


Cover Sheet for Bids <i>(All sections must be completed)</i>			
Name of Strand: Strand A:	<input type="checkbox"/>	Strand B:	<input type="checkbox"/>
		Strand C:	<input checked="" type="checkbox"/>
Name of Lead Institution: University of Sheffield			
Name of Proposed Project: Manuscripts Online: Written Culture from 1000 to 1500			
Name(s) of Project Partners(s) (except commercial sector – see below)			
University of Birmingham, University of Glasgow, University of Leicester, Queen's University Belfast, University of York			
This project involves one or more commercial sector partners NO (delete as appropriate)		Name(s) of any commercial partner company (ies)	
Full Contact Details for Primary Contact:			
Name: Michael Pidd			
Position: HRI Digital Manager			
Email: m.pidd@sheffield.ac.uk			
Tel: 0114 222 6116			
Address: Humanities Research Institute, 34 Gell Street, University of Sheffield, Sheffield S3 2QY			
Length of Project: 15 months			
Project Start Date: 1st November 2011		Project End Date: 31st January 2013	
Total Funding Requested from JISC:		£109,612	
Total Institutional Contributions:		£19,811	
Outline Project Description <i>Manuscripts Online</i> is proposed as a sister site to the JISC-funded <i>Connected Histories</i> (1500-1900) website and will extend the model of data clustering and federated searching developed during this project by providing access to written and early printed primary sources for the period 1000 to 1500 which will be of relevance to researchers in the fields of language, literature and history. However, <i>Manuscripts Online</i> will also address the concerns of its target research community by moving beyond the model of <i>Connected Histories</i> in the following ways: a) it will provide searchable access to resources which are not currently available on the web; b) it will use Natural Language Processing to intelligently identify and tag specific words and phrases for semantic searching, but this process will focus on technical vocabulary of relevance to medievalists, such as illumination terminology, palaeographical features and dialectal forms, in addition to identifying the names of people and places; c) it will enable users to add extensive comments to search result items and blog their discoveries with a view to breaking down the traditional culture of research ownership which persists in the discipline.			
I have looked at the example FOI form at Appendix A and included an FOI form in this bid		YES	
I have read the Funding Call and associated Terms and Conditions of Grant at Appendix B		YES	
For FE institutions only: Please tick this box if you are an FE institution in England, please tick this box to confirm that you meet the eligibility requirement of teaching HE to more than 400 FTE		<input type="checkbox"/>	

Manuscripts Online: Written Culture from 1000 to 1500

1. Background and Rationale

- 1.1** The digitisation of research resources for those studying the written culture of medieval Britain has traditionally arisen from very detailed studies of discrete bodies of evidence in the form of scholarly editions of specific manuscripts and descriptive catalogues. Digitisation of pre-1500 written culture has tended to be manual in the form of full-text transcription or cataloguing, often because of the delicate nature of the documents, the non-standardisation of spelling and the fact that we are dealing with a hand-written culture which still eludes automated recognition. Large scale digitisation rarely exists, except for a few instances such as the British Library's *Online Catalogue of Illuminated Manuscripts* and Gale Cengage's *British Literary Manuscripts Online*, whilst OCR is virtually unknown. Most pre-1500 electronic resources were created by research projects whose aims placed a lot of emphasis upon understanding the codicology, provenance and linguistic traits of a document in addition to the text itself, perhaps because of the overall anonymity of authorship during this period of written culture. As such, a lot of attention was paid to enhancing digital surrogates with the types of paratextual metadata which would assist a project's specific research questions. However, the very nature of this digitisation - small in scale, hand-crafted and driven by research questions - has resulted in a significant body of research data which now lies disconnected and largely unused in the form of independent websites. Yet the medieval research community has a pressing need for linked resources in order to fully realise the investment which has been made by funders and scholars over many years. Although many of these resources were created by scholars who were following a traditional publishing model which placed emphasis upon the self-contained work of scholarship, a new generation of scholars expect research data to be open, accessible and available for re-use.
- 1.2** *Manuscripts Online* is proposed as a sister site to the JISC-funded *Connected Histories* website (<http://www.connectedhistories.org>) and will extend the model of data clustering and federated searching which was developed during the earlier project. It will also build upon the lessons which were learnt and capitalise on the methodologies and processes which were developed. Whereas *Connected Histories* provides federated searching of distributed historical resources from 1500 to 1900, *Manuscripts Online* aims to provide federated searching of written and early printed primary sources for the period 1000 to 1500 which will be of relevance to researchers studying language, literature and history. However, *Manuscripts Online* will also address the concerns of its target research community by moving beyond the model of *Connected Histories* in the following ways: a) it will provide searchable access to resources which are not currently available on the web; b) it will use Natural Language Processing to intelligently identify and tag specific words and phrases for semantic searching, but this process will focus on technical vocabulary of relevance to medievalists, such as illumination terminology, palaeographical features and dialectal forms, in addition to identifying the names of people and places; c) it will enable users to add extensive comments to search result items and blog their discoveries with a view to breaking down the traditional culture of research ownership which persists in the discipline.
- 1.3** It is because of the need to address these three community-specific objectives that funding is sought, rather than simply extending the existing *Connected Histories* service to encompass the 500 years prior to 1500. *Manuscripts Online* seeks to address the specific problems of providing federated searching for primary resources of this period which are not so prevalent from 1500 onwards: the resources are handwritten; spelling is not standardised; the alphabet contains non-Latin characters and abbreviation marks; the texts can be in Anglo-Saxon, Anglo-Norman, French and Latin as well as Middle English; there is a focus upon the materiality of the written document in addition to its text; disciplinary boundaries between historians, linguists and literary scholars tend to be more blurred.
- 1.4** The resources identified for inclusion in this project have been selected because of their quality, importance for research and their representativeness of the primary sources which exist in a digitised format for this period. The ability to search and access these distributed primary sources in a structured and consistent way will transform research and teaching in the United Kingdom and North America as well as in Europe where there is a shared written culture during the medieval period. It will enable the HE research community (academics and postgraduates, within the UK and internationally) to address more effectively research questions such as the provenance of the Canterbury Tales manuscripts, the rise of English and the transformation of British society at this crucial period in our national narrative. It will improve the teaching of English literature, language and history at tertiary and undergraduate level by enabling students to build the technical knowledge which is a prerequisite to understanding written and early printed culture. Crucially, *Manuscripts Online* will provide an

API which will enable users and IT professionals to build other web services which capitalise on the single point of access to these datasets, such as corpus building systems, GIS services for historical and linguistic mapping, and interactive learning modules. As with *Connected Histories*, *Manuscripts Online* will grow beyond the period of funding using our existing infrastructure and sustainability models.

2. Proposal

2.1 The *Manuscripts Online* website will be developed and hosted by the Humanities Research Institute (HRI) at the University of Sheffield, under the direction of an Editorial Group which will comprise six members of the Medieval Manuscripts Research Consortium (MMRC) from the Universities of Birmingham, Glasgow, Leicester, Sheffield, York and Queen's University Belfast. The MMRC is a group which actively promotes the exchange of knowledge and capacity-building within the subject domain amongst academics, postgraduates and undergraduates through workshops, meetings and research training such as *Quadrivium* (<http://www.arts.gla.ac.uk/quadrivium>). The Editorial Group will provide vital guidance in addressing problems specific to the clustering of electronic resources for this period and oversee the dissemination and development of the web service within their research communities.

2.2 During the funded period, *Manuscripts Online* will incorporate the following distributed primary sources:¹

- **AHRC-funded datasets:** 1) *Manuscripts of the West Midlands* (Wendy Scase; Birmingham); 2) *Production and Use of English Manuscripts: 1060 to 1220* (Orietta Da Rold; Leicester); 3) *Imagining History: Perspectives on Late Medieval Vernacular Historiography* (John Thompson; Queen's, Belfast); 4) *Geographies of Orthodoxy: Mapping Pseudo-Bonaventuran Lives of Christ, 1350-1550* (John Thompson; Queen's, Belfast); 5) *An Inventory of Scripts and Spellings in Eleventh-Century English* (Donald Scragg and Alexander Rumble; Manchester); 6) *Late Medieval English Scribes* (Linne Mooney; York); 7) *The Blake Canterbury Tales* (Norman Blake; Sheffield); 8) *The Auchinleck Manuscript* (Alison Wiggins and David Burnley; Sheffield). Full-text transcriptions and databases; publicly available.
- **Online Catalogue of Illuminated Manuscripts** (British Library) describing the codicology, palaeography, illumination and provenance of 2,000 illuminated manuscripts originating in England, Wales, Scotland and Ireland. A descriptive catalogue with accompanying images; database; publicly available.
- **Early English Text Society** (Boydell and Brewer) comprising approx. 433 key scholarly editions of medieval works which exist in a digital format but only available via a print-on-demand service. Users will be able to search the texts but will be directed to the publishers' ordering page. Full-text transcriptions.
- **British History Online** (Institute of Historical Research, Univ. of London) comprising approximately 38,000 documents ranging from administrative and ecclesiastical history to economic and intellectual history (full-text transcriptions and databases). Includes the *Close Rolls* which is available by subscription only.
- **British Literary Manuscripts Online: Medieval & Renaissance** (Gale Cengage) comprising c.500,000 pages of searchable metadata with accompanying digital facsimile images. Database; available by subscription only.
- **The Cause Papers** (Borthwick Institute, Univ. of York) of the Church Courts of the diocese of York. 524 documents fall within the period 1000 to 1500. A descriptive catalogue with accompanying images. XML-based database; publicly available.
- **Early English Books Online** (Historic Books Platform and ProQuest) comprising metadata and digital facsimile images of 782 printed volumes between the year 1473 to 1500. We will store durable URIs for both the HBP and the ProQuest versions to guarantee that all subscribed institutions can access the content. Database; available by subscription only.
- **EEBO Text Creation Partnership** (Bodleian Library and Univ. of Michigan) comprising approx. 136 full-text transcriptions of the *Early English Books Online* volumes, to be accessed via the Historic Books Platform and ProQuest in conjunction with *EEBO*. Available by subscription only.
- **The National Archives**, comprising catalogue data for all documents dating between 1000 and 1500 from collections such as the State Papers, records of the Admiralty, Chancery and Exchequer, the Court of the King's Bench and Petitions and Seals. Descriptive catalogues (databases); publicly available.
- **Taxatio** (Jeff Denton; Univ. of Sheffield) comprising detailed records of the assessment (known as a *taxatio*) of English and Welsh ecclesiastical wealth undertaken in 1291-2. Database of over 15,000 records covering every religious benefice; publicly available.
- **Middle English Dictionary** (Paul Schaffner, Univ. of Michigan) is the authoritative reference work for Middle English from 1100-1500, comprising over 15,000 entries with citations. Database; publicly available.

¹ For URLs to datasets, see <http://www.hrionline.ac.uk/urls-for-datasets>

- **Compendium of Middle English Prose and Verse** (Paul Schaffner, Univ. of Michigan) comprises 146 full-text transcriptions of literary and administrative works, including many out-of-print volumes of the *Early English Text Society*. Publicly available.
 - **Middle English Texts Series Online** (Universities of Rochester and Michigan) comprising 421 annotated editions of key literary works for teaching and research, with 53 editions forthcoming, prepared by TEAMS (The Consortium for the Teaching of the Middle Ages). Full-text transcriptions; publicly available.
 - **Parker on the Web** (Stanford University and Corpus Christi College Cambridge), comprising high resolution images and detailed cataloguing of 559 manuscripts, presented within an online environment which includes annotation tools. Database; available by subscription only.
 - **Europa Inventa** (University of Western Australia); descriptive catalogue of medieval manuscripts held within Australian institutions. Database; publicly available.
- 2.3** It is impossible to estimate the quantity of data which will be made available at this stage, although many of the resources comprise over a million words. Overall *Manuscripts Online* will provide access to a wealth of primary source texts which are central to the study of English language, literature and history during the middle ages, ranging from small, AHRC-funded editions to large cataloguing projects and including resources which are freely available to the public and available via subscription.
- 2.4** The project will not require content providers to modify their data, repository or web service. Building upon lessons learnt from the *Connected Histories* project, *Manuscripts Online* will implement a form of clustering and distributed, semantic searching which is non-invasive, using the following methodology:
- 2.4.1** The project and each content provider will sign a 'Material Transfer Agreement' which sets out the terms of what the project can and cannot do with an individual dataset. The content provider will then supply the HRI with a copy of its text-based data in a format which contains the highest level of structure (so if a transcription was originally encoded in SGML and then converted to XHTML for display on the web we will ask for the SGML version).
 - 2.4.2** Upon receiving the data the HRI will conduct a technical audit in order to establish its structure, character encoding and the method for URL construction (so that users can be directed from the search result to the full document on the content provider's actual website). More than the *Connected Histories* project, understanding the character encoding will be critical for accurate searching and representation of the search results because the presence of non-Latin characters within texts of this period means that editors and transcribers have represented them using a variety of methods, such as notation, customised fonts and, latterly, Unicode.
 - 2.4.3** The data will be analysed and tagged using a Natural Language Processing (NLP) technique known as *automated entity recognition*. This process uses word context combined with controlled vocabularies to intelligently identify words and phrases which belong to particular categories of knowledge. For example, we will seek to identify the names of people and places, but also technical terminology relating to dialect, codicology and illumination. The HRI will build upon the NLP algorithms already developed as part of the *Connected Histories* project.
 - 2.4.4** Additionally, the different editorial approaches taken to representing non-Latin characters and abbreviation marks will be harmonised to a standard representation using machine-transferrable character entities (eg. þ and &yogh;) so that users can search these characters consistently.
 - 2.4.5** The results of the NLP analysis will be verified by the Editorial Group, after which the HRI will generate RDF profiles of the data for public re-use, Lucene indexes of the NLP-processed data for structured searching and Lucene full-text indexes for free-text keyword searching.
 - 2.4.6** The indexes will be hosted on the *Manuscripts Online* site for use by the search engine but users will be directed to the live datasets when viewing the full text of individual results. In the case of commercial sites not accessible without a subscription, the search facility will point to the location of the relevant material, without delivering the full protected content.
 - 2.4.7** Each search result will include a link which will enable users to download the RDF profile for each document whilst each dataset will be accompanied by a high level description of the resource covering areas such as the scope of the resource, the technical methods employed in creating the resource and information about the project and content provider.
- 2.5** The entire procedure for indexing distributed sources, once established, will be systematised as a semi-automated process, as we have done with *Connected Histories*. This means that the process of analysing new datasets using NLP and then indexing them for inclusion within the service becomes much easier once the

algorithms for these types of datasets have been established, with only occasional modification of the algorithms being required if a dataset exhibits an unusual data structure or character encoding.

- 2.6** End users will be able to explore 15 collections of resources in the first instance, differentiated by resource type, time, subject, language, provenance and accessibility (eg. publicly available or available via subscription), with further resources to be added beyond the period of JISC-funding (see section 8 below). Users will be able to conduct full-text keyword searching across the resources, using filters to limit the body of materials to be searched such as resource type and time period. Full-text searching will be available irrespective of whether the resource is a fully transcribed text, a catalogue or a database, even though the information which constitutes each resource type will be different. For example, an end user will be able to ask the question: show me all the documents which a) contain the word "thwitel", b) was thought to have been written between the years 1400 and 1410 and c) are concerned with merchandise. The search engine will draw upon lexicographical indexes provided by the Dictionary of Middle English to identify and thus query words and spellings which are similar, in an attempt to overcome issues of non-standardised spelling during this period. Thus a search for "thwitel" (which means a knife or dagger) will also retrieve documents containing "thwitelle", "thwitil", "thwetil", "twitel", "twhitel", "thikil" and "twikill". In addition to full-text searching, the end user will be able to combine this with structured searching using data which has been identified and indexed as part of the *automated entity recognition* process (NLP). This structured searching will enable users to search for documents which have references to specific people and places, but also documents which have similar physical and linguistic features. For example, an end user will be able to ask the question: show me all documents which are a) thought to have originated from the West Midlands (i.e. dialectally) and b) include historiated initials as part of the illumination. Each result will include a snippet of the relevant text with search terms highlighted, but users will be directed to the content provider's actual website when wishing to view the full document. Building upon expertise from the JISC-funded *Locating London's Past* project, and where metadata permits it, there will also be an option to have search results plotted on Google Maps using the Google Maps API. The mapping of search results will use the four Shepherd maps of medieval Britain for the periods 1087-1154, 1200-1450 and 1455-1494,² showing key towns and regions, overlaid onto Google Maps. The mapping feature will be valuable to researchers because it will enable them to visualise results such as the relationship between the provenance of documents (where they were created and owned) and the dialects in which they were written (which possibly indicates where the scribe came from).
- 2.7** In addition to the core functionality of search, *Manuscripts Online* will provide a number of Web 2.0 features in order to build community, break down the traditional culture of research ownership which persists in the discipline, and add value to the data. These Web 2.0 features will consist of:
- Tools for sharing and annotating the search results and the individual documents.
 - A citation generator which includes clean, 'cool' URIs in order to encourage consistent citation of both the *Manuscripts Online* website and the resources to which it provides access. The generator will create an accurate bibliographic citation which the end user can paste into their essay or article.
 - Blogging of search results, in which users will be encouraged to register with *Manuscripts Online* and blog their discoveries as a modern incarnation of *Notes and Queries*. This will enable small but meaningful observations to be transmitted to the research community, given that none of the resources within *Manuscripts Online* has ever been easily scrutinised before within the context of one another. The need to register for this 'self-publishing' facility will also enable the HRI to solicit valuable but anonymous user information such as nationality and institutional affiliation in addition to data gathered via Google Analytics.
 - A public API with accompanying documentation will enable other service developers to make use of *Manuscripts Online*'s search engine when designing their own PC-based services or mobile apps. The API is a core deliverable because the intention of this project is to develop a solid, sustainable service upon which other, value-adding services can be developed by third parties. The API will be developed in conformity with the *Connected Histories* API, which the HRI plans to release in 2012, so that third-party services can easily pull in data from both websites thereby representing a chronological range of 1000 to 1900. This project will include the HRI working with technical staff at The National Archives and the University of Stanford, USA (*Parker on the Web*) to explore use of the API within the context of their existing services.
- 2.8** Project deliverables will include the *Manuscripts Online* search engine and website, public availability of the API and conference presentations which will publicise the project and generate interest from other, potential content providers beyond the period of funding. The Editorial Group will be responsible for the project's

² See Shepherd, William R. *Historical Atlas*, (New York: Barnes and Noble, 1929)

dissemination strategy and will write four articles for academic journals (eg. *Literary & Linguistic Computing*, *Journal of the Early Book Society*, *Speculum* and *Journal of British Studies*) in addition to presentations at four international conferences which are key fora for research in the language, literature and history of written and early printed culture in the middle ages: The Early Book Society Conference (St. Andrews), The International Congress on Medieval Studies (Western Michigan University), The International Medieval Congress (Leeds) and The New Chaucer Society Congress (Portland, Oregon). The Group will also host two Quadrivium workshops which will use the website to deliver research training to postgraduates and a one-day conference.

3. Workplan

2011		2012											2013	
N	D	J	F	M	A	M	J	J	A	S	O	N	D	J
WP1														
WP2														
	WP3													
					WP4									
							WP5							
		WP6												
							WP7							
												WP8		
												WP9		

WP 1: Project start-up and legal agreements

WP 2: Search engine design and build

WP 3: Data Bundle #1 (AHRC-funded datasets, *Middle English Dictionary*, *Compendium of Middle English*, *Middle English Texts Series*, *Europa Inventa*)

WP 4: Data Bundle #2 (*British History Online*, *EEBO*, *EEBO-TCP*, *British Literary Manuscripts*, *Parker on the Web*)

WP 5: Data Bundle #3 (*The National Archives*, *Catalogue of Illuminated Manuscripts*, *Early English Text Society*)

Left until the last in anticipation of content licences taking longer to arrange.

WP 6: User interface design and development

WP 7: Web 2.0 features and geographical mapping using Google Maps

WP 8: Public API development and documentation

WP 9: Public user testing and project evaluation

A detailed workplan showing tasks within the Work Packages:

	Tasks	Deliverables
1. Nov '11	<ul style="list-style-type: none"> - Project meeting #1. - WP1: Prepare and submit project plan to JISC. - WP1: Begin legal agreements between partner institutions, and with participating websites. - WP1: Create project website and JISC project page. - WP1: Establish Stakeholder Panel. - WP2: Data and process models established. 	<ul style="list-style-type: none"> 1. Project plan 2. Project website 3. Collaboration Agreement
2. Dec '11	<ul style="list-style-type: none"> - WP1: Tendering process for visual design agency. - WP2: Search engine specification. - WP3: Begin gazetteer and dictionary building for the NLP. 	<ul style="list-style-type: none"> 4. Search engine specification 5. WP1 completed
3. Jan '12	<ul style="list-style-type: none"> - Project meeting #2. - WP6: User focus group #1 - WP3: Define and test the NLP algorithms using Data Bundle #1. - WP3: Data Bundle #1 analysed, processed and indexed. 	
4. Feb '12	<ul style="list-style-type: none"> - WP3: NP evaluated and Data Bundle #1 continued. - Quadrivium workshop #1 	
5. Mar '12	<ul style="list-style-type: none"> - Project meeting #3. - Stakeholder Panel meeting #1. - WP6: Visual and interactive designs for review (1st draft). - WP2: Search engine build. 	<ul style="list-style-type: none"> 6. First version of interface designs

	- WP3: Data Bundle #1 continued.	
6. Apr '12	- WP3: Search engine testing using Data Bundle #1. - WP6: Visual and interactive designs 2nd draft for review. - WP3: Resource descriptions prepared for Data Bundle #1	7. Prototype search interface (WP2) 8. Data Bundle #1 (WP3) completed
7. May '12	- Project meeting #4. - WP4: Refine and test the NLP algorithms using Data Bundle #2. - WP4: Data Bundle #2 analysed, processed and indexed. - WP6: Visual and interactive design applied (website build)	9. Final version of interface designs 10. Prototype website (WP6)
8. Jun '12	- Project meeting #5. - WP6: User focus group #2. - WP7: Specification for Web 2.0 and mapping features. - WP4: NLP evaluated and Data Bundle #2 continued.	
9. Jul '12	- Website and search engine revisions in response to user focus group #2. - WP4: Search engine testing using Data Bundle #2. - WP4: Resource descriptions prepared for Data Bundle #2 - WP7: Scans of the Shepherd maps prepared.	11. Data Bundle #2 (WP4) completed
10. Aug '12	- Project meeting #6. - WP5: Refine and test the NLP algorithms using Data Bundle #3. - WP5: Data Bundle #3 analysed, processed and indexed. - WP7: Geographical mapping implemented and tested.	
11. Sep '12	- WP5: NLP evaluated and Data Bundle #3 continued. - WP7: Revisions to geographical mapping.	
12. Oct '12	- Project meeting #7. - Stakeholder Panel meeting #2. - WP5: Resource descriptions prepared for Data Bundle #3 - WP7: Web 2.0 features implemented and tested. - WP9: User testing with wider audience.	12. Data Bundle #3 (WP5) completed
13. Nov '12	- WP7: Revisions to Web 2.0 features - WP8: Specification and implementation of the public API. - WP9: User testing with wider audience. - Quadrivium workshop #2 - One-day conference (Leicester)	13. Web 2.0 features and mapping (WP7)
14. Dec '12	- Project meeting #8. - WP8: Testing and full documentation of the public API. - WP9: User testing with wider audience. - WP9: Final report drafted.	14. API publicly available (WP8)
15. Jan '13	- Full launch of website with press publicity. - WP9: Completion of final report.	15. Final website 16. Final report

4. Project Management

4.1 The allocation of responsibilities and funding within the project will be laid out in a formal agreement between the six partner institutions to be signed in the first month of the project, covering the period of the project, subject to rolling renewal thereafter.

4.2 Overall direction of the project will be the responsibility of the HRI Digital Manager, Michael Pidd, with Dr Orietta Da Rold, in direct consultation with the Editorial Group: Prof. Linne Mooney, Prof. Wendy Scase, Prof. Jeremy Smith, Dr Estelle Stubbs and Prof. John Thompson. Pidd will also oversee the HRI technical staff. Day-to-day project management will be performed by Dr Sharon Howard or Keira Borrill, depending upon availability, reporting to the Project Director.

4.3 Technical development will be undertaken by the Humanities Research Institute (HRI). The HRI uses Scrum-style project management for research projects in which the technical development optimised to apply new techniques and functionality to diverse datasets. The visual and interactive design of the website will be sub-contracted to a commercial web design agency after a tendering process. The design agency will direct the user focus groups and provide the HRI with XHTML design templates for the final pages.

- 4.4** Development of the resource descriptions and organisation of the one-day conference and two *Quadrivium* workshops will be the responsibility of Da Rold on behalf of the Editorial Group whilst evaluation and sign-off of the NLP for each resource will be the responsibility of the entire Group.
- 4.5** Eight meetings will be held (every two months) by all project staff, rotating between the partner institutions. Two of these meetings will include usability sessions with typical end users (academics and students) and so the rotation between institutions is intended to guarantee a variety of participants at each session.
- 4.6** A Stakeholder Panel, comprising representatives of participating websites, digital humanities specialists and subject specialists, will meet twice during the course of the project, in months 5 and 12, in order to assist with the design, guarantee quality and ensure alignment with the intellectual objectives of similar sites such as *Connected Histories*. The following individuals will be invited to participate:
- Prof. Tim Hitchcock (Univ. of Hertfordshire; digital humanities expert responsible for *Connected Histories*).
 - Dr Jane Winters (Univ. of London; Head of Publications for *British History Online* and responsible for *Connected Histories*; medieval historian).
 - Prof. Vincent Gillespie (Univ. of Oxford; J.R.R. Tolkien Professor of English Literature and Language; Executive Secretary of the Early English Text Society) or Caroline Palmer (Editorial Director, Boydell & Brewer).
 - Dr Ian Johnson (Univ. of St Andrews; medievalist).
 - Aleksandr Drozdov (The National Archives; Enterprise Architect; technology expert).
 - Dr Kathleen Doyle (British Library; Curator of Illuminated Manuscripts).
 - Prof. Andrew Prescott (Kings' College London; Head of DDH; medieval historian).
 - Dr Toby Burrows (Univ. of Western Australia; Manager of the eResearch Support and Digital Developments Unit; medievalist).
 - Representative from JISC.

5. Technical Standards

- 5.1** The Natural Language Processing algorithms will be written as Java applications using the *NetBeans* IDE. The algorithms will build upon those already developed by the HRI for *Connected Histories*. The process involves building gazetteers and controlled vocabularies (such as a dictionary of codicological features) and then using syntax, word adjacency and pattern recognition to identify possible matches, even if the entity does not actually appear in the gazetteer or dictionary. This technique is generally known in Natural Language Processing as *named entity recognition* and will be used on unstructured text; i.e. text in which entities are not already explicitly identified. So NLP would not be required for databases in which codicological features are already identifiable from the record structure. The search engine will be written using Apache Lucene and communication between the search form on the website's front-end and the Lucene indexes will use JSP in the form of an application programming interface (API), making requests using HTTP GET and returning results in an XML format for transformation and on-screen rendering. The value of the API approach to communication between the website's front-end and Lucene back-end is that the principle of making the data accessible via an API is built into the very core of the system. The API will be documented and made publicly available for use by third parties. Although the *Manuscripts Online* website will use a generic XML format for returning results, third parties will be able to request that results be returned in TEI XML, RDF or JSON. The website's static pages and visual design will be written in XHTML, CSS and JSP. All technology used by this project will conform to open standards and will be accompanied by comprehensive documentation (line-by-line code commenting accompanied by build and maintenance guidance).
- 5.2** The website and all data will be maintained on two mirrored servers at the HRI. We will use the servers which were originally funded by the JISC for the *Connected Histories* project in order to capitalise on the original investment in this infrastructure. The servers are 'virtual', and so we are able to expand data storage and processing capabilities as required (a key consideration for services in which content will continue growing beyond the initial period of funding). The cost of sustaining this infrastructure over the longer term is described in section 8 below. The use of two servers ensures that there is always a live backup through mirroring that will permit maintenance of the site without interruption to the public service, and enable load balancing. Offsite backup of all data is done daily by the University's computing services. All programming code and associated data files are checked into a CVS repository (Subversion), which is also regularly backed up.

6. IPR

6.1 A 'Material Transfer Agreement' (content licence) will be agreed with each content provider, outlining the terms and conditions under which we can use their data. MTAs will be signed by the content providers and the Univ. of Sheffield only, rather than all other project partners, in order to speed up the process of exchanging contracts. All data generated by the project will be owned jointly by the project partners. A collaboration agreement to cover this shared IPR and the management of the project as a whole, will be completed by the end of month one of the project.

6.2 All code generated by the project will be available as Open Source and governed by a Creative Commons Licence. In addition to a publicly available API, RDF profiles of each search result will be available for download.

7. Risk Analysis

7.1 Pr=Probability (1-5), Se=Severity (1-5), Sc=Score (Pr x Se)

Risk	Pr	Se	Sc	Action
Staffing - illness or unavailability	2	1	2	The HRI employs four technical officers each of whom could undertake the implementation of this project. Further HRI staff would be deployed as necessary. Da Rold and Pidd can increase their commitment should either of them become unavailable
Failure to design a working search facility	1	5	5	Failure to meet this objective could be due to a number of reasons: intellectual approach, methodology or project management. The HRI has significant experience with developing these types of search facilities, and has already implemented them in a working form for the <i>Connected Histories</i> project. The size and experience of the development team and Editorial Group will mitigate these risks.
Failure to implement the public website	1	5	5	The process of designing and implementing a website is clearly understood, and subject to minimal risk. The HRI has extensive experience of both developing sites and implementing site functionality.
Failure to participate on the part of the creators/owners of other sites.	2	2	4	Early negotiation with IP providers and the existence of clear legal agreements will mitigate this risk. No single resource is essential for this project. The <i>Connected Histories</i> project has provided the development team with considerable experience and it should be noted that no issues concerning content providers have arisen to date.
Failure to adapt our existing NLP algorithms to identify new categories of data	2	1	2	The NLP algorithms have been developed by the HRI and are well understood by the development team. Further, members of the development team have a very good understanding of the subject matter (medieval written culture) and will be supported by a significant body of subject expertise (the Editorial Group). However, in the event that the NLP struggles to accurately identify new classes of entities (such as illumination terminology), the existing capabilities of these algorithms (identifying person and place names) will be deemed to add considerable value to the data by the community.

8. Sustainability

8.1 As with *Connected Histories*, the federated nature of the *Manuscripts Online* site creates two issues for long term sustainability: the sustainability of the *Manuscripts Online* site itself (search engine, semantic data/indexes and the Natural Language Processing algorithms) and the sustainability of the content repositories upon which it draws.

8.2 The HRI will host the completed website and search facility. As a digital humanities centre within the University of Sheffield's Faculty of Arts & Humanities, the HRI already hosts and maintains a large number of complex websites and datasets and the sustainability of a service such as *Manuscripts Online* is part of its core mission.

8.3 The full, publicly accessible versions of the datasets will be sustained by the content providers' own business models, some of which are commercial while others are publicly funded. Many of the repositories are owned by individual academics who cannot necessarily guarantee the long-term support of their institution and in the event of a resource becoming permanently offline the HRI will negotiate to re-host the data. However, one should emphasise that, as with *Connected Histories*, the federated nature of *Manuscripts Online* ensures that

the unavailability of an individual content provider's website or repository will not endanger the *Manuscripts Online* website as a whole.

- 8.4** The federated content model also means that there is a minimum overhead for sustainability of the *Manuscripts Online* website itself. The long term sustainability of a service such as this has been greatly informed by a charging formula which the HRI developed and currently implements for *Connected Histories*. This formula is intended to cover the costs of maintaining and growing the server infrastructure as well as the costs associated with adding further content on a regular basis beyond the initial period of funding (every six months). The HRI rents dedicated server infrastructure from the University's Computing Services, and so its costs are real annual charges.
- 8.5** The charging formula for post-project content providers covers the cost of analysing, processing (NLP) and indexing the data (£580), arranging the MTA and authoring a resource description page for the website (£340); plus the cost of physically storing the data which begins at £100 for data under 1 GB. Physical storage is calculated using the pre-indexed size of the data. The aim of the charging formula is to cover the costs of our predicted server requirements but also ensure that inclusion with *Manuscripts Online* is not financially prohibitive for small datasets. As a one-off cost to the content provider (typically at £10,20 + VAT) our model is premised on the service generating more income per annum than the actual, combined cost of storage for new datasets in order to cover storage costs incurred by datasets which have been added in previous years. The HRI reserves the right to revise this charging policy in the future. However, if no income were to be forthcoming at all the HRI would seek to maintain *Manuscripts Online* from other budgets as part of its core mission.

9. Dissemination and Impact

- 9.1** One of the key objectives of *Manuscripts Online* is to raise awareness of discrete electronic resources which are under-used or which have become forgotten over time. The advantage of the federated search model is that one only has to build brand awareness of a single site in order to increase the impact of all participating content. Its value can be seen in the willingness of commercial sites such as Cengage and ProQuest to participate, because the federated search facility essentially becomes another marketing tool for their subscriptions. The HRI has a lot of experience in disseminating knowledge about the sites which it develops and hosts and the HRI works closely with the media relations office at the University. Further, the Medieval Manuscripts Research Consortium (MMRC) who comprise the project's Editorial Group, actively disseminates knowledge and promotes capacity-building within the discipline by hosting workshops for academics, postgraduates and undergraduates, developing training resources and representing UK HE within *Carmen*, the worldwide medieval network.
- 9.2** The impact of this resource will be felt among academic researchers, postgraduate and undergraduate students, librarians, archivists and professionals within the heritage sector. There are a considerable number of lecturers and postgraduate students in early and late medieval English history, language and literature in the UK for whom *Manuscripts Online* would become a key research resource. Further, there are approximately 139,695 undergraduate students taking modules which involve the study of primary sources dating from the early to late medieval period.³ This is in addition to the considerable research interest which exists amongst overseas scholars and students for whom English is a common heritage or an influence upon the written culture of their own nation (for example, there is a shared written culture between Britain and continental Europe during the medieval period). The USA alone has over 4,068 lecturers in medieval English history, language and literature⁴ and the value of *Manuscripts Online* internationally can be seen in the number of USA-based datasets. In preparation for this submission the MMRC surveyed medievalists in the UK and USA via conferences and were overwhelmed with endorsements as the survey went viral via newsgroups.
- 9.3** To reach these academic audiences a dissemination plan will be implemented from the start of the project, overseen by the Editorial Group: announcements will be made at conferences and *Carmen* meetings, four articles will be written for academic journals and two Quadrivium training workshops will be held for doctoral students in addition to participating in the JISC's own information events. Upon the completion of the funded phase of *Manuscripts Online*, the Editorial Group will co-ordinate a publicity strategy with the media offices of their institutions with a view to publicising the story in the popular press and interest magazines such as *BBC History*. Beyond the funded period a central aspect of our ongoing dissemination plan will be six-monthly updates to *Manuscripts Online*. These updates will consist of adding new resources from more content providers and thus generating renewed interest within the research community.

³ Higher Education Statistics Agency. Figures for 2009/10 covering History, Philosophy, English and Archaeology.

⁴ Based on subscriptions to the Medieval Academy of America by senior, tenured academics.

10. Budget

10.1 Although the project comprises six institutional partners, four partners have agreed to contribute their directly allocated costs, amounting to £8,433. In addition, Leicester will contribute 20% and Sheffield 50% of its directly allocated costs.

Directly Incurred Staff	Nov 2011 - March 2012	April 2012 - March 2013	TOTAL £
Total Directly Incurred Staff (A)	£5,739	£22,956	£28,695
Non-Staff			
Travel and expenses: 8 meetings x 8 people @112.50 train and lunch; advisory meetings (2 meetings x 13 people @ 150 train and lunch; 4x international conferences @ £1,000 each	£900	£12,052	£12,952
Technical work by HRI (129 days @ 290.60 per day)	£7,498	£29,991	£37,489
Website visual design (to tender)	£0	9,600	£9,600
Dissemination budget: 2x workshops and 1x conference @ £2,000 each	£0	£6,000	£6,000
Total Directly Incurred Non-Staff (B)	£8,398	£57,643	£66,041
Directly Incurred Total (A+B=C) (C)	£14,137	£80,599	£94,736
Directly Allocated			
Estates (Sheffield)	£93	£370	£463
Estates (Leicester)	£162	£649	£811
Directly Allocated Total (D)	£4,760	£19,019	£23,779
Indirects: University of Sheffield	£429	£1,714	£2,143
Indirects: University of Leicester	£935	£3,738	£4,673
Indirects: University of York	£234	£938	£1,172
Indirects: University of Birmingham	£245	£978	£1,223
Indirects: University of Glasgow	£211	£845	£1,056
Indirects: Queen's University Belfast	£128	£513	£641
Indirect Costs (E)	£2,182	£8,726	£10,908
Total Project Cost (C+D+E)	£21,079	£108,344	£129,423
Amount Requested from JISC	£16,863	£92,749	£109,612
Institutional Contributions	£4,216	£15,595	£19,811
Percentage Contributions over the life of the project	Partners 15.3%	JISC 84.7 %	Total 100%
No. FTEs used to calculate indirect and estates charges	No FTEs 7	Which Staff: Howard, Pidd, Da Rold, Scase, Mooney, Smith, Thompson	

11. Previous Experience

- 11.1 Michael Pidd** (Project Director, Univ. of Sheffield) will be responsible for managing all technical aspects of this project in line with the Editorial Group's vision. As HRI Digital Manager for the Humanities Research Institute he has 17 years experience in all aspects of the planning, management and delivery of ICT-based research projects in the Arts and Humanities, including *Connected Histories* which the HRI continues to support and develop.
- 11.2 Dr Orietta Da Rold** (Project Director, Univ. of Leicester) is a lecturer in medieval literature, specialising in manuscript and textual studies. She has more than nine years experience developing electronic resources on projects funded by the AHRC, ESF, the Bibliographical Society and AMARC, including Director of EMProject. She is the editor and co-editor of five books and the author of 18 articles. Orietta will have specific responsibility for developing the project's resource descriptions and workshops.
- 11.3 Prof. Linne Mooney** (Univ. of York) is Professor of Medieval English Palaeography, specialising in late medieval English literature, manuscripts and scribes. Linne was recently PI of an AHRC major research grant, *Late Medieval English Scribes*, and she is the author, co-author and editor of six books and over 50 articles.
- 11.4 Prof. Wendy Scase** (Univ. of Birmingham) is Geoffrey Shepherd Professor of Medieval English Literature. She is director of the AHRC-funded Vernon Manuscript Project and Manuscripts of the West Midlands Project. She co-organised an ESF Exploratory Workshop on Applying Semantic Web Technologies to Medieval Manuscript Research in 2009 and recently presented at the ESF-COST Strategic Workshop on Safeguard of Cultural Heritage (Univ. of Florence, 2011). She is author, editor and co-editor of 16 books and author of 37 articles and chapters.
- 11.5 Prof. Jeremy Smith** (Univ. of Glasgow) is Professor of English Philology, specialising in English historical linguistics, the history of Scots and English in Scotland, and medieval English and Scottish textual cultures. Jeremy has had two major AHRC research grants recently, including the *Editing Burns Project*, and he is the author of 6 books (2 co-authored), 3 edited collections, 35 chapters and over 16 articles.
- 11.6 Dr Estelle Stubbs** (Univ. of Sheffield) is a Research Fellow in the School of English, having been Research Associate for the AHRC-funded *Late Medieval English Scribes* which shows her analyses of the hands of more than 400 scribes using 17,000 thumbnail images. Estelle has worked on the development of electronic resources for the medieval period for 17 years and has considerable experience in manuscript transcription, codicological analysis and description. She is co-author of *Scribes and the City* with Prof. Mooney (forthcoming).
- 11.7 Prof. John Thompson** (Queen's University Belfast) is Chair of English Textual Cultures, specialising in early book history and the history of printing and publishing, English literary production and reception (c. 1300-1600), anglophone Ireland and textual editing. He has been the Director of two large AHRC research projects, *Geographies of Orthodoxy* and *Imagining History*, in addition to authoring *Imagining the Book* and 15 chapters for co-authored volumes.
- 11.8 Dr Sharon Howard** (Project Manager, Univ. of Sheffield) served as project manager on the JISC-funded *Connected Histories*, *Crime in the Community* and *Locating London's Past* projects. Both her technical and historical expertise and experience in project management make her the ideal manager for this project.
- 11.9 The Humanities Research Institute** (HRI) at the University of Sheffield is one of the UK's leading centres for the study and use of digital technology within the Arts and Humanities. As a major research facility within the University, it currently comprises 21 active research projects and nine staff. The HRI has considerable expertise in all aspects of this project, having been involved in the conception, management and delivery of digital humanities research projects since its establishment in 1992. See <http://www.shef.ac.uk/hri>
- 11.10 Katherine Rogers and Jamie McLaughlin** (Univ. of Sheffield) are HRI Technical Developers with considerable experience of delivering large, technically complex projects. They have worked on the technical implementation of over 20 ICT projects in the Arts and Humanities, working closely with academic investigators. Both Kathy and Jamie have considerable expertise in NLP development, XML processing, data mining, API design and web design technologies including Java, JSP, XSLT, XPath and XHTML. Kathy was principally responsible for the technical implementation of the *Connected Histories* site whilst Jamie is currently the technical developer for *Locating London's Past*.

12. Supporting Letters Letters from the British Library, The National Archives, Boydell & Brewer and Parker on the Web could not be provided for the deadline due to staff vacations. However, all offer support in principle and the HRI has strong, existing relationships with them (including proposed technical collaboration on API and NLP development with TNA and Parker on the Web). The BL and TNA are already participating in *Connected Histories*.