



JISC Project Plan

Project Information			
Project Identifier	<i>To be completed by JISC</i>		
Project Title	Manuscripts Online: Written Culture from 1000 to 1500		
Project Hashtag	MSSO		
Start Date	November 2011	End Date	January 2013
Lead Institution	University of Sheffield		
Project Director	Michael Pidd		
Project Manager	Sharon Howard		
Contact email	sharon.howard@sheffield.ac.uk		
Partner Institutions	University of Birmingham, University of Glasgow, University of Leicester, University of York, Queen's University Belfast		
Project Webpage URL	http://www.manuscriptsonline.org		
Programme Name	<i>Content 2011-13</i>		
Programme Manager	Alastair Dunning		

Document Information			
Author(s)	Michael Pidd, Orietta Da Rold, Sharon Howard		
Project Role(s)	Project Directors, Project Manager		
Date	December 2011	Filename	MSSO_projectplan.doc
URL			
Access	This report is for general dissemination		

Document History		
Version	Date	Comments
1.0	15.12.2011	

Table of Contents

NB : This table of contents 'auto-populates' - to update the table of contents – place cursor in the table of contents, right-click your mouse, click 'update field', select appropriate option

1.	Project Overview.....	3
1.1	Project Summary.....	3
1.2	Objectives.....	3
1.3	Anticipated Outputs and Outcomes	3
1.4	Overall Approach	6
1.5	Anticipated Impact.....	10
1.6	Stakeholder Analysis.....	12
1.7	Related Projects.....	12
1.8	Constraints	13
1.9	Assumptions.....	13
1.10	Risk Analysis.....	13
1.11	Technical Development.....	15
1.12	Standards	15
1.13	Intellectual Property Rights	16
2	Project Resources.....	16
2.1	Project Partners.....	16
2.2	Project Management	17
2.3	Project Roles.....	18
2.4	Programme Support.....	19
3	Detailed Project Planning.....	19
3.1	Evaluation Plan	19
3.2	Quality Assurance	19
3.3	Dissemination Plan	24
3.4	Exit and Embedding Plans	26
3.5	Sustainability Plans	27
	Appendices	28
	Appendix A. Project Budget	28
	Appendix B. Workpackages	28

1. Project Overview

1.1 Project Summary

Manuscripts Online will enable users to search an enormous body of online primary resources relating to written and early printed culture in Britain during the period 1000 to 1500.

A single search engine will enable users to undertake sophisticated full-text searching of literary manuscripts, historical documents and early printed books which are located on websites owned by libraries, archives, universities and publishers. Users will be able to search the resources by keyword, but also by specific keyword types, such as person and place name, date, author, scribe, manuscript feature and illumination terminology, thanks to techniques which we are using called *automated entity recognition*. Additionally, users will be able to visualise search results using maps of medieval Britain and add their own annotations to the data for public consumption, thereby building a knowledge base around this critical mass of primary source materials.

Automated entity recognition is a Natural Language Processing technique within information science whereby algorithms are able to intelligently identify the occurrences of specific types of words, such as names, concepts and terminology, using three methods: dictionaries (such as a historical gazetteer of place names), lexical pattern matching and syntactic context.

Manuscripts Online will be of interest to researchers and students in the fields of medieval English language, literature and history and it will become a sister site to the JISC-funded Connected Histories website (<http://www.connectedhistories.org>) which already provides similar search services for the period 1500-1900.

Manuscripts Online is funded by the JISC and supported by the Humanities Research Institute at the University of Sheffield and specialists in medieval studies at the universities of Leicester, Birmingham, Glasgow, York and Queen's University Belfast.

More information about *Manuscripts Online* can be found at <http://www.manuscriptsonline.org>.

1.2 Objectives

1.2.1 *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.

1.2.2 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 in addition to identifying the names of people and places, it will also focus on identifying different language instances (such as Latin, Anglo-Saxon and Anglo-Norman, as distinct from Middle English); 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; d) over the longer term, it will explore how value can be added to search through the use of 'activity data'.

1.2.3 *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.

1.2.4 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.

1.2.5 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.

1.2.6 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.

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

Data Bundle 1

- **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. *NB Needs to be in first bundle because of its role in the search methodology.*
- **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) *Middle English Grammar Project* (Jeremy Smith; Glasgow)
 - 6) *Late Medieval English Scribes* (Linne Mooney; York);
 - 7) *The Norman Blake Editions of the Canterbury Tales* (Norman Blake; Sheffield);
 - 8) *The Auchinleck Manuscript* (Alison Wiggins and David Burnley; Sheffield). Full-text transcriptions and databases.
- **Europa Inventa** (University of Western Australia); descriptive catalogue of medieval manuscripts held within Australian institutions.
- **The Cause Papers** of the Church Courts of the diocese of York (Borthwick Institute, Univ. of York) A descriptive catalogue with accompanying images. 524 documents fall within the period 1000 to 1500.
- **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.

Data Bundle 2

- **British History Online** (Institute of Historical Research, Univ. of London) Full-text transcriptions and databases comprising approximately 38,000 documents ranging from administrative and ecclesiastical history to economic and intellectual history.
- **British Literary Manuscripts Online: Medieval & Renaissance** (Gale Cengage) Database comprising c.500,000 pages of searchable metadata with accompanying digital facsimile images.
- **Early English Books Online** (Historic Books Platform and ProQuest) Database 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.
- **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*.
- **Parker on the Web** (Stanford University and Corpus Christi College Cambridge), comprising high resolution images and detailed cataloguing of 559 manuscripts.

Data Bundle 3

- **The National Archives**, Descriptive catalogues 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.
- **Online Catalogue of Illuminated Manuscripts** (British Library) A descriptive catalogue describing the codicology, palaeography, illumination and provenance of 2,000 illuminated manuscripts originating in England, Wales, Scotland and Ireland.
- **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*.
- **Middle English Texts Series Online** (Universities of Rochester and Michigan) 421 annotated editions of key literary works for teaching and research, with 53 editions forthcoming.

1.3 Anticipated Outputs and Outcomes

Output / Outcome Type (e.g. report, publication, software, knowledge built)	Brief Description
Website	A publicly available website with search engine, user tagging/commenting, blogging and mapping features.
API	Publicly documented API. The website itself will illustrate the API in action.
Documentation	Documentation, historical background and help texts (approximately 5,000 words)
Events	Two Quadrivium workshops which will use the website to deliver research training to postgraduates and a one-day conference
Report	Final Report, documenting project outcomes, the results of the evaluation, and pathways for further development of the resource.
Publications	Five articles in peer-reviewed journals. See section 3.3 below.
Dissemination	Presentations at six conferences (five international). See section 3.3 below.

1.4 Overall Approach

1.4.1 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:

1.4.2 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).

1.4.3 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.

1.4.4 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.

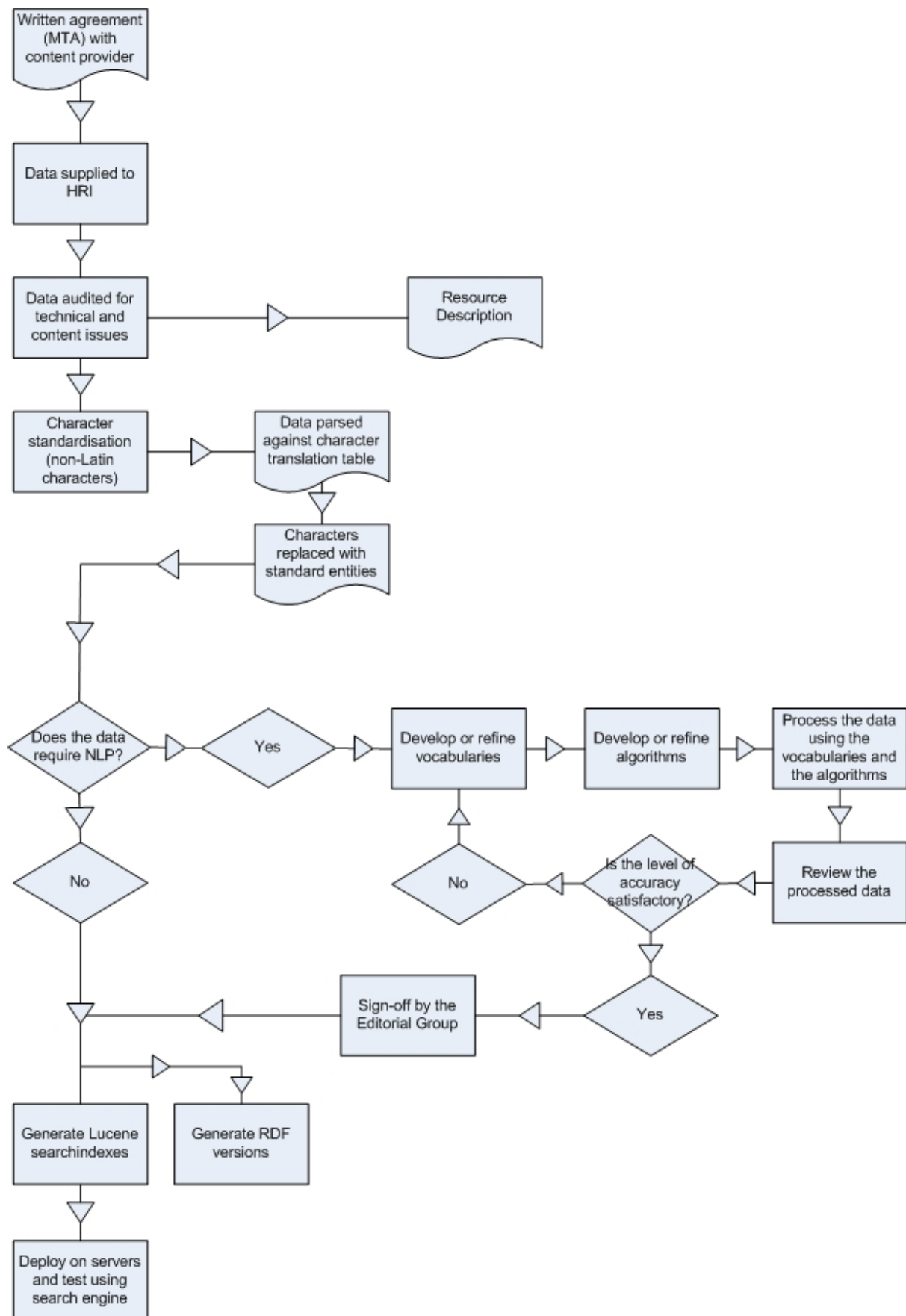
1.4.5 The data will be analysed and tagged using a Natural Language Processing (NLP) technique known as *automated entity recognition*. This process uses word context combined with controlled vocabularies to intelligently identify words and phrases which belong to particular categories of knowledge. For example, we will seek to identify the names of people and places, but also different language instances, such as Latin, Anglo-Saxon and Anglo-Norman. The HRI will build upon the NLP algorithms already developed as part of the *Connected Histories* project.

1.4.6 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.

1.4.7 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.

1.4.8 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.

1.4.9 Process Flowchart #1 – Data Preparation – version 1 (13 Nov 2011)

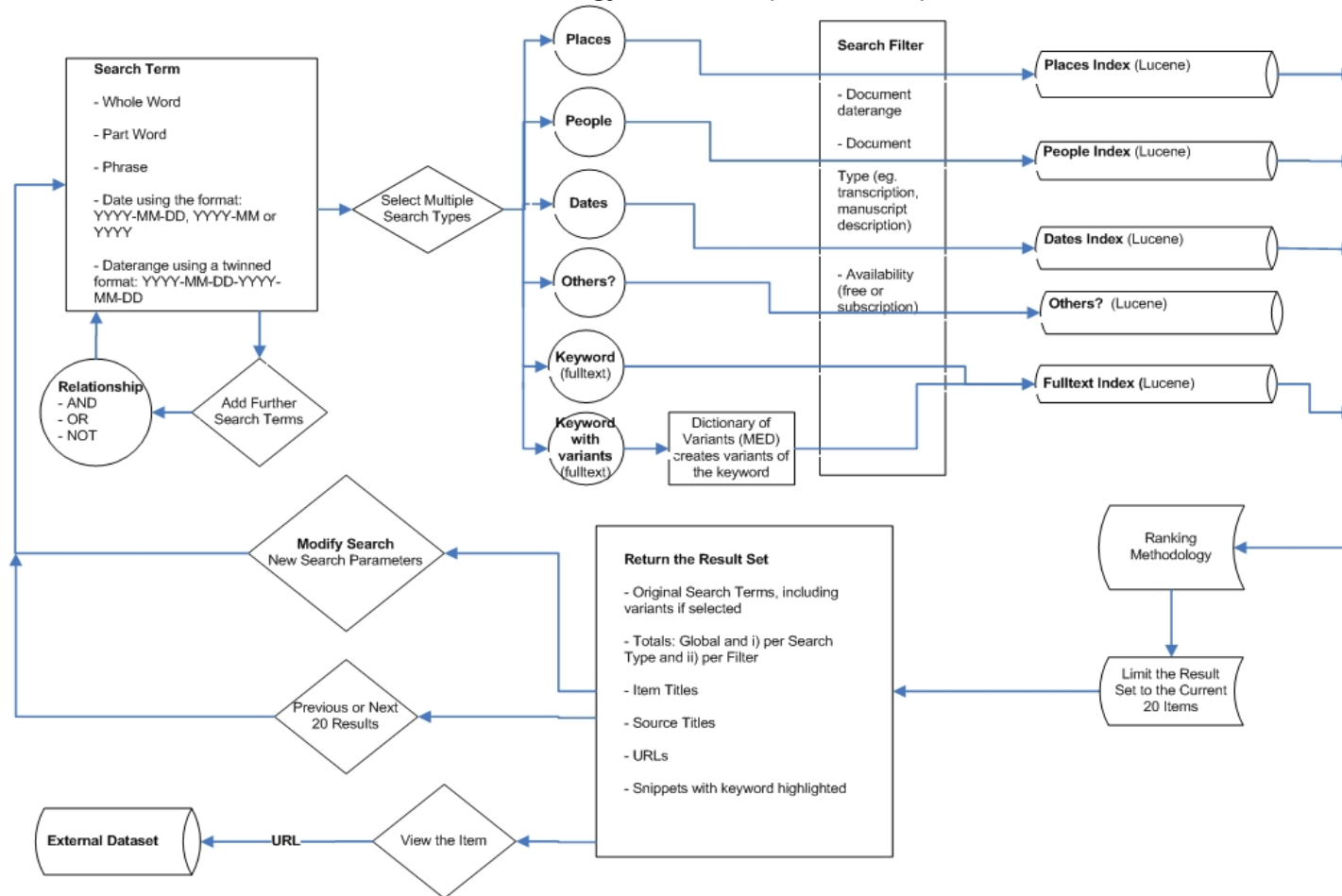


1.4.10 The entire procedure for indexing distributed sources, once established, will be systematised as a semi-automated process, as we have done with *Connected Histories*. This means that the process of analysing new datasets using NLP and then indexing them for inclusion within the service becomes 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.

1.4.11 End users will be able to explore 21 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. 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. 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.

1.4.12 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 draw upon different languages such as Latin, Anglo-Saxon and Anglo Norman. 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.

1.4.13 Process Flowchart #1 – Search Methodology – version 1 (13 Nov 2011)



1.4.14 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,[1] 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).

1.4.15 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.

1.5 Anticipated Impact

1.5.1 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. 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. The USA alone has over 4,068 lecturers in medieval English history, language and literature and the value of *Manuscripts Online* internationally can be seen in the number of USA-based datasets.

We intend to measure impact in the following ways:

- Building a user community throughout the life of the project (rather than at the end) by inviting people to join a 'testers mailing list' and gain early access to test releases of the site. Test releases will be accompanied by an online user survey.
- Tracking usage via Google Analytics.

- Building a citation generator into the system to guarantee accurate citation in research papers.
- Social bookmarking for tracking citations/mentions (Delicious.com).

Impact Area	Anticipated Impact Description
maintain research excellence	<p>1. The project will assist the research reputations of all partner institutions (particularly in the current economic climate) and hopefully strengthen applications for RCUK and EU funding.</p> <p>2. Within the partner institutions and within the research community more widely, Manuscripts Online will undoubtedly improve access to online primary sources and facilitate new research.</p> <p>3. The integrated searching and Web 2.0 features should hopefully break down the 'silo' mentality which the discipline can be prone to. It should certainly help to bridge disciplinary divide between history and language/literature, thereby enabling key research questions to be explored more collaboratively.</p>
maintain teaching & learning excellence	<p>4. The project should strengthen the partner institutions' 'offer' to prospective undergraduates, demonstrating that they are key departments for the study of the discipline, especially for Russell Group institutions which emphasise research-led teaching.</p> <p>5. Within the partner institutions and within the research community more widely, Manuscripts Online should make the teaching of medieval literature and history easier because it will be easier for teachers to pull together exemplar materials from a range of sources.</p> <p>6. In particular, we hope that it will help to build research capacity in the field of medieval studies by inspiring a new generation. This is a key activity of the Quadrivium Workshops.</p>
be more effective/save money	<p>7. There is an argument that less time spent searching for distributed resources increases productivity.</p> <p>8. Manuscript studies is particularly prone to the need for researchers to make speculative trips to archives in search of evidence. One could argue that the more evidence we provide in one place, in a digital form, the easier it is for researchers to focus their research trips.</p>
have a positive impact on wider society	
be ready for technology needs in the future	<p>9. The data standards which we are using, the API, and the act of essentially aggregating key research resources should open up opportunities for the data to be re-used in the future. We cannot predict what devices will use our data or how, but good practice in data standards and data access (APIs) should position us for technology changes.</p> <p>10. APIs are all well and good, but technologists and organisations need to know that they are there. Publicity for our API should hopefully encourage people to think about new services.</p>

1.6 Stakeholder Analysis

Stakeholder	Interest / stake	Importance (H/M/L)
HE research community (academics and post graduates)	This search facility will improve the discovery of relevant primary source material, allowing new types of structured searching and analysis to facilitate innovative research. These user needs are critical. Manuscript studies in particular is traditionally hampered by the need to have knowledge of a wide range of sources which are often small, discrete and distributed.	H
HE undergraduates	Undergraduates frequently need access to medieval english language, literature and history sources for their primary source based dissertations, special subjects, and essays. This will become a major resource for advanced level undergraduate english language, literature and history teaching and learning in the UK and the rest of the English speaking world.	H
The digital humanities community, in particular those creating digital resources.	The resource will provide a new way to drive usage to a range of frequently under-used primary sources and historical websites, creating a new ecology of data within which all kinds of medieval english language, literature and history websites can thrive. For those in other disciplines, it will provide a model for the creation of new search facilities for distributed electronic resources. The API should permit easy re-use of the website's search technology and underlying data for the development of new services.	H
Schools and tertiary students and educators.	The resource should make the teaching of medieval english language, literature and history much easier and, in particular, will help to build capacity in the discipline at an early age. For example, it will be easier to teach fundamental principles of medieval english language and literature such as non-standardised spelling, literacy and the transmission of texts and texts/knowledge.	M

1.7 Related Projects

1.7.1 Primarily *Connected Histories*, but also international data aggregation services such as *Europeana Regia*, *NINES*, *18thConnect* and *CERL* (Consortium of European Research Libraries). Internally within the HRI, there is a related project to provide a single point of access for information about HRI APIs, including demonstrators of how our APIs could be re-used within different types of services (mobile as well as desktop). In the case of *Europa Regia*, which is thematically related to *Manuscripts Online*, we hope to make use of their API and thus facilitate a two-way exchange of data.

1.8 Constraints

1.8.1 The project must be completed within 15 months, and the amount of staff time available is limited: Sharon Howard has 50% of her time available, and the HRI has budgeted for 129 days of technical work. All the work must be completed within these time constraints. The project must capitalise and build upon existing knowledge, techniques, resource and infrastructure.

1.9 Assumptions

1.9.1 The project has been designed on the basis that the work can be done within the staff time available as outlined under 'Constraints'.

1.9.2 We have assumed that the datasets, for which we have outline permission to include in the project and post on the website, will actually be delivered in good time.

1.9.3 We have allocated £9600 to website design and assume that will be sufficient.

1.9.4 We have assumed that our NLP techniques can be extended to other types of entities.

1.9.5 We have assumed that communication with TNA's own API (the only method by which we are permitted to incorporate their data within the project) will produce no significant overhead on the performance of our search methodology. The lessons learnt will be applied to *Connected Histories*.

1.10 Risk Analysis

Risk Description	Probability (P) 1 – 5 (1 = low 5 = high)	Severity (S) 1 – 5 (1 = low 5 = high)	Risk Score (PxS)	Detail of action to be taken (mitigation / reduction / transfer / acceptance)
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 Connected Histories project. The size and experience of the development team and EditorialGroup 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	2	2	4	Early negotiation with IP providers and the existence of clear legal agreements will mitigate this risk. No single resource is

creators/owners of other sites.				essential for this project. The Connected Histories 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. Exploring this risk is a research activity from our perspective, and therefore welcome.
Failure to secure necessary permissions from the Middle English Dictionary (or additional dictionaries)	2	3	6	The MED plays a key part in the search methodology's ability to handle spelling variants within documents of this period and will be invoked by all search requests. If MED do not grant permission for their data to be used in this way the search methodology will be reduced in its effectiveness. The risk is reduced by good relationships with the content provider, but ultimately it would only lower the effectiveness of search, rather than jeopardise it.
Poor performance of the TNA API within the project's search infrastructure. For TNA political reasons their inclusion within the project will require a different methodological approach: fetching live data in response to search queries via their own API and then combining it with search results acquired via the project's API.	2	1	3	As explained above, no single resource is essential to this project. However, evaluating this risk is a research activity from our perspective, and so welcome.

1.11 Technical Development

1.11.1 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.

1.11.2 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.

1.11.3 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).

1.11.4 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.

1.12 Standards

Name of standard or specification	Version	Notes
XHTML	1.0	Website front-end
CSS	2.0	Website front-end
JSP	2.0	Server-side processing of search queries, results and user-related data
Apache Lucene		Search engine and indexing
XML		Data exchange format for returning results from Apache Lucene to the JSP scripts (via the API)
TEI XML	P5	Possible data return format for the API (an option for future developers)
RDF		Possible data return format for the API (an option for future developers). Also visible on the public front-end
JSON		Possible data return format for the API (an option for future developers)

Java	6.0	Server-side API
GoogleMapsAPI	3.0	API for an underlying, third-party mapping service
HTTP GET		RESTful search request protocol (via the API)
MySQL	5.x	Storage of user data relating to Web 2.0 and blogging/commenting/tagging features

1.13 Intellectual Property Rights

1.13.1 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.

1.13.2 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 two of the project.

1.13.3 All code generated by the project will be available as Open Source and governed by a Creative Commons Licence.

1.13.4 In addition to a publicly available API, RDF profiles of each search result will be available for download and reuse by researchers.

1.13.5 The final report and substantial web content (i.e. background pages and web design) will be jointly owned by the project partners and made available free of charge for non-commercial use.

2 Project Resources

2.1 Project Partners

Humanities Research Institute, University of Sheffield, lead institution

Main contact: Michael Pidd

Roles: Project management, technical development, data retrieval, analysis and processing, sustainability

University of Leicester

Main contact: Dr Orietta Da Rold

Roles: dissemination (lead), quality control, subject specialism, Editorial Group (commissioning post-project resources for inclusion in the service)

University of Sheffield

Main contact: Dr Estelle Stubbs

Roles: dissemination, quality control, subject specialism, Editorial Group (commissioning post-project resources for inclusion in the service)

University of York

Main contact: Professor Linne Mooney

Roles: dissemination, quality control, subject specialism, Editorial Group (commissioning post-project resources for inclusion in the service)

Project Identifier: MSSO
Version: 1.0
Contact: Sharon Howard
Date: December 2011

University of Birmingham

Main contact: Professor Wendy Scase

Roles: dissemination, quality control, subject specialism, Editorial Group (commissioning post-project resources for inclusion in the service)

University of Glasgow

Main contact: Professor Jeremy Smith

Roles: dissemination, quality control, subject specialism, Editorial Group (commissioning post-project resources for inclusion in the service)

Queen's University Belfast

Main contact: Professor John Thompson

Roles: dissemination, quality control, subject specialism, Editorial Group (commissioning post-project resources for inclusion in the service)

2.2 Project Management

2.2.1 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.

2.2.2 All staff designated to work on the project are currently in post. Our risk assessment indicates that suitable alternative staff are available should any become ill or otherwise unavailable.

2.2.3 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, reporting to the Project Director.

2.2.4 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.

2.2.5 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.

2.2.6 Eight meetings will be held at Sheffield (every two months) by all project staff, rotating between the partner institutions.

2.2.7 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. Peter Ainsworth (Univ. of Sheffield; French medievalist and digital humanities expert).
- 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).
- Peter White (Senior Product Manager, ProQuest).
- Dr Elmer Mittler (Europa Regia)
- Dr Adam Fraquhar (Head of Digital Library, British Library)
- Alastair Dunning (JISC).

2.3 Project Roles

Team Member Name	Role	Contact Details	Days per week to be spent on the project
Michael Pidd	Project Director responsible for managing all technical aspects of this project	Humanities Research Institute, University of Sheffield m.pidd@shef.ac.uk	0.5
Dr Orietta Da Rold	Project Director responsible for developing the project's resource descriptions, workshops and dissemination strategy	University of Leicester odr1@leicester.ac.uk	0.5
Prof. Linne Mooney	Co-Investigator, member of Editorial Group (dissemination, quality control, future commissioning)	University of York linne.mooney@york.ac.uk	0.1
Prof. Wendy Scase	Co-Investigator, member of Editorial Group (dissemination, quality control, future commissioning)	University of Birmingham w.l.scase@bham.ac.uk	0.1
Prof. Jeremy Smith	Co-Investigator, member of Editorial Group (dissemination, quality control, future commissioning)	University of Glasgow Jeremy.Smith@glasgow.ac.uk	0.1
Dr Estelle Stubbs	Co-Investigator, member of Editorial Group (dissemination, quality control, future commissioning)	University of Sheffield e.stubbs@shef.ac.uk	0.1
Prof. John Thompson	Co-Investigator, member of Editorial Group (dissemination, quality control, future commissioning)	Queen's University Belfast j.thompson@qub.ac.uk	0.1
Dr Sharon Howard	Project Manager	HRI, Sheffield sharon.howard@shef.ac.uk	2.5
Katherine Rogers	Technical Developer (NLP, data processing, API, search and interface)	HRI, Sheffield k.m.rogers@shef.ac.uk	2.5
Jamie McLaughlin	Technical Developer (API documentation and elaboration)	HRI, Sheffield j.mclaughlin@shef.ac.uk	0.5

2.4 Programme Support

2.4.1 As with Connected Histories, support is requested in relation to garnering the enthusiastic support of participating websites. Further advice on issues of IPR, sustainability, and dissemination will also be sought, as will the programme managers' participation in the stakeholder panel meetings in project months 5 and 12.

3 Detailed Project Planning

3.1 Evaluation Plan

Timing	Factor to Evaluate	Questions to Address	Method(s)	Measure of Success
Weekly	Formative: overall progress	Is the project on track?	Monitoring by the PI. Project meetings. Oversight from JISC.	Meeting each milestone within the planned month.
	Formative: user engagement	Are user's sufficiently engaged with the project?	Online surveys. 2x user focus groups.	Survey responses to be completed by 80% of subscribers to the mailing list for user testers. Clear outcomes from the user focus groups.
	Formative: visual and interactive design	Do users like/agree with what we are doing?	Online surveys. 2x user focus groups. Peer review by the Stakeholder group.	Positive responses to the surveys. Constructive responses from the user focus groups and stakeholder meetings.
	Formative: NLP accuracy	Is the NLP achieving an acceptable level of accuracy?	Quality Assurance by the Editorial Group.	To be determined by the Editorial Group
Monthly	Formative: dissemination plan	Does the wider community know about our project?	Subscriptions to the mailing list for user testers.	A steady increase of new users joining the mailing list during the life of the project: 10 new testers per month.
	Formative: relationship with JISC	Is the project contributing to the overall Programme?	Regular review of other project activities within the Programme. Self-assessment at project meetings.	Alignment with the Programme and other projects within it, based on feedback from the JISC.
Weekly	Formative: technical research	Are we learning new things (eg. in our development of NLP)?	Project technical meetings.	Positive feedback from technical developers concerning

				challenging areas.
	Summative: overall success	Have we clustered thematically-related content?	Final peer review meeting by the Stakeholder Group. Feedback from users. A post-launch user survey Web logs.	Positive feedback from the Stakeholder Group. User requests to contribute additional data. Positive responses to the survey. Repeat visits to the site and significant site activity during individual visits.
	Summative: technical design	Apart from the facility to cross-search datasets, does our search add value over and above what is available via the individual resource websites?	Final peer review meeting by the Stakeholder Group. Feedback from users. A post-launch user survey	Positive feedback from the Stakeholder Group. Positive responses to the survey.
	Summative: technical research	Have we advanced our NLP knowledge and expertise?	Interview the developers as part of the annual Staff Review and Development Scheme. HRI strategic review.	Positive feedback from technical developers concerning challenging areas. Confidence to address similarly challenging areas in the future.
	Summative: relationship building	Have we built new relationships with public and commercial third parties?	HRI strategic review.	Number of new relationships. Plans for ongoing collaborations in other areas.
	Summative: limited representation of medieval studies within the faculty at Sheffield	Has the project increased interest in medieval studies within the partner institutions, but particularly at the University of Sheffield (which has low level interest)?	Assess the development by existing Faculty staff of teaching modules and research applications via our existing seats on Faculty committees. Track the appointment of new teaching staff by Faculty. Web logs	Number of new teaching modules and research applications in development. Number of new staff appointed. Repeat visits to the site and significant site activity during individual visits by users from within the project's partner institutions.
	Summative:	Is the server	Daily load testing	<i>Nagios</i> problem

	technical infrastructure	infrastructure holding up?	and CPU/stroage monitoring using <i>Nagios</i>	reports to be generated no more than four times per month, of which 100% of reports to be WARNING and 0% of reports to be CRITICAL.
	Summative: sustainability	Is our business model right?	Annual re-assessment of the HRI's charging formula (developed for <i>Connected Histories</i>).	Number of new resources offered for inclusion. Income for the first two years following the end of JISC funding to cover the infrastructure costs for all current data for 10 years.
	Summative: sustainability	Is there interest in contributing additonal resources by the wider community?	Responses to a public request to the academic community	Number of new resources offered for inclusion.
	Summative: API	Is the API being used by third parties?	API tracking using call-back functions	Number of API deployments. Number of API requests.
	Summative: project management	What lessons have we learnt overall and how could we improve?	Exit meeting (by skype) involving all project members. HRI strategic review. Interview the developers as part of the annual Staff Review and Development Scheme.	Positive feedback from the project team, technical developers and the JISC.

3.2 Quality Assurance

3.2.1 The HRI will employ Scrum-style project management. These general quality principles will be applied to each cycle of development:

3.2.2 At the beginning of each cycle of development, the outputs of that cycle will be fully specified in order that test scenarios can be produced for those outputs.

3.2.3 Code will be produced in Java, using best practice object-orientated techniques and the application of patterns. Code inspections will take place within the HRI to assure compliance with this criteria.

3.2.4 Code will be fully documented using Java's javadoc facility and line comments. Code inspection and the production of javadoc documentation will ensure compliance.

3.2.5 A complete change history of code will be maintained. Code will be checked into *Subversion* version control and change comments logged, so that a complete history of changes can be viewed.

3.2.6 Where XML is produced, it will be validated against an appropriate DTD or schema. A test harness will be created to ensure that all XML has been validated and the results of running this test harness recorded.

3.2.7 NLP output will be validated by the Editorial Group, who will develop a formal evaluation methodology in month 2 covering matters such as the target level of accuracy and sign-off. NLP output which falls beneath the target level of accuracy will be re-processed in order to improve quality.

3.2.8 Further testing will take place at the end of each cycle to ensure that the product functions correctly as a component of the larger software system. This may include usability testing and testing of any application programming interfaces.

3.2.9 Decisions about requirements, changes to requirements, testing and other project management information will be recorded using the *Basecamp* project management tool to fully document the history of the project development.

Output / Outcome Name	NLP processing of datasets	
When will QA be carried out?	Who will carry out the QA work?	What QA methods / measures will be used?
After NLP processing of each dataset, prior to indexing.	Three members of the Editorial Group will analyse a sample each.	Statistical analysis of true positives, false positives, true negatives and false negatives across three samples for each dataset requiring NLP.

Output / Outcome Name	Search methodology	
When will QA be carried out?	Who will carry out the QA work?	What QA methods / measures will be used?
Months 6, 9, 12	Kathy Rogers	Unit and integration testing with issue tracking (as each Data Bundle is added)
Months 6, 9, 12	Mike Pidd, Orietta Da Rold, Sharon Howard	User testing with scenarios and issue tracking (as each Data Bundle is added)
Month 8	Kathy Rogers	System testing with issue tracking (upon completion of the interface)
Months 8, 12, 13	Editorial Group	User testing with scenarios and issue tracking (during User Focus Group #2 and Quadrivium Workshop #2)

Output / Outcome Name	API development	
When will QA be carried out?	Who will carry out the QA work?	What QA methods / measures will be used?
Months 6, 9, 13	Kathy Rogers	Unit and integration testing with issue tracking (during search engine build)
Month 13	Kathy Rogers and Jamie McLaughlin	System testing with issue tracking (after specification and development of public API features)
Month 14	Jamie McLaughlin, Parker on the Web (Stanford)	User testing with scenarios and issue tracking (after completion of documentation)

Output / Outcome Name	Website usability and testing	
When will QA be carried out?	Who will carry out the QA work?	What QA methods / measures will be used?
Month 3	Design agency	Consultation on visual and interactive designs (during User Focus Group #1)
Month 5	All project members	Review of visual and interactive designs - draft 1
Month 5	All project members	Review of visual and interactive designs - draft 1 (during Stakeholder Panel #1)
Month 6	All project members	Review of visual and interactive designs - draft 2
Month 8	Editorial Group	User testing with issue tracking
Month 8	Design agency	User testing with scenarios (during User Focus Group #2)
Month 12	All project members	User testing with scenarios (during Stakeholder Panel #2)
Month 13	Editorial Group	User testing with scenarios (during Quadrivium Workshop #2)

Output / Outcome Name	Mapping features	
When will QA be carried out?	Who will carry out the QA work?	What QA methods / measures will be used?
Month 8	Editorial Group	Specification review and scenario testing (during User Focus Group #2)
Month 10	Kathy Rogers	Unit and integration testing with issue tracking
Months 12 and 13	All project members	User Testing with scenarios (during Stakeholder Panel #2 and Quadrivium Workshop #2)

Output / Outcome Name	Web 2.0 features	
When will QA be carried out?	Who will carry out the QA work?	What QA methods / measures will be used?
Month 8	Editorial Group	Specification review and scenario testing (during User Focus Group #2)
Month 12	Kathy Rogers	Unit and integration testing with issue tracking
Months 12 and 13	All project members	User Testing with scenarios (during Stakeholder Panel #2 and Quadrivium Workshop #2)

Output / Outcome Name	Server performance and resiliency	
When will QA be carried out?	Who will carry out the QA work?	What QA methods / measures will be used?
Months 6, 9 and 12	Kathy Rogers	Load testing with issue tracking (as each Data Bundle is added)
Monh 12	Kathy Rogers	Load testing with issue tracking (when TNA is added)
Months 12, 13, 14 and 15	Kathy Rogers	Load testing with issue tracking (throughout WP9 and at weekly HRI technical meetings following launch)

3.3 Dissemination Plan

Timing	Dissemination Activity	Audience	Purpose	Key Message
Feb 2012	Quadrivium workshop 1 (Jeremy Smith)	Postgraduate students, postdoctoral and early career researchers	Raise awareness; inform; educate; engage	This is what we are going to do. Do you like the way in which it will look and function?
Mar 2012	Publicity material, Medieval Academy Meeting in St Louis (Wendy Scase)	Medievalists; HE and FE teachers and students; librarians; curators; archivists; independent scholars; international	Raise awareness; inform; engage	This is what we are doing and how we are doing it. Do you want to contribute in the future?
May 2012	Lecture, University of Tier, Germany (Orietta Da Rold)	Internaional scholarly community, undergraduate and posgraduate students	Raise awareness; inform; engage	This is what we are doing and how we are doing it. Do you want to contribute in the future?
Jun 2012	Paper, Writing Europe Before 1470: A Colloquium, Bergen, Norway (Estelle Stubbs and Orietta Da Rold)	International scholarly community	Raise awareness; inform; engage	This is what we are doing and how we are doing it. Do you want to contribute in the future?
Jul 2012	Publicity material, New Chaucer Society Congress in Portland, Oregon	Medievalists; HE and FE teachers and students; librarians; curators; archivists; independent scholars; international	Raise awareness; inform; engage	This is what we are doing and how we are doing it. Do you want to contribute in the future?

Sep 2012	Paper, Germany (Wendy Scase)	International scholarly community	Raise awareness; inform; engage	This is what we are doing and how we are doing it. Do you want to contribute in the future?
Sep 2012	Paper, Sheffield (Estelle Stubbs)	Palaeography, medievalists and early modernists; HE teachers and students	Raise awareness; inform; engage	This is what we are doing and how we are doing it. Do you want to contribute in the future?
Sep 2012	Paper and publicity materials, Digital Humanities Congress, Sheffield (Michael Pidd)	Digital humanities; international	Inform; engage; educate	This is why APIs are next step after good data
Nov 2012	Quadrivium workshop 2, Sheffield (Orietta Da Rold and Estelle Stubbs)	Postgraduate students, postdoctoral and early career researchers	Raise awareness; inform; educate; engage	Is this useful for you? If not, what should we change?
Dec 2012	API Workshop, Sheffield (Michael Pidd)	Digital humanities, technologists, information scientists, JISC community	Inform; educate; engage	How can we use APIs in the digital humanities?
Jan 2013	Project Conference, Leicester (Orietta Da Rold)	Scholarly community; digital; stakeholders	Promote; engage	It's here! Please use and contribute.
Jan 2013	Final report	JISC Community	Inform; educate	What did we do and what did we learn from it?
Jul 2013	Paper/demonstration, Early Book Society, St Andrews (Michael Pidd)	Medievalists; HE and FE teachers and students; librarians; curators; archivists; independent scholars; international	Promote; engage	It's here! Please use and contribute.
Dec 2012	Article in <i>Digital Medievalist</i> (co-ordinated by Wendy Scase)	Digital humanities; scholarly community; international		
Sep 2012	Article in <i>Medieval Academy of America Newsletter</i> (co-ordinated by Orietta Da Rold)	Medievalists; International		

Jan 2013	Article in <i>Journal of the Early Book Society</i> (co-ordinated by Estelle Stubbs)	Medievalists; Scholarly community; manuscript scholars		
Jan 2013	Article in <i>Computing in the Humanities</i> (co-ordinated by Orietta Da Rold)	Digital humanities; International; Scholarly community		
Mar 2013	Article in <i>Working Papers in the Digital Humanities</i> (forthcoming HRI online journal) by Katherine Rogers and Sharon Howard	Digital humanities; international		

3.3.1 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.

3.3.2 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.

3.3.3 To reach the intended 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.

3.3.4 Upon the completion of the funded phase of *Manuscripts Online*, the Editorial Group will coordinate 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*.

3.3.5 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.

3.4 Exit and Embedding Plans

Project Outputs/Outcomes	Action for Take-up & Embedding	Action for Exit
Manuscripts Online website	The service needs to become the <i>de facto</i> source for anyone doing research or teaching in this area. Therefore, its profile within the community needs to be maintained	The project will move from a development phase to a service phase which is supported by a renewal of the Collaboration Agreement, implementation of

	through a dissemination plan which continues beyond the life of the project, a rolling programme of data ingestion and by encouraging citation.	the charging model and continued planning by the HRI and Editorial Group for twice yearly additions to the content.
Public API	In addition to advertising the API via the <i>Manuscripts Online</i> website and appropriate gatherings (eg. developer workshops and humanities conferences) the HRI will develop a strategy during 2012 for improving take-up and innovative use of its APIs.	The public API will be fully documented and maintained over the long term.
Academic articles	These need to be written and referenced so as to illustrate how the new search facility can be used imaginatively.	Publishing in high level academic journals across the range of relevant sub-disciplines.
Final report	This needs to be publicised as being available to a wide audience of educationalists and the research community.	An informal programme of personal distribution; posting on HRI and partners' websites and through social media

3.5 Sustainability Plans

Project Outputs	Why Sustainable	Scenarios for Taking Forward	Issues to Address
<i>Manuscripts Online</i> website and search	The site is designed to generate sufficient independent income to allow updating.	We will seek to add ten or more new resources add each year, generating an adequate ongoing income to cover infrastructure costs. Income projections are based on a forecasting model developed by the HRI during the <i>Connected Histories</i> project. A charging formula is then derived from this model.	1) Will the <i>Connected Histories</i> charging formula be transferable to <i>Manuscripts Online</i> ? 2) Given that the majority of datasets which will be contributed in the post-funding phase are likely to come from ex-funded research projects, will content providers be able to meet the costs?
Public API	The API will be a documented protocol rather than a physical object and so it can exist while ever <i>Manuscripts Online</i> exists.	We will be keen to see the API used by other third party services and so during 2012 the HRI intends to devote time to a) developing a single point of access for all HRI APIs; b) publicising its APIs; c) inviting students and others to think up innovative data mashups using our APIs and the APIs of other data providers; d) educating the community in the use of APIs through a workshop.	None at present.

3.5.1 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.

3.5.2 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.

3.5.3 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.

3.5.4 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.

3.5.5 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.

Appendices

Appendix A. Project Budget

Appendix B. Workpackages



JISC WORK PACKAGE

WORKPACKAGES	Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
		N	D	J	F	M	A	M	J	J	A	S	O	N	D	J									
1: Project start up																									
2: Search engine																									
3: Data bundle 1																									
4: Data bundle 2																									
5: Data bundle 3																									
6: User interface																									
7: Web 2.0 and mapping																									
8: Public API																									
9: Public user testing																									

Project start date: 1 November 2011

Project completion date: 31 January 2013

Duration: 15 months

Project Name:

Version:

Date:

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
YEAR 1					
WORKPACKAGE 1: Project start-up and legal agreements					
<i><u>Objective:</u></i>					
1. Prepare and submit project plan, work plan, and detailed budget to JISC.	11/11	12/11	Project plan		MP, SH, OD
2. Complete legal agreements between partner institutions, and with participating websites.	11/11	12/11	Collaboration agreement		SH
3. Create project website and JISC website project page.	11/11	11/11	Project website		SH
4. Agree dates for all future project meetings					SH
5. Establish stakeholder panel.	11/11	12/11			MP
6. Set up project management systems: Basecamp; GoogleDocs accounts for everyone; Resource Descriptions Database; MTA tracking	11/11	11/11			SH, MP
7. Agree specification for the visual design ITT	12/11	12/11			MP, OD
8. Tendering process for visual design.	12/11	12/11			SH
9. Proceed with MTA negotiations	12/11	08/12			MP, SH
10. Dissemination Plan formalised	12/11	12/11			OD
11. User testers recruited via mailing list	12/11	01/12			SH, OD
12. Agree programme of blog updates	12/11	01/12			MP, OD, SH
13. Agree dates for user focus groups	12/11	12/11			MP, SH, OD, WS

Project Name:

Version:

Date:

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
WORKPACKAGE 2: Search engine design and build					
<u>Objective:</u>					
14. Establish data and process models.	11/11	11/11			KR, MP
15. Search engine specification.	12/11	12/11	Search engine specification		KR, MP, EG
16. Build search engine.	03/12	03/12	Prototype search interface.		KR
WORKPACKAGE 3: Data bundle 1 (AHRC datasets, Middle English Dictionary, Compendium of Middle English, Middle English Texts, Europa Inventa, Taxatio, Cause Papers)					
<u>Objective:</u>					
17. Build gazetteer and dictionary for NLP.	12/11	01/12			KR, OD
18. Define and test NLP algorithms.	01/12	01/12			KR
19. Analyse, process and index data.	01/12	04/12			KR
20. Evaluate NLP	02/12	04/12			SH, OD
21. Use data to test search engine (unit and integration testing).	04/12	04/12			KR
22. Load testing	04/12	04/12			KR
23. User testing	04/12	04/12			MP, OD, SH
24. Prepare resource descriptions for data.	04/12	04/12		Data Bundle#1 completed	OD

Project Name:

Version:

Date:

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
YEAR 2					
WORKPACKAGE 4: Data bundle 2 (British History Onlie, EEBO, EEBO-TCP, British Literary MSS, Parker on the Web)					
<u>Objective:</u>					
25. Refine and test NLP algorithms.	05/12	05/12			KR
26. Analyse, process and index data.	05/12	07/12			KR
27. Evaluate NLP	06/12	07/12			SH, OD
28. Unit and integration testing with the search engine	07/12	07/12			KR
29. Load testing					KR
30. User testing					MP, OD, SH
31. Prepare resource descriptions for data.	07/12	07/12		Data Bundle#2 completed	OD
WORKPACKAGE 5: Data bundle 3 (The National Archives, Catalogue of Illuminated MSS)					
<u>Objective:</u>					
32. Refine and test NLP algorithms.	08/12	08/12			KR
33. Analyse, process and index data.	08/12	10/12			KR

Project Name:

Version:

Date:

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
34. Evaluate NLP.	09/12	10/12			OD, SH
35. Unit and integration testing with the search engine	09/12	09/12			KR
36. Load testing	09/12	09/12			KR
37. User testing	09/12	09/12			MP, OD, SH
38. Resource descriptions	10/12	10/12		Data Bundle#3 completed	OD
WORKPACKAGE 6: User interface design and development					
<u>Objective:</u>					
39. User focus group 1 (Birmingham)	01/12	01/12			WS, design agency
40. First draft of visual and interactive designs for review.	03/12	04/12	First version of interface designs		all, design agency
41. Second draft of visual and interactive designs for review.	04/12	05/12	Second version of interface designs		all, design agency
42. Visual and interactive design applied (website build)	05/12	05/12	Final version of interface designs		KR
43. System testing	06/12	06/12			KR
44. Internal user testing with scenarios	06/12	07/12			EG
45. User focus group 2 (sheffield)	06/12	06/12	Website with search engine, data and user interface	User interface and search engine completed	OD, SH

Project Name:

Version:

Date:

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
WORKPACKAGE 7: Web 2.0 features and geographical mapping					
<u>Objective:</u>					
46. Specification for Web 2.0 and mapping features	06/12	06/12	Specification for Web 2.0 and mapping features		KR, MP, OD, SH
47. Prepare scans of the Shepherd maps	07/12	07/12			SH
48. Implement mapping.	08/12	09/12			KR
49. Unit and integration testing of map features	08/12	09/12			KR
50. Revisions to mapping.	09/12	09/12			KR
51. Implement Web 2.0 features	10/12	10/12			KR
52. Unit and integration testing of map features	10/12	11/12			KR
53. Internal user testing with scenarios	10/12	11/12			EG
54. Revisions to Web 2.0 features	11/12	11/12	Web 2.0 and mapping features.	Web 2.0 and mapping features completed	KR
WORKPACKAGE 8: Public API development and documentation					
<u>Objective:</u>					
55. Specification and implementation of the public	11/12	12/12			KR, JM

Project Name:

Version:

Date:

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
API					
56. Full system test	11/12	11/12			JM, SH
57. Load testing	11/12	12/12			KR
58. Documentation	12/12	12/12	Publicly available API	Public API completed	KR, JM, Stanford
WORKPACKAGE 9: Public user testing and project evaluation					
<u>Objective:</u>					
59. User testing with wider audience (including scenarios).	10/12	12/12		User testing completed	SH
60. Load testing	12/12	01/13			KR
61. Write final report	12/12	01/13	Final report	Full launch of website with press publicity.	MP, OD, SH

Project Name:

Version:

Date:

Workpackage and activity	Earliest start date	Latest completion date	Outputs (clearly indicate deliverables & reports in bold)	Milestone	Responsibility
--------------------------	---------------------	------------------------	--	-----------	----------------

Members of Project Team:

Name	Role
Michael Pidd (MP)	Project Director responsible for managing all technical aspects of this project
Orietta Da Rold (OD)	Project Director responsible for developing the project's resource descriptions, workshops and dissemination strategy
Sharon Howard (SH)	Project Manager
Katherine Rogers (KR)	Technical Developer (NLP, data processing, API, search and interface)
Jamie McLaughlin (JM)	Technical Developer (API documentation and elaboration)
Wendy Scase (WS, EG)	Co-Investigator, member of Editorial Group (dissemination, quality control, future commissioning)
Linne Mooney (LM, EG)	Co-Investigator, member of Editorial Group (dissemination, quality control, future commissioning)
Estelle Stubbs (ES, EG)	Co-Investigator, member of Editorial Group (dissemination, quality control, future commissioning)
JohnThompson (JT, EG)	Co-Investigator, member of Editorial Group (dissemination, quality control, future commissioning)
Jeremy Smith (JS, EG)	Co-Investigator, member of Editorial Group (dissemination, quality control, future commissioning)