IIPC netpreserve.org INTERNATIONAL INTERNET PRESERVATION CONSORTIUM BnF ina iM INTERNET MEMORY FOUNDATION

General Assembly 2014: Open Workshops (22nd and 23rd May 2014)

Details of these workshops are given on the following pages.

THURSDAY 22nd MAY

14:00-17:00 – Crawl engineers and operators workshop
Moderators: Bert Wendland (Bibliothèque nationale de France), Kristinn Sigurðsson (National Library of Iceland) and Roger Coram (British Library)
Location: Belvédère (grey side)

14:00-17:00 – Curator tools fair
Moderators: Abbie Grotke (Library of Congress) and Nicholas Taylor (Stanford University Libraries)
Location: Belvédère (red side)

FRIDAY 23rd MAY

9:30-12:30 – Curating web archives: who cares for content?
Moderator: Louise Fauduet (Bibliothèque nationale de France)
Location: Salle des commissions 70

9:30-12:30 – Live archiving workshop
Moderator: Claude Mussou and Thomas Drugeon (Institut national de l'audiovisuel)
14:00-17:00 – Hyphe: crawling web archives
Moderator: Paul Girard (Medialab, SciencesPo)
Location: Mezzanine salle P
NB - The Live Archiving and Hyphe workshops are related, however participants may choose to attend only one or the other. Places are limited so we may not be able to accommodate all requests for these workshops.

9:30-17:00 – Open Wayback developers group
Moderator: Helen Hockx-Yu (British Library)
Location: Salle des commissions 4
This is a working meeting of the developers of the OpenWayback project. Developers interested in joining the effort are welcome to this meeting and should contact Helen Hockx-Yu (Helen.Hockx-Yu@bl.uk) and/or Kristinn Sigurðsson (kristinn@landsbokasafn.is).

14:00-17:00 - **NetarchiveSuite workshop (project participants only)**
Moderator: Sara Aubry (Bibliothèque nationale de France)
Location: Salle des commissions 70
This is a working meeting of the users of NetarchiveSuite. It is a time to review each institution's current projects and development priorities related to this software.

Crawl engineers and operators workshop
Thursday 24th May 2014, 14:00-17:00

This workshop is intended for crawl operators with at least some previous experience of running crawls who would like to explore ways of performing more complex crawls to deal with specific problems. The workshop will begin with case studies that focus on highlighting tools and solutions that are already available but may not be well known. Following this participants will be asked to present issues and difficulties they may have encountered and the group will discuss them to try to identify solutions. The workshop will be principally based on Heritrix, covering both version 1 and 3.

14:00 – Case studies

- Roger Coram (British Library)
- Kristinn Sigurðsson (National Library of Iceland)
- Bert Wendland (Bibliothèque nationale de France)

15:30 – Short break

15:45 – Interactive session

Curator tools fair
Thursday 24th May 2014, 14:00-17:00

The WIkipedia list of web archiving initiatives suggests that there is remarkable consistency in the predominant solutions that web archiving organizations are using to capture web content (Heritrix), store web content (WARC), and provide access to web content (Wayback). This apparent standardization belies the diversity of workflow systems and wrappers that are necessary to conduct web archiving at programmatic scale. What curator tools are IIPC member organizations using and to solve what workflow challenges?

Adoption of and improvements to shared curator tools benefit current and prospective users of those tools, increase the efficiency and effectiveness of our operations, allow us to involve a greater range of staff in our own organizations in web archiving, and make web archiving a more accessible activity for a greater number of organizations, thereby growing the community of practice and better assuring the preservation of the Web.

Web archivists organized into three panels -- end-to-end service providers, end-to-end local platforms, and specialized tools -- will each provide an overview of a curator tool used or provided by their organization. The discussion will be framed around the Web Archiving Life Cycle Model as a mechanism for facilitating a high-level comparison of what aspects of the web archiving workflow each tool does or does not address. The panels will be followed by open demonstrations of the curator tools, during which attendees can freely circulate around the room.

Session participants will gain a better understanding of how different curator tools handle similar (and diverse) workflow requirements, for cross-pollination in feature development or consideration of whether other platforms might address local challenges. Participants will also gain a better understanding of the workflow capabilities of each of the systems in a way specifically oriented toward a high-level comparison. Secondary expected outcomes for the session are improvement to the IIPC's tools and software documentation and the Web Archiving Life Cycle Model itself.

The session will also raise questions about the different curator roles and skills involved in web archiving, which will feed into the discussions in the workshop "Curating web archives: who cares for content?" on Friday morning.

Schedule

14:00 – Welcome/Introduction to Curator Tools Fair (Abbie Grotke, Nicholas Taylor)
14:10 – Panel: End-to-end service providers

- Archive-It (Lori Donovan)
- PageFreezer (John Jansen)
- CDL Web Archiving Service (Rosalie Lack)
- Archivethe.net (Chloé Martin, Leïla Medjkoune)

14:40 – Panel: End-to-end local platforms

- Web Curator Tool (Nicola Bingham)
- Netarchive Suite (Sabine Schostag, Mar Pérez)
- DigiBoard (Abbie Grotke)
- BCWeb (Géraline Camile)

15:05 – Panel: Specialized tools

- Annotation and Curation Tool (Peter Webster)
- UNT Nomination Tool (Mark Phillips)
- QA presentation (Brenda Reyes)

15:30 – Short break - Islandora WARC Solution Pack (Nick Ruest, video)

15:45 – Discussion on tools vs. lifecycle
16:15 – Demonstrations
17:00 – Ends

Curating web archives: who cares for content?
Friday 23rd May 2014, 9:30-12:30

DESCRIPTION: Organizations archiving the web have each set up their own structures for dealing with the technical challenges of crawling, preserving and accessing it; there have been multiple occasions through IIPC to exchange best practices on the technical aspects and plan tool developments. Readers and researchers have also been the subject of much scrutiny. This workshop is aimed at other practitioners who may have had fewer opportunities to discuss their trade: the people in charge of dealing with the content issues of web archiving.

More specifically, the goal of this workshop is to discuss who deals with the web archives as collections across the various institutions which crawl the web, and how. There are people involved in developing and using web archives who are neither technicians dealing with operations and progress in hardware and software, nor researchers building knowledge out of them; people who are the necessary bridge between these two, and are often a bit of both by trade. Who are these web curators, whether they are originally librarians, archivists, or IT specialists? What do they do that others don't — select sites, control quality, help users navigate content? What are their skills? Can one be trained to become a web archives curator?

Moreover, institutions have set up their web archiving teams differently: some mix IT and information science specialists in one specific task force; others rely on existing structures to find resources based on different areas of knowledge; still others use a combination of the two, depending on the importance of general and targeted crawling. This workshop will allow participants to compare their arrangements and how they fit their needs.

The French National Library's interest in holding such a workshop is twofold:
- The library has set up a network of "corresponding officers" to help breach the divide between its IT and support services, where the web archiving coordination is located, and its collection services, which curate content across different media. This framework is almost over 10 years old, and it seems like the time is right to reflect upon it in light of other organizations' choices and experience.
- Since 2010, the library has been testing a new way of tracking the influence of its increasing digital collections on the way its librarians do their job, ORHION (observatoire des Organisations et Ressources Humaines sous

l'Impact Opérationnel du Numérique: Organization and Human Resources under Digital Influence). ORHION is an informal “observatory” of the changes in the librarians’ practices and skills. Its members, working in the field of digitization, digital preservation, or acquisition of digital material, are eager to compare and contrast the changes in tasks and methods brought by web curating activities with lessons from other organizations.

Participants are welcome to present the make-up of their own organization, how it came to be, and how it aids their goals in web archives curation.

BENEFITS: At the end of the workshop, participants will have leads to refine the manner in which they curate web collections, by comparing the ways they organize their teams and the skill sets they look for in web content managers, to other web archiving teams or to teams in charge of other kinds of digital material.

The discussion will reveal clues about the evolution of the missions and skills of the librarian, archivist or curator in the near future, to deal with such challenges as describing, preserving and disseminating digital collections with the same level of service to which users are accustomed, and in a way that befits the particular characteristics of these materials.

To know more about ORHION and its activities, please have a look at:
- Watching the library change, making the library change? An observatory of digital influence on organizations and skills at the Bibliothèque nationale de France, http://conference.ifla.org/past/2012/150-clatin-en.pdf
- The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France, ijdc.net/index.php/ijdc/article/download/175/244

Schedule

9:30 – Introduction (Louise Fauduet, Bibliothèque nationale de France)

9:45 - Sub-session 1: Who are the content curators? What are their collections? What do they do? Web archives in the day to day work of content curators.

- Nicholas Taylor (Stanford)
- Michael Neubert (Library of Congress)
- Discussion

10:35 – Break

10:50 – Sub-session 2: Who are the content curators ? Where do they work and with whom? Web curation within institutions. Dealing with researchers.

- Barbara Signori (National Library of Switzerland)
- Mark Phillips (University of North Texas)
- Discussion

11:40 – Sub-session 3: Who are the content curators? How did they get there? How do they learn their trade? Training and information of curators.

- Annick Le Follic (BnF)
- Discussion

12:20 – Conclusions
12:30 – Ends

Live Archiving Workshop
Friday 23rd May 2014, 9:30-12:30

The live archiving workshop (LAW) will explore live archiving benefits and methods by demonstrating the Live Archiving Proxy (LAP) tool built at Ina for the IIPC community.

The ambition of the workshop is to cover a large spectrum of live archiving techniques, ranging from a simple manual web browser for specific archiving of rich interaction web contents to fully customized specialized bots (rapid prototyping, automated web toolkits, crawler integration, QA, etc.). We shall also try exploring other paths by sharing experiences and ideas among participants, possibly guiding future developments.

This workshop is targeted for web curators, technicians, crawler designers and researchers and will be jointly animated by Ina crawl engineers and curators and the crawl team from the Paris Sciences-Po Medialab. The archive built during the workshop as a result of participant's experiments will be made available in both WARC and DAFF formats to IIPC members and workshops attendees.

The archive will also be used in the Hyphe workshop in the afternoon.

Possible work groups include:

- Manual crawling of interactive or complex content such as web documentaries
- Deploying and administring live archiving tools as a centralized hub for crawling
- Developing specialized crawling tools based on the PhantomJS framework
- Using Hyphe as a discovery and archiving tool

Hyphe: crawling web archives
Friday 23rd May 2014, 14:00-17:00

The workshop will present Hyphe[1], a research-driven crawler designed for Humanities researchers or librarians. It will briefly present the core functions and methodological issues of the tool (Which documents is a web corpus based on? How to select and collect these documents from the web?) and propose to try to use it on existing web archives, including those created during the Live Archiving Workshop earlier in the day.

1 - Which documents is a Web corpus based on?
As a source of digital traces, the Web creates opportunities that we do not want to miss. However we have to understand its structure as a network of documents. It is both a hierarchy of linked resources (HTML content) and a network of locations (HTTP/URLs).

What is the right granularity of documents?

Researchers do not want to work at the level of web pages. We see the Web as an information system where pages are documents. When we use the Web as a field of studies, we want to aggregate groups of pages to reflect an actor’s presence on the Web. We typically think in terms of websites.

Each researcher should be able to set his own way of aggregating web pages. In a given study, we may want to consider Sciences Po pages as a whole, because we consider the university as an institution. While in another study, we may want to break the institution into multiple entities (research centers, library, administrative services...) because we consider it as a network. For these reasons, we propose the notion of « web entity ».

2 - How to sieve the web?
Building a web corpus involves defining web entities as well as discovering new resources. Researchers need a method to identify relevant contents depending on their research topics. We propose a method to build a

[1] See a demo at http://jiminy.medialab.sciencespo.fr/hyphedemo/ and the source code at https://github.com/medialab/HypertextCorpusInitiative/

corpus following the medium's principles (its hyperlink structure), and compatible with the methodological constraints of humanities.

Research driven crawling[2]

We define the research driven crawling as a strategy where the researcher makes every important decision. That is why we only use a "distance0" crawler: it only follows links within a web entity.

The crawler automatically harvests the pages of a given web entity, but does not explore those new to the corpus. It only provides a list of discovered entities, which the researcher alone can decide to extend the corpus with. A prospection feature proposes the most linked web entities around the corpus as candidates for extending it.

3 - Crawling the archives

Hyphe is made to give researchers the possibility to build consistent web corpus. What works on the live web might be useful also to help researchers using the archived web. Using the proxy developed by the archiving institutions for users to browse the archive, we can use our crawler to build corpus directly on the archive. We will have to find a way to set/handle the date/version of the archive we want to crawl from the tool. We will then create a corpus on the archive and compare different version of the same corpus at different time on a topological point of view.

This workshop will also be the occasion for Hyphe development team to imagine what should be done in the future to integrate different versions of a same corpus in the tool.

[2] We share the paternity of this expression with Julien Masanès from the Internet Memory Foundation who proposed us this expression when we presented our intentions in the first user workshop of the LAWA project in Paris http://www.lawaproject.eu/index.php/news/1st_lawa_user_workshop