



## Project Document Cover Sheet

Project Information			
<b>Project Acronym</b>	WWWoH		
<b>Project Title</b>	World Wide Web of Humanities		
<b>Start Date</b>	1 April 2008	<b>End Date</b>	31 March 2009
<b>Lead Institution</b>	University of Oxford		
<b>Project Director</b>	Eric T. Meyer		
<b>Project Manager &amp; contact details</b>	Kris Carpenter Negulescu, Internet Archive Director, Web Group kcarpenter@archive.org 415.561-6799, ext 1		
<b>Partner Institutions</b>	Hanzo Ltd., London Internet Archive, San Francisco		
<b>Project Web URL</b>	General: <a href="http://www.oii.ox.ac.uk/research/project.cfm?id=48">http://www.oii.ox.ac.uk/research/project.cfm?id=48</a> Collection: <a href="http://ehumanities.archive.org/collections/wwone/">http://ehumanities.archive.org/collections/wwone/</a>		
<b>Programme Name (and number)</b>	JISC/NEH Transatlantic Digitisation Collaboration Programme		
<b>Programme Manager</b>	Alastair Dunning		

Document Name			
<b>Document Title</b>	Project Plan		
<b>Reporting Period</b>	<i>n/a</i>		
<b>Author(s) &amp; project role</b>	Eric T. Meyer, Project Director		
<b>Date</b>	25 April 2008	<b>Filename</b>	WWWoH_ProjectPlan.pdf
<b>URL</b>	<i>n/a</i>		
<b>Access</b>	X Project and JISC internal	<input type="checkbox"/> General dissemination	

Document History		
Version	Date	Comments
0.9	16 April 2008	First draft - ETM
1.0	20 April 2008	Second Draft - KCN
2.0	9 May 2008	Third Draft – KCN, ETM

## Table of contents

Project Document Cover Sheet .....	1
1 Overview of Project.....	3
1.1 Background .....	3
1.2 Aims and Objectives.....	4
1.3 Overall Approach.....	5
1.4 Methodology and standards .....	6
1.5 Project Outputs.....	10
1.6 Project Outcomes.....	11
1.7 Stakeholder Analysis.....	11
1.8 Risk Analysis .....	11
1.9 Standards .....	12
1.10 Technical Development.....	13
1.11 Intellectual Property Rights .....	13
2 Project Resources .....	14
2.1 Project Partners.....	14
2.2 Project Management .....	14
2.3 Budget.....	17
3 Detailed Project Planning .....	18
3.1 Workpackages / GANTT chart .....	18
3.2 Detailed Schedule .....	19
3.3 Evaluation & Quality Plan.....	25
3.4 Dissemination Plan.....	26
3.5 Exit and Sustainability Plans .....	26
4 JISC Website Template for Projects .....	27



## 1 Overview of Project

### 1.1 Background

The World Wide Web is enormous and is in constant flux, making more web content lost to time than is currently accessible via the live Web. The growing body of archived web material available to researchers is immensely valuable as a record of modern society, but there is often little supporting infrastructure, process or trusted methods available to facilitate domain specific Internet research at a large scale, especially research that takes account of temporal elements. e-Researchers are often expected to individually assemble the necessary data and tools needed to analyze the web without the direct aid of computer scientists. This can be cost prohibitive in terms of resources, necessary expertise, and time. And yet, researchers and academics are actively engaging in Internet research in isolation and in combination with traditional offline methods.

There are at least three general types of tools critical to performing e-Research: tools for data collection and curation, for text analysis, and for hyperlink analysis. The proposal outlined here is mainly focused on creating and assembling a suite of open source tools in the first of these areas, and will employ third party tools for the analysis that these curation and search tools will enable. The aim of this project is to begin to support new methodologies for Internet research built around large collections of web data, using automated tools to extract, index, and analyze the data. This effort will focus on e-Humanities, but a major value of this effort is not just to create an e-Humanities archive, but to create methods of building similar archives on a wide variety of issues, and to tie this project to other projects currently underway at the Oxford Internet Institute and elsewhere. This project will also create novel research interfaces, one browser-based and one or more programmatic interfaces, to the first of many scholarly e-Humanities web collections.

As a way to build and test the tools proposed here, the project will build a focused research collection built around the topic of **World War I**. This particular case was selected for a variety of reasons that make it a compelling test of the methods we are developing. First, the materials available on WWI represent a well-rounded set of humanities materials that will allow us to test the tools against a variety of types of documents and resources. WWI collections on the web include materials that fall under the topics of history, journalism, art, art history, advertising, literature, poetry, political science, military history and others. The types of materials that have been digitized also cover a range of challenges that will allow robust testing of our approach, including multiple formats (text, images of documents, photos, audio), multiple languages (English, German, etc.), and spanning multiple types of collections (government sources, personal collections, university collections, museum collections). Additionally, the topic of WWI is relevant for this transatlantic cooperation because materials on this time period exist both in the UK and in the US, as well as in other countries around the world involved in the conflict. We also think that a collection focused on WWI would have relevance not only for scholars, but also for teachers of secondary and HE/FE courses and for amateurs interested in WWI. Finally, the materials from this time period are out of copyright, which removes much of the legal complexity that could potentially arise in a more contemporary collection.

The World War I collection will include up to 250 million Uniform Resource Identifiers (URIs) which will be assembled from existing data in the Internet Archive as well as new data

collected from the web based on a number of seeds and crawls. This will be done by curatorially selecting source materials and by exploiting the Internet Archive to undertake a hyperlink analysis of the domain, with a particular focus on the US and UK. The collection will be designed to help researchers and policy makers gain an understanding both of the state of the art of e-Humanities and of historical trends and developments regarding World War I collections and studies. The Internet Archive is a unique resource that permits doing historical analysis retroactively, tracing networks over time. In this way, the proposed project will make an important contribution to charting the dynamics of online knowledge in the humanities.

This project represents a new collaboration between the OII, Hanzo and the Internet Archive. Each of the partners has expertise in complementary areas. The fact that this proposal has allowed us to initiate this new collaboration between leading academic, non-profit and private organizations in the US and the UK is an indication that the goal of the program, to create additional research strength by fostering US/UK collaborations, will be furthered in this case. Should this proposal be successful, we envision additional collaboration in the future between the partners in this proposal and also among other partners to collaborate on other, related projects in the area of e-Research and Cyberinfrastructure.

The project will culminate in a public workshop to be held in Oxford, UK which is scheduled for Thursday, March 19, 2009. This event will allow us to demonstrate the tools developed, present the test collection/s, and discuss the lessons learned from the project. Invitees for this workshop will include domain experts (humanities scholars), digital humanities researchers and e-Researchers, curators of digital collections, representatives of funding bodies, and other interested parties.

## **1.2 Aims and Objectives**

The primary aims of this project are to build a set of tools that support data collection and curation of collections of websites to support research, and to build a topical collection as a test of the tools and proof of the concept. We have identified several target audiences for these tools and the collected materials. First, humanities scholars, particularly those interested in the selected topic of World War I, will be able to use the collection this project generates and that will be made available at the Internet Archive. Second, curators and others involved in digitisation projects and collections development will be able to use the methods developed in this project for enhanced resource discovery, federated search, and for building collections on a variety of topics. Third, funding bodies such as JISC and NEH will be able to use and recommend the tools developed here both as a way to indirectly measure impact of web-based collections and to assess developments in this area over time. Finally, researchers interested in the development of web resources, trends in e-Research, uses of the Internet for formal and informal scholarly communication, and social network analysis will be able to use these data and tools to answer a plethora of new questions about the digital humanities domain.

In this proposal, the term 'web collections' is used to describe collections of archived web sites. Both the Internet Archive and Hanzo have extensive experience in web archiving, and are prominent players internationally in the creation of web collections such as those described in this proposal, including the largest of all web collections, the Internet Archive's Web collection accessible via the WayBack Machine.

Key objectives for this project include:

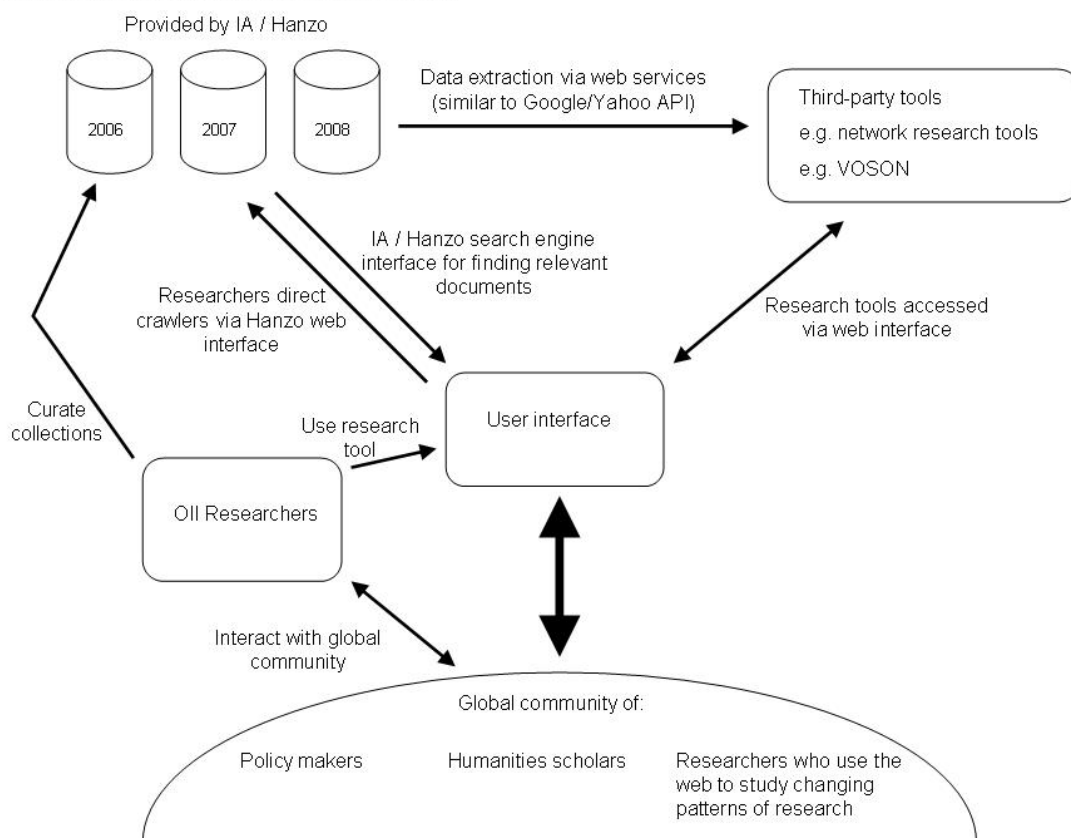
1. To make a significant contribution to the development of an e-Infrastructure enabling research using data from the World Wide Web (WWW), with a particular focus on the

- creation of data resources (web collections) that are (a) focused on particular areas of research; (b) facilitate inter-temporal and historical internet research.
2. To address significant technical and operational problems associated with the creation of appropriate data resources for web research: (a) indexing and analyzing the textual parts of web collections of moderate to large size (i.e., 100's of millions of objects, including metadata and structural information such as citations and hyperlinks); (b) enabling researchers to have appropriate input into the crawler activity i.e. focused crawling (c) unifying access to these data resources through reusable, scalable, open source federated search services and connecting (via web services) the data resources to a wider e-Infrastructure.
  3. To provide two interfaces into the data resource: a web interface so researchers can access web documents via a full-text search engine and a web services interface facilitating programmatic access (i.e. via an API), thus enabling the data resource to be a component of a wider e-Infrastructure facilitating research using web data.
  4. To demonstrate the usefulness of this framework for creating data resources via the creation of a web collection focused on developments in e-Humanities, which will be guided and curated at the OII, and used as input into research.
  5. To conduct innovative research into the role of the humanities in e-Research (relevant to both online research methods and to research policy), with OII researchers accessing the data resource directly via both the web interface described above and a wider e-Infrastructure that is being developed in parallel (thus enabling the use of third-party research tools beyond the OII).

### **1.3 Overall Approach**

The Internet Archive and Hanzo will assemble existing open source technologies, that have been optimized for operation at a scale of 100's of millions of web objects, into a framework that enables developers and technicians to extract, index, and analyze the textual information associated with these objects including metadata and structural information (e.g., citations and hyperlinks) as well as to unify access to web resources through federation and through other novel means. All services and methods will be enabled using open source tools. Within the limited scope and resources of this project, it will not be able to go beyond an initial assessment of e-Humanities research. But this will provide a foundation for future research initiatives well beyond those articulated within this proposal. The overall approach of this project as described here is illustrated in Figure 1.

Indexed (focused) collections of web documents on particular topics (e.g., e-Research in the Humanities)



**Figure 1: Schematic of general approach**

## 1.4 Methodology and standards

IA/H/OII methodologies will not diverge from established standards and best practices with the exception of the novel means with which we intend to combine data formats, open source tools, best practices, and methods to create a new framework for e-Humanities research.

### Assembly of the Collection

The collection will be assembled by the OII in conjunction with IA through an iterative process. The details of this process are documented in the table of workpackages (see pg. 18). In short, the process will begin with a set of key resources developed by OII in consultation with domain experts in the humanities, particularly those with expertise in World War I collections. Note that we have every expectation that this topical focus will prove fruitful; however, we have also identified a secondary collection (on the topic of Tibetan resources) in case we run into insurmountable problems with a focus on WWI. The first list of seeds will be provided to IA, with particularly important resources indicated. IA will then extract the historical webpages from the Internet Archive of the seeds and the pages to which they link. In the case of the designated important resources, if data is not already present in the IA's collection, a new web crawl will be initiated to gather current data. These seeds will then be traced backwards through time in the IA collection to identify current and former links to related materials on the web. IA will then provide a list of new candidate seeds to OII (based on the link extraction) for verification by OII experts in consultation with domain experts. At each step of this process, IA will report to OII the current size of the

collection with respect to the overall target of 250 million objects. Once the complete collection is assembled, IA will extract metadata from the records and index the files. This initial assembly is projected to be completed by the end of August 2008.

### **Automated Metadata Extraction**

With unstructured data such as web pages, it is useful to apply automated techniques to extract and populate metadata records per seed, i.e. per site, in a collection, if not for every capture contained within the collection.

We will not limit in any way the meta data formats that could be supported by the research framework we are designing and deploying for the collection. However, for practical application we have selected an XML format to house the extracted metadata and Researcher-defined collection meta data. For this project we intend to use the Metadata Object Description Schema (MODS): MODS is a schema for a bibliographic element set that may be used for a variety of purposes. As an XML schema, MODS is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. MODS is expressed using the [XML schema language](#) of the [World Wide Web Consortium](#). The standard is maintained by the [Network Development and MARC Standards Office](#) of the Library of Congress (LoC) with input from users.

IA intends to extract metadata using this defined criteria with the goal of populating a single, zipped XML file, parseable by site, formatted in the proposed schema template and zipped using a Unicode compliant zipper. Zipped XML files will be stored in WARC files for preservation of the metadata. The schema will be populated with extracted metadata and OII researcher-provided metadata as described below.

Extracted metadata will likely include:

1. <title>: HTML title tag, from first capture.
2. <dateCaptured encoding="iso8601" point="start">: First capture date, expressed in YYYYMMDD format.
3. <dateCaptured encoding="iso8601" point="end"> Last capture date, expressed in YYYYMMDD format.
4. <physicalDescription><internetMediaType>: By site, each kind MIME type from total capture
5. <abstract>: HTML META Description/Abstract, from first capture
6. <topic>: HTML META Keywords/http-equiv, from first capture.
7. <location>: Wayback resource page for the URL.
8. <recordCreationDate>:Date Record Created in YYYYMMDD format.

Other metadata standards, such as [Dublin Core Metadata standard](#), are likely to be applied and used during the course of this project. We also intend to experiment with [ORE/ATOM](#) for this collection as are applicable and as resources and time permit.

### **Link Extraction & Analysis**

IA plans to use Hadoop, an open source implementation of the Google distributed file system, to facilitate link extraction, canonicalization of source and target URLs, assignment of unique identifiers, and generation of web graphs for this collection. Page Rank analysis may also be applied to the data for the purposes of prioritising new harvests, filtering robots.txt files from materials, etc.

Output of each analysis will be used to identify additional source materials for inclusion in the collection as well as to identify trends for more detailed research.

### **Organization of Materials**

The captures included in the e-Humanities web collection and their associated metadata will be stored in in ARC/WARC files. The WARC/ARC files contain the actual archived documents (html, gif, jpeg, ps, etc.) each preceded by some header information about the document. These archived files are individually compressed and accessible. Each WARC/ARC file has a corresponding DAT file. The DAT files contain meta-information about each document; outward links that the document contains, the document file format, the document size, etc. WARC/ARC and DAT files will be indexed with CDX files for access via browse tools such as the Wayback Machine. CDX files, extended to include link data, and will support access to this collection via programmatic interfaces for the purposes of analysis as well as retrieval. Full text indexes will also be generated to support traditional keyword based search of the collection.

ARC file documentation - <http://www.archive.org/web/researcher/ArcFileFormat.php>

WARC file documentation – The [WARC file format](#) has been submitted to ISO as a proposed file format standard to facilitate the storage, curation, maintenance and migration of web archival materials. The proposed standard is in its final stages of review and is expected to be approved in May 2008. Once approved, the WARC file format will replace the ARC. Archival crawlers like Heritrix have already integrated support for WARC. Additional tools and resources for file management and conversion from ARC to WARC, etc., will be available in early to mid-2008. Hanzo is leading the effort to establish a framework for the creation of WARC tools as well as an initial suite of solutions for WARC/ARC file management, conversion, and manipulation.

DAT file documentation - [http://www.archive.org/web/researcher/dat\\_file\\_format.php](http://www.archive.org/web/researcher/dat_file_format.php)

CDX file documentation - [http://www.archive.org/web/researcher/cdx\\_file\\_format.php](http://www.archive.org/web/researcher/cdx_file_format.php)

DAT/CDX file legend - [http://www.archive.org/web/researcher/cdx\\_legend.php](http://www.archive.org/web/researcher/cdx_legend.php)

Example ARC files:

<http://archive-crawler.sourceforge.net/ARC-SAMPLE-20060928223931-00000-gojoblack.arc.gz>

### **Access to Materials**

Standard end-user, browser based interfaces to the e-Humanities collection will be generated using a dedicated implementation of the open source Wayback machine, two open source full text search engines – Nutch and one other to be defined during the course of this project-, and an open source tool for federating searches. The collections to be federated will be defined during the course of this project.

IA intends to deploy OAI-PMH, OpenSearch, and Web services standard interfaces to the meta data records and full text indexes of the e-Humanities web collection assembled during this project. IA also plans to extend OAI-PMH data provider services to include the e-Humanities web collection for the purpose of integration of results with an open source, federated search tool.

IA currently supports OAI-PMH and Open Search for all web collections created using the Archive-It subscription service. The publicly accessible Katrina and Tsunami collections also

support interfaces designed to enable federation with complimentary collections. The National Library of Australia has used IA's collections to demonstrate federation of IA collections with their national domain harvests but the service is currently only accessible from within their reading rooms. To view an example of an Archive-It collection, go to <http://www.archive-it.org/> to search or browse by institution.

### **Storage Maintenance and Protection of Data**

IA proposes to store maintain, migrate, and preserve fully functional master files and metadata for this collection. IA has evolved preservation and migration policies that are cost effective and scalable to large collections. The preservation strategy includes storage of multiple copies of files on separate servers as well as geographic distribution of copies.

Primary and secondary copies of the materials will be stored and preserved as a collection by IA. One or more access copies of the collection materials will also be generated in support of researcher interfaces to the collection hosted by IA. Access to the collection may or may not be controlled via a user account and login. Individual materials may be available for download but the collection as a whole will not be.

Additional preservation copies of the collection materials will be transmitted to the Library of Alexandria in Egypt and to an agreed upon repository in the UK (currently in negotiation with the Oxford Research Archive (ORA)). An additional replica will be provided to Hanzo in the UK to support development of the search engine framework.

IA's technical migration strategy for digital materials is two-fold: Hardware and file access formats.

- Since its inception in 1996, IA has migrated its collections three times – from tape to disk, and twice more to later generation disk systems. The most recent migration occurred in Fall 2006. Accordingly, technology migration has occurred to and from platforms operating at very large scale.

File access formats: IA attempts to maintain accessibility of files by generating derivative formats from the original record and by evolving the access tools used to recreate the original end-user experience. For example, videos harvested as MPEG4 might be derived to create flash files of those objects. Similarly, the open source Wayback Machine employs coding techniques to make transition from older less sophisticated html to today's, distributed web publishing formats more fluid and less prone to leakage onto the live web. Additionally, we plan to preserve the extracted text and index segments used to create full text search of this collection as items within the IA general repository.

### **Technology and functional components**

The following list of technologies and functional components have been identified for inclusion in the platform. Each will play a unique role in the overall research framework:

1. Open Source Wayback Machine (Java): This tool will be customized and integrated with search services to enable replay of the archival web materials. This tool includes an XML interface. A web services interface will be added by the end of 2008. There may be other tools integrated and customized to query and replay specific files types and/or support specific aspects of the research workflow.
2. Search engine framework: This framework will enable researchers to apply diverse tools and workflows to the analysis of a collection by enabling full text search of the resources using one or more distinct engines. For this project we will integrate NutchWAX/Nutch and one other open source search engine into the framework for application to the WWOne collection. Hanzo will lead the development of the framework.

3. NutchWAX: This tool is one of two search engines that will be used to index the resulting research collection for full text search. New versions of NutchWAX will be available in May and Oct 2008. IA will update the WWOne collection to take advantage of any new features or advances made to this tool and its underlying components. Nutch and Hadoop open source packages are two of the underlying components. The former includes support for the OpenSearch protocol which will be used to demonstrate integration with other full text search services of relevant web archives.
4. A second search engine (to be selected by Hanzo by end of June 2008) will be used to index the collection for full text search.
5. Data extraction: This function will be supported by a set of processes that IA will automate as part of this project.
6. Link extraction: This function will be supported by a set of processes that IA will automate as part of this project.
7. Link graphing: IA has packaged with Heritrix, a separate open source, package for link graphing and analysis using Hadoop. Other link graphing tools may be integrated into researcher workflow prototypes.
8. Meta data extraction: This function will be supported by a set of processes that IA will automate as part of this project.
9. Project Home/Login/Researcher workflow prototypes: The technology used to support the workflow and researcher interfaces will be agreed upon by all participants during the course of the project.

Ease of use for the researcher is perhaps the most challenging issue we aim to address. Each of the stages described above are complex. We hope to simplify each through automation, packaging, documentation and unified interfaces but these are lofty goals to tackle in a single year.

We will not address the following issues as they represent current limitations regarding research of the web and the use of web archives to facilitate research:

- the relevancy of search results and the optimization of full text search services
- the replay of web resources of various types or ages
- the temporal challenges associated with crawled materials
- the elimination and/or reduction of spam except as it pertains to data extraction from an existing archive

There are a number of factors that will influence the success of this project:

- The speed with which we can define and compile the collection will determine the amount of time we have to spend on the creation and iteration of interfaces and workflows.
- The ability to document requirements and to assemble a prototyping environment to facilitate rapid iteration in a distributed and collaborative manner.
- Ability to deploy, test and iterate collaboratively and across time zones

## **1.5 Project Outputs**

1. An initial list of seeds important to the proposed research (OII)
2. A final list of seeds included in the collection – curated & mined (All)
3. WWI collection of 100+ mil resources harvested over a period of 12 years, including a separate store of all links extracted from these resources (IA)
4. MODs records for each seed in the collection (IA)
5. Collection indexed by NutchWAX for full text search (IA)
6. Collection indexed for browsing via Wayback (IA)

7. Collection indexed via a search engine yet to be named using the SE Framework for full text search (Hanzo)
8. Web services API to the collection (IA and Hanzo)
9. OpenSearch API to the collection (IA)
10. OAI-PMH interface to the collection (IA)
11. Search Engine framework v1.0 that can be integrated with open source WARC tool set (Hanzo)
12. Web UI/dashboard for humanities researchers to manage the workflow of collection assembly and analysis and replay of resources Proof of Concept (IA & Hanzo)
13. Documentation of researcher interfaces and researcher tools/workflows (IA and Hanzo)
14. Final Report (OII, IA, Hanzo)

## 1.6 Project Outcomes

We intend this project to contribute the following:

- Prototyping environments that enable students, faculty, and researchers to experiment with tools, workflows and api's to assemble and analyse web collections over time.
- Use cases to be used by teaching and learning communities that illustrate the emerging challenges and opportunities presented by the study of digital humanities va the Web.
- New insights into the evolution of digital humanities research

## 1.7 Stakeholder Analysis

Stakeholder	Interest / stake	Importance
JISC	Funding body	High
NEH	Funding body	High
Oxford Internet Institute	Project partner	Medium
Hanzo Ltd	Project partner	Medium
Internet Archive	Project partner	Medium
e-Humanities projects	Information provider	Medium
Humanities researchers	User group	Medium

## 1.8 Risk Analysis

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
Staffing				
1. Loss of staff	3	3	9	Have engaged a diverse group of participants, multiple from each partner. If one departs, others can assume his/her responsibilities.
Organisational				
1. Cost over-run	2	3	6	Regular review of budget spending with finance officer
2. Schedule over-run	2	3	6	Monthly progress meetings and contingency plans built

				into timetable of deliverables
Technical				
1. Software failures	2	4	8	The promise of automation may fall short of the mark. We will monitor this carefully on a monthly basis. If this appears to be the case, we will ensure that each step in the workflow is distinct and can be deployed as individual tools even if not as an integrated end to end workflow.
2. Insufficient access to web resources	3	4	12	Aim to assemble the initial collection by the end of June, enabling us to spend 9 months on the integration of components and work flows.
External suppliers				n/a
Legal	2	4	8	Services offered to researchers will need to be gated using a login if the full contents of the collection are to be made available for research. Alternately, robots.txt and manual exclusion files will be used to filter content from "open" or "public" researcher interfaces.  Individual resources and the collection in its entirety will not be available for replication by researchers.

## 1.9 Standards

Name of standard or specification	Version	Notes
MODs	V3.3	<a href="http://www.loc.gov/standards/mods/">http://www.loc.gov/standards/mods/</a>
OAI-PMH	V2	<a href="http://www.openarchives.org/pmh/">http://www.openarchives.org/pmh/</a>
XML API		<a href="http://archive-access.sourceforge.net/projects/wayback/administrator_manual.html">http://archive-access.sourceforge.net/projects/wayback/administrator_manual.html</a>
WARC/ARC	Pending ISO standard	<a href="http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717">http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717</a>
OpenSearch Protocol	V1.1	<a href="http://www.opensearch.org/Home">http://www.opensearch.org/Home</a>

## **1.10 Technical Development**

There are two critical approaches to technical development that will help ensure the success of the project:

1. Sound code lifecycle management with an emphasis on use cases and usage requirements

Although the teams will not be adopting Extreme Programming practices for this project, some of the fundamental tenets will be applied, including coding driven by users, use cases, and usage requirements. Technology will not be developed and deployed for technology sake but in support of a community of users with specific needs and objectives. The seasoned engineers assigned to this project will also ensure that the resulting services and framework are flexible enough to be adapted as needs change and evolve.

Hanzo and IA are prepared to independently implement some portions of the workflow in parallel and to test and validate the contributions of the other to the project.

2. A commitment to supporting standard interfaces and open source tools to enable researcher workflows.

There is a long history of prior application of open source tools and coding techniques by these teams.

## **1.11 Intellectual Property Rights**

Any code developed by IA to extend tools or components assembled to support researcher workflows will be licensed under LGPL or Apache licenses for distribution as open source packages.

Any researcher interfaces hosted by IA will be offered at the discretion of IA under terms defined by IA.

Any code developed by Hanzo to create or extend tools or components that might be assembled to support researcher workflows will be licensed under an Apache license and will be available for distribution as open source packages.

Any researcher interfaces hosted by Hanzo will be offered at the discretion of Hanzo under terms defined by Hanzo.

Contents of the WWI collection are subject to legal restrictions based on the region where the materials are hosted. Any publicly available services will respect robots.txt and manual exclusions requested by content owners. Any password protected access for e-humanities researchers will be granted full rights to view and analyze resources individually and as a collection but replication of the materials is not permitted except for those materials harvested from the live Web during the course of this project.

## 2 Project Resources

### 2.1 Project Partners

#### **Oxford Internet Institute (OII)**

OII is responsible for defining the initial scope of the WWOne collection, communicating researcher requirements for the framework and workflow, facilitating data analysis, and reporting on the results at the conclusion of the project.

#### **Internet Archive (IA)**

IA is responsible for assembling and indexing the collection, extracting meta data, hosting and providing access to the collection, preserving the data, implementing, customizing, and supporting interfaces to the collection for researchers, students, and the general public.

#### **Hanzo Archives (Hanzo)**

Hanzo is responsible for hosting and indexing a replica of the collection, prototyping, and customizing interfaces to the collection in support of defined workflows that meet the needs of researchers and for the development and delivery of the search engine framework.

### 2.2 Project Management

The overall project will be led by Eric Meyer (OII), Project Director. He will also provide all documentation to JISC. Kris Carpenter Negulescu will serve as Project Manager and will lead all NEH related reporting and communication requirements.

Progress will be reviewed monthly – each partner will be responsible for producing a report of their activities during the prior month and for describing the proposed activities for the next month. Any issues raised regarding the completion of milestones will be discussed during a monthly teleconference call. Recommendations may come from any participant in the project but ultimately, decisions will be made &/or validated by a committee of three members comprised of one member from each of the participating institutions (Kris Carpenter Negulescu(IA), Mark Middleton(Hanzo), and Eric Meyer (OII)).

Regular e-mail will also be used to communicate among the project's team members using the team e-mail list, [WWWoH@jiscmail.ac.uk](mailto:WWWoH@jiscmail.ac.uk).

#### Schedule of Skype conference calls with Breeze/WebEx collaboration

All meetings to be held on Thursdays at 4:00 pm (UK) / 8:00 am (California) on the 2<sup>nd</sup> Thursday of each month through March 2009, with additional meetings at two-week intervals during the start up phase.

2008: May 22

June 5 & June 19

July 10

August 14

Sept 11

Oct 9

Nov 13

Dec 11

2009: Jan 8

Feb 12

Mar 12

Initial milestones were defined and ratified at the project team meeting in San Francisco on May 1<sup>st</sup> & 2<sup>nd</sup> and are reflected in the workpackage plan on page 18.

Project Team members include:

### **Oxford Internet Institute**

#### Main Investigators

- Professor William H. Dutton, Director of the OII, Professor of Internet Studies and Fellow of Balliol College, Oxford.
- Dr. Eric T. Meyer, Research Fellow, OII
- Dr. Ralph Schroeder, James Martin Research Fellow, OII

#### Other Personnel

- Christine Madsen, Research Assistant, OII

#### Advisors

- Dr. Robert Ackland, Research Associate at OII and Fellow in the Research School of Social Sciences at the Australian National University

**Prof. William H. DUTTON** is Director of the Oxford Internet Institute, Professor of Internet Studies, University of Oxford, and Fellow of Balliol College, Oxford. He will serve in an advisory role throughout the early stages of the project, and will contribute heavily to the final report for the project.

**Dr. Eric T. MEYER** is Research Fellow at the Oxford Internet Institute, Oxford. He will serve as Project Director, responsible for ensuring the overall progress of the project. Meyer will also be the primary person responsible for JISC reporting requirements.

**Dr. Ralph SCHROEDER** is James Martin Research Fellow at the Oxford Internet Institute, Oxford. He will work with Madsen to oversee the development of the collection and help to establish contact with appropriate domain experts. He will also be involved in the analytical portions of the grant as the collection is analysed from an e-Research perspective.

**Christine MADSEN** is Research Assistant at the Oxford Internet Institute, Oxford. Madsen is a librarian who is expert in the development of efficient, replicable methods for the creation of comprehensive, subject-based digital resources. She will be responsible for working with the assembly of the collection, and for interfacing with humanities experts.

### **Hanzo Archives**

**Mark MIDDLETON** is CEO of Hanzo. He is a software & technology entrepreneur with an international blue-chip company background, with over nine years of technology and business leadership experience and 18 years experience in the software and computer industry. He will oversee the main Hanzo work on developing tools for working with web collections.

**Younes HAFRI** is an engineer at Hanzo who will be working extensively on the development of the search engine framework for web collections.

### **Internet Archive**

#### Main Investigators

- Vinay Goel, Crawl & Web Analytics Engineer
- Brad Tofel, Senior Software Engineer, Wayback
- Aaron Binns, Senior Software Engineer, Search

Other Personnel

- Kris Carpenter Negulescu, Director Web Group
- Gordon Mohr, Chief Technologist Web Group
- John Lee, Senior Operations Engineer

**Kris Carpenter NEGULESCU** is Director of the Web Group. She is project manager for this project.

**Gordon MOHR** is the Chief Technologist & Development Manager, Web Group. Gordon will oversee the architecture and framework approach developed and deployed by IA for this project.

**John LEE** is the Web Group Operations Manager. He will be responsible for capacity planning, storage and data management for the WWWoH collection. He will also coordinate ops support for deployed services and researcher interfaces.

**Brad TOFEL** is the lead developer of the open source Wayback Machine. He will be extending and configuring Wayback to support the researcher workflow requirements.

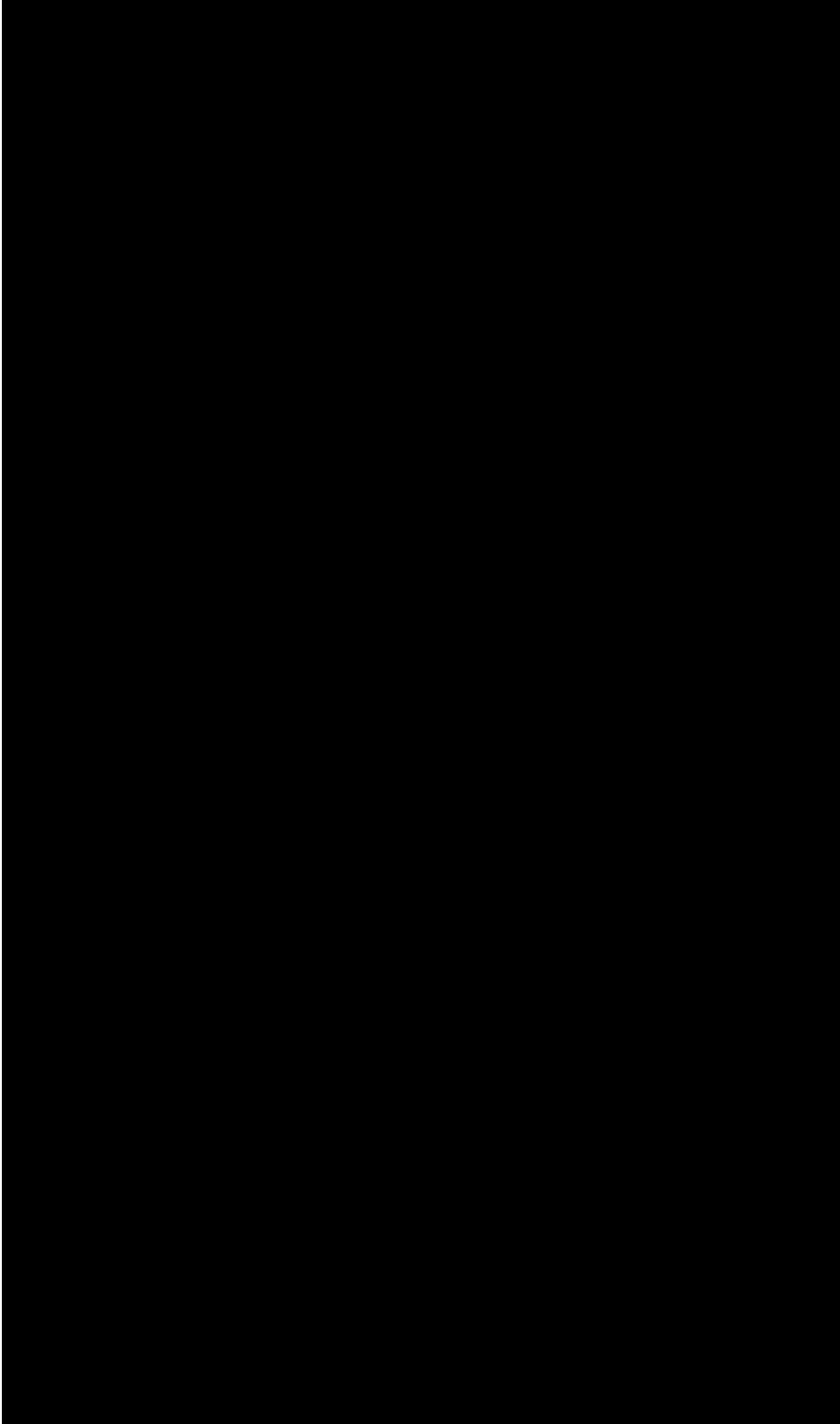
**Aaron BINNS** is a senior developer who leads NutchWAX development.

**Vinay GOEL** is a software engineer responsible for running focused web crawls for Internet Archive partners, for the deployment of web archive access tools, and for automated data extraction and analysis of web collections.

No training is required to support this project.

### **2.3 Budget**

Note: There are no changes from the originally proposed budget.



### 3 Detailed Project Planning

#### 3.1 Workpackages / GANTT chart

WORKPACKAGES	Month	1	2	3	4	5	6	7	8	9	10	11	12
		A p r	M a y	J u n	J u l	A u g	S e p	O c t	N o v	D e c	J a n	F e b	M a r
1: Assembly													
2: Interfaces													
3: Analysis Toolkits													
4: End User Prototypes													
5: Dissemination													

Project start date: April 1, 2008  
 Project completion date: March 31, 2008  
 Duration: 12 months

### 3.2 Detailed Schedule

(Key deliverables in <b>bold</b> )	Earliest start date	Latest completion date	Outputs	Milestone	Responsibility
<b>YEAR 1</b>					
<b>WORKPACKAGE 1: ASSEMBLY</b>					
<i>Objective:</i> Assembly of the collection/s (from live crawl and data extraction using seeds derived from domain experts and hyperlink analysis of the collection domain)					
1. Identification of seed websites on topically focused research collection/s	1 May 2008	1 June 2008	Iterative list developed and provided by OII to IA over the course of 4 weeks, with feedback from IA on the complete size of the collection based on the seeds and additional required material. Key seeds will be identified.	Seed list completed	OII
2. Extraction of seeds from the Internet Archive	25 May 2008	30 June 2008	Collection as a set of extracted ARC files	ARC files done	IA
3. If required: Live crawls	1 June 2008	1 Aug 2008	If required: live crawling of web for inclusion of key seeds not available in Internet Archive	ARC files done for live crawls	IA
4. Link extraction and analysis	15 June 2008	15 July 2008	Complete selection of candidate seeds per year going back through the IA collection.	Candidate seed list	IA
5. Domain expert validation of extracted links	20 June 2008	25 July 2008	Have domain experts look at a small list of candidates and validate the inclusion of extracted links	Augmented seed list	OII
6. Extraction of additional seeds from augmented list	25 July 2008	4 Aug 2008	Final extraction of additional seeds	Final list of 250 M	IA

(Key deliverables in <b>bold</b> )	Earliest start date	Latest completion date	Outputs	Milestone	Responsibility
				documents	
7. Metadata extraction from seeds	1 Aug 2008	8 Aug 2008	Extract metadata from seeds to populate the MODs records	Populated MODs records	IA
8. Full text indexing	9 Aug 2008	31 August 2008	Text searchable collection across the years 1996-2007 for all seeds	Collection	IA
9. Link extraction from augmented seeds	1 Aug 2008	8 Aug 2008	Complete list for analysts of link data from the collection	Link file for analysis	IA
<b>WORKPACKAGE 2: INTERFACES</b>					
<i>Objective:</i> Interfaces to the Collections (custom Wayback Machine interfaces, full text search via NutchWAX, and deployed including support for OAI-PMH OpenSearch protocol and XML API to collection)					
<b>10. Browser interface to the collection for end users, specifically e-Humanities scholars and curators</b>	September 2008	March 2009	Access interface to the collection for e-Humanities scholars (only includes post collection creation tasks such as search, browse of the collection)	Interface	IA
11. XML API to the collection	September 2008	March 2009	Experimental interface for integrating workflow and analysis tools	Interface	IA
12. OAI-PMH interface to the collection	September 2008	March 2009	Enables scholars to harvest metadata from the collection	Interface	IA
13. OpenSearch protocol support	September 2008	March 2009	Enables federated search with other relevant collections	Interface	IA
14. Web Services API to the collection	December 2008	March 2009	Experimental interface for integrating workflow and analysis tools	Interface	IA
<b>WORKPACKAGE 3: ANALYSIS TOOLKITS</b>					

(Key deliverables in <b>bold</b> )	Earliest start date	Latest completion date	Outputs	Milestone	Responsibility
<u>Objective</u> : Enable research using third party tools.					
15. Identify & evaluate existing open source tools for workflow and analysis	June 2008	August 2008	Experts will evaluate existing open source analysis tools as candidates for inclusion in end user prototypes. Tools will be considered for all phases of the workflows to be prototyped, including workflow creation and management software commonly in use by the e-humanities and e-research communities.	List of existing tools to integrate with end user prototypes, classified by use case and stage of workflow.	Oll, Hanzo
16. Search Engine Framework Beta	May 2008	Oct 2008	A beta version of the search engine framework will be available for download and will include support for index of an arbitrary collection of WARC files.	Beta software package available for download and installation locally.	Hanzo
17. Search Tools v1	May 2008	November 2008	The first version of the search engine framework will be available for download and will include support for index of an arbitrary collection of WARC files as well as integration with basic analysis tools. This code will be available for installation anywhere and will complement the hosted services deployed by IA.	Software package available for download and installation locally.	Hanzo
18. Analysis Toolkit (locally deployed) API	November	January 2009	The Search Tools will be extended to	API to the	Hanzo

(Key deliverables in <b>bold</b> )	Earliest start date	Latest completion date	Outputs	Milestone	Responsibility
	2008		include an API that results in a prototyping platform upon which we can experiment with researcher interfaces and workflows.	Search Tools v1	
<b>WORKPACKAGE 4: END-USER PROTOTYPES</b>					
<u>Objective:</u> Build end-user prototypes that support use of the tools and collections appropriate to the target audiences.					
19. Define research questions	June 2008	Sept 2008	Definition of specific research questions to use to illustrate how the platform, tools, and workflows might be applied to the WWOne collection as well as how the process might be replicated on an entirely different collection or subset of the Web and digitized humanities materials. At least one question per profile, i.e. per end user, will be defined and documented.	List of research questions per actor and/or type of research	OII
20. Develop experimental prototype for the e-Humanities scholar/curator	August 2008	March 2009	Creation, implementation, and evolution of browser-based tools and interfaces to the WWOne collection that facilitate e-humanities curation and research.	Hosted service for e-Humanities curators and researchers	IA, OII
21. Develop experimental prototype for the e-researcher	August 2008	March 2009	Creation, implementation, and evolution of programmatic and browser-based interfaces to the	Hosted service for e-	IA, OII

(Key deliverables in <b>bold</b> )	Earliest start date	Latest completion date	Outputs	Milestone	Responsibility
			WWOne collection, integrated with existing tools that facilitate e-research.	researchers	
<b>22. Develop experimental prototype for locally hosted solution</b>	December 2008	March 2009	Creation, implementation, and evolution of programmatic interfaces to an arbitrary collection of WARC files, integrated with existing tools that facilitate e-research.	Downloadable software toolkit	Hanzo, OII
<b>WORKPACKAGE 5: DISSEMINATION</b>					
<u>Objective:</u> Disseminate materials from the project					
<b>23. Interim Reports to JISC and NEH</b>	August 2008	September 2008	Creation of interim report to JISC	Report	OII, IA
<b>24. Project website:</b> <a href="http://www.oii.ox.ac.uk/research/project.cfm?id=48">http://www.oii.ox.ac.uk/research/project.cfm?id=48</a>	May 2008	March 2009	Project website will include material on the project, publications/presentations relating to the project, and other materials of interest to the e-Humanities and e-Research communities.	Website	OII
<b>25. Copy of archive delivered to a European archive</b>	February 2009	March 2009	Replication of collection to a location in Europe	Replicated archive	IA, Hanzo, OII
<b>26. Collection website housed at IA under:</b> <a href="http://ehumanities.archive.org/collections/">http://ehumanities.archive.org/collections/</a>	August 2008	March 2009	Collection website that is searchable and usable as per the technical details listed in other workpackages.	Website	IA
<b>27. Presentations and publications</b>	June 2008	After end of project	Sample plans for presentations and publications: iPres, London, September 2008 Web Histories Conference, Denmark, October 2008 CNI, Washington, DC, December 2008	Presentations and publications	OII, IA, Hanzo

(Key deliverables in <b>bold</b> )	Earliest start date	Latest completion date	<b>Outputs</b>	<b>Milestone</b>	<b>Responsibility</b>
			Digital Curation Conference, Edinburgh, December 2008 iConf, Chapel Hill, NC, 2009		
28. Final Reports to JISC and NEH	February 2009	March 2009	Final summative report of the project, including report of deliverables, budget reports, and summary of findings. We will also include information on lessons learned from this collaboration and recommendations for future programs.	Report	OII, IA, Hanzo
<b>29. Demo event in the UK</b>	March 19, 2009	March 19, 2009	Demo event at Oxford, including all project partners, funders, digital humanities curators, domain experts, and other interested parties.	Workshop	OII, IA, Hanzo

### 3.3 Evaluation & Quality Plan

Key deliverables (indicated in **bold** in the table above) will all be evaluated as shown below. These include workpackage items #10, 20, 22, 26, 27, 29.

Timing	Key deliverable / Factor to Evaluate	Questions to Address	Method(s)	Measure of Success
Sept 08 and repeated in Feb 09	10: Browser interface to the collection for end users, specifically e-Humanities scholars and curators	1. Ease of Use of the researcher dashboard/workflow tools for analysis 2. Is the framework sufficiently documented and easy to use for "non technical researchers	User interviews and review of framework and interfaces	Researcher is able to access the collection and find items of scholarly interest.
Ongoing from Nov 08	20 & 22: Develop experimental prototype for the e-Humanities scholar/curator and Develop experimental prototype for locally hosted solution	1. Simplicity and ease of use. 2. Usefulness of creating focused collection 3. Progress towards replication of process	Demonstrate multiple implementations  Online focus group / survey of domain users	Positive responses on simplicity, usefulness, and general direction of progress
Sept 08	26: Collection website housed at IA under: <a href="http://ehumanities.archive.org/collections/">http://ehumanities.archive.org/collections/</a>	1. Seed scope, breadth and depth	Online report created using reporting tools	Metrics demonstrating comprehensiveness and consistency
Ongoing	27: Presentations and publications	Outreach, education about the project, dissemination of results	Peer review	Engagement with relevant audiences
March 19, 2009	29: Demo event in the UK	Demonstrate an end to end prototype of the e-Humanities researcher platform	Public event with discussion	Usefulness to target audience and successful demonstration

### **3.4 Dissemination Plan**

See Work package 5 in the detailed schedule that starts on page 19 of this document. In addition to the outputs listed there, we have also established a low-volume listserv which will be used to announce major outputs of the project and to invite attendees to the final event on March 19, 2008. The e-mail address of this list is **WWWoH-Announce@jiscmail.co.uk** and Madsen will keep this list updated with new contacts acquired over the course of the project.

### **3.5 Exit and Sustainability Plans**

<b>Project Outputs</b>	<b>Action for Take-up &amp; Embedding</b>	<b>Action for Exit</b>
WWI collection and interfaces	Get the word out through conferences, online pubs, blogs, etc regarding the availability of the collection and analysis tools to encourage other researchers to examine the collection and methodologies used to assemble it.	Project Whitepaper Data preservation of collection materials Tools maintenance (IA/Hanzo) Beta Researcher API program (IA) Ongoing public access to the WWI collection (IA) Replication of collection in Europe
Analysis Toolkit	Same as above	Tools maintenance (IA/Hanzo)
Ongoing e-humanties research	Same as above	OII, Hanzo & IA plan to work together even after the conclusion of this project to continue improving e-research tools, frameworks, and methodologies



## 4 JISC Website Template for Projects

To be completed by the Projects	
Project Title	World Wide Web of Humanities (WWWoH)
Project website address	General: <a href="http://www.oii.ox.ac.uk/research/project.cfm?id=48">http://www.oii.ox.ac.uk/research/project.cfm?id=48</a> Collection: <a href="http://ehumanities.archive.org/collections/wwone/">http://ehumanities.archive.org/collections/wwone/</a>
Start date	1 April 2008
End date	31 March 2009
Overview	This project aims to establish a framework for e-Humanities (also called Digital Humanities) research using available open source tools and technologies and archived web content to create novel research interfaces to the first of many, scholarly, e-Humanities web collections.
Aims and objectives	<p>The project aims to assemble a World War One collection using current and historic web data and to create and/or integrate a suite of open source tools and standard interfaces that enable researchers to assemble, index and analyze subsets of web resources derived from archives and the live web as research collections.</p> <p>The project also intends to demonstrate a prototype of the e-Humanities researcher/curator framework by applying the tools, methodologies, and techniques to a study of the World War One e-Humanities resources using web data harvested from 1996 to the present and by integrating existing WWI digitized humanities collections.</p> <p>The project will also produce a new set of open source tools called the search Engine framework for distribution under an Apache license.</p>
Project methodology	<p>OII will be leading the definition and scope of the e-Humanities research.</p> <p>Hanzo will be developing and deploying the search engine framework, prototyping interfaces to the collection, and hosting a replica of the collection and integrated researcher workflow</p>

	<p>prototypes.</p> <p>IA will be assembling and analyzing the collection, indexing it for full text search and for browse via the Wayback machine, and will be prototyping researcher/curator workflows. IA will also expose XML and web services API's to the collection along with support for OAI-PMH and OpenSearch protocols.</p>
Anticipated outputs and outcomes	<p>A collection of 100-250 mil URIs that characterize the growth in discussion, dissemination, and evaluation of WWI on the web.</p> <p>A platform for assembling focussed research collections from general archives and the live web, extracting links text and metadata, indexing and analyzing these resources, as well as unifying digitized collections.</p> <p>A Final Report on the research and e-Humanities research prototyping environment</p>
Technology / Standards used (if applicable)	<p>WARC          MODs          XML API, Web services API          OAI-PMH          OpenSearch</p>
Project Manager & Team	<p>Kris Carpenter Negulescu          Internet Archive, Web group          kcarpenter@archive.org          415.561-6799, ext 1</p>
Project Team	<p>Oxford Internet Institute          Hanzo Archives          Internet Archive</p>
Lead Institution	<p><b>Oxford Internet Institute</b>, <a href="http://www.oii.ox.ac.uk">www.oii.ox.ac.uk</a></p>
Project partners	<p><b>Hanzo Archives</b>, <a href="http://www.hanzoarchives.com">www.hanzoarchives.com</a>  <b>Internet Archive</b>, <a href="http://www.archive.org">www.archive.org</a></p>
<b>To be completed by Programme Managers</b>	
JISC programme	
JISC theme(s)	
JISC Programme Manager	
JISC Programme Director	
Related projects	