Digital Humanities: Birds of a Feather session

eResearch NZ 2019

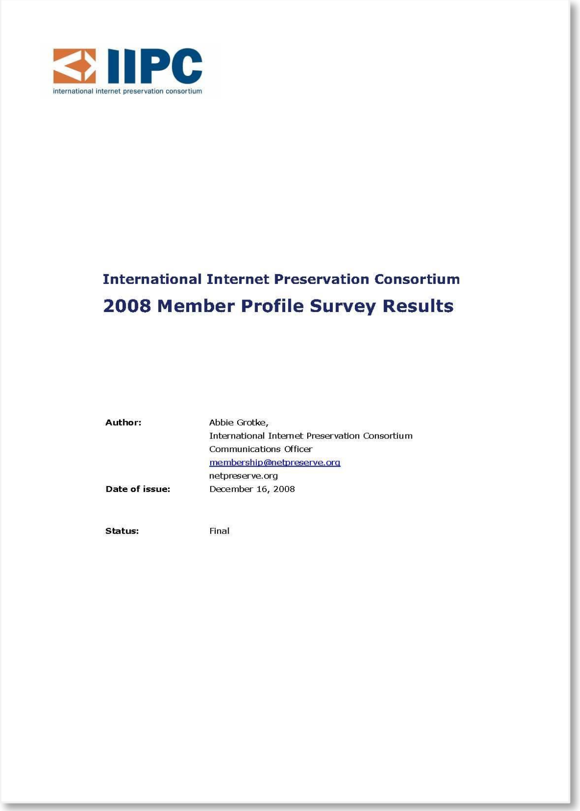
Auckland 18-20 February 2019

Steve Knight, Programme Director, Preservation Research & Consultancy, National Library of New Zealand
steve.knight@dia.govt.nz

Today – 3.30 – 4.30

3.30	Introduction and welcome
3.35	What w ewant to have a look at today
3.40	A whip round the room
3.45	Three brief commentaries
4.00	Group exercise
4.20	Wrap-up and next steps

Landscape of web archiving activities



IIPC

international internet preservation consortium

International Internet Preservation Consortium

2008 Member Profile Survey Results

Author: Abbie Grotke,
International Internet Preservation Consortium
Communications Officer
membership@netpreserve.org
netpreserve.org

Date of issue: December 16, 2008

Status: Final

IIPC Survey, 2008
IIPC Survey, 2018

Web Archiving in Europe

INTERNET MEMORY FOUNDATION

INTERNET MEMORY SURVEY

- Sent in December 2010
- 385 questionnaires sent to 364 independent institutions
- 37 European countries
- Translated in 5 languages (English, German, French, Italian and Spanish)

SOME SURVEYS ON WEB ARCHIVING

- February 2011, Arquivo da Web Portuguesa/Portuguese Web Archive (to be published): http://www.arquivo.pt
- December 2010, University of North Texas: http://research.library.unt.edu/[illegible]/images/2/29/[illegible]_survey_report_[illegible]_14dec2010.pdf
- April 2010, National Diet Library: www.ndl.go.jp/en/cdnlao/meetings/pdf/report_[illegible]_doc3.pdf
- December 2008, International Internet Preservation Consortium (IIPC): http://netpreserve.org/publications/reports.php?id=005
- July, 2007, Koninklijke Bibliotheek/ National Library of the Netherlands: http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/documenten/[illegible]_UserSurvey_Webarchive_EN.pdf

A survey provided by the Internet Memory Foundation, 2010

Context

In the framework of the European Research Project Living Web Archives -LiWA-, a three-year project funded by the European Commission through the Seventh Research Framework Programme under Project No 216267), the Internet Memory Foundation carried out a survey on Web archiving among European and International institutions. The aim of this survey was to draw up an inventory of Web archiving institutions and to have a clearer understanding of problems encountered in the field of Internet archiving.

Key figures of results

Survey Participants

A representative panel of Web archiving in Europe based on 74 answers.
Mostly composed by Audiovisual archives (25,7%), National libraries (23%) and Institutional archives (12,2%).

Maturity of Web archiving program

30%
19%
5%
23%
7%
1%
7%
5%
● Yes, fully operational
● Yes, operational but still experimenting
● Yes, just starting a Web Archiving project
● No, but we plan to do it
● No, no funding
● No, not in our mandate
● No, an other institution is already in charge
● answered question

41 institutions have a Web archiving program: fully operational, experimenting, and starting. And 17 institutions, which do not have a current Web archiving program, plan to have one in the next three years.

Legal context

50% do Web archiving protected by a law enacted or passed

Selection policy

Most of the institutions operate selective crawls (71%) and 30,8% focus on their own websites. Only 23% of this panel run domain crawls.

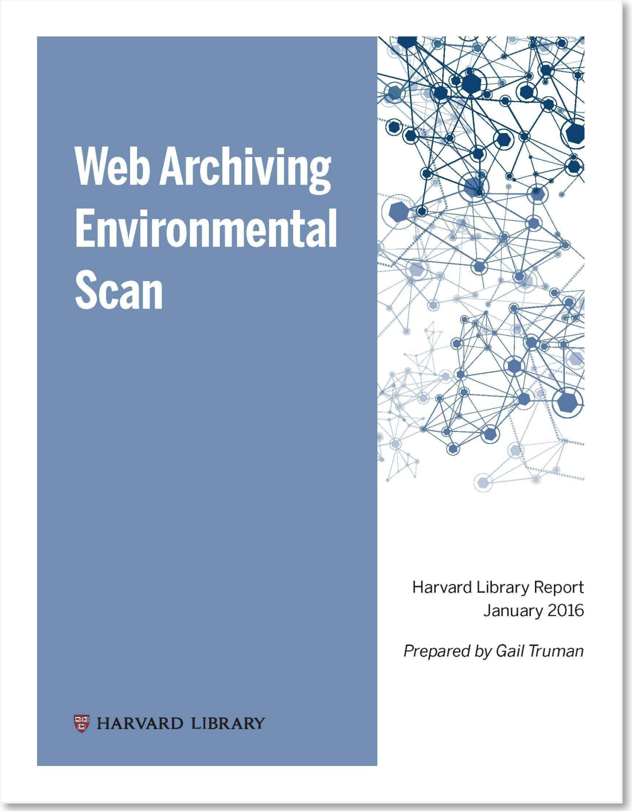
Management

To fit with their Web archiving project, this panel develops their own strategy and skills. 64% operate their Web archive in house versus 36% which outsource the service.
In majority they use Heritrix (80%) and manage their collection with a curator tool with an in house solution for 35% of them. 35% use open source solution as NAS, WCT or Pandora...

Access

User can browse Web archive by URL (68%), by searching by keyword (70%) and through thematic collection (65%).

Internet Memory Foundation, 2010



Web Archiving Environmental Scan

Harvard Library Report
January 2016

Prepared by Gail Truman

HARVARD LIBRARY

Harvard Library, 2016

National Digital Stewardship Alliance
Web Archiving Surveys
2011, 2013, 2016, 2018

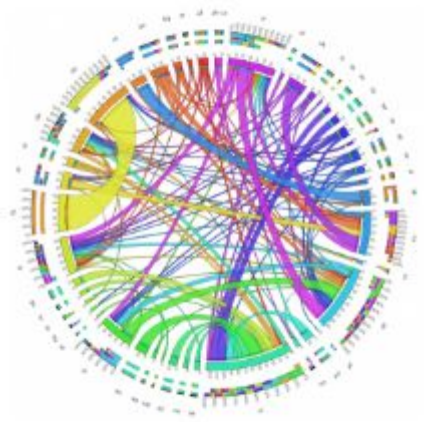
Born Digital Legal Deposit Policies and Practices, 2017

promise

Preserving Online Multiple Information: Towards a Belgian Strategy, 2018

Case studies

LINK ANALYSIS



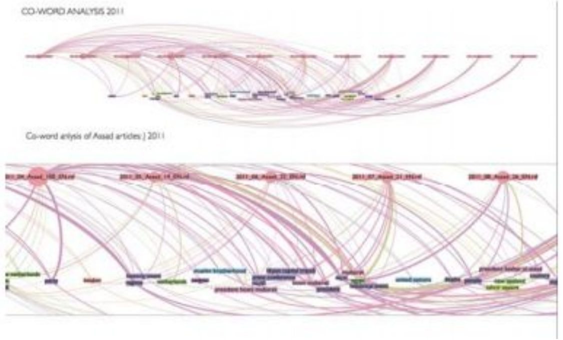
Hannes Mühleisen

Visualization of linking between websites of different languages,

Babel 2012 Web Language Connections

https://github.com/norvigaward/2012-naward25/wiki/Babel-2012---Web-Language-Connections

LINK ANALYSIS



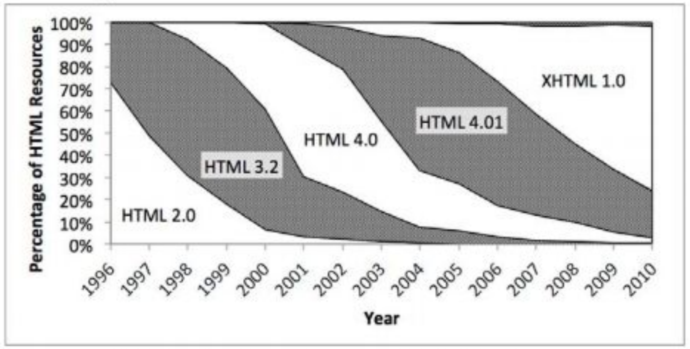
N-gram Search, UK Web Archive (British Library)

Searching the (News) Archives, Web Archive Retrieval Tools (University of Amsterdam)

Sentiment Analysis and the Reception of the Liverpool Poets, Helen Taylor (Royal Holloway)

Footprint of the world top companies, Alexandre Marah and Ferenc Szabó (University of Tewnte)

ANALYSIS OF TECHNOLOGY TRENDS



HTML Version Usage Over Time, UK Web Archive

Format Profile, UK Web Archive (British Library)

Web Data Commons, Christian Christian Bizer et al. (University of Mannheim & Karlsruhe Institute of Technology) http://webdatacommons.org/

An analysis of the use of JavaScript libraries on the web, Dennis Pallett et al. (University of Twente)

Emily Reynolds. *Web Archiving Use Cases*, 2013: netpreserve.org/use-cases

Researching web archives

SURVEYS, REPORTS, PRESENTATIONS (selected)**:**

- Maria-Dorina Costea. **Report on the scholarly use of web archives**. NetLab. 2018.
- Emily Maemura. **Explorations of Netarkivet: Preliminary Findings**. 2018.
- Janne Nielsen, **Using web archives in research**, NetLab 2017.
- Web Archives as Scholarly Sources: Issues, Practices and Perspectives.
- Harriet Riley & Mark Crookston. **Awareness and use of the New Zealand Web Archive**. 2015.
- Meghan Dougherty, Eric T. Meyer, Christine McCarthy Madsen, Charles van den Heuvel, Arthur Thomas, and Sally Wyatt. **Researcher Engagement with Web Archives: State of the Art**. JISC Report, 2010.
- Miguel Costa & Mario J Silva**.** **Understanding the Information Needs of Web Archive Users**. 2010.
- Jane Winters. **Big UK Domain Data for the Arts and Humanities**. IIPC WAC 2015.
- Helen Hockx-Yu. **Up close and personal - Researchers and the UK Web Archive Project**. IIPC WAC 2011.
- Jefferson Bailey. **Program Models for Research Services**. IIPC WAC 2016.

PROJECTS, NETWORKS, SERVICES, CASE STUDIES (examples):

- **ARS: Archive-It Research Services**
- **AU**: **Archives Unleashed Project**
- **RESAW.eu**: **Research Infrastructure for the Study of Archived Web Materials**, 2012-
- **BnF**: **Néonaute** 2018
- **ATI & UKWA**: Detecting semantic shift in large corpora, 2018
- **Arquivo.pt**: **Investiga XXI**, 2017 & **Awards**, 2018
- Niels Brügger and Ralph Schroeder, eds., **The Web as History: Using Web Archives to Understand the Past and the Present.** London: UCL Press. 2017.
- **BUDDAH**: **Big UK Domain Data for the Arts and Humanities**, 2013-4 □ **SHINE**
- **WebART**: **Web Archive Retrieval Tools, 2012-2016**

Program Models for Research Services, 2016 – lessons learned

- Researchers don’t always know what they want
- Researchers default to wanting access to raw/all data
- Researchers will have varying levels of technical resources or support
- Address upfront issues of technical proficiency, nonarchive technical support and/or methodological stuff
- Will require reference/resources to explain and contextualize web archive tools and processes
- More data doesn’t equal better analysis

- Focus on derivation, portability, and access
- Focus on scalable partnerships & decentralization
- Research support expectations often != with available resources or services
- Research methodologies (conceptual, practical, technical) often != with data, collecting, tools
- Service models or death (though yet to emerge for most data-driven LAM-ish research)

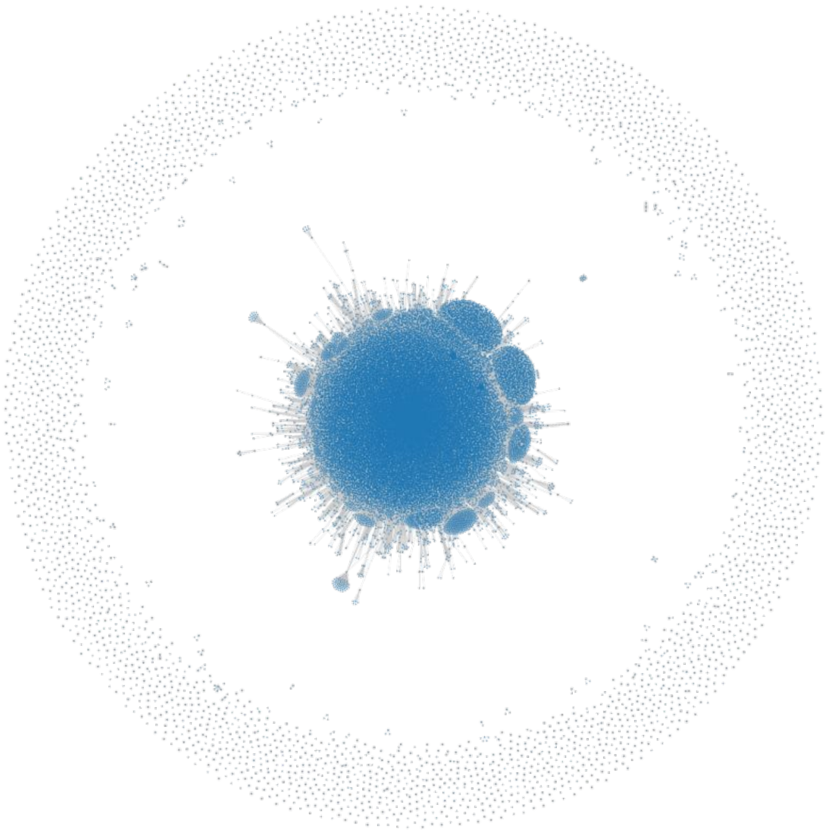
Jefferson Bailey. **Program Models for Research Services**. IIPC WAC 2016.

Use cases: Buddah

Big UK Domain Data for the Arts and Humanities

Institute of Historical Research, UoL
Oxford Internet Institute
Aarhus University
British Library

Topics
disability
Euroscepticism
public archaeology
Ministry of Defence
companies, corporations
reception of Beat literature
French expatriates in London
poetry community & networks



Rainer Simon: Visualisation of links between websites from the UK crawled during 1996,

Project team

Jane Winters
Helen Hockx-Yu
Eric Meyer
Ralph Schroeder
Niels Brügger
Jonathan Blaney
Andrew Jackson
Peter Webster

Data: the UK web domain crawl 1996-2013

buddah.projects.history.ac.uk

Arquivo.pt

- **DIOGO DUARTE**, The Study of punk culture through the Portuguese Web Archive
- **DIOGO SILVA DA CUNHA**, Today’s news to be forgotten tomorrow?
- **RICARDO BASÍLIO**, Memory of the online presence of a Faculty: an exhibition

netpreserveblog.wordpress.com/tag/investiga-xxi

Questions to dhn researchers

DIGITAL HUMANIORA I NORDEN
DIGITAL HUMANITIES IN THE NORDIC COUNTRIES

DATA

- What to archive (seed selection)?
- Thematic crawls – mainstream, “alternative” or both?
- Types of datasets: raw data / WARC files?
- Crawling: frequency, depth, width, completeness (formatting, images, etc.) – is documentation concerning this of interest to researchers? If so, how detailed? In the form of an overview or more closely attached to the harvested material?

CONTENT

- How should we focus on validity and reliability of the archives?
- How should we document the way our archives (and single crawls) have been constructed?

TOOLS

- Our tools or your tools?
- Are there any tools you want us to make available which would be useful without stepping into legal issues (e.g. N-gram services)?

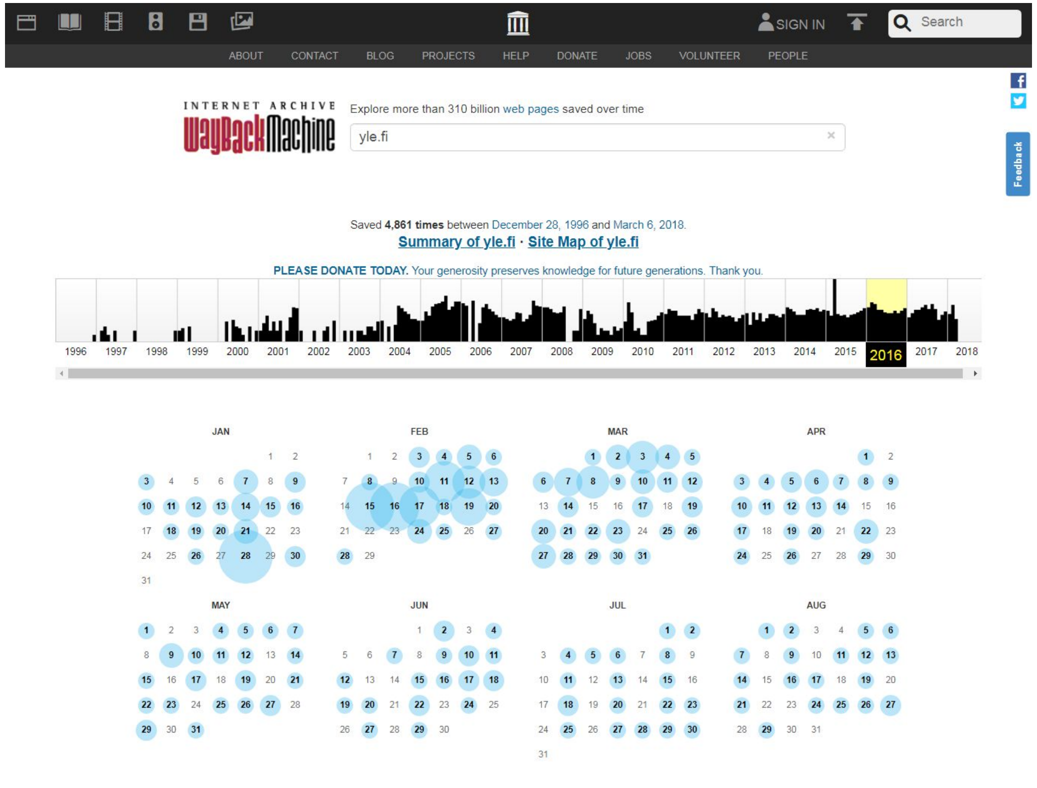
Vefsafn.is
Suomalainen verkkoarkisto
Norsk Nettarkiv
Kulturarw3
netarkivet.dk

WEB ARCHIVES:
WHAT´S IN THEM FOR DIGITAL HUMANISTS?

PANEL ON WEB ARCHIVING IN THE NORDIC COUNTRIES

National Library of Finland, **Lassi Lager**
National Library of Norway, **John Erik Halse**
National Library of Sweden, **Pär Nilsson**
The Royal Danish Library, **Caroline Nyvang**

Looking for Brexit in the Finnish web archive
DHN2018: Hacking the News



Explore more than 310 billion web pages saved over time

yle.fi

Saved **4,861 times** between December 28, 1996 and March 6, 2018.

Summary of yle.fi · **Site Map of yle.fi**

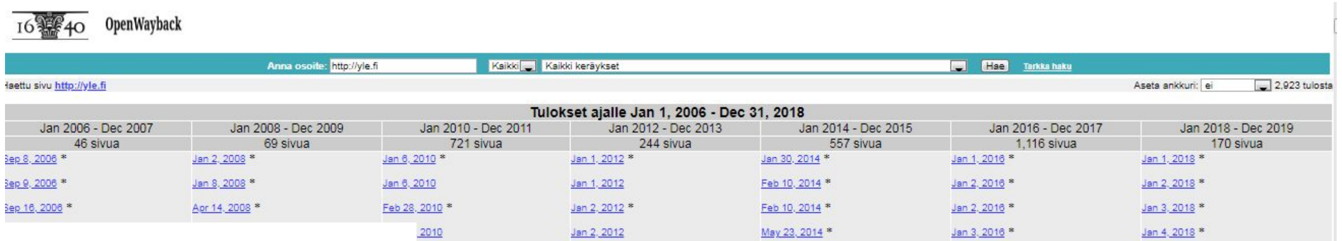
PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.

Front page of yle.fi: captures in IA vs FWA (Brexit)

http://verkkoarkisto.kansalliskirjasto.fi/wayback/*/yle.fi

132 captures (FI) vs. 68 (IA)

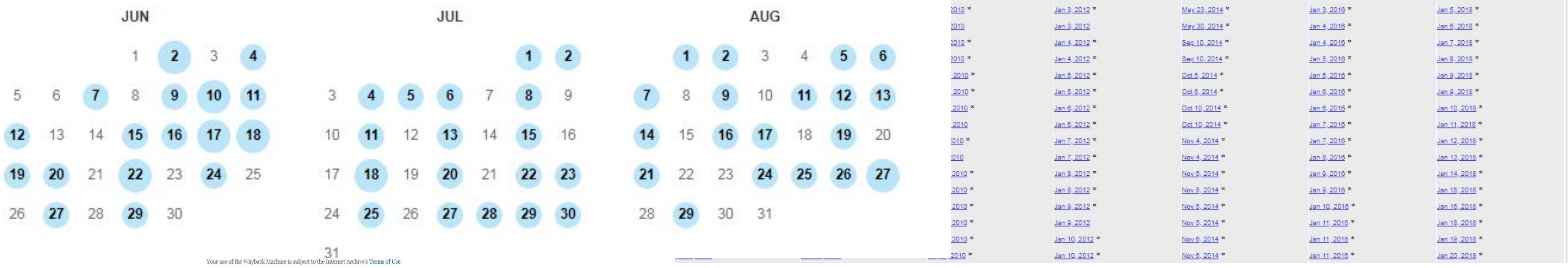
WARC files: 377MB extracted for this timelapse via AUT for yle.fi (+ 3 newspapers: satakunnankansa.fi, aamulehti.fi, pohjolansanomat.fi)



Anna osoite: http://yle.fi   Kaikki   Kaikki keräykset

Haettu sivu http://yle.fi

Aseta ankkuri: ei   2,923 tulosta



JUN

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

JUL

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

AUG

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

Tulokset ajalle Jan 1, 2006 - Dec 31, 2018

Jan 2006 - Dec 2007	Jan 2008 - Dec 2009	Jan 2010 - Dec 2011	Jan 2012 - Dec 2013	Jan 2014 - Dec 2015	Jan 2016 - Dec 2017	Jan 2018 - Dec 2019
46 sivua	69 sivua	721 sivua	244 sivua	557 sivua	1,116 sivua	170 sivua
Sep 8, 2006 *	Jan 2, 2008 *	Jan 6, 2010 *	Jan 1, 2012 *	Jan 30, 2014 *	Jan 1, 2016 *	Jan 1, 2018 *
Sep 9, 2006 *	Jan 8, 2008 *	Jan 6, 2010	Jan 1, 2012	Feb 10, 2014 *	Jan 2, 2016 *	Jan 2, 2018 *
Sep 16, 2006 *	Apr 14, 2008 *	Feb 28, 2010 *	Jan 2, 2012 *	Feb 10, 2014 *	Jan 2, 2016 *	Jan 3, 2018 *
		2010	Jan 2, 2012	May 23, 2014 *	Jan 3, 2016 *	Jan 4, 2018 *
		2010 *	Jan 3, 2012 *	May 23, 2014 *	Jan 3, 2016 *	Jan 5, 2018 *
		2010	Jan 3, 2012	May 30, 2014 *	Jan 4, 2016 *	Jan 6, 2018 *
		2010 *	Jan 4, 2012 *	Sep 10, 2014 *	Jan 4, 2016 *	Jan 7, 2018 *
		2010 *	Jan 4, 2012 *	Sep 10, 2014 *	Jan 5, 2016 *	Jan 8, 2018 *
		2010 *	Jan 5, 2012 *	Oct 5, 2014 *	Jan 5, 2016 *	Jan 9, 2018 *
		2010 *	Jan 5, 2012 *	Oct 6, 2014 *	Jan 6, 2016 *	Jan 9, 2018 *
		2010 *	Jan 6, 2012 *	Oct 10, 2014 *	Jan 6, 2016 *	Jan 10, 2018 *
		2010	Jan 6, 2012 *	Oct 10, 2014 *	Jan 7, 2016 *	Jan 11, 2018 *
		010 *	Jan 7, 2012 *	Nov 4, 2014 *	Jan 7, 2016 *	Jan 12, 2018 *
		010	Jan 7, 2012 *	Nov 4, 2014 *	Jan 8, 2016 *	Jan 13, 2018 *
		2010 *	Jan 8, 2012 *	Nov 5, 2014 *	Jan 9, 2016 *	Jan 14, 2018 *
		2010 *	Jan 8, 2012 *	Nov 5, 2014 *	Jan 9, 2016 *	Jan 15, 2018 *
		2010 *	Jan 9, 2012 *	Nov 5, 2014 *	Jan 10, 2016 *	Jan 16, 2018 *
		2010 *	Jan 9, 2012	Nov 5, 2014 *	Jan 11, 2016 *	Jan 18, 2018 *
		2010 *	Jan 10, 2012 *	Nov 6, 2014 *	Jan 11, 2016 *	Jan 19, 2018 *
		2010 *	Jan 10, 2012 *	Nov 6, 2014 *	Jan 11, 2016 *	Jan 20, 2018 *

JISC Report, 2010: Key points

□ gap between the ***potential* community of researchers** who have good reason to engage with creating, using, analysing and sharing web archives, and **the *actual* (generally still small) community of researchers currently doing so**

- **Encourage the creation of communities** that increase the accessibility and usability of web archiving tools
- Efforts should be made to diversify tool and interface development beyond preservation and into use
- Tools should be developed that are able to execute query searches over multiple web archives
- Standards, protocols and methods of quality control are need for interoperability, but not at the cost of flexibility
- The possibilities of web archives should be communicated to a much broader research community

Meghan Dougherty, Eric T. Meyer, Christine McCarthy Madsen, Charles van den Heuvel, Arthur Thomas, and Sally Wyatt. **Researcher Engagement with Web Archives: State of the Art**. JISC Report, 2010.

RESEARCH WORKING GROUP

http://bit.ly/iipcRWG

Scope

- The Research Working Group seeks to promote the use of web archives and IIPC collections among researchers, share information about web archiving research projects at IIPC member organisations, including workflows and lessons learnt, and facilitate ways for dissemination and discussion of use cases.

- The group will collaborate with relevant research communities to help make the IIPC collaborative collections and IIPC member collections available to researchers. The Research Working Group will include not only IIPC members but it will be open to researchers.

The toolkit

Platform/UI	**Extraction**	**Textual analysis**	**Graphing**
Shine https://github.com/ukwa/shine	Web-archive discovery https://github.com/ukwa/webarchive-discovery	Spacy https://spacy.io/	Gephi https://gephi.org/
SolrWayback https://github.com/netarchivesuite/solrwayback	Archives Unleashed Toolkit https://github.com/archivesunleashed/aut	Sem https://github.com/YoannDupont/SEM	SigmaJS http://sigmajs.org/
Warclight https://github.com/archivesunleashed/warclight	Web-archive indexing https://github.com/ikreymer/webarchive-indexing	Open NER http://www.opener-project.eu/	
UKWA UI https://github.com/ukwa/ukwa-ui		Stanford CoreNLP https://stanfordnlp.github.io/CoreNLP/ner.html	
AUT Cloud https://github.com/archivesunleashed/aut		Voyant Tools https://voyant-tools.org/	
		http://tapor.ca/home	
		https://cloud.archivesunleashed.org/derivatives/text-sentiment	
		https://cloud.archivesunleashed.org/derivatives/text-antconc	