

The State of the Art and the Future(s)  
of  
**Web Archiving**

Eric T. Meyer

with Arthur Thomas & Ralph Schroeder

Oxford Internet Institute 2011

# Learning from the Live Web

Blurring the distinction  
between  
the live web  
and the historical web

# Challenges

# The Web in a Changing World

Just a few big events so far in 2011...

Massive Floods in Queensland  
US Congresswoman Gabrielle Giffords  
Anti-Government Protests  
Anti-Government Protests  
Anti-Government Protests  
Earthquakes in Christchurch  
Fighting in Libya to remove Qaddafi  
Earthquakes, tsunamis, and nuclear accidents  
Royal Wedding of William and Kate  
Death of Osama bin Laden

## Royal wedding: Prince William and Kate Middleton marry



Prince William and Kate Middleton

the royal wedding

TOP STORIES

## US mourns Arizona shooting victims



The BBC says they were

Officials say at floods

10 January 2011 Last updated at 18:56

US President Barack Obama has led a nationwide tribute to the six people killed in a shooting in Arizona which left a congresswoman

# The Web in a Changing World

Immediate: Set up ways for researchers to trigger collection, increased frequency of crawls, etc.

Developing: Use triggers, like RSS, to find and archive changing content, particularly emerging fast-changing content

Long term: Develop algorithms that follow trends to trigger archiving

# Server Logs

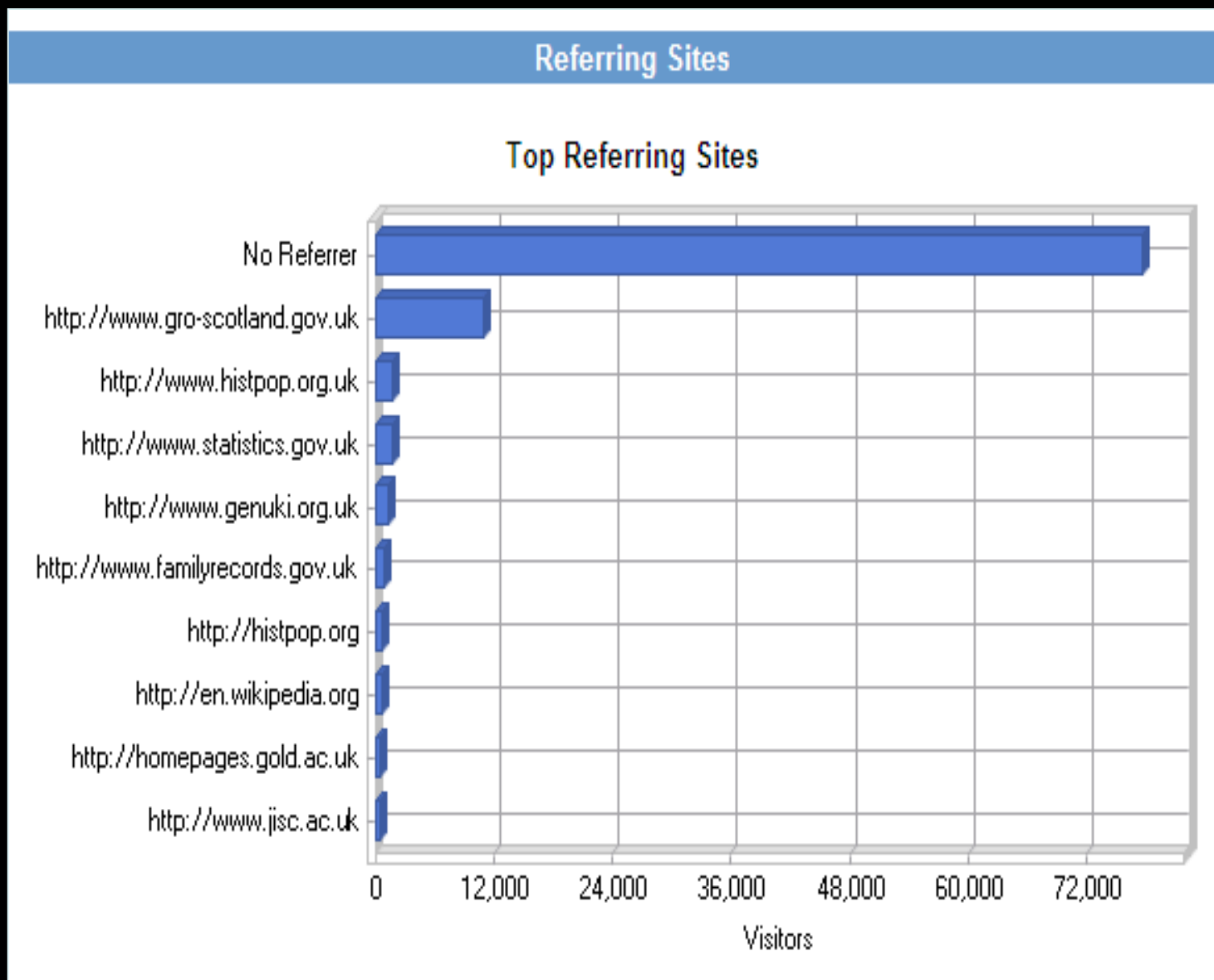


Figure 20. Sample log file data for the site histpop.org. Source: Meyer, et al. (2009)

# Server Logs

Immediate: Archive your own logs

Long term: Allow logs and analytics to be donated and associated with web archives

Ambitious: Archive *web traffic itself*

# The archive-in-a-box

There is a change from looking at past pages, to looking at collections, and collections of collections

- Shift towards data analysis, rather than reference
  - Data set, rather than reference collection
  - e-Research, data sharing, linked data, APIs

# Challenges to Discuss

We have just touched on a few...

National Webs and the International Web

The Context of Content

Places, Real and Virtual

Annotating the Archive

Saving the Links

Understanding the Structure of the Web

How Do Ideas Proliferate?

What about the Illicit Web?

A Digital Footprint: Remembering and Forgetting

Apps and APIs

Web of Data and the Internet of Things

# Reports

**<http://ssrn.com/abstract=1830025>**

*Web Archives: The Future(s)*

Meyer, E.T., Thomas, A., Schroeder, R. (2011). London: IIPC.

**<http://ssrn.com/abstract=1714997>**

*Researcher Engagement with Web Archives: State of the Art*

Dougherty, M., Meyer, E.T., Madsen, C., van den Heuvel, C., Thomas, A., Wyatt, S. (2010). London: JISC.

<http://ie-repository.jisc.ac.uk/544/>

**<http://ssrn.com/abstract=1715000>**

*Researcher Engagement with Web Archives: Challenges and Opportunities for Investment*

Thomas, A., Meyer, E.T., Dougherty, M., van den Heuvel, C., Madsen, C., Wyatt, S. (2010). London: JISC.

<http://ie-repository.jisc.ac.uk/543/>

# Recommended Reading

<http://www.timeshighereducation.co.uk/story.asp?sectioncode=26&storycode=417285&c=1>



The screenshot shows the top portion of a news article. At the top left is the logo for 'THE Times Higher Education', with 'THE' in large, colorful letters (red, purple, blue) and 'Times Higher Education' in smaller text below. To the right of the logo is a navigation bar with links: HOME | CONTACT | COMMENT | NEWS | CULTURE | RANKINGS. Below this is another bar with EXTRAS | ADVERTISE. The main headline is 'Memory failure detected' in a large, bold, black font. Below the headline is a social media sharing bar with a Facebook 'Like' button (showing 7 likes) and a Twitter button (showing 47 retweets). The date '1 September 2011' is displayed in red. The lead paragraph reads: 'A coalition of the willing is battling legal, logistical and technical obstacles to archive the riches of the mercurial World Wide Web for the benefit of future scholars. Zoë Corbyn reports'. The first sentence of the article body is visible: 'It is 2031 and a researcher wants to study what London's bloggers were saying about the riots taking place in their city in 2011. Many of the relevant websites have long since disappeared, so she turns'.

**THE**  
Times  
Higher  
Education

HOME | CONTACT | COMMENT | NEWS | CULTURE | RANKINGS

EXTRAS | ADVERTISE

## Memory failure detected

 Like    47

1 September 2011

**A coalition of the willing is battling legal, logistical and technical obstacles to archive the riches of the mercurial World Wide Web for the benefit of future scholars. Zoë Corbyn reports**

It is 2031 and a researcher wants to study what London's bloggers were saying about the riots taking place in their city in 2011. Many of the relevant websites have long since disappeared, so she turns



**Eric T. Meyer**

[eric.meyer@oii.ox.ac.uk](mailto:eric.meyer@oii.ox.ac.uk)

<http://people.oii.ox.ac.uk/meyer>

**Arthur Thomas**

[arthur.thomas@oii.ox.ac.uk](mailto:arthur.thomas@oii.ox.ac.uk)

**Ralph Schroeder**

[ralph.schroeder@oii.ox.ac.uk](mailto:ralph.schroeder@oii.ox.ac.uk)

<http://people.oii.ox.ac.uk/schroeder>

With funding from:

**JISC**



And additional support from:

