Manuscripts Online: Written Culture from 1000 to 1500

Final Report by Michael Pidd, Orietta Da Rold and Katherine Rogers

Project Team

Dr Orietta Da Rold (Project Leader, Univ. of Leicester)
Matthew Groves (Developer, Univ. of Sheffield)
Dr Sharon Howard (Project Manager, Univ. of Sheffield)
Jamie McLaughlin (Developer, Univ. of Sheffield)
Michael Pidd (Project Leader, Univ. of Sheffield)
Katherine Rogers (Lead Developer, Univ. of Sheffield)

Editorial Board

Synopsis

Manuscripts Online (http://www.manuscriptsonline.org) provides an intuitive interface which enables users to search a significant body of online primary resources relating to written and early printed culture in Britain during the period 1000 to 1500. This final report explores some of the challenges encountered during the development of the website and reflects upon the lessons learnt.

1. Introduction

Manuscripts Online (http://www.manuscriptsonline.org) provides an intuitive interface which enables users to search a significant body of online primary resources relating to written and early printed culture in Britain during the period 1000 to 1500. Users are able to search up to 21 free and subscription-based resources simultaneously by keyword or a specific category of information: person name, place name, date range and may also filter their results by language, source date etc. The user is presented with the results of their search organised by resource/collection, comprising the document title and a text snippet which shows their keyword in context. When the user clicks on a link they are directed to the full text on the content provider's own website. *Manuscripts Online* builds upon the model of the JISC-funded *Connected Histories* (http://www.connectedhistories.org) which provides federated searching of online primary sources from 1500 to 1900. However, *Manuscripts Online* goes further than *Connected Histories* by attempting to solve some of the search challenges which are particular to primary sources for this period: non-standardised spellings and the use of non-Latin characters such as 'thorn' and 'yogh' (although these did not simply disappear from use in 1500).

Manuscripts Online was developed by a project team comprising the Humanities Research Institute at the University of Sheffield and the School of English at the University of Leicester, supported by an Editorial Board of eminent colleagues in manuscript studies at the universities of Birmingham, Glasgow, Sheffield, York and Queen's University Belfast. The visual design was conceived by Mickey and Mallory Limited (http://www.mickeyandmallory.com).

2. Context

Medievalists have been quick to recognise the potential of digital media for their research, creating and circulating full text transcriptions and descriptive catalogues of primary sources alongside tools which improve historical, linguistic and palaeographical analysis. However, many of these resources have been created using different methods and standards, they are scattered across the web and their existence tends to be known only to specialised research groups within the medieval manuscripts community. Fortunately the web now enables us to pull these resources together in a way which enables relatively consistent searching across the underlying data. The use of dictionaries and lookup lists enables users to search for words irrespective of spelling variation whilst Natural Language Processing techniques are able to automatically identify categories of data such as person names, place names, dates, document references and different languages, thereby enabling a more forensic approach to searching

3. Challenges

The hand-crafted, specialised nature of online medieval resources presented us with a number of challenges when it came to developing a clustering methodology for *Manuscripts Online*:

- How do we pull together such a diverse range of resources when some of them are freely available, some are only available through subscription and some are poorly maintained?
- How do we enable users to search consistently across a body of data when non-Latin characters have been represented in different ways, spelling is not standardised and different languages are used?
- How do we encourage a culture of collaboration and sharing within the manuscript studies research community?

3.1 Gathering the Content

A key aspect of this project was our relationship with the content providers, which included individual academics, large organisations such as The British Library and The National Archives, and publishers such as ProQuest and Gale Cengage Learning. The project undoubtedly benefited from relationships and trust which had been built during the *Connected Histories* project.

We had a dedicated Project Manager who was responsible for managing these relationships: communicating our aims, negotiating the Material Transfer Agreements (content licences) and arranging access to the data. We also scheduled our data processing workflow into three ‘data bundles’, each bundle representing our estimation of how fast or lengthy the negotiations with the content providers would be. For example, all data owned by the project's partner institutions were scheduled as Bundle #1, being the easiest to acquire consent to use.

3.2 Searching Consistently

Providing users with a consistent search experience was the greatest challenge for this project. Materials for this period use un-standardised spellings (eg. church, chirch, chirche, cherche, churche, kirk and kirke), non-Latin characters (eg. ð, þ and 3) and a range of languages (e.g. English, Latin and

Anglo-Norman French). Further, content creators used a variety of methods for representing non-Latin characters, largely due to the age of some of the datasets which predate widespread standard such as Unicode. These problems meant that it would be difficult for users to acquire a useful body of results when undertaking a search. Our solution was to harmonise the presence of non-Latin characters to a consistent form of electronic representation and then provide users with the facility to include variant spellings in their search. When this option is checked, the search keyword is parsed against the *Middle English Dictionary* and our own character substitution patterns in order to generate alternative spellings.

More information about our variant search techniques can be found on the website: http://www.manuscriptsonline.org/technical

3.3 Adding Structure to Search

We developed Natural Language Processing algorithms, specifically a technique called *automated entity recognition*, in order to help us identify and automatically tag different categories of data. We sought to capture different languages, person names, place names, dates and document references. The techniques involved gazetteer lookups and grammar-based rules.

The development work for identifying person names, place names and dates had been largely established during *Connected Histories*. We used a statistical and dictionary-based approach to attempt to identify Latin and French phrases, to enable the user to search, for example, for "benefice" but only within a Latin context. We also decided that we would not attempt to distinguish between Old English, Middle English and Modern English (the presence of Modern English is particularly prevalent in resources that comprise manuscript descriptions). The boundaries between these phases in the development of English are not clear and it would be impossible in some cases to say that such-and-such a word was Middle English rather than Modern English.

More information about our Natural Language Processing techniques can be found on the website: http://www.manuscriptsonline.org/technical

3.4 The Way of API

Manuscripts Online has a Web API at the heart of its architecture, communicating between the user interface and the search engine. A Web API (Application Programming Interface) is a protocol for communicating between two systems. For example, any web page which includes a Google Map is using Google's Web API. The web page is located on one server and it is fetching map data from Google's server. Many projects cite the development of an API as one of the means of facilitating greater data re-use by third parties. However, the Manuscripts Online API is integral to the system architecture and so although we too have documented it for others to use, it is the benefit of an API-based approach to in-house development which has already proved to be useful. By using an API for communication between the interface and the search engine - essentially bridging two separate processes - it has proved much easier for different personnel to work on different components and quicker to resolve bugs and other issues. Further, we hope that the API will assist with the sustainability of the site long term, because the user interface and/or the search engine can be modified with minimal impact on one another.

Technical documentation for using the *Manuscripts Online* Web API can be found on the website: http://www.manuscriptsonline.org/api

3.5 Changing the Culture

During the *Manuscripts Online* project we attempted to make tentative steps towards changing the culture of manuscript studies. Despite this research community being an early adopter of digital techniques - converging image digitisation and text analysis tools to reinvent the 'critical edition' - manuscript studies appears to remain a conservative discipline on matters concerning user generated content. For example, we originally explored an idea whereby users would be able to assign vocabulary to dialectal regions, using crowd sourcing techniques to judge the results, and thereby develop a community-generated linguistic atlas of Middle English. However, medievalists on the team felt that users of the website would not have the knowledge or expertise to make these judgements. Similarly, there were concerns about introducing a feature whereby users would be able to explore the search paths of other users, for fear of making public an individual's research agenda.

Our solution was twofold: we introduced a facility for creating comments and storing search pathways which can be made public or private; and we developed a mapping feature whereby users can plot their comments on Google Maps if the comments have a geographical significance. Both the commenting and the mapping features have no instructions dictating what constitutes an appropriate contribution, so long as contributions are not offensive. In other words, rather than trying to predict what users might wish to do with the site, we have simply provided them with some tools for generating content and we will now leave them to it.

4. Sustainability

The sustainability plan for *Manuscripts Online* will mirror the plan which is already being successfully implemented for *Connected Histories*. The sustainability plan is as follows:

- *Manuscripts Online* will be hosted, maintained and enlarged by the Humanities Research Institute (HRI) at The University of Sheffield, in partnership with the project team, for as long as the HRI is permitted to do so.
- Resources which have been added to *Manuscripts Online* during the JISC-funded phase are considered to be 'foundation resources', to remain a part of the site while ever the site remains in existence unless the content providers determine otherwise.
- Further resources will be added to *Manuscripts Online* beyond the period of JISC funding, with a view to enlarging the site's content and its value to research, by disseminating information about the site's purpose and value to the wider research community. Additional resources will be included in the site subject to a fee.

More information about contributing a resource to *Manuscripts Online* can be found on the website: http://www.manuscriptsonline.org/participate

5. Impact

We first became aware of the project's potential for impact whilst approaching Gale Cengage Learning for permission to include one of their resources in the site: *British Literary Manuscripts Online, Medieval and Renaissance*. Gale were keen to include this subscription-based resource in order to raise its profile within the manuscript studies community. We made enquiries with fellow medievalists and none of them had heard of it.

Our immediate plans are to embed *Manuscripts Online* within its native research community: medieval studies. This is an international community with good representation in the Higher Education subjects of English Language, Literature and History (particularly in the UK and North America) as well as within the libraries and archives sector. We intend to achieve this by implementing search engine optimisation techniques, *Wikipedia* article links and a 'critical mass' effect.

The main search engine optimisation technique we will use involves the following process: we top slice the 1,000 highest ranking person names and place names from the *Manuscripts Online* search index; programmatically run searches for each of these 1,000 keywords; generate static search result pages from these searches and expose them to search engines. This process makes it easier for search engines such as Google to crawl our data.

Our 'critical mass' effect involves us advertising *Manuscripts Online* and seeking to attract more and more content providers. Eventually, anyone proposing a research project or product development which involves the creation of a digital resource for medieval studies will hopefully view inclusion within *Manuscripts Online* as the *de facto* activity for helping to improve the exposure and impact of their digital resource.

Manuscripts Online will also have institutional impact for us, based on the experience of sites such as *Connected Histories*. This impact will be primarily a business benefit to the HRI and the departments of our partner institutions: Leicester, Birmingham, York, Glasgow and Queen's University Belfast. It will enable us to open up new research agendas with colleagues who become aware that they can cross-search these previously unconnected bodies of data; it will be included in the Research Excellence Framework (REF) for our institutions; it will be part of the 'offer' presented to new, prospective students as a way of demonstrating how our institutions are research leaders in this area; and it will be embedded within our postgraduate modules, including the University of Sheffield's Faculty modules on the digital humanities.

6. Conclusion

The important success factors for our project were the relationships which we developed with our content providers (the largest and most important group of stakeholders in our view), the research expertise of our Editorial Board and an agile development methodology. We feel that we built up a lot of trust with our content providers and that they were fully supportive of the project's ambitions. This reduced a significant risk in the project: failure to secure permissions to incorporate third-party datasets. This success factor is all thanks to our Project Manager, Sharon Howard. Our Editorial Board ensured that *Manuscripts Online* remained focussed on the needs of the research community and appreciative of the discipline's research methods and intellectual values. The Board comprised Wendy Scase, Linne Mooney, Estelle Stubbs, John Thompson and Jeremy Smith. Further, although the auditing, processing and inclusion of the datasets became by necessity a pipeline process, the development of the search engine and the user interface remained agile in its design (thanks largely to the API approach). As such, we were able to change things relatively easily when ever the Editorial Board identified problems. Our developers were Katherine Rogers (lead developer), Matthew Groves and Jamie McLaughlin. Their expertise was, unquestionably, the most important success factor for the project, as it tends to be for most of our projects! Thank you everyone.