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

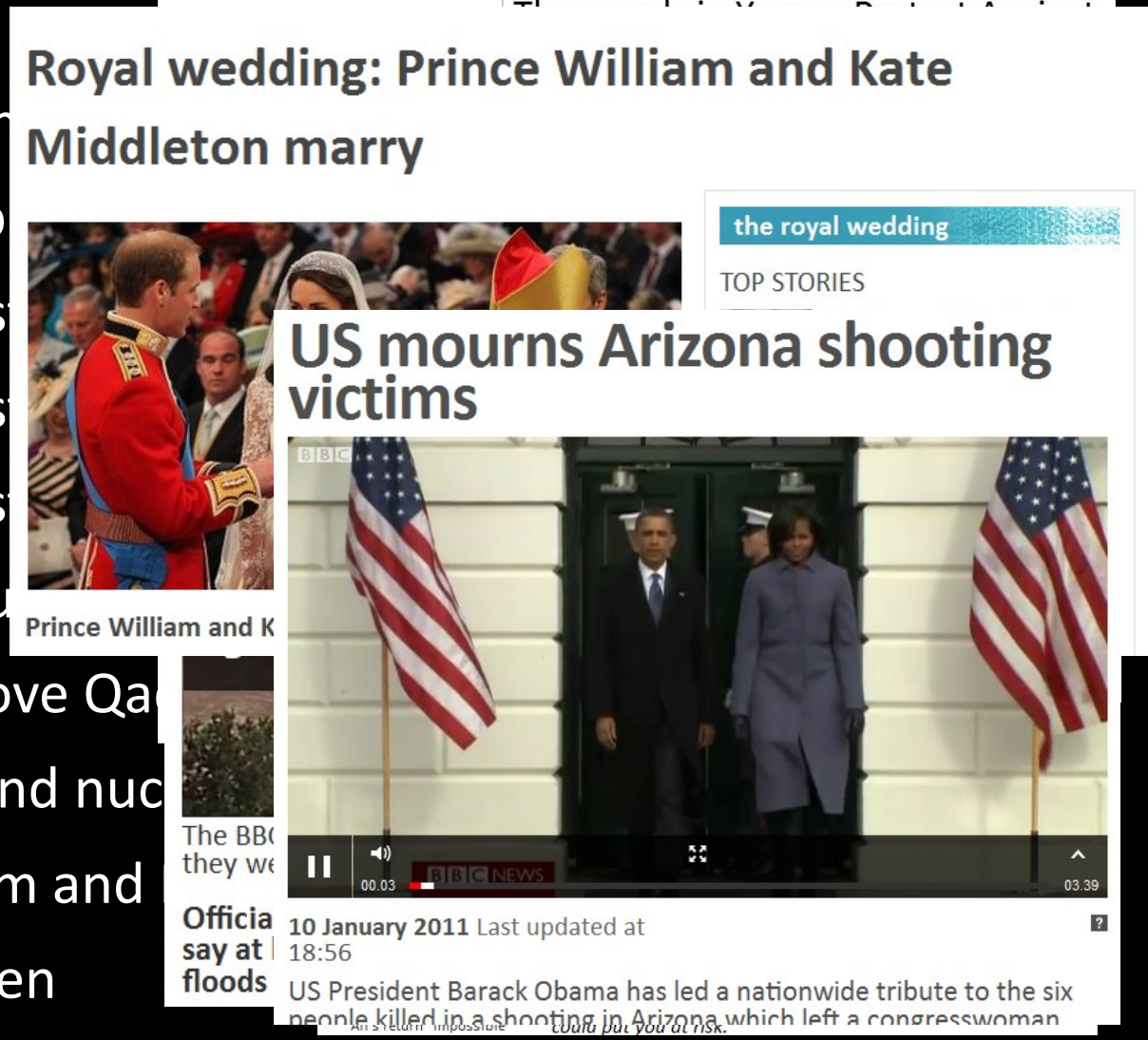
and the historical web

Challenges

The Web in a Changing World

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

- Massive Floods in Queen
- US Congresswoman Gab
- Anti-Government Protes
- Anti-Government Protes
- Anti-Government Protes
- Earthquakes in Christchu
- Fighting in Libya to remove Qa
- Earthquakes, tsunami, and nuc
- Royal Wedding of William and
- Death of Osama bin Laden



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

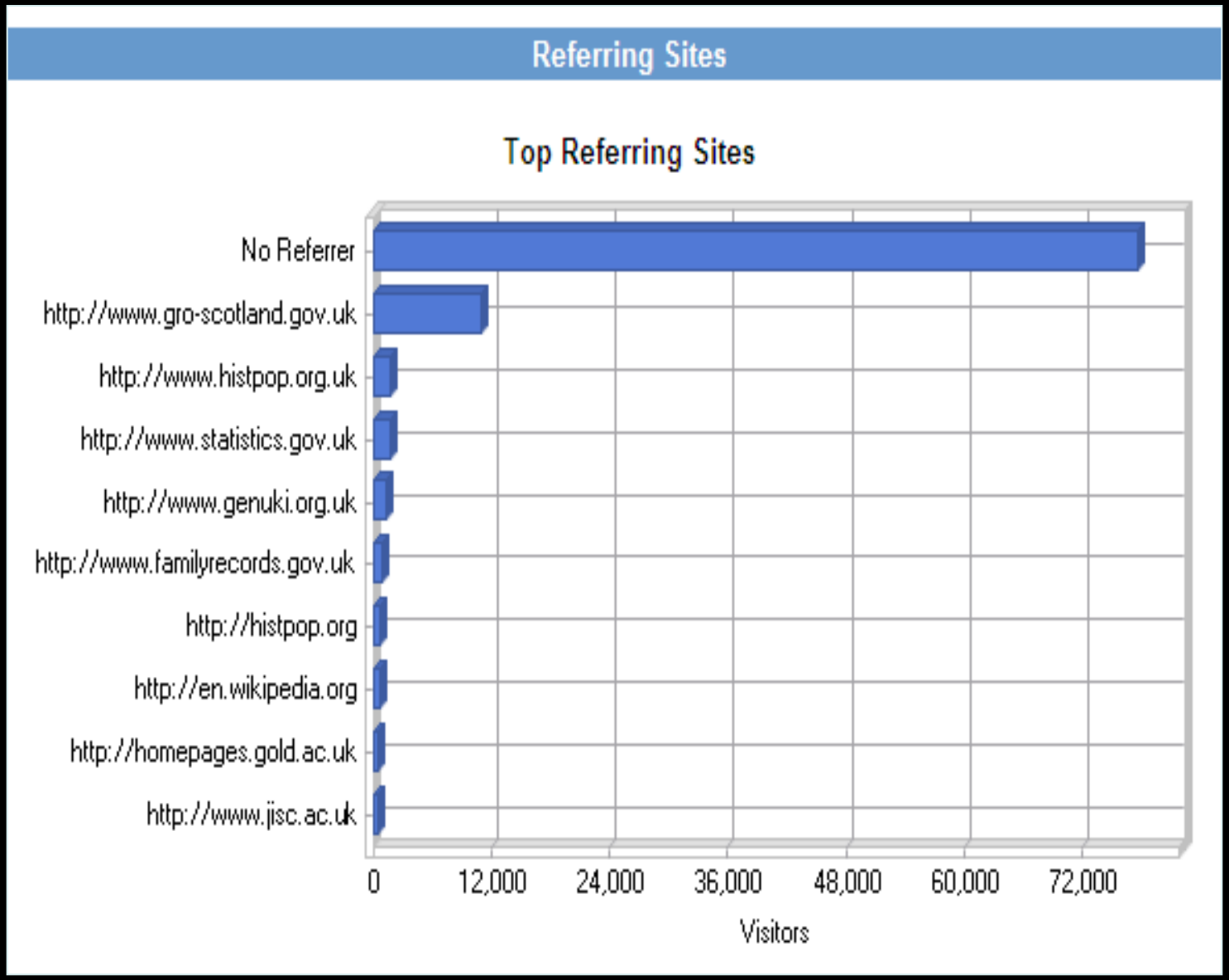


Figure 20. Sample log file data for the site histpop.org. Source: Meyer, et al. (2009)

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

oii oxford
oii internet institute
oii university of oxford

Eric T. Meyer
eric.meyer@oii.ox.ac.uk
http://people.oii.ox.ac.uk/meyer

Arthur Thomas
arthur.thomas@oii.ox.ac.uk

Ralph Schroeder
ralph.schroeder@oii.ox.ac.uk
http://people.oii.ox.ac.uk/schroeder

With funding from:

JISC

netpreserve.org
international internet preservation consortium

And additional support from:

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL