



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*

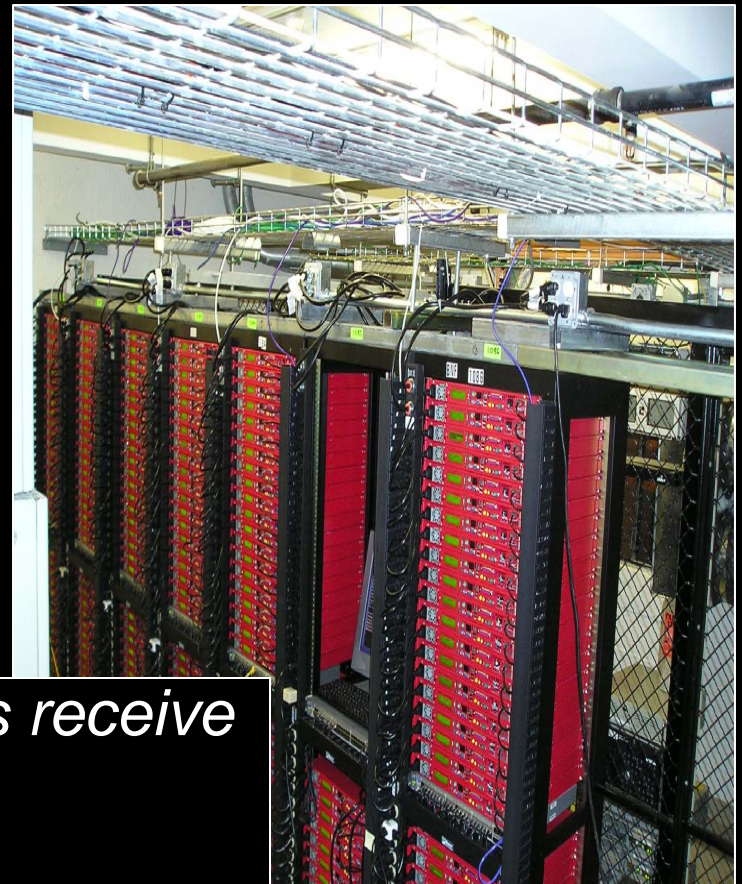
missing links: the enduring web, 21 July 2009, British Library Conference Centre, London



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



IA Web Archives

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.



Discipline Specific Data Extraction from Longitudinal Web Archives: The WWWoH Case Study

<http://wwwwoh-access.archive.org/wwwoh/>

World Wide Web of Humanities

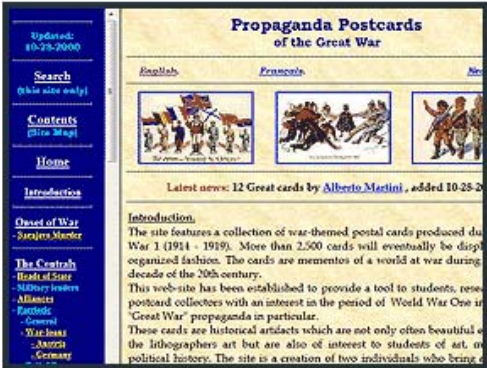
WW I & WWII

Home WWI WWII APIs About

Search by: **Keyword** Web Address

Keyword:


WWI Highlight



Browse a web site cataloging more than 2500 postcards published during WWI. This site was archived in Dec 2000.

[Visit Site »](#)

WWII Highlight



Browse a web site in English or in Polish dedicated to the history of solving the Enigma Code. This site was archived in Dec 2002.

[Visit Site »](#)

WWI Link List »

Browse list of popular WWI links in alphabetical order.

WWII Link List »

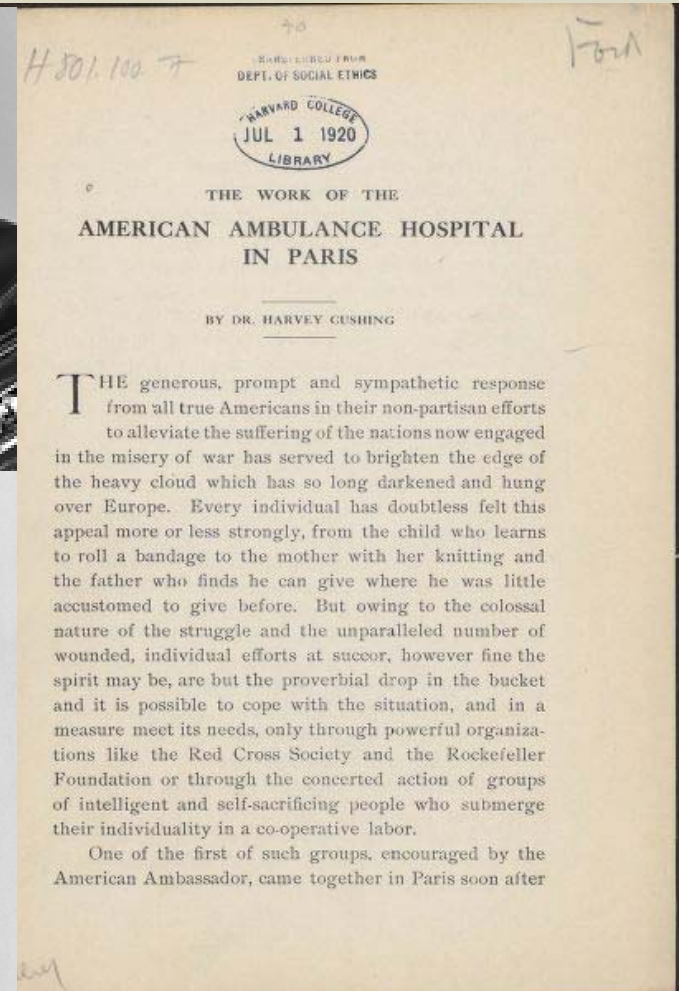
Browse list of popular WWII links in alphabetical order.

This service was brought to you by the [Internet Archive](#) with the generous support of the [National Endowment for the Humanities \(NEH\)](#) and the [Joint Information Systems Committee \(JISC\)](#). [Terms of Use/Privacy & Copyright Policy](#)



Why WWI and WWII?

Well-rounded set of materials



Slide courtesy of Christine Madsen, OII



Many branches of the humanities

History

Journalism

Art

Art history

Advertising

Literature

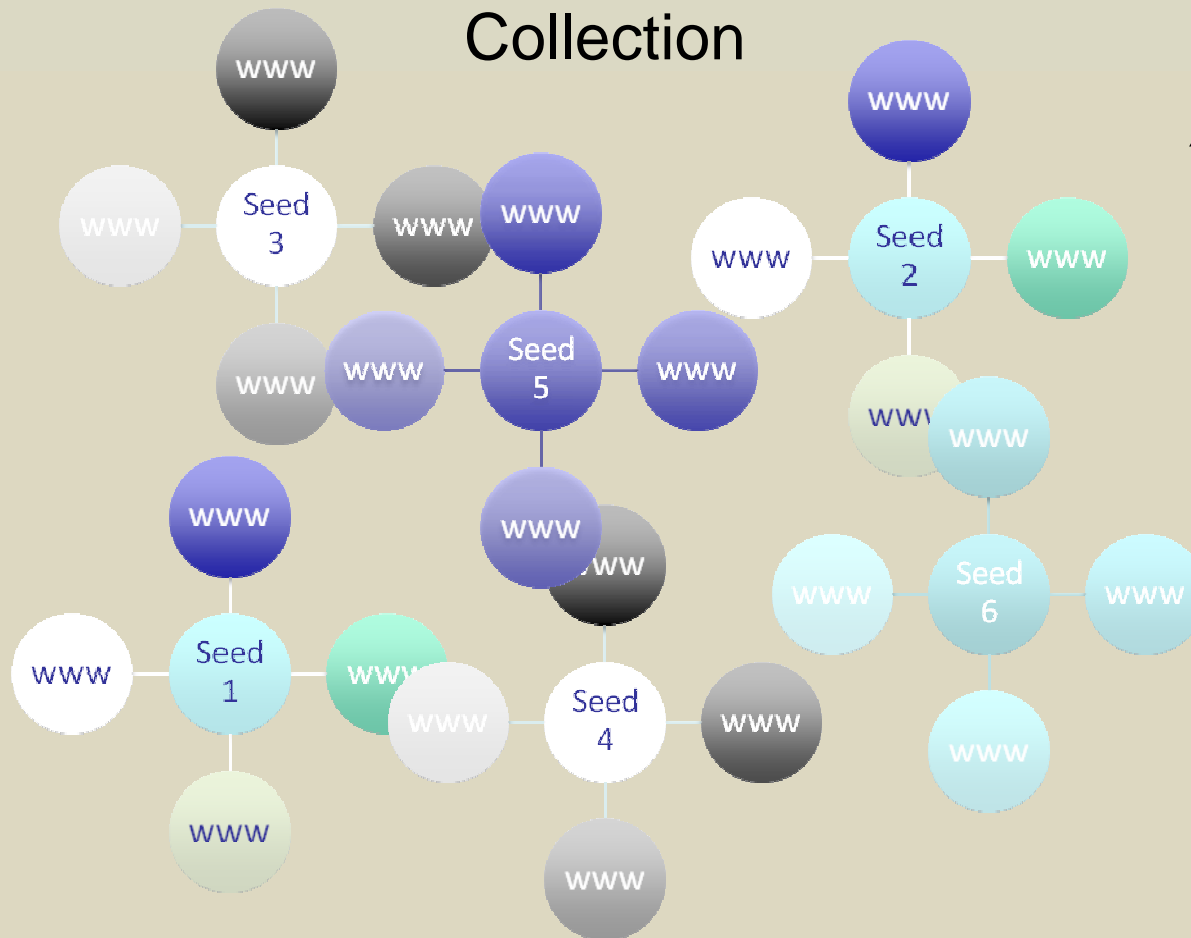
Poetry

Political
science

Military
history



Expanded Collection



*A **seed** is also a web site from which additional sites can be discovered via the hyperlinks of the site*



Started with WWI



Too small (under 1,000,000 pages / object)

Target was 250 million



Building the Collection

Expanded to WWII



Final collection: 6,575,509 unique URLs

Slide courtesy of Christine Madsen, OII



Table 1: Initial collection sizes

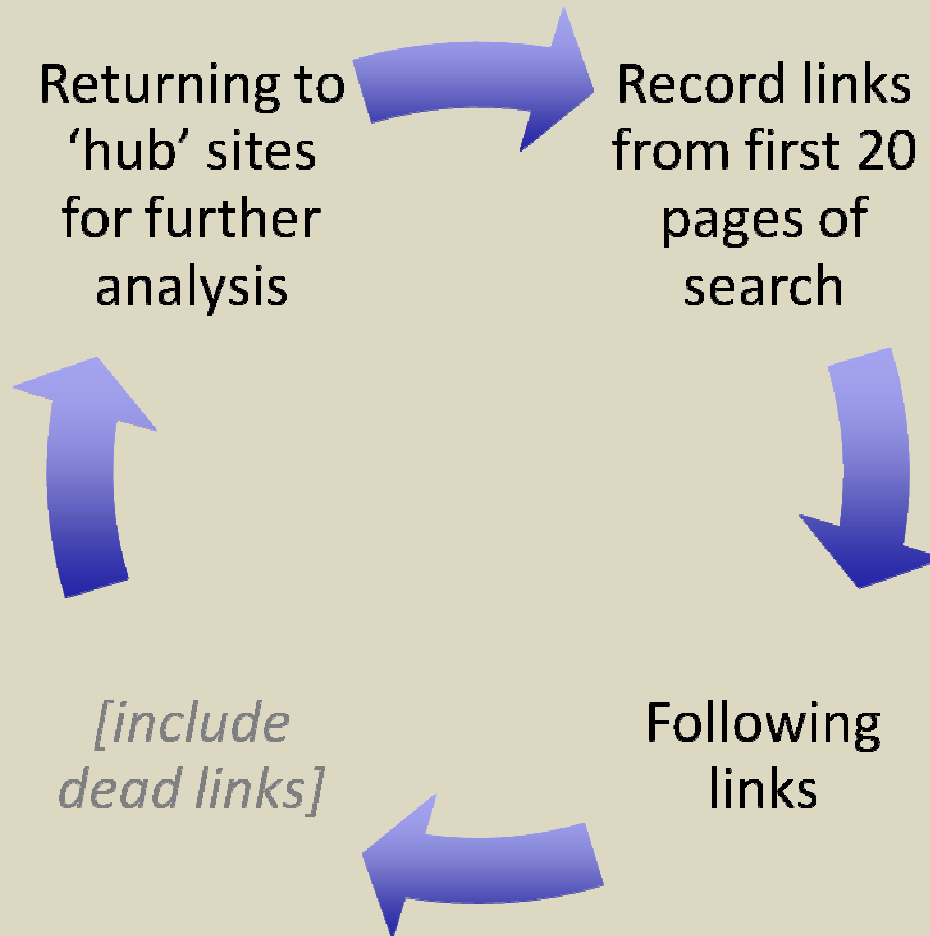
	World War I	World War II	Common to both WWI and WWII	Not seeded	Total
Number of seeds	2,263	2,592	29	-	4,884
Number of unique hosts	906	1,475	149	7,137	9,667
Number of unique URLs	2,312,937	3,160,408	624,610	477,554	6,575,509
Number of captures	8,424,630	13,320,354	858,736	2,428,434	25,032,154
Number 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.



Building the Collection





Dealing with illogical or flat directory structures

~~www.eyewitnesstohistory.com/ <= don't want whole site~~

www.eyewitnesstohistory.com/blitzkrieg.htm
www.eyewitnesstohistory.com/dday.html
www.eyewitnesstohistory.com/midway.htm
www.eyewitnesstohistory.com/airbattle.htm
www.eyewitnesstohistory.com/dunkirk.htm
www.eyewitnesstohistory.com/francesurrenders.htm



WWI: Example

Viewing version 8 of 23
21:35:09 Apr 27, 2004



Web Address:

All

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

WWII: Example

Viewing version 16 of 82
8:14:32 May 30, 2005

Sep 2002
Jan 2008

Web Address: All


INDIANA UNIVERSITY · BLOOMINGTON
SCHOOL OF JOURNALISM

?

Academics

?

?

Summer Workshops

?

People & Groups

Gallery

Resources

Contact the School

Search this Site

Announcements

May 16:
Faculty news: Cookman featured in new mag
[more...](#)

May 16:
Faculty news: Stocking presents paper at Cornell
[more...](#)

May 04:
Faculty news: Wisconsin honors Nord
[more...](#)

April 19:
Faculty news: Polsgrove judges SDX awards
[more...](#)

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.)


Column 1



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.

Column 2



Killing Is All That Matters

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

WWII: Example

Viewing version 7 of 41
13:19:02 Mar 10, 2005

Jul 2002 Nov 2007

Web Address: All Go



INDIANA UNIVERSITY · BLOOMINGTON

SCHOOL OF JOURNALISM

[Academics](#)
[People & Groups](#)
[Gallery](#)
[Resources](#)
[Contact the School](#)

Search this Site

Google Search

Announcements

Mar 8:
Faculty news: Fargo earns 'Top Faculty Paper'
[more...](#)

Mar 7:
Faculty news: Stocking, Holstein to present paper at science conference
[more...](#)

Mar 7:
Faculty news: Johnson's book reviews in three publications
[more...](#)

Mar 3:
Faculty news: Weaver co-authors article

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.

<http://www.woh.hanzoarchives.com/>

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 (WWWoH) project, funded by NEH and JISC. For more information on Search Tools, see <http://code.google.com/p/search-tools/>.

Copyright © 2009. Hanzo Archives Limited

HANZO  **ARCHIVES**

JISC

NEH

Internet Archive

Internet Archive



University of
Oxford

HANZO  **ARCHIVES**



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

Search Tools Demonstration

World Wide Web of Humanities

Search results for :**+"doomed youth" +url:(+bbc)"**

1

4 results

BBC - History - Wilfred Owen biography

2002-04-09 T 15:08 UTC | text/html | View Live

BBC - History - Wilfred Owen biography `td {font-size: 10pt} // CATEGORIES TV RADIO COMMUNICATE.....poetry of the war, including Dulce et Decorum est and Anthem for Doomed Youth . BR Go further Read a year-by-year summary of events of World War One...`
<http://www.bbc.co.uk:80/history/war/wwone/owen.shtml>

BBC - collective - the times london film festival preview 2004

2004-10-25 T 03:12 UTC | text/html | View Live

BBC - collective - the times london film festival preview 2004 `@import url(/includes/tbenh.css) ; Home TV Radio Talk Where?Live A-Z?Index Monday.....included. Greg Araki, provides another beautiful and uncompromising vision of doomed youth in the gay love story, Mysterious Skin . Two must-see...`
<http://www.bbc.co.uk:80/dna/collective/ff>

BBC - collective - the times london film festival preview 2004

2004-10-18 T 20:20 UTC | text/html | View Live

BBC - collective - the times london film festival preview 2004 `@import url(/includes/tbenh.css) ; Home TV Radio Talk Where?Live A-Z?Index Monday.....included. Greg Araki, provides another beautiful and uncompromising vision of doomed youth in the gay love story, Mysterious Skin . Two must-see...`
<http://www.bbc.co.uk:80/dna/collective/ff>

BBC - h2g2 - Benjamin Britten's War Requiem

2005-03-15 T 06:31 UTC | text/html | View Live

BBC - h2g2 - Benjamin Britten's War Requiem `body {margin:0px;} .nav,A.nav,A.nav:link,A.nav:visited {color:#FFFF00;text-decoration:none;} A,A.....shrill, demented choirs of wailing shells... - Wilfred Owen, 'Anthem for Doomed Youth' The 20th Century saw a great renaissance of British composers...`
<http://www.bbc.co.uk:80/dna/ww2/A735734>

HANZO ARCHIVES

Search Tools Demonstration | World Wide Web of Humanities...

WWWoH

+"doomed youth"

+url:(+bbc)

Mime Types

Instances

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

Domain level 1

TLD	Instances
uk	4

Domain level 2

TLD	Instances
co.uk	4

Copyright © 2009. Hanzo Archives Limited

Search Tools Demonstration | World Wide Web of Humanities... |Choice Graph

WWWoH

+"doomed youth"

Search

(close)

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

(close)

Guess

Select

This graphical analysis application uses the following graphing systems:

Guess : <http://graphexploration.cond.org/>



Graphviz : <http://www.graphviz.org/>



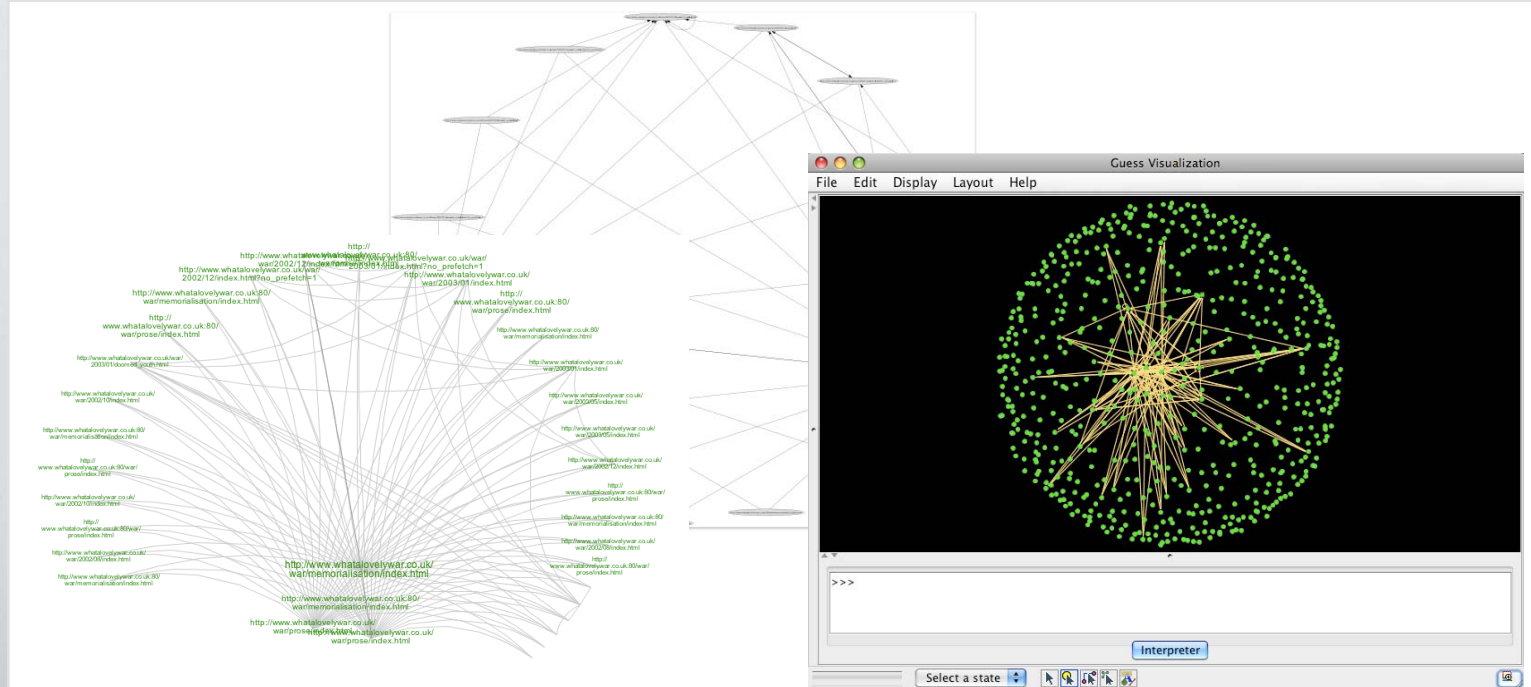
HyperGraph : <http://hypergraph.sourceforge.net/>



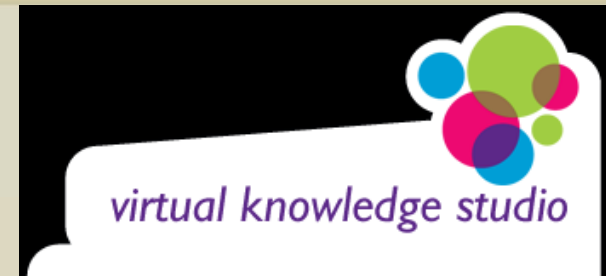
Copyright © 2009. Hanzo Archives Limited

HANZO ARCHIVES

Graphing Tools

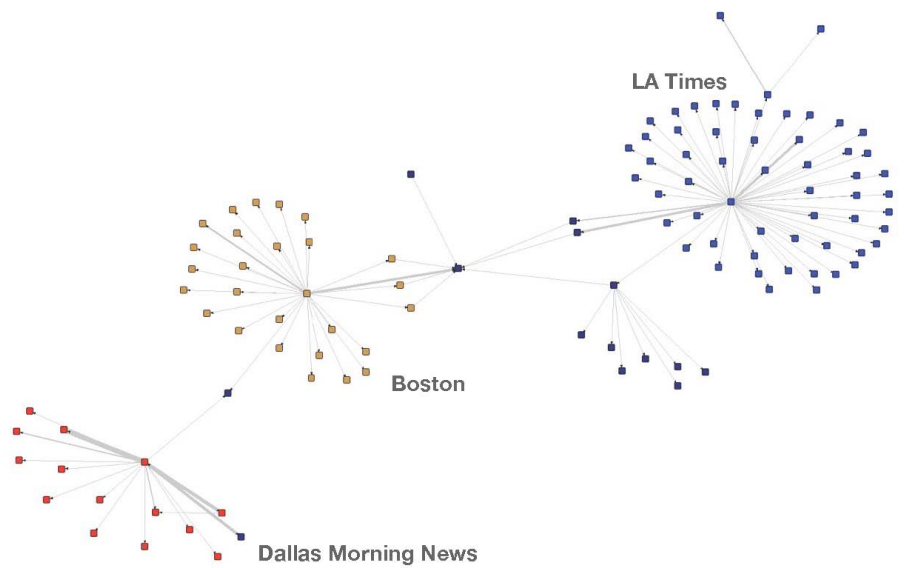


But what of the future?

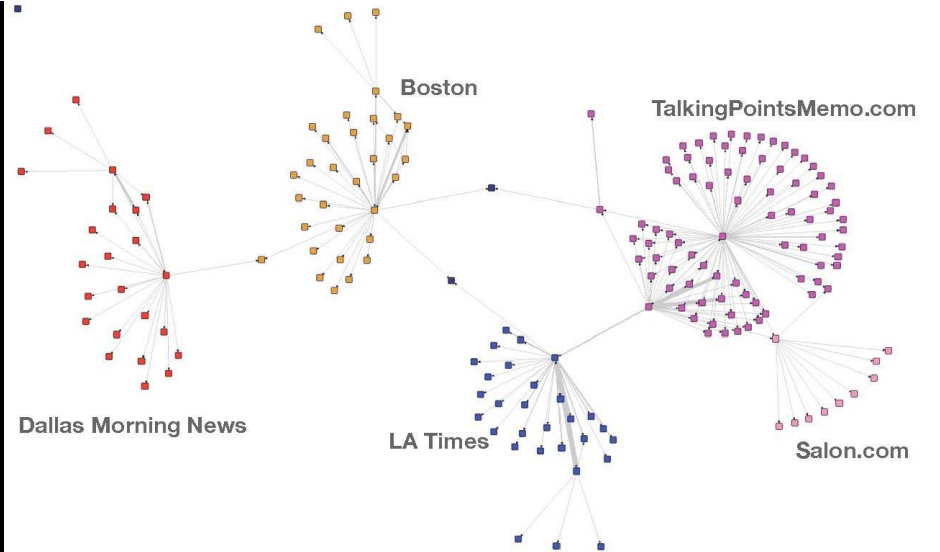




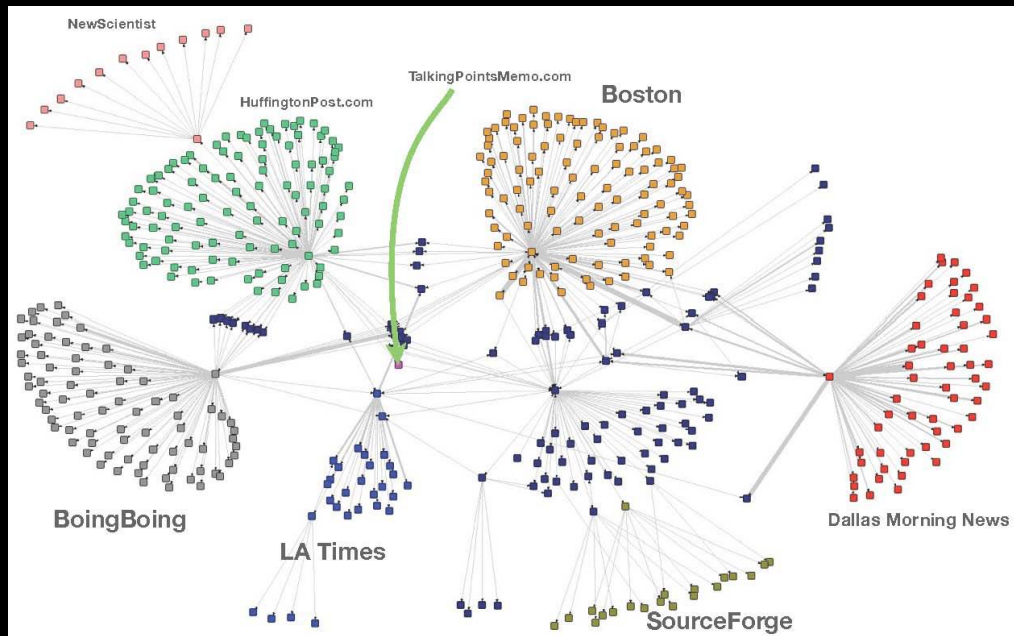
- Annotating web archives
- Selecting content to extract from archives
 - Efficiently extracting data from IA
- Analysing large bodies of time-series data
 - Involvement of domain experts (webometrics, SNA, e-Research, etc.)
 - Move from snapshots to more continuous data (within technical limitations)
 - Both outlinks and inlinks
- Sharing results in meaningful ways



1998



2002



2006

Source:
Matt Weber & Peter Monge,
University of Southern
California



oxford internet institute university of oxford

“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>



NATIONAL ENDOWMENT FOR THE HUMANITIES



Oxford e-Social Science Project

