Workshop on Web Archiving

MODULE 2: EXISTING WEB ARCHIVES

Niels Brügger
Asger Harlung

Workshop AU
16.08.2018

netlab.dk

Module 2: Existing Web Collections

A short introduction to existing web archives

- The Danish Netarkivet
- Internet Archive
- Library of Congress
- Other web archives

netarkivet.dk indsamler og bevarer den danske del af internettet

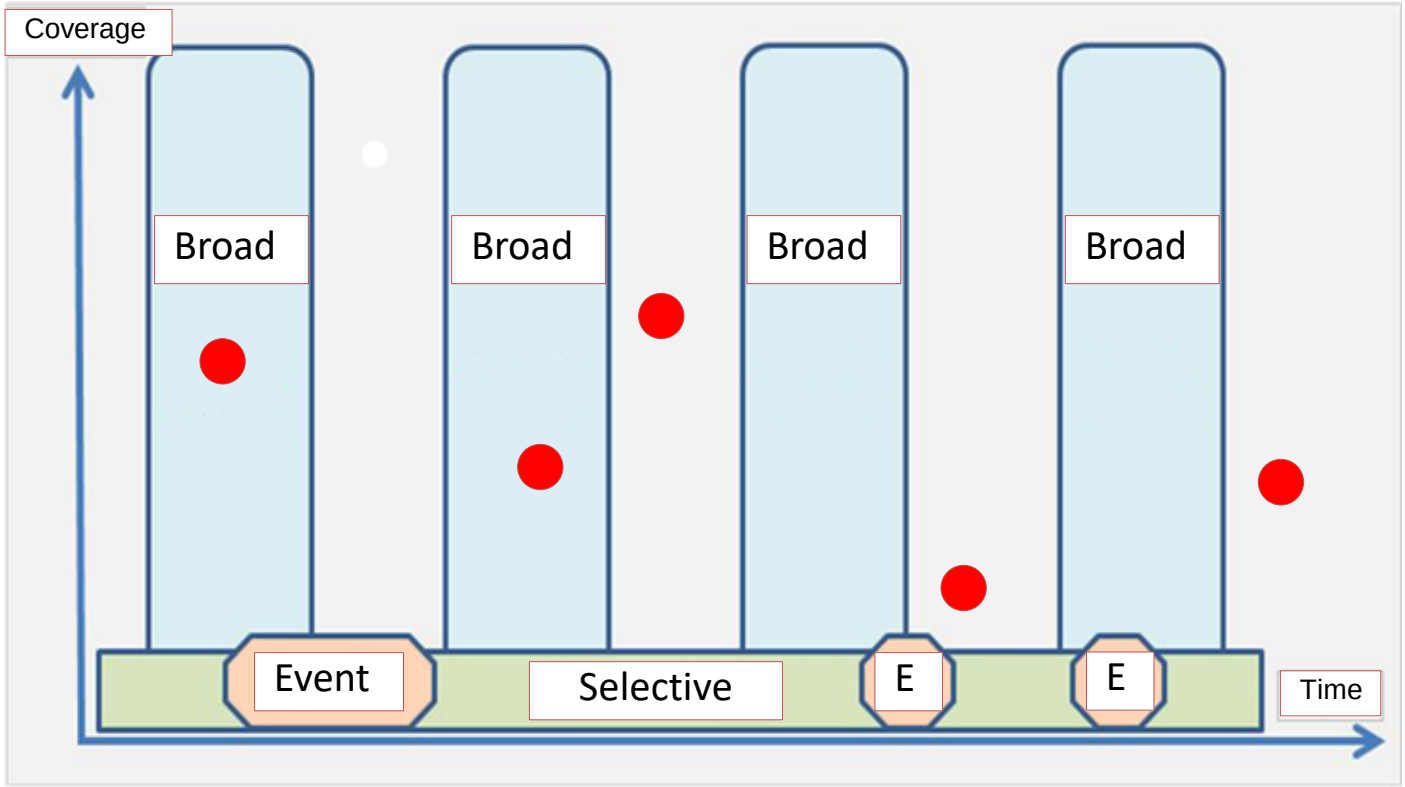
- Netarkivet is run by the Royal Library (national Library of Denmark, Copenhagen).
- The Danish part of the Internet is defined as cultural heritage in the Legal Deposit Act (Act no. 1439 of 22.12.2004), effective from June 1st, 2005
- The ”Danish part of the Internet” = all Internet content in Danish or meant for Danes → the top level domain .dk and danica (e.g. sites in Danish or addressing Danes on other domains such as .com, .eu, .nu, etc.)
- .dk domain names: 607.000 in July 2005, 960.000 in January 2013
- Dead .dk domains from July 2005 to January 2013: 741.838
- 2011: Roughly 222 TB; 6 m objects, most common file types are html, jpeg, gif and png
- 2013: Most common file types are html, jpeg, pdf and mp4 (video)
- 2014: On July 27 the data in Netarkivet amounted to 501 TB
- 2015: On November 15 the data comprised 654 TB

Netarkivet

2005 →

Strategies:

- Broad/bulk
- Selective
- Event
- Special



From http://netarkivet.dk/om-netarkivet

Netarkivet

Important because:

The largest and most comprehensive collection of the .dk domain, and sites relevant to Denmark and Danish culture (”danica”).

Legal restrictions on access and use.

Access and terms of use: http://netarkivet.dk/adgang/

Url: https://netarkiv-wayback.kb.dk/vpn/index.html

Nyt interface på vej: http://193.6.201.202/solrwayback/

The Internet Archive:

- american non-profit
- from 1996
- not based on national legislation
- in general based on cumulative archiving, following hyperlinks from what was already archived
- the worlds largest collection of archived web
- more than 491 billion web pages, collects app. 1 billion pages per week
- quality is erratic — often only top level(s)
- heterogenious collection, no overall strategy, including donations…

NetLab

The Internet Archive

Important because:

The world's first and largest archive, accessible to all.

Content and depth varies, and pages are hidden upon request from site owners, but the amount of content remains larger than any other archive.

Url: https://archive.org/

Library of Congress web archive:

- from 2000
- curated, topic based and selective collections
- harvested by the Internet Archive (not Archive-It)
- 763TB

LIBRARY OF CONGRESS

Sort By Title Go View Gallery Go

Brazilian Presidential Election 2010 Web Archive

Collection Period: September 2010 to January 2011. The Brazilian elections were held October 3, 2010. The president, Luiz Inácio Lula da Silva, was not allowed by law to ...

View 48 Items

Burma/Myanmar General Election 2010 Web Archive

Collection Period: May 2010 to January 2011. The November 7, 2010 Burma/Myanmar elections provided for under the 2008 constitution were watched by the world as critical to a ...

View 34 Items

Crisis in Darfur 2006 Web Archive

Collection Period: February 20, 2006 to December 5, 2006. Described by the United Nations as the world's worst humanitarian crisis, the current conflict in Sudan's western region of ...

View 220 Items

Egypt 2008 Web Archive

Collection Period: April 12, 2008 to May 31, 2008. In 2008 Egypt witnessed a remarkable experience rich in political and democratic practices, and the sites captured back then ...

View 12 Items

Indian General Election 2009 Web Archive

Collection Period: April 9, 2009 to June 29, 2009. This collection of websites documents the 2009 Indian General Elections. As the world's most populated democratic nation with the largest ...

View 59 Items

Indonesian General Election 2009 Web Archive

Collection Period: March 10, 2009 to October 21, 2009. This collection of websites documents the Indonesian general elections held in 2009. The collection covers three separate rounds of elections ...

View 80 Items

Iraq War 2003 Web Archive

Collection Period: March 13, 2003 to June 16, 2003 and December 31, 2004 to July 21, 2009. On March 20, 2003, the United States began military action against ...

View 232 Items

Laotian General Election 2011 Web Archive

Collection Period: June 2011 to November 2011. Laos, one of the world's few remaining communist regimes, holds general elections for the National Assembly every five years. Parliamentary Election ...

View 17 Items

Library of Congress Web Archive

Important because:

Open for all, many ready examples of well curated collections.

Url: https://www.loc.gov/websites/collections/

Other Web Archives

IIPC Member Archives

http://netpreserve.org/about-us/members/

List of Web archiving initiatives,

https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

Truman, G. (2016). WebArchiving Environmental Scan. Harvard Library Report.http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314