The future of researching the past of the Internet

Eric T. Meyer
Oxford Internet Institute
University of Oxford

With contributions from colleagues at:
Oxford Internet Institute
Internet Archive
Hanzo Archives
University of Southern California

The Internet Archive is...
A digital library of ~4 petabytes of information

- Web Pages
- Educational Courseware
- Films & Videos
- Music & Spoken Word
- Books & Texts
- Software
- Images

The Archive’s combined collections receive over 6 mil downloads a day!

www.archive.org

Slide courtesy of: Molly Bragg, Partner Specialist, Web Group, The Internet Archive
1.6+ petabytes of primary data (compressed)

- 150+ billion URIs, culled from 85+ million sites, harvested from 1996 to the present
- Includes captures from every domain
- Encompasses content in over 40 languages
- As of 2009, IA will add ½ petabyte to 1 petabyte of data to these collections each year.

Slide courtesy of: Molly Bragg, Partner Specialist, Web Group, The Internet Archive
Why WWI and WWII?

Well-rounded set of materials

---

THE generous, prompt, and sympathetic response from all true Americans in their nonpartisan efforts to alleviate the suffering of the nations now engaged in the misery of war has served to brighten the edge of the heavy cloud which has so long darkened and hung over Europe. Every individual has doubtless felt this appeal more or less strongly, from the child who learns to roll a bandage to the mother with her knitting and the father who finds he can give where he was little accustomed to give before. But owing to the colossal nature of the struggle and the unparalleled number of wounded, individual efforts at succor, however fine the spirit may be, are but the proverbial drop in the bucket and it is possible to cope with the situation, and in a measure meet its needs, only through powerful organizations like the Red Cross Society and the Rockefeller Foundation or through the concerted action of groups of intelligent and self-sacrificing people who submerge their individuality in a co-operative labor.

One of the first of such groups, encouraged by the American Ambassador, came together in Paris soon after...
Why WWI and WWII?

Many branches of the humanities

- History
- Journalism
- Art
- Art history
- Advertising
- Literature
- Poetry
- Political science
- Military history

Slide courtesy of Christine Madsen, OII
A seed is also a web site from which additional sites can be discovered via the hyperlinks of the site.
Building the Collection

Started with WWI

Too small (under 1,000,000 pages / object)
Target was 250 million

Slide courtesy of Christine Madsen, OII
Building the Collection

Expanded to WWII

Final collection: 6,575,509 unique URLs

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

Total size of compressed ARC data = 240GB
Building the Collection

Returning to ‘hub’ sites for further analysis

Record links from first 20 pages of search

[include dead links]

Following links

Slide courtesy of Christine Madsen, OII
Dealing with illogical or flat directory structures

www.eyewitnesshistory.com/ <= don’t want whole site

www.eyewitnesshistory.com/blitzkrieg.htm
www.eyewitnesshistory.com/dday.html
www.eyewitnesshistory.com/midway.htm
www.eyewitnesshistory.com/airbattle.htm
www.eyewitnesshistory.com/dunkirk.htm
www.eyewitnesshistory.com/francesurrenders.htm
True Stories of Life as a Doughboy in World War I

By James E. Darst, former Lieutenant of 341st Machine Gun Battalion, 89th Division

(Contributed By The Family of James E. Darst)

Lt. Darst tells how the 89th Division Lived, Suffered, Laughed and Fought on Soil of France
The Wartime Columns of Ernie Pyle

For many journalists, Ernest Taylor Pyle, an Indiana native better known as “Ernie,” continues to be an icon of excellence decades after his death at the hands of a Japanese machine-gunner in World War II. For the last ten years of his life he wrote feature columns six times a week, primarily for Scripps-Howard newspapers. As his fame increased during the war, other newspapers, including weekly ones, published Pyle’s work.

In 1944 Ernie Pyle won a Pulitzer Prize for his stories about the ordinary soldiers fighting in World War II.

At this website you will find a selection of his wartime columns in both written and spoken versions. In addition, you will find some pictures of Pyle and tips on where you can find more information about him.

We welcome your comments about the site and stories you might have to tell about meeting Pyle or reading Pyle’s columns.

(These columns are reprinted with the permission of the Scripps Howard Foundation.)

A Dreadful Masterpiece

Pyle wrote this column nearly a year before the United States entered World War II. It describes the awe he felt as he watched the German air attacks on London.

Killing Is All That Matters

In this column, Pyle explains how servicemen going into battle will be changed by the experience.
A Dreadful Masterpiece

Pyle wrote this column nearly a year before the United States entered World War II. It describes the awe he felt as he watched the German air attacks on London.

Listen to this column read by Owen V. Johnson, Associate Professor, School of Journalism, Indiana University

- 5 minutes, 33 seconds
- 6.3 megabyte filesize

London, December 30, 1940-Someday when peace has returned to this odd world I want to come to London again and stand on a certain balcony on a moonlit night and look down upon the peaceful silver curve of the Thames with its dark bridges.

And standing there, I want to tell somebody who has never seen it how London looked on a certain night in the holiday season of the year 1940.

For on that night this old, old city - even though I must bite my tongue in shame for saying it - was the most beautiful sight I have ever seen.

It was a night when London was ringed and stabbed with fire.
Search Tools Demonstration

World Wide Web of Humanities

This application demonstrates Hanzo’s open source Search Tools as a foundation for search and analytical applications using web archive files. This application was developed in collaboration with the Oxford Internet Institute and Internet Archive. The content comprises of a comprehensive collection of archived humanities research websites on World War I and World War II, collected as part of the World Wide Web of Humanities (WWWh) project, funded by NEH and JISC. For more information on Search Tools, see http://code.google.com/p/search-tools/.
Migrating ARC to WARC

- Data extracted from IA in ARC files
- Hanzo WARC Tools and Search Tools projects combined enabled us to migrate ARC to WARC files (WARC is the new ISO standard):
  - Some challenges: broken ARCs, scale, etc.
  - 3,264 WARC files
WWWoH

Mime Types for: +"doomed youth" +url:(+bbc)

WWWoH

Top Level Domains for: +"doomed youth" +url:(+bbc)

Domain level 1

TL D Instance

uk 4

domain level 2

TL D Instance

co.uk 4

Graph for: +"doomed youth" +url:(+bbc)

This graphical analysis application uses the following graphing systems:

Guess : [Link to graphexploration.cond.org]

Graphviz : [Link to www.graphviz.org]

HyperGraph : [Link to hypergraph.sourceforge.net]
Graphing Tools
But what of the future?
Challenges for the Future

• Annotating web archives

• Selecting content to extract from archives
  – Efficiently extracting data from IA

• Analysing large bodies of time-series data
  – Involvement of domain experts (webometrics, SNA, e-Research, etc.)
  – Move from snapshots to more continuous data (within technical limitations)
    – Both outlinks and inlinks

• Sharing results in meaningful ways
Source:
Matt Weber & Peter Monge,
University of Southern California
“Humanities on the Web: Is it working?” conference
Date: Thursday, 19 March 2009, 10-4
Location: Oxford University, Oxford, UK
Webcast URL: http://webcast.oii.ox.ac.uk/?view=Webcast&ID=20090319_275
Slide URL: http://www.slideshare.net/etmeyer/WWWoH

Final report to JISC for WWWoH:
http://www.jisc.ac.uk/media/documents/programmes/digitisation/humanitiesfinalreport.pdf
Oxford Internet Institute
University of Oxford

Eric T. Meyer, Ph.D.
Research Fellow
eric.meyer@oii.ox.ac.uk
http://people.oii.ox.ac.uk/meyer