Workshop: Symmetrical Web Archiving With Webrecorder

Dragan Espenschied & Ilya Kreymer
Rhizome
235 Bowery
NY 10002, USA
+1-212-219-1288
{firstname.lastname}@rhizome.org

ABSTRACT

This paper describes a **workshop** for the novel, open source web archiving tool *Webrecorder*. Until now, web archiving has mainly been thought to be synonymous with “spidering” or “crawling,” meaning that a very basic, simulated version of a web browser travels paths of links and storing what it encounters, based on a certain set of rules.

Webrecorder introduces a new web archiving concept, *symmetrical archiving*, which makes use of actual browsers and actual user behavior to archive the web, as well. The software stack used for accessing or replaying archived material is exactly the same as during the capturing process. This allows for unprecedented fidelity in web archiving, , enabling the preservation of items embedded complex, dynamic web applications, while keeping their whole, interactive context as well as any user specific content.

This new approach to web archiving requires new ways of working within institutions; the proposed workshop serves as an introduction to symmetrical archiving, using Webrecorder’s emulation-based browsers, defining object boundaries, and transitioning from or augmenting crawler-based archives.

Keywords

web archiving; symmetrical archiving; emulation; collection management; appraisal; access; open source; institutions

1. THE NEED FOR SYMMETRICAL ARCHIVING

Current web archiving tools are based on the assumption that resources are published within a continuous, two-dimensional system, based on a **location**—the URL—and the **time** the resource was accessed.

The reality of the web has changed, as early as the introduction of the Netscape Navigator browser and Netscape Enterprise Server in 1994: The new server allowed for session-based, personalized content being served to users, client-side scripting in the form of Java applets and Javascript turned browsers into virtual machines that could execute complex behaviors. Although many of those innovations did not follow any rule book or standard, they have effectively made the web culturally and economically relevant, transforming it into the technically and culturally dominant medium it is today. At the same time, current web archiving practices and available tools do not sufficiently acknowledge the web being a complex software environment rather than a document retrieval system, making it impossible to create web archives that reflect for example current practices of cross-embedding, personalized services, web applications etc…

While many big data tools exist to analyze information from the web on a large scale, this is missing the way real users actually consume and create material on the web: the affect is only created with the relationships and contexts staying intact.

Taking on this challenge, *Symmetrical Archiving* assumes that the product of any web archiving activity is highly dependent on the actual activities carried out, and the technological context it is happening in. Resources that are only created in the moment when they’re accessed require archivists to be conscious about this activity.

Symmetrical Archiving means that for capture and access the same code is executed. Within the open source web archiving platform *Webrecorder*, looking at a live web site, an archived site, or recording a new site, is the same process: real interactions—carried out manually or via automation—are executed via a real browser on live and/or recorded material, with all data flowing through *Webrecorder* software, with a recording component optionally working.

This has implications for how object boundaries are defined, quality assurance is carried out and collections are structured.

2. EMULATED BROWSERS

A key component of Webrecorder is a growing library of real browsers—from current version of Firefox and Chrome to legacy versions of Internet Explorer, Safari, Netscape and Mosaic—that can be used to record and access web archives. This produces the highest possible web archiving fidelity: during recording, data being requested by for example Java, Adobe Flash, Shockwave or Real Media browser plug-ins can be captured; during playback, the same resources can be re-performed within exactly the same software environment they were recorded in, ensuring long term access to complex web materials.

The emulated browsers are presented to users as a hosted emulation service based on carefully pre-configured software environments that can be brought up in arbitrary number on-demand and are accessible via any modern web browser. No special configuration or complicated security precautions are required. The service is a new emulation framework designed specifically for web browsers, which also powers our previous effort, http://oldweb.today/.

3. TARGET AUDIENCE

The workshop is targeted at archiving professionals from all kinds of memory institutions that either are already engaged in web archiving or are planning to start a web archiving program. No prior knowledge of web archiving or certain kinds of tools are necessary.

Collection managers, curators and institutional infrastructure providers (IT) are welcome to join as well.

4. WORKSHOP PROGRAM

The following topic will be discussed in the workshop:

1. Introduction to symmetrical archiving: In-depth discussion of the concept and its consequences for web archiving practice
2. Introduction to Webrecorder, comparing it with existing web archiving tools
3. Creating a collection, from initial curation to publishing, including choosing the right emulated environment
4. Managing collections
5. Using Webrecorder as a service or deploying it within an institution
6. Advanced users: customizing Webrecorder, emulated environments, using Webrecorder components

The workshop can accommodate up to 18 participants and lasts 90 minutes. At the end of the workshop, users will create their own web archives, which they can choose to keep in Webrecorder hosted service or transfer to a different storage medium as best meets their needs.

5. ABOUT WEBRECORDER

Webrecorder is part of Rhizome's digital preservation program.

Rhizome's digital preservation program supports social memory for internet users and networked cultures through the creation of free and open source software tools that foster decentralized and vernacular archives, while ensuring the growth of and continuing public access to the Rhizome ArtBase, a collection of 2,000+ born-digital artworks.

Rhizome is a non-profit organization that commissions, exhibits, preserves, and creates critical discussion around art engaged with digital culture. It is located in New York and an affiliate to the New Museum.

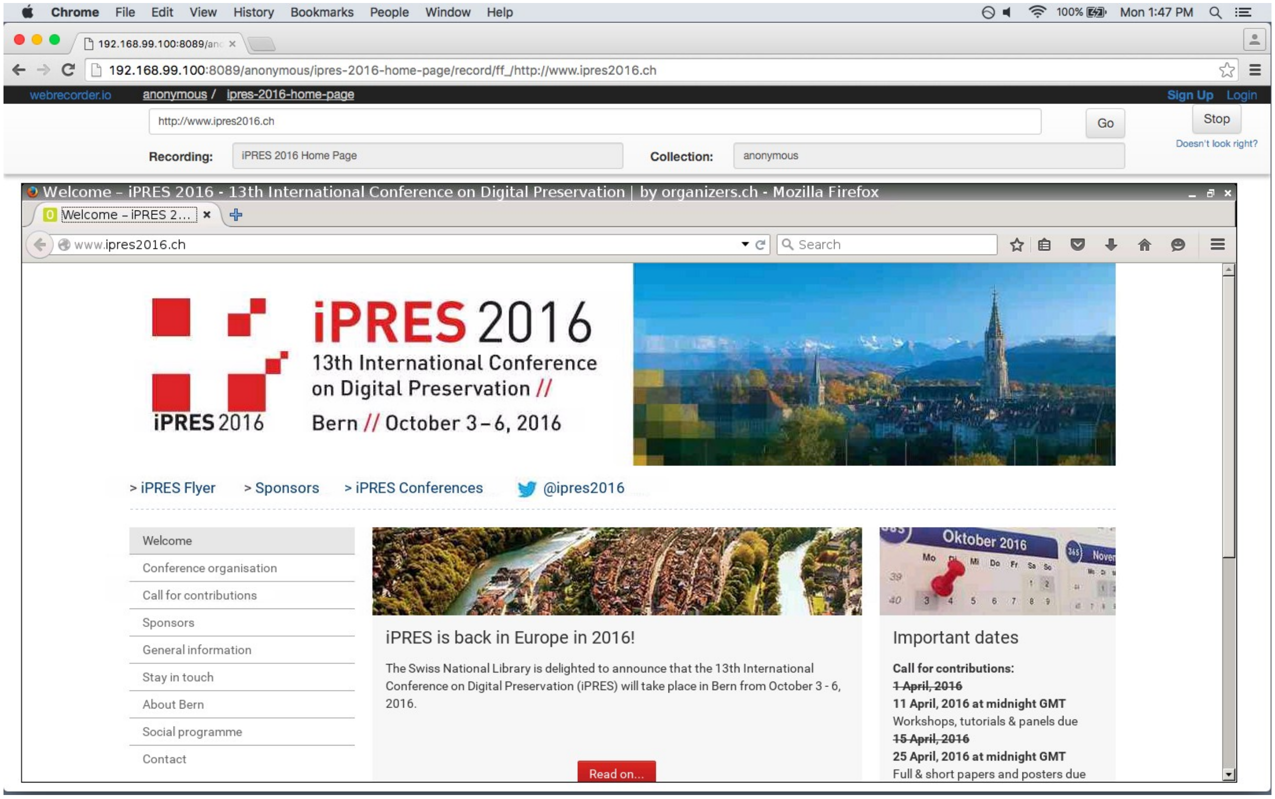
Webrecorder is free, open source software, available at http://github.com/webrecorder/ and as a hosted service at http://webrecorder.io.

Workshop personel:

Ilya Kreymer has previously worked as a software engineer for the Internet Archive and is now the lead developer of Webrecorder at Rhizome.

Dragan Espenschied is the head of Rhizome's Digital Preservation program.

The Webrecorder project receives significant support from the Andrew W. Mellon Foundation.



192.168.99.100:8089/anonymous/ipres-2016-home-page/record/ff_/http://www.ipres2016.ch

http://www.ipres2016.ch

Recording: iPRES 2016 Home Page

Collection: anonymous

Welcome – iPRES 2016 - 13th International Conference on Digital Preservation | by organizers.ch - Mozilla Firefox

www.ipres2016.ch

iPRES is back in Europe in 2016!

The Swiss National Library is delighted to announce that the 13th International Conference on Digital Preservation (iPRES) will take place in Bern from October 3 - 6, 2016.

Important dates

Call for contributions:
~~1 April, 2016~~
11 April, 2016 at midnight GMT
Workshops, tutorials & panels due
~~15 April, 2016~~
25 April, 2016 at midnight GMT
Full & short papers and posters due

Illustration 1: Webrecorder in action, recording the iPRES2016 site in an emulated Firefox browser usable within the user's native Chrome browser.