Panel: Preserving Web Archives: One Size Fits All?

Abstract

Panel from members of the Preservation Working Group of the IIPC (International Internet Preservation Consortium): Libor Coufal, Andrea Goethals, Gina Jones, Clément Oury and David Pearson, moderated by Pamela Armstrong
The IIPC is made up of about forty institutions that collect web content for heritage purpose. Its Preservation Working Group brings together international experts who work on policies, practices and resources in support of preserving the content and accessibility of web archives.
IIPC institutions have globally similar goals and generally share the same harvesting and access tools. There are however differences between us: difference of status (libraries or archives, national or local…), legal mandate, collection size… Do we even have the same understandings of what we are doing?
The questions which the panel will address are: is there a best, a unique approach to preserve web archives? How institutional differences can affect our preservation planning decisions? And where can international collaboration help us preserving our archived collections?"

Statements

Andrea Goethals, Harvard University Library – The goal of web archiving is to acquire web content for preservation - how we do this is just a means to an end. Just as the Web itself continues to morph in large ways, so will our methods of capturing it change over time. While we tend to capture by crawlers today, we should consider any and all techniques that allow us to capture the Web content we want to preserve. We should always try to preserve documentation about how content was acquired.

Libor Coufal, The National Library of the Czech Republic – The National Library of the Czech Republic is trying to archive the „Czech“ web but in fact, we are only able to archive frozen snapshots of it and even that far from being complete or perfect. Ideally, we would like to preserve the current web-like look and feel of the archive as much as possible but we are aware that a more pragmatic approach might turn out to be necessary in the future. Current preservation strategies and tools are not mature yet, certainly not ready for web archives needs. The best we can do at the moment is to make sure the stuff survives for the near future and keep optimistic about the humankind finding a solution to preserving its digital heritage.

Clément Oury, French national Library – According to the philosophy of the legal deposit, we need to collect and give access to the collections in the form in which they were presented to the public. On the other hand, we cannot pretend to give access to the “web”, which is a living space, but only to web archives, i.e. frozen artefacts produced by our crawl engine. So we need to preserve all information about the activity of the robot – the “producer” in the archivist sense: from this perspective, it is as important to take care of configuration or log files as of web content, in order to let future researchers and users be able to understand how the collection was built".

Gina Jones, Library of Congress – The Library of Congress's web archives consist of any and all content on the web that the selecting librarians consider of value and worth preserving. The ephemera range from websites to social media sites and beyond, including such sites as twitter and facebook. The fact that web content is selected for the LC web archives begins our preservation process. As content is collected, we replicate it to multiple locations and do bit preservation. Further preservation steps may be taken, but not on the whole of the archive. We're hoping that the preservation fairy will wave her magic wand so we never have to migrate or provide an emulator to this content-soon to be 200 terabytes.

David Pearson, National Library of Australia – The NLA currently does not have a preferred option on which preservation strategies we will use to preserve our Web Archives. But we do know that whatever it is, we will need to use tools that currently are not up to the task. All of the current preservation solutions assume that we can identify how the content is encapsulated and that we have plans in place to do something about it. The NLA also takes the view that “methods” and “solutions” are meaningless without the context of what needs to be achieved”.

Contact details

Pam Armstrong
Gestionnaire / Manager
Bureau des Normes et des services de dépôt numérique / Digital Repository Services and Standards Office
Bibliothèque et Archives Canada / Library and Archives Canada
550, boul. de la Cité, Gatineau, QC
Canada K1A 0N4
Téléphone / Telephone 819-953-7118
Télécopieur / Facsimile 819- 934-4422
Gouvernement du Canada / Government of Canada
www.collectionscanada.gc.ca
pam.armstrong@lac-bac.gc.ca

Libor Coufal
WebArchiv project manager
National Library of the Czech Republic
Klementinum 190, Prague 1, the Czech Republic
T: ?221663-256
F: ?221663-301
libor.coufal@nkp.cz

Andrea Goethals
Digital Preservation and Repository Services Manager
Harvard University Library Office for Information Systems
90 Mt. Auburn Street
Cambridge, MA 02138
phone: (617) 495-3724
andrea_goethals@harvard.edu

Gina Jones
Digital Media Project Coordinator
Web Archiving Team
Office of Strategic Initiatives
Library of Congress
gjon@loc.gov

Clément Oury
Web archive preservation manager
Legal Deposit Department
Bibliothèque nationale de France (French National Library)
Quai François-Mauriac
75706 Paris Cedex 13
Phone 33 (0)1 53 79 46 93
clement.oury@bnf.fr

David Pearson
Acting Director
Web Archiving & Digital Preservation
National Library of Australia
(02) 6262 1570
dapearso@nla.gov.au