Digitizing Correspondence

JISC

Digitizing Correspondence Workshop Report

The Digitising Correspondence Workshop was held at CELL on 17 September 2009. Its aim was to bring together academics, editors of print and electronic editions and archivists to discuss issues relating to the production and promotion of electronic correspondence collections. The workshop was funded by JISC and its aim was to promote the development of an integrated approach to the digitization of letter collections, in order to advance the JISC priority of creating a critical mass of correspondence and to link previously unassociated materials. It aimed to clarify the scholarly requirements of researchers and the concerns of archivists within a framework of creating more comprehensive and accessible collections of correspondence.

Hosts

Lisa Jardine	Director, CELL
Jan Broadway	Technical Director, CELL
Robyn Adams	Senior Research Office, CELL
Matthew Symonds	Research Officer, CELL
Tessa Whitehouse	Postgraduate reporter
Lizzy Williamson	Postgraduate reporter

Attendees

Nadine Akkerman	University of Leiden
Melanie Bigold	University of Cardiff
James Brown	University of Oxford
Joanna Corden	Royal Society
Nicola Court	Royal Society
Helen Dampier	Leeds Metropolitan University
James Daybell	University of Plymouth
Johanna Harris	Université de Genève
Arnold Hunt	British Library
Samuli Kaislaniemi	University of Helsinki
Anouk Lang	University of Birmingham
Adam Mosley	Swansea University
Daniel Starza-Smith	UCL
Filipo de Vivo	Birkbeck
Helen Watt	Aberystwyth University

Workshop Structure

The workshop was structured to allow flexibility, so that the round table discussion could address the interests and needs of the participants. The presentations were intentionally kept short, to allow the maximum time for roundtable discussion. It had been intended that towards the end of the workshop the participants would breakout into smaller groups, but it was decided on the day that this would be an unwelcome interruption to the lively general discussion.

A.M.

Welcome & introductions (Lisa Jardine / Jan Broadway)
Presentations:
- A brief survey of the field (Jan Broadway)
- Case study of the Bodley project (Robyn Adams)

Round table discussion

P.M.

Presentations:
- Timelines and the Bodley project (Matt Symonds)
- Geospatial Data (Samuli Kaislaniemi)
- Searching, data visualization and sharing resources (Jan Broadway)

Conclusion & close

Survey of the Field

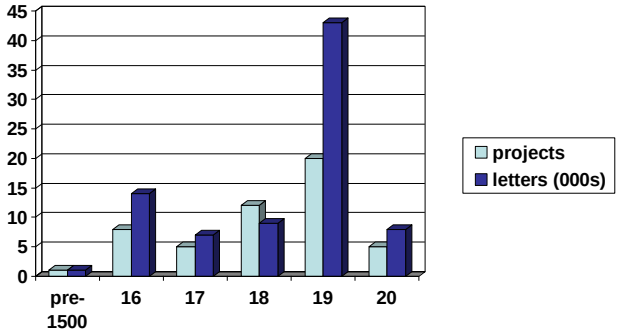
In preparation for the workshop participants were asked to explore the outputs of a range of digitized correspondence projects. The list of projects was composed by searching through Google and Intute using keywords such as correspondence and letters, with some sites being added as a result of personal suggestions. Some are editions, others are archival collections and some are linked to printed editions. Some projects were excluded from consideration, because links within their sites no longer worked. It was also decided to exclude projects that required users to download viewers or other software to their own machines or required a subscription. The final list included 38 projects:

Adams Family Papers
Papers of Sir Joseph Banks
Beethoven-Haus Bonn Digital Archives
Bevan Family Letters
Diplomatic Correspondence of Thomas Bodley
Boulton & Watt Papers
Breadalbane Letters
Emma Spaulding Bryant Letters
Canadian Letters and Images
Carlyle Letters Online
Cartas Desconhecidas, Unknown Letters
Darwin Correspondence Project
Roger Fenton's letters from the Crimea
Correspondence of William Henry Fox Talbot
Vincent Van Gogh
Great War Archive: letters
Letters of William Herle
Leigh Hunt Online
Correspondence of Constantijn Huygens
I remain: A Digital Archive
Jefferson Digital Archive
Linnaean Correspondence
Livingstone online
MacDonald Family Letters
Medici Archive Project
Medieval Women's Latin Letters
John Muir Correspondence
John Murray Archive
Florence Nightingale Letters
Correspondence of William of Orange
Letters of Philip II of Spain
Portsmouth and Macclesfield Collections
Spenser Letters
Spy Letters of the American Revolution
Mark Twain Project Online
Correspondence of James McNeill Whistler
William Wordsworth: Electronic Manuscripts
WWI: Experiences of an English Soldier

The number of letters included in these projects varies greatly, as does the amount of contextual and supporting material. The majority of projects (around two thirds) are based on the correspondence of a specific individual. WWI: Experiences of an English Soldier and William Wordsworth: Electronic Manuscripts represent interesting attempts to make archival letters accessible. Spy Letters of the American Revolution provides an example of the provision of teaching materials, albeit within an American context. Compiling the list proved to be more difficult than expected. This raised the question of how important visibility is to a digital project and there was discussion of how projects achieve that.

A consideration of those projects that were excluded, because links no longer worked raised the issue of long-term sustainability and how that might be achieved. Some of these projects date

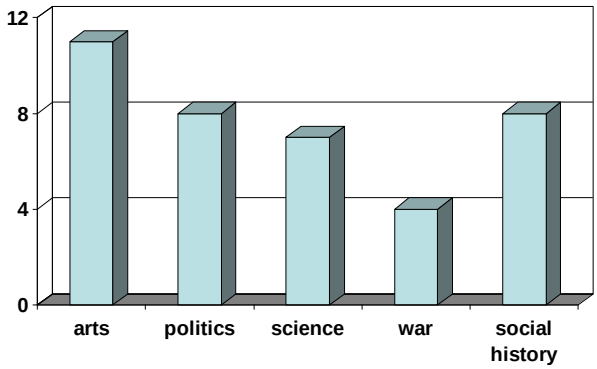
from the 1990s and present interfaces that were cutting edge at the time. There was some discussion as to whether a digital edition requires updating in a way that a printed book does not. We also considered whether the interface tells us something important about the content in the online context as well as the printed.



Coverage by period

This chart shows the coverage by period of the projects and very roughly of letters. Some projects cover more than one century – although where the period extends into a new century by only a few years, this was not counted. The number of letters included in the projects varies greatly, from as few as 10 to over 12,000. It is sometimes difficult to be certain how many letters are actually in a collection, where it is not explicitly stated. A quarter (24%) of projects include fewer than 50 letters – although this includes the Diplomatic Correspondence of Thomas Bodley, which will eventually include around 1,000 letters. Four projects from our sample included over 10,000 letters.

Availability and accessibility of material is probably significant here. Apart from the Correspondence of William of Orange, the largest collection in our sample, the 16th century collections all have fewer than 350 items – although the Bodley will grow. Of the five projects that only cover the 18th century, four have more than 1,000 letters (the other has just ten). Of fourteen projects covering just the 19th century, six have fewer than 50 letters, while five have more than 2,000 (and three more than 10,000). The dominance of the 19th century is largely due to collections relating to individuals involved in the creative arts, the largest collections being Correspondence of William Henry Fox Talbot, Carlyle Letters Online and Correspondence of James McNeill Whistler. The dreaded subject of copyright and commercial value will presumably influence the speed of increase in the availability of 20th century material.

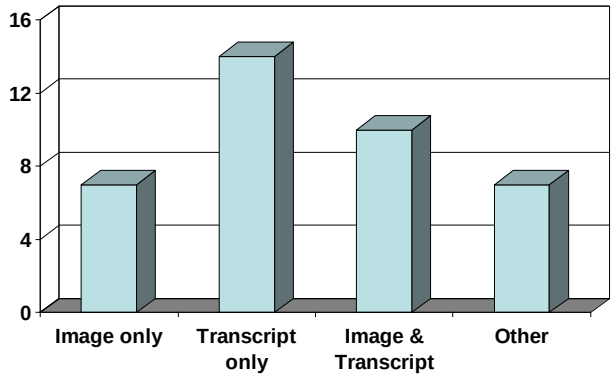


Coverage by subject

The majority (84%) of projects relate to an individual or family, while 16% represent more general collections. This chart maps coverage of the project by very rough subject categories.

- Arts covers artists, poets, composers, authors, photographers and publishers.
- Politics covers kings, presidents and diplomacy (incorporating spying).
- Science predominantly covers the collections of individual scientists, although Portsmouth and Macclesfield Collections covers 61 individual scientists. The Boulton & Watt Papers relating to Cornish tin mining were included in this category.
- War covers the American Revolution and twentieth century wars only.
- Social history represented a catchall category for all the other collections – family archives, Medieval Women's Latin Letters, Florence Nightingale Letters etc.

Obviously collections of letters cover subjects beyond the main domain of the archive. A scientist may see a play or attend a concert, an artist may discuss a family wedding or a husband complain about the state of the economy. There was some discussion of how far projects could and should cater for people who come to projects with interests other than those which drove the initial creation of the collection.



Composition of collections

Finally we considered the composition of the outputs produced by the projects.

- Image & transcript = all letters include both an image and a full transcript.
- Other = covers those with only some images or only some transcripts or some letters being summarized or extracted. Beethoven-Haus Bonn Digital Archives also includes sound recordings.

The majority of images are provided as JPEGs of varying resolutions, although the Letters of Philip II of Spain and Florence Nightingale Letters provide GIFs and Correspondence of William of Orange pdfs. The majority do not have a colour profile. Leigh Hunt Online states that 600dpi TIFFs are available. It was seen as potentially useful to provide access to higher resolution images offline, but it did have ongoing maintenance issues. Around 15% of the images are taken from microfilms and around 10% are digital facsimiles of a printed book. There was some discussion of the various image formats and participants were advised that advice on creating and using digital images is available from JISC Digital Media.

It was noted that the majority of the transcripts are presented as HTML – some generated interactively and some presumably offline from XML. Breadalbane Letters 1548-1583, Spenser Letters and Correspondence of Constantijn Huygens provide their transcriptions as PDFs. WWI: Experiences of an English Soldier presents letters through a daily blog, which it was felt might be

used to make academic outputs more accessible. Reference was made to the blog of Pepys diary. It was noted that Great War Archive: letters comprised items submitted by members of the public and transcripts were provided in a variety of forms, including text and word processed files. It was agreed that the general trend is very much towards the use of XML and there was some discussion of the desirability or otherwise of using the TEI Guidelines when transcribing manuscript correspondence.

It was agreed that it is always good practice to publish a transcription policy. Where editions are composed of transcripts only, one is essential since the reader has no access to images from which the editorial practices might be deduced. Around a third of transcription-only editions in the sample had no policy. Roger Fenton's letters from the Crimea states: 'This website publishes faithful transcripts of letters'and WWI: Experiences of an English Soldier declares : 'I have edited nothing. The spellings and grammar are exactly as Harry wrote them'. Both these statements raise important issues for textual editors. Correspondence of James McNeill Whistler rather charmingly states that they 'reproduce the text as written, including punctuation, capitalisation and errors of pelling, grammar and foreign accents'. It was noted that transcription policies vary widely. Some projects attempt to preserve the materiality of the original, while others standardize 'what does not contribute to the intellectual content of the letters'. It was felt to be most important that the policy should be clearly stated, rather than to attempt to proscribe what approach should be adopted.

Providing full text transcriptions is obviously time consuming and the Joseph Banks archive explains in some detail why they preferred to concentrate on indexing. We discussed briefly how far is it reasonable to expect readers to be able to read images. This clearly depends to a large extent on the period in question and the quality of the handwriting.

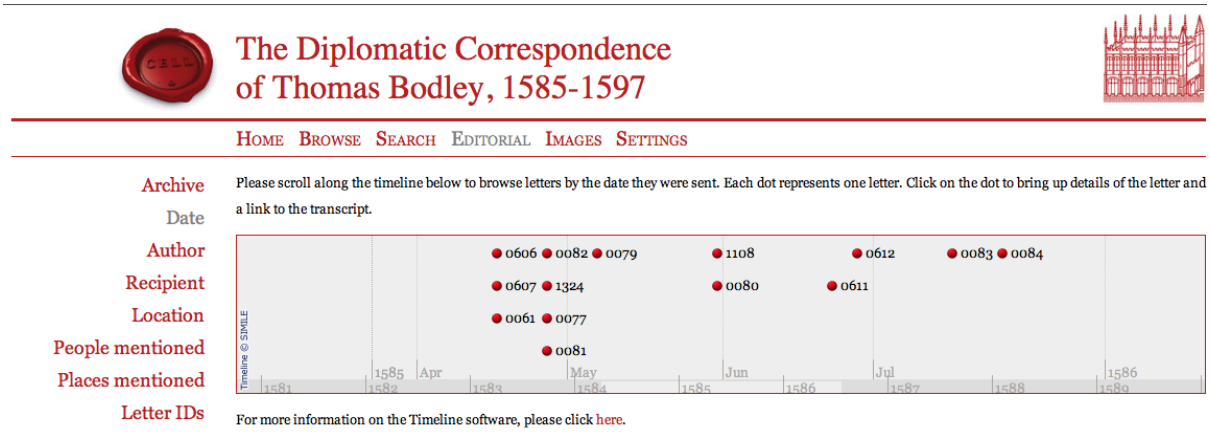
Robyn Adams then interactively demonstrated aspects of Diplomatic Correspondence of Thomas Bodley, explaining in particular the use of the Transcriber's Workbench[1] software to ensure consistency in the XML tagging and the provision of customizable transcription options on the site. This was followed by a general roundtable discussion based on the participants own exploration of the projects surveyed.

Visualizing and Searching Correspondence Collections

To start the afternoon session Matthew Symonds described the way in which the interface of the Diplomatic Correspondence of Thomas Bodley would be enhanced as the number of letters in the on-line edition increased. Eventually the correspondence would include around a thousand separate items and the presentation of the indexes would be overwhelmed by their volume. The enhancements would provide not only easier navigation, but also provide an overview of the collection as a whole.

This was demonstrated by the example of a timeline, implemented using the Simile timeline widget. Matthew explained that, although customizing the widget required making modifications to the JavaScript, this did not require extensive programming skills and there was guidance available to guide you through the process. Matthew also demonstrated how the periodicity of the timeline could be modified, to reflect the fluctuating volume of correspondence over time.

[1] Developed for CELL by Mind Magic Ltd., this software relieves transcribers from the need to learn XML and ensures that tagging is consistent.



Screenshot of the Bodley timeline pilot

Samuli Kaislaniemi of the Research Unit for Variation, Contacts and Change in English (VARIENG) at the University of Helsinki then gave a presentation concerning the use of geospatial data. Samuli works on the correspondence of the East India Company and has experimented with the use of Google maps to display the motions of the fleet. His presentation considered what information concerning correspondence could be mapped geographically:

- the location of authors and recipients;
- the route taken by a letter;
- the movements of correspondents.

All this data includes a time component, which we should be able to map digitally more effectively than in print.

Samuli had analyzed the geographical information provided by the 38 projects that had formed the basis for the morning's discussion.

- Contextual geographical information, such as maps, not tagged to letters:
    - Various types, 4 sites
- Letters tagged with locations; geographical indexes:
    - Origins only, 6 sites
    - Origins & targets, 2 sites
    - Tagged but not searchable/sortable (!), 2 sites
- Letters put on a map
    - Interactive, 1 site

He explained the advisability of employing software to capture the data, if dealing with an extensive correspondence and provided examples from his own work.

- Manually in Google Maps
    - EIC sailing routes
- Semi-automatically, with *Spreadsheet mapper*
    - Richard Cocks
    - Corpus of Early English Correspondence

Samuli also utilised the tools provided by Powerpoint to demonstrate the potential use of animation. Finally Samuli considered ways of mapping geospatial information that did not depend

upon maps, showing various examples from different sources. This led to a lively discussion of the potential for and hazards of data visualization.

For the final presentation of the day Jan Broadway considered the ways in which search facilities could be provided, offered some of her own thoughts on data visualization and concluded with some thoughts on how virtual correspondence editions might be created. She identified four methods for searching collections:

- free text search;
- keywords;
- encoding/semantic web;
- natural language understanding.

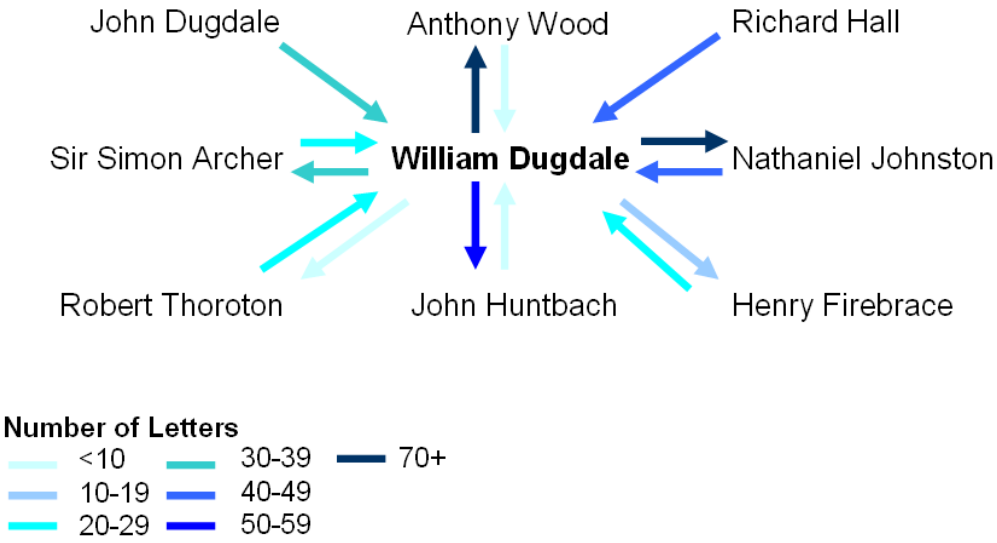
Free text searching can be easily supplied using Google, if full transcriptions are available in HTML. Jan explained how at CELL the XML produced by projects is processed to provide dictionaries, which are used as the basis for searching. This allows for expanded and unexpanded forms of words and could also allow for translations of foreign text to be included in the search. Jan observed that free text searching had limitations, but some users would expect it and it ought to be provided if possible.

The Papers of Sir Joseph Banks chose not to include transcripts of letters within their project, but to add index terms or keywords to the descriptions of the letters. It was agreed that there were potential problems with this, since it presupposes that the indexer knows what the potential reader is looking for. Jan suggested that, if keywords are used, it is a good idea to have a list of the words and their meaning available.

Whether creating separate index fields within a catalogue entry or encoding transcripts some search terms are relatively straightforward e.g. people and places. Keywords are problematic – the catalogue of the Bacon correspondence as originally received by CELL had keywords applied to each letter – this was not incorporated into the online version of the catalogue, since around 1,000 letters had produced over 600 keywords excluding personal and placenames of which the vast majority occurred only once or twice and many of them were different ways of expressing the same or related concepts. Jan described how she began to develop a conceptual hierarchy of key terms for the Bacon project – identifying Six Clerks, Hanaper, Lord Chancellor as sub-terms within Chancery, for example. This work stalled when the development of the semantic web promised to provide a methodology and tools for the formal description of concepts, terms, and relationships within a given knowledge domain. The Leigh Hunt Online project is extracting data as Resource Description Framework (RDF) tags to enable it to be integrated with the larger NINES (Networked Infrastructure for Nineteenth Century Electronic Scholarship) as 'part of the push towards the semantic web experience'. From Jan's personal encounters with projects utilizing this method, she had gained the impression that the sheer overhead of semantic web encodingwas discouraging its take-up. She also felt that the technology was still comparatively immature and that it would be wise for projects to be cautious in commiting resources to its use. James Brown explained that his own Cultures of Knowledge project was using RDF encoding, which will provide useful data for other projects.

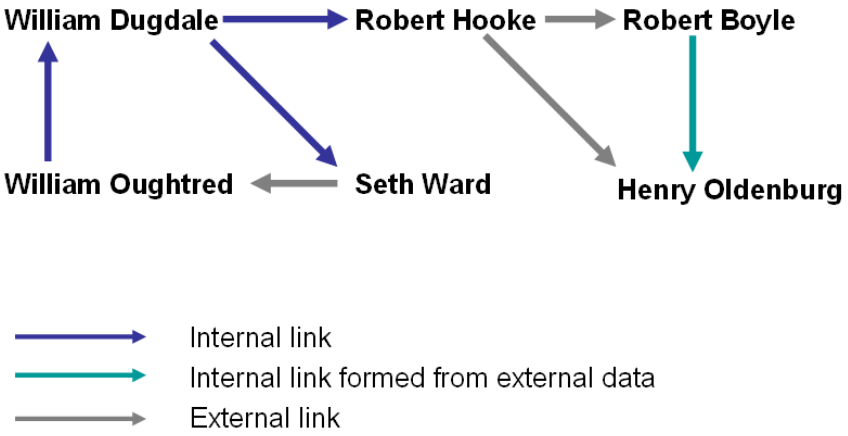
Jan spoke briefly about natural language understanding and how Developing natural language understanding of modern texts and speech is a major outstanding problem in the field of artificial intelligence. Historical texts as ever present particular problems of language and context. However, Jan reported on the collaboration between IBM and the 1641 Depositions Project at Trinity College Dublin aimed at developing intelligent search facilities based on cutting-edge technology.

Jan then turned to the issues of identifying people and places. Anyone who has tried to either index or markup a collection of letters for correspondents, places of origin and destination and people and places mentioned will know that it is not a simple task and invariably takes longer than expected. Variant spellings, oblique references and difficulties in differentiating between individuals all complicate the task of compiling a biographical index. Few projects can identify all correspondents, let alone all references within letters. One of the things that surprised Jan was how few projects had an easily accessible index of correspondents. Places unlike people do stay put – but like people they change their names over time and may occur in correspondence under local or familiar names that don't appear on maps. For modern data EDINA reports that placenames can be resolved to geospatial data with a high degree of accuracy. The method relies on a gazetteer and there is currently no good gazetteer of historical data for the British Isles generally available, although a good deal of work has been done in this area. There was some discussion of the possibility of developing a reliable source to which projects could contribute via the web and access programmatically to acquire information in a standard form for incorporation in their own sites – a historical people and places wiki. This is an area where co-operation between projects could yield real benefits to the community.



A view of William Dugdale's correspondence network

Returning to the subject of data visualization, which had been initiated by Samuli, Jan presented her own primitive visualization of the data for the correspondence of William Dugdale, in which the darker the arrow the greater the volume of correspondence. If the two arrows are close in colour, it is likely that most of the correspondence survives – if they are quite different (or one is missing) then one side of the correspondence is probably lost. Jan explained that this simple visualization had helped her to appreciate the nature of the surviving evidence for the network better than looking at the raw numbers in her database. A sophisticated visualization might use animation to map the evolution of a network and changes in the volume of data available over time. Social network analysis may be the province of sociologists, but Jan suggested that researchers in other disciplines might be able to utilize some of the tools they have developed to enhance their own work – and put the results online to assist their readers to understand a correspondence collection.



A model for a correspondence meta-edition

Finally, Jan explained that as a researcher her interest is predominantly in groups with shared interests rather than individuals. Consequently, it has long been her hope that a model of co-operating projects might be developed, where meta-editions would select subsets of letters from different letter collections. So to conclude she considered the possibility of virtual recreations of correspondence networks.

She had modelled a small network based on her own work on William Dugdale's correspondence and its intersection with that of the post-Restoration scientists represented on the Portsmouth & Macclesfield site. It offered a small example of how correspondence projects are particularly likely to have potential for links beyond their own parameters. For the purposes of the demonstration it was assumed that only one document is known linking individual members of the network. The blue arrows indicated internal links – which in the demonstration went to a new Powerpoint slide, but on the web would go to a page within the site. The grey arrows indicated links to the Portsmouth & Macclesfield site, which opened in a new browser window – this allowed the links to be made, but users might find it confusing to navigate different sites. The turquoise arrow envisaged a situation where the meta-edition was able to send a request for information to the external site, format that information according to its own requirements and present it to the user. This could be implemented either through calls to a remote database or an agreed XML format. Jan suggested that creating virtual correspondence networks seems an obvious way to link previously unassociated materials, which was a JISC priority that the workshop was intended to address.

The workshop concluded with a vigorous discussion of various issues that had arisen during the day. It was strongly felt that the participants would benefit from further opportunities to pool ideas and experience. CELL undertook to consider how they could further this aim through the development of the Digitizing Correspondence strand of their activities.

Jan Broadway