

Final Report

Project Name	World Wide Web of Humanities
Project Websites	http://www.oii.ox.ac.uk/research/project.cfm?id=48 http://wwwoh-access.archive.org/wwwoh/ http://wwwoh.hanzoarchives.com/
Report compiled by	Eric T. Meyer, Kris Carpenter, Mark Middleton
Contact Person:	Eric T. Meyer (eric.meyer@oii.ox.ac.uk)
Date	28 June 2009

Table of Contents

1 Executive Summary 4
2 Background 5
3 Aims & Objectives 6
4 Implementation 7
4.1 Methodology 8
4.2 Assembly of the Collection 8
4.3 Automated Metadata Extraction 10
4.4 Link Extraction & Analysis 10
4.5 Organization of Materials 11
4.6 Access to Materials 11
4.7 Storage Maintenance and Protection of Data 13
5 Outputs and Results 15
5.1 Outcomes 15
6 Implications & Recommendations 16
6.1 Building Collections of Web Sites 16
6.2 Analysis of Content 16
6.3 User Scenarios / Research Questions 16
7 Conclusions 19
8 References 20
9 Appendix: Implementation Schedule 21

Acknowledgements

The WWWoH project was a collaboration among the Internet Archive, The Oxford Internet Institute, and Hanzo Archives.

Oxford Internet Institute (OII)

OII was responsible for defining the initial scope of the WWI and WWII collections, communicating researcher requirements for the framework and workflows, facilitating data analysis, and reporting on the results at the conclusion of the project. The Oxford Internet Institute and the Australian National University contributed expertise in web research and led the curation of collections for application and use by humanities scholars.

Internet Archive (IA)

IA was responsible for assembling and indexing the collection, extracting metadata, 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 was 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 an independent access application incorporating the search engine framework and prototype analytical applications.

Both the Internet Archive and Hanzo applied extensive experience in web archiving and in the creation of web collections, including the largest of all web collections, the Internet Archive's Web collection accessible via the WayBack Machine (www.archive.org), to extract, assemble, analyze and make accessible archival data germane to this project.

The following individuals were instrumental in this project:

- Kris Carpenter, Internet Archive (IA, Principal Investigator US & Project Manager)
- Dr. Eric T. Meyer, Oxford Internet Institute (OII, Principal Investigator UK)

Hanzo Archives

- Mark Middleton
- Younes Hafri

Internet Archive

- Vinay Goel
- Brad Tofel
- Aaron Binns
- Molly Bragg
- Gordon Mohr

Oxford Internet Institute, University of Oxford

- Dr. Robert Ackland, Australian National University (ANU) & OII Research Associate
- Professor William H. Dutton
- Christine Madsen
- Dr. Ralph Schroeder

1 Executive Summary

The World Wide Web is enormous and is in constant flux, with 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 important aspects of modern society. But, there are few tools available to facilitate research using archived web materials. Humanities researchers are expected to individually assemble research data and e-Research tools needed for analysis. This can be cost-prohibitive in terms of resources and time, and more importantly can dissuade researchers from undertaking relevant research simply due to the technical barriers involved.

This one-year project addressed this gap by establishing a framework for e-Humanities (also called digital humanities) research using available open source tools and technologies and archived web content. The project created novel research interfaces to the first of many scholarly e-Humanities web collections. Within the context of this project, the term 'web collections' was used to describe collections of archived websites and 'born digital' content.

During the twelve months of the World Wide Web of Humanities project, the project team assembled two collections of web sites focused on World Wars I & II, drawn primarily from the Internet Archive and enhanced with new crawls to complete the collections. These sample collections were compiled to help illustrate researcher needs and requirements, and in anticipation of further development of tools for working with the very large volumes of data housed by digital archives such as the Internet Archive. The team identified a range of research questions that such archival web collections could be expected to be used to answer, from the point of view of humanities scholars, of internet researchers, and of programmers who work with web archives. These research questions were then used to inform the design and implementation of search tools and collection interfaces.

The key goals of the project have been successfully met, and we have made a number of relevant observations based on this experience that will be of great value to similar efforts in the future. Among the key lessons learned from this research project are:

1) The importance of including participants with a strong personal research interest in developing a given collection will be key in future efforts, as this involvement enhances the likelihood of having user-friendly collections and interfaces. Collections that are built with strong involvement of a champion or community with domain expertise can, for instance, speed the curation stage of similar projects.
2) Available tools to facilitate e-humanities research are still either too hard to use, i.e. they require collaboration with an engineer, or they are missing entirely from the workflow limiting options for the creation of a user friendly platform easy enough for non-technical humanities researchers to adopt and use on a regular basis. Based on the findings and experience gained in this project, it is now clearer which direction these efforts will need to take.

2 Background

There is 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. 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.

The Internet Archive (IA), which maintains the world's largest collection of web pages, offers the possibility of access to a large collection of materials that are of interest to researchers. The IA Wayback Machine contains 1.5 petabytes of data compressed, representing 150 billion snapshots of websites dating from 1996. However, the default public interface to the Wayback Machine is primarily designed to allow access to specific websites and pages by entering individual URIs. Researchers interested in studying collections of material, however, need to be able to access materials in greater volume and to be able to manipulate and analyse those materials in a variety of ways.

There are at least three general types of tools critical to performing research using archival material from the Internet Archive: tools for data collection, curation and search; for text analysis; and for hyperlink analysis. This project mainly focused on creating and assembling a suite of open source tools in the first of these areas, and employed third party tools for the analysis that these curation and search tools enabled. The aim of this project was 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 focused on e-Humanities, but a major value of this effort was not just to create a specific 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 created novel research interfaces, both browser-based and programmatic, to the first many scholarly web collections.

As a way to build and test the tools constructed, the team assembled two research collections around the topics of **World War I** and **World War II**. These particular cases were selected for a variety of reasons that make them a compelling test of the methods developed. First, the materials available represented a well-rounded set of humanities materials that allowed us to test the tools against a variety of types of documents and resources. WWI and WWII 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 allowed for 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 topics of WWI and WWII were relevant for this transatlantic cooperation because materials on this time period exist in the UK and in the US, as well as in other countries around the world involved in these conflicts. We also believed that collections focused on the world wars have relevance not only for scholars, but also for teachers of secondary and HE/FE courses and for amateurs interested in WWI and WWII. Finally, many of the materials from this time period are out of copyright, which removed much of the legal complexity that could have potentially arisen had we compiled a more contemporary collection.

When we began the project, we believed the World War I collection might include up to 250 million Uniform Resource Identifiers (URIs) 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 would have been sufficient for our research purposes. However, the size of the initial WWI collection was significantly smaller than anticipated (i.e. barely 1 million records) so the WWII collection was added to increase the overall scope and scale of the materials. This was done curatorially, by 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. Rules for inclusion of materials were strict, biasing relevance over breadth and depth of coverage that resulted in some gaps in the source materials but avoided inclusion of non-relevant materials from other wars, for example. The collections were 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 and World War II collections and studies.

This project culminated in a public workshop held in Oxford, UK on Thursday, March 19, 2009. The event allowed us to demonstrate the tools developed, present the test collections, and discuss the lessons learned from the project with over 100 attendees, including domain experts (humanities scholars), digital humanities researchers and e-Researchers, curators of digital collections, representatives of funding bodies, and other interested parties. Given the success of this project and the positive response to the event, we envision additional collaboration in the future between the partners and among others to collaborate on related projects in the area of e-Research and cyberinfrastructure.

3 Aims & Objectives

The primary aims of this project were to build a set of tools that supported data assembly 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 identified several target audiences for the tools and the collected materials. First, humanities scholars, particularly those interested in the selected topics of World War I and World War II, were targeted to use the collections and tools this project generated. This is a key lesson we have taken away from this project: the importance of including participants with a strong personal research interest in developing a given collection, rather than selecting collections based on technical considerations. This lesson is already informing future plans for collaboration, as we are planning collections that are built with strong involvement of a champion with domain expertise.

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 are now able to use these data and tools to answer a plethora of questions about the digital humanities domain. Nevertheless, an important finding of the research is that there are still additional elements of a user-friendly workflow that need to be developed before this becomes a seamless process. For instance, the selection process is still quite heavily manual and will need to be made more lightweight and user-friendly in the future.

Our primary objectives for the project were 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 assembly of data resources (web collections) that are (a) focused on particular areas of research and that (b) facilitate inter-temporal and historical internet research. We provided two distinct interfaces into the data resource: two web interfaces that allow researchers to 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. We have been successful in these objectives, and furthermore intend to continue to pursue this path through additional grant applications and other mechanisms for collaboration moving forward.

4 Implementation

All partners were intimately involved in the successful outcome of this project.

Following an intensive requirements gathering and documentation phase in month one of this project, the plan of work proceeded in roughly three stages. First, IA and OII collaborated on the curation and assembly of the e-Humanities collections, encompassing resources drawn from archived data sets and focused harvests performed as part of this research. During this stage IA replicated the collection materials to Hanzo in support of collection indexing and analysis. In parallel, Hanzo designed and developed a search framework written in C, as well as the adaptation of Ferret to work within this resulting framework and ensure continuity with available libraries and tools for storage, management, migration, and manipulation of WARC/ARC file formats.

Second, IA began to process and analyse the collection with guidance and oversight from OII. This included automated metadata extraction, unification of automated and researcher defined metadata, validation of the wide CDX file format, as well as link graphing and analysis. In parallel, IA and Hanzo indexed the collections for full text search and browser based navigation modes.

Finally IA/HA/OII collaborated on the development and deployment of researcher interfaces to the collections. The browser based user interface designed to support definition of and access to research collections was very basic and should be considered a proof of concept demonstrating the underlying capabilities of the framework vs. a perfectly designed user interface for e-Research. Emphasis instead was placed on the development and definition of programmatic interfaces (OpenSearch, and Web services) that can be leveraged for analysis and retrieval of materials as well as federation with other collections and repositories. Distribution of preservation copies of collection materials to IA global preservation partners was part of the deliverables of the final stage of the project.

At the end of the initial phase of this project, we successfully identified 4884 seed web sites, as detailed in Table 1. This resulted in the subsequent extraction of ARC files from the Internet Archive.

Table 1: Initial collection sizes

	World War I	World War II	Common to both WWI and WWII	Not seeded	Total
Number of seeds	2,263	2,592	29	-	4,884
Number of unique hosts	906	1,475	149	7,137	9,667
Number of unique URLs	2,312,937	3,160,408	624,610	477,554	6,575,509
Number of captures	8,424,630	13,320,354	858,736	2,428,434	25,032,154
Number of links	143,017,686	252,153,151	49,262,548	20,709,600	465,142,985

Total size of compressed ARC data = 240GB
Source: Internet Archive captures between May 1996 and Aug 2008.

The initial goal in building these collections was to select high-quality web resources on the topic of World War I. Early on in the seed selection process though, we realized that our initial selection policy would not result in the targeted 100 million or so pages. In order to more closely approximate the target size (to test the viability of the search tools) the scope of the collection was expanded to include World War II, and the selection criteria were changed to favour broader inclusiveness over selectiveness. This is a valuable finding from this research: although the web is assumed to be almost endless, once specific topics are targeted, the actual number of unique websites can quite quickly shrink to a more usable set that enables meaningful research.

Once scoping was complete, the work of migrating the data from ARC to WARC format and indexing of the data began in earnest. By October of 2008 the materials had been assembled and indexed and work on the workflows and interfaces commenced.

As the development of the search tools and user interface progressed, researchers at the OII developed a set of user scenarios and potential research questions that guided the needs assessment of the end products. These use cases and questions covered a broad spectrum of disciplines, from humanities researchers to social scientists to computer scientists. This process also helped the project team to refine the access and outreach strategy used to deliver these tools to a wide audience.

4.1 Methodology

The Internet Archive and Hanzo Archives assembled existing open source technologies, that have been optimized for operation at a scale of 10's of millions to 100's of millions of web objects, into a framework that enabled 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 novel means. All services and methods were enabled using open source tools. In this way, the team was able to undertake an initial assessment of e-Humanities research and provide a foundation for future research initiatives.

IA/HA/OII methodologies did not diverge from established standards and best practices with the exception of the novel means with which we combined data formats, open source tools, best practices, and methods.

4.2 Assembly of the Collection

The collections were assembled by the OII in conjunction with IA through an iterative process. The details of this process are documented in the table of work packages (see the Appendix). In short, the process began with a set of key resources ("seed sites") identified by OII in consultation with domain experts in the humanities, particularly those with expertise in World War I and II collections where feasible. As the goal was to gather a collection of archived web sites, links to sites that no longer exist were also recorded. These dead links, which appear to be useless on the live web, represent one advantage to this collection method: if the Internet Archive includes archived versions of these pages, they can still be included in the collection. This represents an improvement over the native interface to the Internet Archive's Wayback Machine, which requires users to type in a URL and then select from various snapshots of those pages collected over time.

The next step was to generalize the URLs in order to maximize the number of pages in the collection. For each URL copied, references to specific pages were removed and the URL truncated to the root site or most logical directory. For example, it was logical to conclude that the entirety of http://www.greatwar.co.uk was on topic, so all references to specific pages, such as

http://www.greatwar.co.uk/westfront/Somme/index.htm were removed and replaced with http://www.greatwar.co.uk. Duplicate sites were removed automatically. Many collections of materials—in particular those from universities, archives, and libraries—were not resident on unique domains. In these cases, the URL could only be truncated as far back as the directory containing the relevant materials. For example: http://memory.loc.gov/ammem/collections/maps/wwii/index.html to http://memory.loc.gov/ammem/collections/maps/wwii/.

Another important lesson from this research has to do with the ways in which material is organized on web servers: illogical directory structures were often encountered and required considerable manual work for inclusion in the collection. EyeWitnesstoHistory.com contains first person accounts of historical events and contains almost fifty pages dedicated to the First and Second World Wars. Each page file sits in the root directory, though, and so needed to be provided individually. The entire site (http://www.eyewitnesstohistory.com/) could not be included because only a fraction of it falls within the scope of the collection, therefore individual pages (http://www.eyewitnesstohistory.com/blitzkrieg.htm, ../dday.html, etc.) had to be recorded.

Although this process may seem to result in an almost infinite number of sites, it became clear that after gathering several hundred seeds, most of the resulting sites identified were redundant. At that point, more precise search terms were selected and the process re-initiated. Narrower topic searches were commonly either biographical (Hitler, Churchill, etc.), event-based (Battle of Midway, Guadalcanal campaign, surrender of Japan), or based around on subjects that while technically broader in scope, are commonly associated with one of the two wars, (holocaust, Allies, home front.) Because of the time consuming nature of the collection-building process, a decision was made to focus the foreign-language part of the collection on German sites; with the idea that it would be more useful to have one language with a deep collection than many with shallow ones. (Sites identified in other languages were included, but not sought after.) Native German-speakers were consulted and helped design a search strategy to maximize the number of resulting German sites. This strategy took into account local conventions on not speaking only of World War II (zweiter Weltkrieg), for example, but more commonly of the period in which the Nazis ruled (Nazizeit) or Zeit des Nationalsozialismus, the period of National Socialism. This approach illustrated the need for localization, not just translation, when building a collection of sites in other languages.

As the topics for collection development were narrowed, the collection of seed sites continued to grow, but there were several content areas that remained difficult to include. A majority of the material from museums, libraries, and archives was not findable using the subject searches mentioned above. Most of this material was identified using targeted searches of domains likely to contain relevant content. Many of these institutions use local databases to deliver content that are not publicly indexed by common search engines. The New York Public Library has an extensive digital collection of photographs, over 2,100 of which are relevant to one of the world wars. These materials can only by located by first going to NYPL's site. Similarly, Harvard University has a collection of almost one thousand digitized pamphlets from World War I. They can only be found by searching in the library's union catalogue. In each of these cases, knowing that the materials exist—or might exist—is a prerequisite for being able to find them. But even when located, materials in databases remained problematic. There is usually no directory structure that can capture a number of items at once, nor are the URLs generated by database searches commonly stable. URLs to the Harvard materials, for example (http://pds.lib.harvard.edu/pds/view/7845178) only provide access to the first page of the multi-page objects. While NYPL does provide stable

URLs for the objects in its database, these need to be identified within each bibliographic record in order to be added to the seed list.

This difficulty has pointed to a challenge for future research: building tools for dynamically assembling and maintaining web archives focused on a topic. While building such a tool is not within the scope of this project, some project members are engaged in parallel efforts which we hope to be able to leverage for just such a purpose.

The first list of seeds was provided to IA, with particularly important resources indicated. IA then extracted the historical web pages from the Internet Archive of the seeds and the pages to which they linked within the same domain. In the case of the designated important resources, if data was not already present in IA's global web archive, a new web crawl was initiated to gather the current data. These seeds were then traced backwards through time in the IA global web archive to identify current and former links to related materials on the web. IA provided a list of new candidate seeds to OII (based on the link extraction) for verification by OII experts. At each step of this process, IA reported to OII the current size of the collection. Once the complete collection was assembled, IA extracted metadata from the records and indexed the files. Given the relatively small scope of the WWI collection, the team decided to add a WWII collection and to compare the two. The same methodology described above for WWI was applied to the creation of the WWII collection.

4.3 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 seed usually equates to per site, in a collection, if not for every capture contained within the collection.

We did not limit in any way the metadata formats that could be supported by the research framework we designed and deployed. However, for practical application we selected an XML format to house the extracted metadata and Researcher-defined collection metadata. For this project we used 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 analyzed each unique seed and newly discovered domain within each collection. IA used scripts to automatically extract and record metadata using this defined criteria with the goal of populating a single, zipped XML file, parseable by site, formatted in the MODS schema template and zipped using a Unicode compliant zipper. Zipped XML files were stored in WARC files for preservation of the metadata.

4.4 Link Extraction & Analysis

IA used 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 each collection. Output of each analysis was used to identify additional source materials (URIs) for inclusion in the collections as well as to identify trends for more detailed research. The results of this exercise lead to additional important findings from this

project.

The team experienced significant challenges automating the identification of subject-matter resources of interest and their extraction. Tools missing from the workflow that fall outside the scope of this project made the initial scoping of the collection difficult. Either the process was biased toward inclusion and too much material, including irrelevant content, was extracted or the process was biased toward manual curation resulting in the inadvertent exclusion of germane materials. For example, you can choose to extract materials for all outbound links identified per seed or to only harvest those that are manually identified as relevant. Ultimately, we adopted a hybrid approach with emphasis on the latter vs. the former – quality over comprehensiveness. This meant that some materials that should have been extracted were not and that odd gaps emerged in the end user navigational experience of some web pages included in the collections. In addition, the tools that were available during the project required engineering skills to utilize, making them infeasible for application and use by the typical e-humanities scholar. The development of easy to use scoping tools to create logical subsets of existing archives is a prime area of future research to be explored. These tools need to be able to inspect both inbound and outbound link data.

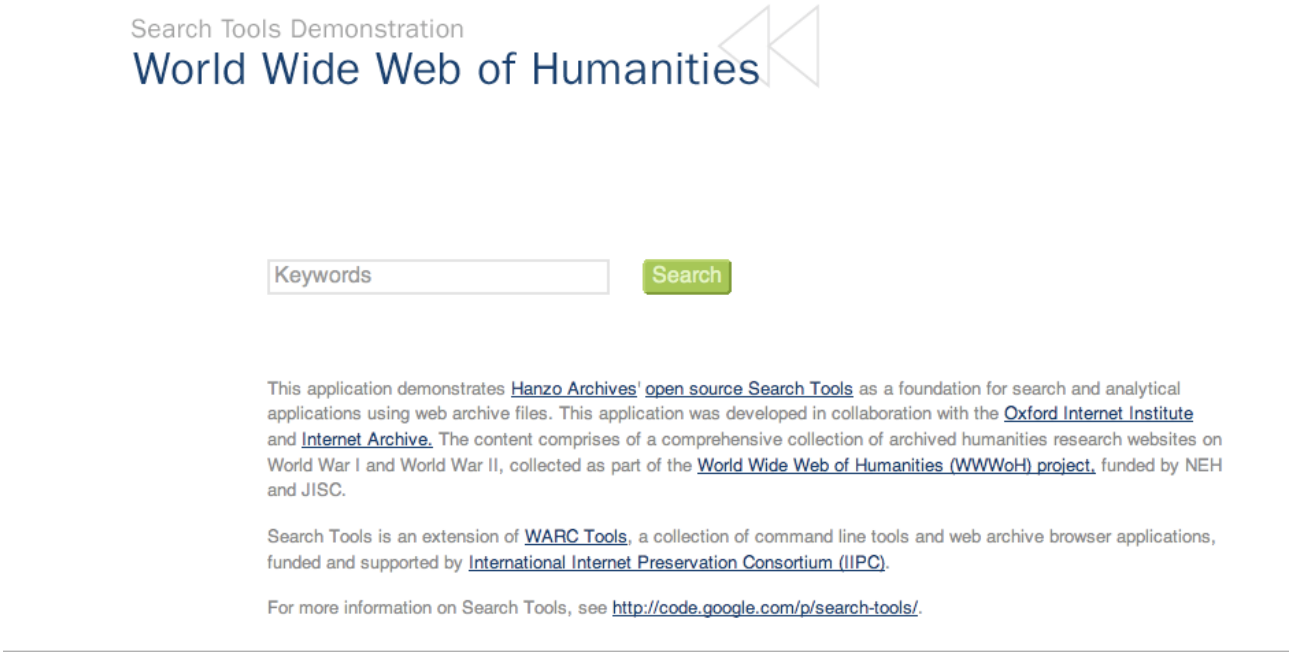
4.5 Organization of Materials

The captures included in the e-Humanities web collection and their associated metadata were stored in ARC and 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 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 files record and store this meta-information natively.

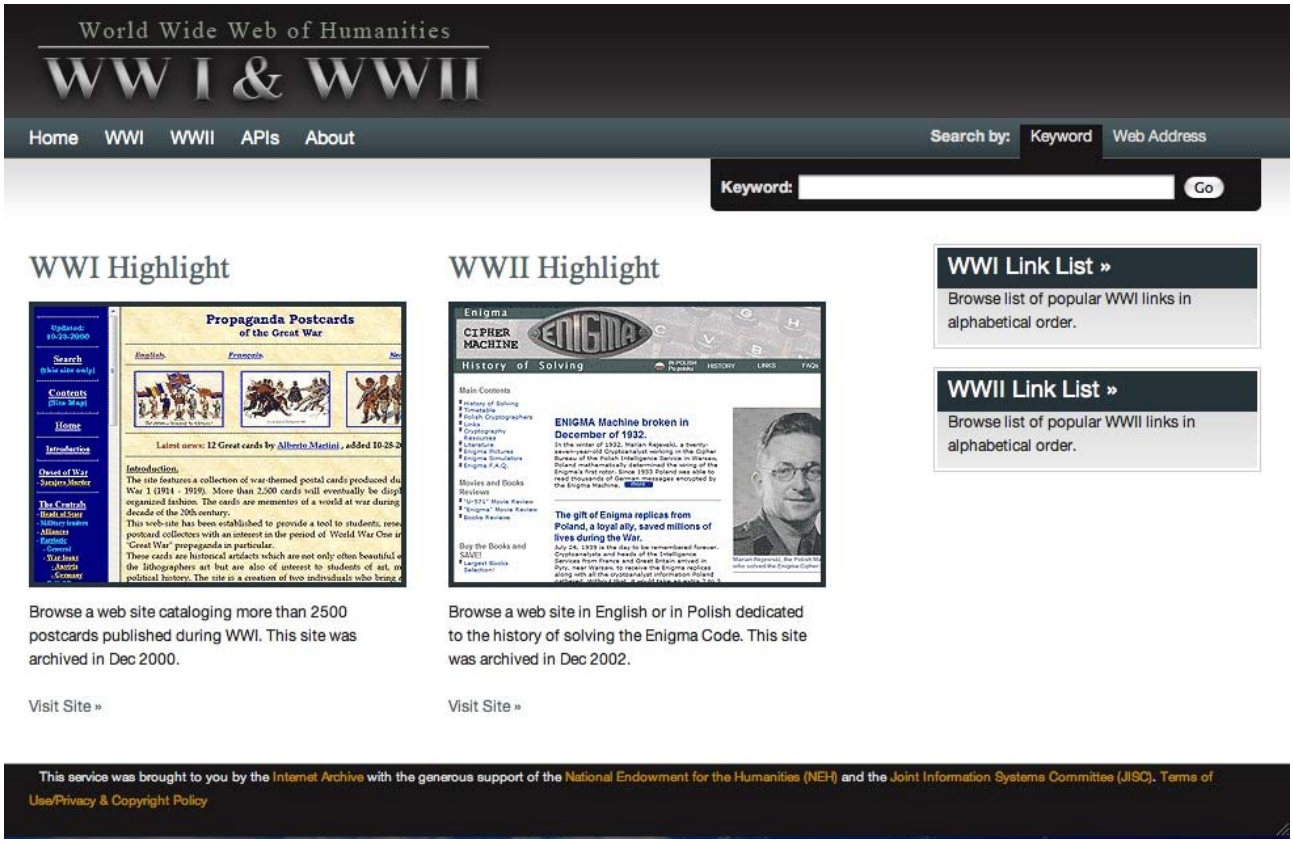
The WARC and ARC/DAT files were indexed for access via browse tools such as the Wayback Machine, and via full text indexes using NutchWAX and Ferret open source search engine packages to enable traditional keyword based search of the collections.

4.6 Access to Materials

Standard end-user, browser based interfaces to the e-Humanities collections were generated using a dedicated implementation of the open source Wayback machine and a new web archive browser derived from Hanzo's open source WARC Tools, two open source full text search engines – Nutch and Ferret – and a suite of open source tools for analyzing archival web data, including a general analytics API and example analytical applications, such as frequency tables for domains, MIME types, countries and a range of link graphing tools. In the future, it will be important to implement additional tools with which one might federate other e-humanities collections.



http://wwwoh.hanzoarchives.com/



http://wwwoh-access.archive.org/wwwoh/

IA also deployed OpenSearch, and XML and XSLT and other standard Web services interfaces to the meta data records and full text indexes of the e-Humanities web collections assembled during this project.

Hanzo deployed web services interfaces to the metadata records and full text indexes of the e-

Humanities web collections assembled during this project as well as open source data analytic tools.

Independent implementation of two distinct interfaces enabled us to design and explore two visions of the solutions that could help advance e-humanities research. Each interface incorporated a unique tool set, distinct workflows, and separate data storage formats, and development platforms, but each accessed the same source data set. The interfaces also targeted the needs of slightly different audiences. Hanzo built a suite of interfaces designed for an e-humanities researcher and incorporated visualization tools specifically desired by internet researchers. IA built an interface for e-humanities researchers and experimented with featured content that might appeal to an educator seeking supporting materials for curriculum development or to your average "arm chair" historian. We only regret that we did not have more time to gather more feedback.

Most of the planned enhancements to the IA interfaces involve improvements to the software used to browse and search the archival web materials. Improvements stemming from research and development in these areas will be applied to the interface on a regular basis to improve the quality of the end user experience and the collection. Any materials identified for inclusion in the data set will be extracted and added to augment relevant available web resources.

Hanzo are committed to the ongoing development of the software application developed and are actively seeking funds to enable this. New features and refinements in the browse and search functionality of the system will be uploaded to the public source code repository on a regular basis.

Hanzo's implementation also addresses the developers of analysis and visualisation tools. There is a new ISO format for web archives, WARC, and the Hanzo tools enables developers to make use of this format in their analysis and visualisation tools. This will hopefully encourage more widespread adoption of the WARC standard beyond the web archiving community.

The long term benefits are that many researchers gather material for analysis and are forced to adopt to the formats used by their tools. By keeping the gathered material in WARC files, the content will be searchable and browsable, and may be donated/submitted to an archive, such as IA for long term preservation. Furthermore, gathering can be carried out with IA's archival crawler, which is freely available.

The IA Open search API is available currently. We will add a page to the site called APIs that explains how to make an open search API call to the collection. To access the other APIs you must have a login. We will include information on the API page regarding how to request API credentials.

4.7 Storage Maintenance and Protection of Data

IA is committed 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.

IA is committed to host the interface and data set on dedicated hardware for the life of the equipment (a period of at least 5 years) to facilitate ongoing access and experimentation. In truth we will maintain the collection forever, it simply won't be housed on dedicated hardware and may

not always be accessible 24x7x365 after this initial r&d period expires.

Hanzo will provide access to the software source code on an ongoing basis via a publicly accessible source code repository. The WWWoH data is hosted as a test of the software, and Hanzo will retain this for a period of 1-2 years as the development of the software progresses.

During this time, Hanzo will provide the data and software to OII to host independently.

Primary and secondary copies of the materials are being stored and preserved as a collection by IA. Access copies of the collection materials have also been generated in support of researcher interfaces to the collections hosted by IA. Access to the collection APIs is controlled via a user account and login. Individual materials are available for download but the collection as a whole is not.

Additional preservation copies of the collection materials were transmitted to the Library of Alexandrina in Egypt and to OII. An additional replica was provided to Hanzo in the UK to support migration of ARC to WARC format and the development of the search engine framework.

IA's technical migration strategy for digital materials is two-fold: Hardware upgrades and file access formats. Since its inception in 1996, IA has migrated its collections four times – from tape to disk, and three more times to later generation disk systems. The most recent migration occurred in Winter 2008/2009. Accordingly, technology migration has occurred to and from platforms operating at very large scale. IA also 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.

Ease of use for the researcher was perhaps the most challenging issue we aimed to address, and there is still room for additional developments in this area.

We also discovered a number of additional issues that are important for the web archiving research community to focus on in upcoming years:

- 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

Other important lessons from this project:

- The speed with which we defined and compiled the collections was longer than planned: it took six months versus three to compile the source materials leaving less time than originally desired for development and iteration of end user interfaces. This lesson will be important to similar projects in the future, and reinforces our finding that curatorial tools need to be significantly developed to support these sorts of activities.
- Documenting requirements and assembling a prototyping environment and a team of domain experts to evaluate the resulting resources and tools will help to facilitate rapid iteration in a distributed and collaborative manner

5 Outputs and Results

The project was completed on schedule and within budget with regard to the detailed schedule of deliverables outlined in the project planning report (see appendix for details).

The core search tools software developed during the course of the project is available for download from http://code.google.com/p/warc-tools/ and http://code.google.com/p/search-tools/ under the open source license Apache License 2.0. The most recent release enables a researcher to take a collection of web archive files, index them and perform sophisticated full-text search and analysis on the collections.

Additionally two web sites that include access to the WWI and WWII collections via API and browser-based interfaces were made available at the conclusion of this project:
http://wwwoh-access.archive.org/wwwoh/
http://wwwoh.hanzoarchives.com/
They are open and available to researchers and the general public.

The final event for the project, held on 19 March 2009, attracted over 100 participants and resulted in lively discussion not only of these resources but of broader issues pertaining to the digital humanities. The slides from the day are available at: http://www.slideshare.net/etmeyer/wwwoh. There is also a webcast of the talks, available at: http://webcast.oii.ox.ac.uk/?view=Webcast&ID=20090319_275

5.1 Outcomes

The initial selection policy the team employed was built around identifying high-quality web resources that fell broadly within the initial chosen topic, 'World War I.' Early on in the seed selection process, though, we realized that this selection policy would not result in anything close to the original target of 100-250 million pages, as the first few passes through the collections yielded barely 1 million pages. In order to expand the sample size to allow the team to test the viability of the search tools, the scope of the collection was expanded to include World War II, and the selection criteria were changed to favour inclusiveness more heavily than selectiveness.

In the end, our collections were smaller than the total possible limits identified by those responsible for the technological implementation. This was a key lesson: even though the data deluge (Hey & Trefethen, 2003) is often identified as a key challenge for researchers across fields, focused collections in the humanities are still relatively unlikely to encompass hundreds of millions of objects. In some ways, this presents an opportunity for humanities scholars; if collections in this area are smaller than the petabyte-sized databases of physics and astronomy, the technical challenges of dealing with the sheer size of the data is much less likely to be a limiting factor in designing the research.

We intended this project to attract extensive involvement by humanities researchers, specifically those with domain expertise in WWI and WWII. We have had increasing attention from and engagement with the relevant research communities, but we suggest that similar projects front-load these efforts to recruit the involvement of scholars with domain specific expertise. In our experience, for future projects this should be a top priority.

The team is proud to have enabled students, faculty, and researchers to experiment with tools, workflows and api's that facilitate the assembly and analysis of focused web collections over time. We also firmly believe this project has contributed to an understanding of the use cases to be used by teaching and learning communities to illustrate the emerging challenges and opportunities

presented by the study of digital humanities via the Web.

6 Implications & Recommendations

One of the exercises undertaken during the course of this research was to identify potential workflows for researchers using collections such as these. These have been partially implemented, but need additional research to make them user-friendly and seamless for non-technical users.

6.1 Building Collections of Web Sites

The process of gathering collections of web sites, for many social scientists and humanities scholars, has in the past been more of a process of gathering lists of citations and links. Most 'web site gathering' has simulated traditional citation gathering. Some researchers use the bookmarking function in their browser to gather lists of sites, while others use tools such as Delicious. The tool that is rapidly becoming the most common, though, is Zotero (http://www.zotero.org.) Zotero is a production of the Center for History and New Media at George Mason University and is a free and open source Firefox extension. It began primarily as a citation tool—similar to traditional reference management software such as EndNote—but with the added ability to detect the presence of a book or article citation in the browser frame. Recent iterations of the software have provided the ability to download and attach pdf and other documents to a citation, and most importantly in this context, to take a "snapshot" of a web page or an entire web site for later reference on or offline. Once downloaded, the snapshot can also be annotated with tags and notes. Zotero is in many respects enabling for the first time the building of actual collections of web sites though a simple browser plug-in.

Another current approach—employed by scholars who are interested in text analysis—is to take advantage of the RSS capabilities of a site. An RSS reader, such as Google Reader, is used gather the text context of a web site. The URL provided by Google Reader can then be used to import the text into a database application for organization and analysis.

6.2 Analysis of Content

Current activities in the analysis of web site content currently fall into two broad categories—qualitative content analysis and social network or link analysis. Qualitative content analysis is usually preceded by the downloading of content, either using the RSS-based method described above, or by copying and pasting text from a site into a word processor. The most popular content analysis tools—Nvivo, Altas.ti, and QDA Miner—are not open source. Tams Analyzer (http://tamsys.sourceforge.net) is one of the few open source qualitative analysis softwares available.

For social network analysis and link analysis, there are several open source and heavily used tool. VOSON (http://voson.anu.edu.au/), LexiURL Crawler, and SocSciBot are all web-based software that engage in web mining and data visualization. Each requires the users input a list of seed sites to begin.

6.3 User Scenarios / Research Questions

Given the nascent state of research on historical collections of web sites, the facilitation of searching collections of web sites will undoubtedly present user scenarios and research questions that could not have been predicted. Below are several sets of user scenarios and research questions that are indicative of where research such as this can lead.

1. The internet scholar

Consider a master's student in a social science program with a background in the humanities. She is studying the coverage of the two world wars on the web. She is computer-literate, but does not have any programming or highly-technical skills. She is good with word-processing programs and spreadsheets, but not much beyond that. She gathers lists of all the web sites she finds by doing topic-related subject searches in Google. In order to get lists of links, she clicks on each link and then copies and pastes it into the spreadsheet.

Questions might be:

- How many of the sites are about WWI vs. WWII?
- What has been the annual count of sites for the last 7 years?
- What is the creation date of the pages? [She would like to identify trends in when the pages were created? Does it cluster? Or has it been a steady building process?
- What languages are the pages in?
- How many of the sites are from educational domains (.ac or .edu) vs. commercial domains?

2. The social network analyst

This example is a doctoral student who is interested in social network analysis. He is interested in whether there are visible networks of web sites that have grown up around WWI and WWII sites and whether there are any noticeable hubs. To identify the seed sites, he starts with subject searches, but has written a program that scrapes the URLs from the first 10 pages of search results. He would like a snapshot of the relationships between (inlinks, outlinks, co-inlinks and co-outlinks) between all of the sites he can find on WWI and WWII every 3 years.

A web archive might contain important information about where amateur or professional historians/genealogists/veterans gather, providing historians with ways of contacting groups or individuals who may be able to help them with their work. Pages/Sites which had provided a popular meeting place, for instance, but which may no longer exist, might help the historian to develop similar resources.

Questions might be:

- Where on the web do historians/genealogists/veterans gather to talk about their research/experiences?
- What kinds of web pages/sites attract contributions from those who experienced the war?
- What (online) methods can be developed to attract such contributions?
- Researchers might also use the web pages to gather information about other researchers' experiences of known archives, revealing strengths and weaknesses, problems or advantages.
- Which are the key archives in this area? How easy is it to extract relevant information from them?
- Have these archives been used for my purposes before?
- Has the release of certain documents/data made specific events/sources/historical interpretations more popular among these groups?

3. The linguist

For this example, consider a linguist who is studying the historical use and understanding of the word 'dictator' from the beginning of world war one to the end of world war two. She would also like to analyze contemporary use of the word in relationship in regard to accounts of both wars. She has used Zotero to take snapshots of a sample of web pages. She has separated the sites (by hand) into two categories – those containing historical documents and those containing contemporary accounts. Within each group of sites, Suzanne would like to

identify the pages that contain the word and those that don't. Within the pages that contain the word, she would like them ranked by word frequency. She is interested in the overall findings and whether the data is from a page that is current or no longer exists is of only anecdotal importance to her research.

Questions might be:

- How are meaningful words clustered by geography? Is the word 'dictator' more likely to occur in pages from certain countries?
- How frequently are politically charged words used in materials in official collections (such as libraries and museums) versus on the open web?
- Are pages that contain certain highly charged words also more likely to link to each other?

4. The personal history scholar

   A great many sites covering the two World Wars contain personal testimony and copies of original sources such as photographs, letters and official documentation. It may be that members of the public who might not think to approach an archive or library with their own story or personal mementos would be more likely to mount details or copies of their mementos online. Often people have responded to sites which invite those who lived through these events to contribute their memories, and people may be more willing to do so in the privacy of their homes via the internet, than (for instance) to take the trouble to attend a more formal local history event. One of the attractions of these sites to historians, therefore, might be that they offer previously unavailable or untapped primary sources.

   Questions might be:

   - How many photographic sources are available on the web for WW1 /2?
   - How many personal reminiscences are available?
   - When were the reminiscences written?
   - When were these sources collected?
   - How reliable are these sources?
   - Which are the most popular sources to present on internet sites?
   - Why might these sources be more prevalent – does this tell us more about society now or society then?
   - Do these sources/accounts challenge the current historiography?
   - To what extent have these sources been overlooked by historians? Why?

5. Scholar interested in historiography

   An archive of websites such as this might give historians the opportunity to survey a different kind of historiography, which might exist largely outside the academy. The two World Wars attract a great deal of interest from amateur historians, and this may be of use to professional historians, documenting areas of interest, and the way that they are presented. You could trace popular events, sources, discussions etc, and map different historical and historiographical trends. You could also trace the impact of the digitisation or release of certain archives or repositories.

   Questions might be:

   - How does the web historiography on the two World Wars compare to that present elsewhere?
   - How can such historiography add to our knowledge of these events?
   - Are these sources more or less subject to historical trends? Is this useful?
   - Which archives/sources have become popular with amateur historians?

A web archive might also allow historians to search through conference web pages for references to people/subjects/papers. This would allow unpublished or ongoing work to be traced, and would also allow historians to make informal approaches to colleagues who have spoken on relevant subjects.

Questions might be:

- How many conferences have been organised on a particular subject/set of resources?
- Who has attended/presented work at these conferences?
- Who is working on this area but not publishing their work?

6. Scholar interested in national/international perspectives
   An archive which contained web material from a variety of different countries would provide useful information for researchers interested in exploring international perspectives on the conflict. Viewing the development of particular pages or sites might allow the historian to understand shifting perspectives or attitudes, and to understand the way in which particular events are viewed by different groups. This would allow historians to approach these issues with greater sensitivity and success (one would hope!).

   Questions might be:

   - How are the World Wars viewed in contemporary (e.g. French/German/Polish/American) society?
   - Have these perspectives shifted in recent years?
   - Do these perspectives differ across social groups?
   - Can we trace these perspectives to particular historical events/experiences?

These sample scenarios are just a few among many potential areas of research. As noted elsewhere in this report, it is very important to recruit subject domain experts, not simply discipline specific scholars, at the beginning of a project such as this one if the resulting collections are going to have a greater chance of having a lasting impact. The assembly of collections such as this remains a time-consuming and labour-intensive process that needs to be streamlined if researchers are likely to engage in this sort of research. This is partly tied to the specificity of much humanities and social science research: an individual scholar or small handful of scholars often need to create and use very specialised collections, but do not currently have adequate tools to enable that activity. Also, there needs to be considerable additional work to leverage and simplify tools for analyzing both inbound and outbound links to web resources at a larger scale, which in turn will inspire new research questions.

7 Conclusions

There is tremendous benefit to researchers from having access to logical subsets of longitudinal web archives, specific to their research interests. Although there are lots of areas of development needed before web research will become an everyday activity of researchers, this project has helped contributed to the pool of available tools and knowledge of best practices that will help move us in this direction.

The opportunity under the JISC-NEH Transatlantic Digitisation Programme to build a new collaboration between the UK and the USA has been a decidedly positive experience. The partners on this project have worked well together, and we are already putting together further applications to develop areas of importance that have emerged as a result of this project.

There are still critical tools missing from the researcher workflows, and some that are still too complicated for use by your average researcher. There is much work to be done in the vein of ease of use and enhancement of the end user experience. Also, involving e-humanities scholars with domain expertise in the creation and evaluation of collections will have a significant impact on the quality of the collections produced and upon the resulting tools and interfaces generated. This project has supported the findings of other work suggesting that the diversity of researchers and research goals is a key challenge to implementing usable researcher workflows: since researchers have so many approaches to research, building workflows that support existing approaches is a key lesson for similar projects.

8 References

Hey, T., & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective. In F. Berman, G. Fox & T. Hey (Eds.), Grid Computing: Making the Global Infrastructure a Reality (pp. 809-824). Hoboken, NJ: John Wiley & Sons, Ltd.

9 Appendix: Implementation Schedule

(Key deliverables in **bold**)	X if completed, or anticipated completion date	**Outputs**	**Milestone**	**Responsibility**
YEAR 1				
WORKPACKAGE 1: ASSEMBLY Objective: Assembly of the collection/s (from live crawl and data extraction using seeds derived from domain experts and hyperlink analysis of the collection domain)				
Identification of seed websites on topically focused research collection/s	X	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
Extraction of seeds from the Internet Archive	X	Collection as a set of extracted ARC files	ARC files done	IA
If required: Live crawls	X	If required: live crawling of web for inclusion of key seeds not available in Internet Archive	ARC files done for live crawls	IA
Link extraction and analysis	X	Complete selection of candidate seeds per year going back through the IA collection.	Candidate seed list	IA
Domain expert validation of extracted links	X	Have domain experts look at a small list of candidates and validate the inclusion of extracted links	Augmented seed list	OII
Extraction of additional seeds from augmented list	X	Final extraction of additional seeds	Final list of	IA

(Key deliverables in **bold**)	X if completed, or anticipated completion date	**Outputs**	**Milestone**	**Responsibility**
			250 M documents	
Metadata extraction from seeds	X	Extract metadata from seeds to populate the MODs records	Populated MODs records	IA
Full text indexing	X	Text searchable collection across the years 1996-2007 for all seeds	Collection	IA
Link extraction from augmented seeds	X	Complete list for analysts of link data from the collection	Link file for analysis	IA
WORKPACKAGE 2: INTERFACES **Objective:** 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				
Browser interface to the collection for end users, specifically e-Humanities scholars and curators	X	Access interface to the collection for e-Humanities scholars (only includes post collection creation tasks such as search, browse of the collection)	Interface	IA
XML API to the collection	X	Experimental interface for integrating workflow and analysis tools	Interface	IA
OAI-PMH interface to the collection	X	Enables scholars to harvest metadata from the collection	Interface	IA
OpenSearch protocol support	X	Enables federated search with other relevant collections	Interface	IA
Web Services API to the collection	X	Experimental interface for integrating workflow and analysis tools	Interface	IA

(Key deliverables in **bold**)	X if completed, or anticipated completion date	**Outputs**	**Milestone**	**Responsibility**
WORKPACKAGE 3: ANALYSIS TOOLKITS **Objective**: Enable research using third party tools.				
Identify & evaluate existing open source tools for workflow and analysis	X	Experts evaluated existing open source analysis tools as candidates for inclusion in end user prototypes. Tools were considered for all phases of the workflows to be prototyped, including workflow creation and management software commonly in use by the e-humanites and e-research communities.	List of existing tools to integrate with end user prototypes, classified by use case and stage of workflow.	OII, Hanzo
Search Engine Framework Beta	X	A beta version of the search engine framework was available for download and included support for index of an arbitrary collection of WARC files.	Beta software package available for download and installation locally.	Hanzo
Search Tools v1	X	The first version of the search engine framework is available for download and included support for index of an arbitrary collection of WARC files as well as integration with basic analysis tools. This code is available for installation anywhere and	Software package available for download and installation locally.	Hanzo

(Key deliverables in **bold**)	X if completed, or anticipated completion date	**Outputs**	**Milestone**	**Responsibility**
		complements the hosted services deployed by IA.		
Analysis Toolkit (locally deployed) API	X	The Search Tools were extended to include an API that results in a prototyping platform upon which were able to experiment with researcher interfaces and workflows.	API to the Search Tools v1	Hanzo
WORKPACKAGE 4: END-USER PROTOTYPES **Objective**: Build end-user prototypes that support use of the tools and collections appropriate to the target audiences.				
Define research questions	X	Definition of specific research questions to use to illustrate how the platform, tools, and workflows might be applied to the collection as well as how the process might be replicated on an entirely different collection or subset of the Web and digitized humanties 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
Develop experimental prototype for the e-Humanities scholar/curator	X	Creation, implementation, and evolution of browser-based tools and interfaces to the collection that	Hosted service for e-Humanities	IA, OII

(Key deliverables in **bold**)	X if completed, or anticipated completion date	**Outputs**	**Milestone**	**Responsibility**
		facilitate e-humanities curation and research.	curators and researchers	
Develop experimental prototype for the e-researcher	X	Creation, implementation, and evolution of programmatic and browser-based interfaces to the collection, integrated with existing tools that facilitate e-research.	Hosted service for e-researchers	IA, OII
Develop experimental prototype for locally hosted solution	X	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
WORKPACKAGE 5: DISSEMINATION **Objective**: Disseminate materials from the project				
Interim Reports to JISC and NEH	X	Creation of interim report to JISC	Report	OII, IA
Project website: http://www.oii.ox.ac.uk/research/project.cfm?id=48	X	Project website includes 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
Copy of archive delivered to a European archive	X	Replication of collection to a location in Europe	Replicated archive	IA, Hanzo, OII
Collection website housed at IA /	X	Collection website that is searchable and usable as per the technical details	Website	IA

(Key deliverables in **bold**)	X if completed, or anticipated completion date	**Outputs**	**Milestone**	**Responsibility**
		listed in other workpackages.		
Presentations and publications	After end of project	Papers accepted to several conferences, presentations scheduled, and publications in process	Presentations and publications	OII, IA, Hanzo
Final Reports to JISC and NEH	X	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
Demo event in the UK	X	Demo event at Oxford, including all project partners, funders, digital humanities curators, domain experts, and other interested parties completed on March 19, 2009	Workshop	OII, IA, Hanzo