protocol, whether they are given explicitly or implicitly in the text itself.

DEEP operates a processing chain in which the volumes are scanned and Optical Character Recognition applied in Belfast, and a manual high-level tagging structure added. Language Technology Group (LTG) in Edinburgh moves this output from Word into structured XML tags with as much accuracy as possible. The XML is then checked via an exhaustive manual editing and proof reading process in Nottingham, before structured according to MADS schema at KCL.

2. THE SURVEY OF ENGLISH PLACE-NAMES

The University of Nottingham has been the home of the English Place-name Society, which undertakes the Survey, since 1967. The Survey results in a comprehensive collection of spellings of place-names in individual English counties from the earliest recorded to about 1900, allowing the linguistic, geographical and historical interpretation of the names themselves. Overall, it is estimated to contain some millions of place-names, and many more name forms, that is historic variants of the names. Some place-names are so ancient that their original language is not obvious. In all cases, it is important for the Survey to establish the etymologies of place-names, as well as authoritatively stating their different forms. This is the SEPN's primary function. There are some gaps and inconsistencies in its coverage: for example the county of Somerset has not yet been surveyed; however the Survey nonetheless represents a vastly comprehensive dataset for the etymological, historical and geographical study of English toponyms.

The amount of information to be encoded is thus extremely large, with eight pages of a Survey volume equating to 83 pages of XML mark up. The Survey contains a complex series of abbreviations, which relate to place-name forms, locations and to the bibliography and manuscripts which it draws on. Forms, dates, and sources are encoded according to a set of editorial conventions which run through the Survey volumes, but they are not necessarily applied with total consistency, and some of these conventions have changed in emphasis over the course of the Survey's history. This adds some complexity to the task to extracting information automatically and encoding it computationally and systematically.

3. FUNCTIONALITIES OF THE DEEP RESOURCE

The DEEP data should be able to support fuzzy searching: a user should be able to go from a global overview of the data to a local one, without having to know which individual name forms they (might) be interested in. There are already some coordinates in the Survey itself, but by adopting a procedure of minting individual persistent URIs for each name form, DEEP will allow each entry in the gazetteer to be associated with whatever geographic information is available for each name form, wherever that information came from, including crowd sourcing (with the caveat that the visual interface via which the data is accessed must make clear what information has come from the Survey as published, and what has come from elsewhere).

Abbreviations and cross-references

The Survey makes extensive use of cross-references both across volumes and internally within volumes, and externally to other sources. The search and navigation functions will need to reflect these references. This in turn means that cross-references must be extracted along with the other classes of significant entity.

Every bibliographic and manuscript source in the Survey has a shortform ID, as does every county. The project is assembling a complete list of Survey abbreviations, with an explanation of what those abbreviations mean. The DEEP content will also link to the database of the Key to English Place-names.

The Survey also makes extensive use of conflation when discussing forms with the same stem and different endings Where successive forms partially share spellings, devices are used to save space on the page, e.g. by not repeating some character strings and so conflating some forms by bracketings, and by the use of hyphens to indicate repeated material. From a text processing point of view this will be challenging. It will be necessary to take the original form, strip it back to the stem and then make the full forms.

Error rates and inconsistencies

The LTG output will be as accurate as possible, but it will not be 100% accurate. It was noted that the Dictionary of the Older Scottish Tongue has faced similar issues. Currently Nottingham proofreads the XML output. It has found that there have been some very minor grammatical and functional errors, but otherwise the accuracy of both the OCR and the XML extraction has been extremely high.

It was noted that the fineness of detail varies in the editions, and that error rates are therefore also likely to differ. Also, some conventions have changed from the early volumes to the current ones. Some elements of the Survey are not simply inconsistent, they are not parsable – forms differ in a way that a human reader cannot distinguish (since the meeting, it has been noted that this also applies to some individual characters, such as the letter <I> and the number 1 in small caps). There are further anomalies: for example it has come to light that macrons are missing from certain forms in the first 19 volumes of the Survey. It was, in any case, agreed that the main focus of the project must be solely to digitise the survey and, where possible, bring consistency to its structure (and correct the most obvious errors), not in any way to question, or seek to improve, its semantic accuracy.

For accurate error analysis and evaluation, it is necessary to have a gold standard. Currently, OCR and XML extraction is quality assured as the digitisation occurs by sampling. Belfast produces a 20 page sample of OCR output, it is run through the Edinburgh extraction tool, and the quality/accuracy of both assessed by all partners following an internal process of QA and consultation between CDDA and LTG. As the DEEP project matures, it will develop a gold standard as more volumes become digitally available. The project will run the system and correct it, rather than define a set of standards in advance. It will not be relying on adaptive technologies, it is all rule based.

In terms of delivering on the project, the component where a high degree of textual accuracy is most important is the gazetteer. It is clear that the historic forms should be at least 99% accurate. The gazetteer is the key deliverable required by JISC.

It was noted that cross checking with resources such as Vision of Britain can be challenging, due to lack of overlap of the content. For example the CHALICE project examined crossovers between the Clergy of the Church of England database and Survey data, and found that it was not always consistent. At present, no cross-checking is done. It was however, noted that it would be desirable to collect together any other lists of places for cross checking purposes.

4. TEXT AND DATA STRUCTURE

The Survey's structure varies across both the historic range of the volumes and the geographic range of the Counties. Etymologies for each name form, and how these should be integrated was discussed. It was noted that etymologies for particular classes of names are frequently indicated by some particular text: for example, a parish's block of field names usually beings with the text '[T]he principal forms in (a) are...'. Such consistent patterns can be detected by automated text processing. It was pointed out that difficulties will arise where one language is described in ways relating to possibly overlapping chronological periods, e.g. Late British/Early Brittonic/Early Welsh etc. It was, however, suggested that it would be interesting to see where and when the Survey employed such descriptions of languages.

Treatment of administrative hierarchies

Resources such as the Vision of Britain[1] have developed consistent methodologies for handling administrative hierarchies, which can be ill-defined (especially before the advent of Ordnance Survey mapping), and which change over time. When mounted in an electronic publication, such polyhierarchies require both a meaningful interface and a flexible data structure.

Integration with Unlock

A key outcome of the project will be a gazetteer containing historic as well as contemporary name forms for the JISC Unlock service. Unlock was originally based on OS data, but has been opened up. It will thus be accessible though an API. Conceptually, it will not be dissimilar to the rendering of GeoNames.

It was noted that resources in other parts of the UK experience similar problems of integration, and that there would be much benefit obtained from the editing infrastructure that DEEP creates being re-usable. Use of such tools could be written in to the workplans of future content-creating funding bids.

5. CROWD SOURCING

The possible contribution of crowd sourcing to the DEEP could be: a) correcting errors or omissions in the OCR or extracted XML; b) providing additional structure to the data or c) adding data. In the

[1] www.vionofbritain.org.uk

latter category, the most useful data that could be crowd sourced is the addition of geodata that is not currently in the Survey. Currently, the only geodata available is at the level of grid references for parishes, but this is not consistent in format or accuracy. It was noted that the DEEP project should be seeking to crowd source data rather than interpretation. The latter is unlikely to be anything other than the purview of expert place-name scholars.

There are varied semi-expert communities around place-name studies, including local historians, archaeologists, local geographers etc. A key challenge for DEEP, at both a technical and an intellectual level, is how they can be engaged with the content creation process. The experience of the Family Names of the UK project suggests that there are many knowledgeable individuals who will volunteer their time and energy.

'Long tail' type crowd sourcing – where large numbers of people make minor contributions or edits to large quantities of data - does not work for this kind of data. It will require a different level of engagement and expertise.

It was suggested that the main value of crowd sourcing was likely to be tying geographic references back to maps (the success of the British Library's Georeferencer project[2] was cited as an example). It was also noted that urban place-names, which are not currently well georeferenced in the Survey, could be a fruitful area for some form of self-selecting crowd-sourcing, where contributors added/associated major street names, urban districts etc. Such work would be of greater interest for adding (much) finer-grained information within the existing DEEP structure rather than identifying 'new' historical variants or colloquialisms. Such work has been carried out successfully in Finland and elsewhere.

Another area where this approach has worked in the past is in colloquial toponyms. The *Location Lingo* project[3] is one such example, and the Ordnance Survey have been working in this area too (although OS is not concerned with historic name forms). Colloquial names can be serious references to a place, or they can be used humorously. Either type might be produced by crowd sourcing. It was suggested that connecting colloquialisms to EPNS records might be an interesting avenue of research.

It will be extremely important for the DEEP project to define very clearly what 'the crowd' is being expected to do: thus, the tasks should be kept simple. Given the diversity of the volumes, and the different complexities of issues they present, it was suggested that the tasks could be County specific. The project is, however, at too early a stage to specify what those tasks might be.

There might be useful information sharing exercises to be done with other projects on crowd sourcing methodology. For example it was noticed that Victoria County History (VCH) is arranged in county offices of people working on county-based contributions. This is a crowd that we could very usefully accessed by DEEP. But that would be an extension to the existing parameters of the project, and would require extra funding to realise.

2 www.bl.uk/maps

3 http://www.englishproject.org

It was noted that, whatever questions are tackled with crowd sourcing and whatever tasks devised, DEEP has a clear responsibility to maintain a clear distinction between the crowd-sourced material and material derived from the Survey volumes.

6. CONNECTIONS WITH OTHER PROJECTS

Vision of Britain

One benefit Vision of Britain has experienced stems from the fact that it is a semantic structure openly exposed to the Web rather than a geographical one. It is therefore exposed to Google. Within the IPR agreement, external sites could point to an online parish-by-parish version of the survey. For example, Vision of Britain could cite individual name form URIs. In this sense, DEEP would resemble the URI-based structure of the Pleiades project.

It was noted that resources such as Vision of Britain were of great interest to genealogists, and that DEEP might well generate significant interest from this quarter. It was reported that the project has already received an approach from an amateur genealogy group in the US which would be interested in the data (as far as licencing agreements allow).

Victoria County History

There are close parallels between VCH and the Survey volumes with which DEEP is concerned. They have similar coverage, and a similar (parish by parish etc) structure; and there is a parallel set of problems. There is scope for expansion on both sides: for example, a VCH author might have identified a manorial name which does not have a reference in the Survey. This could provide an excellent enhancement of the DEEP data. Similarly, the Survey has considerably more documents referred to than VCH. The latter would therefore clearly benefit from referring to the former. In general, it was noted that a parish-level connection between DEEP and the Survey (perhaps supported by machine tags from the former embedded using widgets in the latter) would be very valuable.

Old Bailey Online

A further relevant project is the Old Bailey Online. The OBO has worked with text recognition and would be interested in using DEEP to find names. OBO is already part of Connected Histories. Other resources noted included an IHR project to digitise the Royal Commission on Scheduled Ancient Monuments, and this Domesday Project at KCL (although this uses only modern translations/versions of place-names). All of these overlap chronologically with both the Survey and the VCH.

History of Parliament

Place-names are of particular interest to MPs for a variety of reasons. They are interested in constituency names, the names of areas within constituencies, the names of areas covered by regeneration, and so on. There is also interest in relationship(s) between place-names and political events, such as Sellafield, Windscale, Westland etc: names that, for some reason, become significant. There would also be interest in framing particular questions around place-names, e.g. 'Could you show me employment data back to Date X'. Currently Parliament has little provision for

digital librarianship of placenames or digital gazetteers, yet there are currently ambitious plans for digitisation. It was agreed that the DEEP project should keep a watching brief on this.

Back issues of the *Radio Times* have recently been digitised by the BBC. It was noted that mirco-topnymy could be of interest here – small names mentioned in the *Radio Times* could be linked with 'parent' units in the Survey.

It was noted that there is currently no overriding technical architecture for these projects. The range of possibilities was immense, however it was noted that it was not within the remit of the DEEP project to explore them. DEEP must focus on making the Survey content digitally available and interoperable, within the licencing agreements with the EPNS. It was strongly agreed however that a separate feasibility study would be needed to fully understand and explore the links with key projects.

7. FUTURE CONTENT CREATION

In general, a key part of the Nottingham RA's role will be to gather requirements for future EPNS data gathering, and how DEEP might support these. Many of these tools will be aimed at future generations of Survey compilers, but it is essential to engage current contributors/editors as much as possible and get their input into the formulation of those requirements. It would also be useful to consider how people with statistical or digital expertise could be recruited to support current and future Survey compilation work.

The DEEP project will provide the EPNS with tools to facilitate the creation of born-digital content, which will conform to the DEEP data model, in the future. This is not just important for creating new content, Counties published in the 1920s need to be revised according to modern editorial conventions and standards.

A medium term aim of the Survey is to follow the pattern set by Shropshire Vol. I and /or have a popular dictionary for each County. DEEP could support this work. It could also assist setting up the Survey of Somerset.

It was suggested that one form of commercial exploitation could be to sell, for profit, the methodology developed by DEEP to other countries that wish to compile place-name surveys of their own.