

DIGITAL TOOLS FOR HUMANISTS

SUMMER SCHOOL 2022, University of Pisa

Meaning-making and storytelling in the age of
databases, websites, and social media

Thursday, 16 June 2022

PART 2: Afternoon Session – Web Archives



Dr Seamus Ross,
Professor, Faculty of Information, University of Toronto

- All these Powerpoint Slides contain copyright material. The copyright in the lecture as a whole rests with the instructor. In many cases the copyright in the material within the lecture belongs to the instructor. In other cases it belongs to the University. In some cases the slides contain material the copyright of which belongs to other individuals, institutions, or entities. Where this is the case this material is *used* under one of the exemptions of the Canadian Copyright Act (<https://laws-lois.justice.gc.ca/eng/acts/C-42/index.html> (Links to an external site.)), such as, but not limited to, Fair Dealing (Sections 29, 29.1 and 29.2), and Education Exemptions (Sections 30.04, 29.4(1), 29.4(2), 29.5, 29.6, and 29.7). Where material is used under such exemptions as Copyright Act Sections 29.1 and 29.2 the source of the work and the author are clearly stated on the relevant slides.
- Provision of access to these Powerpoint Slides themselves is done under the exemption granted in Section 30.01 (Communication by Telecommunication). This exemption requires that you delete these slides within 30 days of the end of the course.

➤ Welcome and Introduction

➤ Who am I

➤ Overview of the day

- Lectures in Morning
- Interactive Activities & Experimentation in Afternoon

Timetable & What we will cover

- 09:00 – 10:30 Lecture on Open Data & Databases
- 10:30 – 11:00 Break
- 11:00 – 12:30 Story-telling with a Database (Group Activity)
- 12:30 – 14:00 Lunch
- 14:00 – 15:30 Lecture on Web Archiving and Web Archives
- 15:30 – 16:00 Break
- 16:00 – 17:00 Story-telling with Web Archives
- 17:00 – 17:30 Discussion

HIGH-LEVEL PRESERVATION VIEW



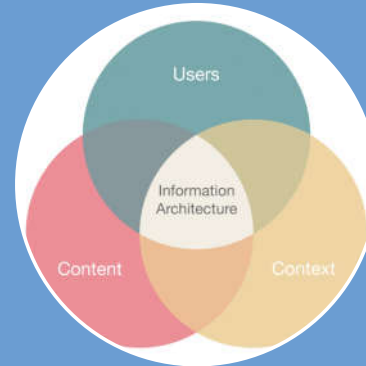
Bit Stream

- (01100101101010010)



Information Content

- (e.g. images, sounds, text)



Context of Information

- (e.g. information architecture, linkages, interrelatedness)



Experience

- (e.g. speed, layout, quality of display device, input device characteristics)

KEEP IN MIND.....PERFORMANCE

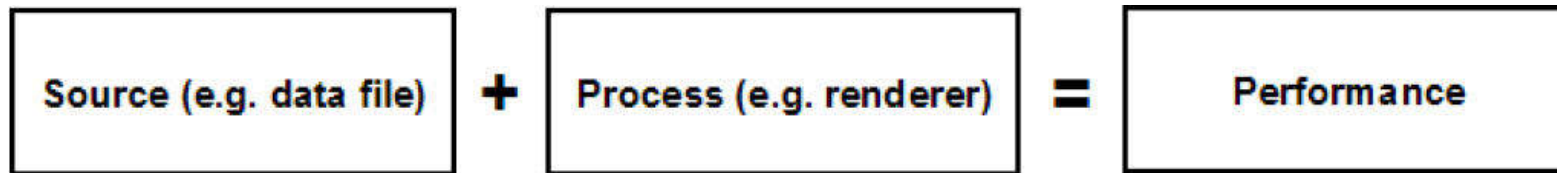


Figure 2.1: National Archives of Australia's Performance Model

Multi-layered performance and semantic intelligibility any many different layers

A relationship between the user of an object and the object itself which is “brokered by software and hardware” (NAA, 2002)

- Example
 - Data represented as magnetic charges on media
 - Interpreted as 1's and 0's and presented as a sequence which to the Operating System appears as a file.
 - File presented to application which performs it.
- **Performance is nuanced and dependent upon a variety of successful performances**

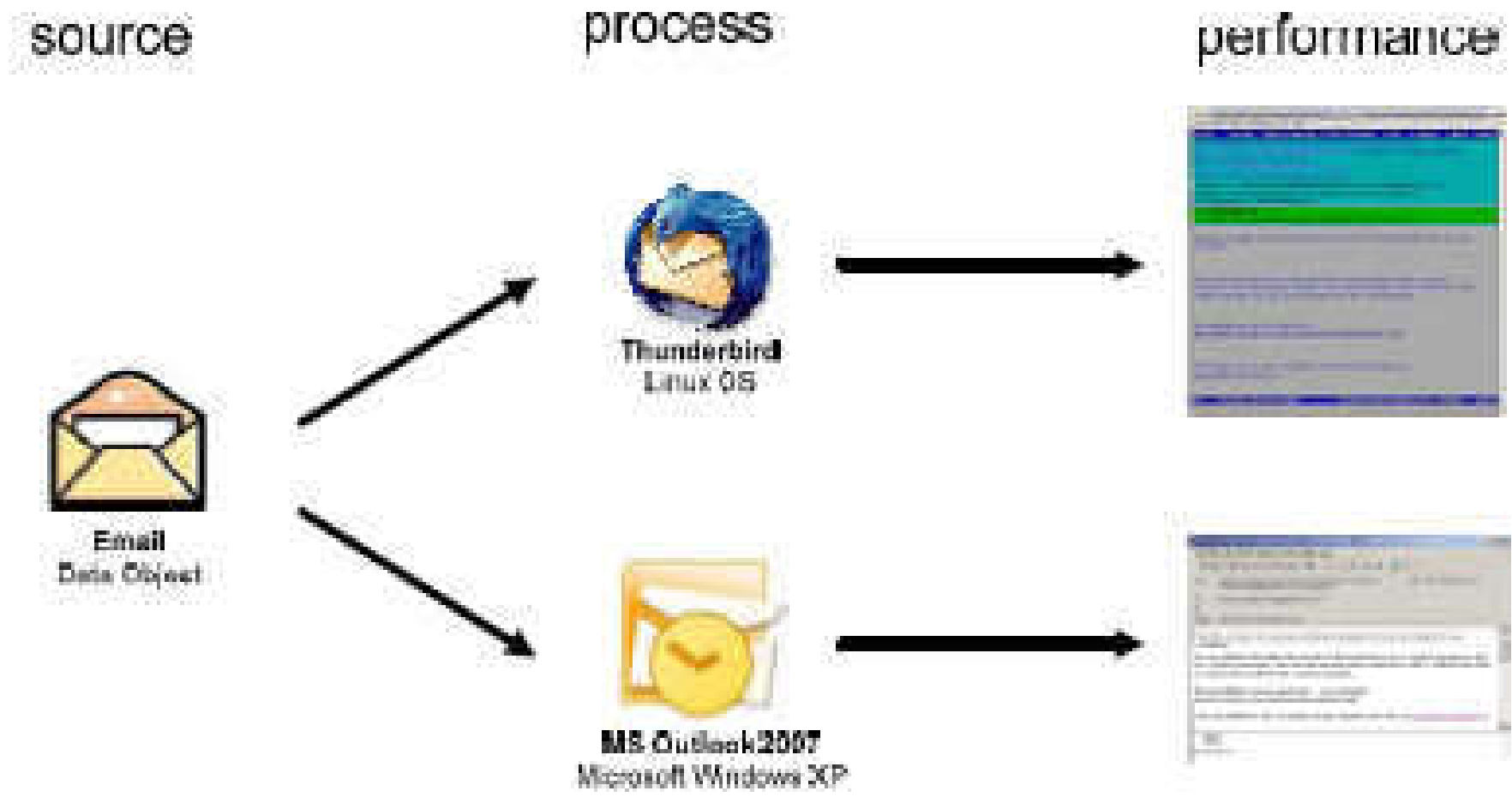
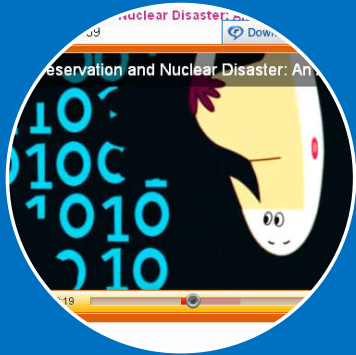


Figure 5 Application of the Performance Model to emails

G. Knight, and L. Montague, 2009, "InSPECT: Final Report", <http://www.significantproperties.org.uk/inspect-finalreport.pdf>, p.27

DIGITAL ENTITIES HAVE



Syntax



Semantics

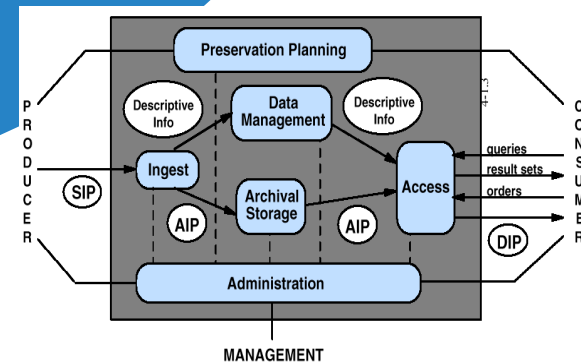


Pragmatics



Abstraction and modelling provides a mechanism to improve understanding and communication about long term preservation and curation of digital assets.

- OAIS Digital Preservation Model
- DCC UK: Digital Curation Lifecycle Model



OAIS Model & Example

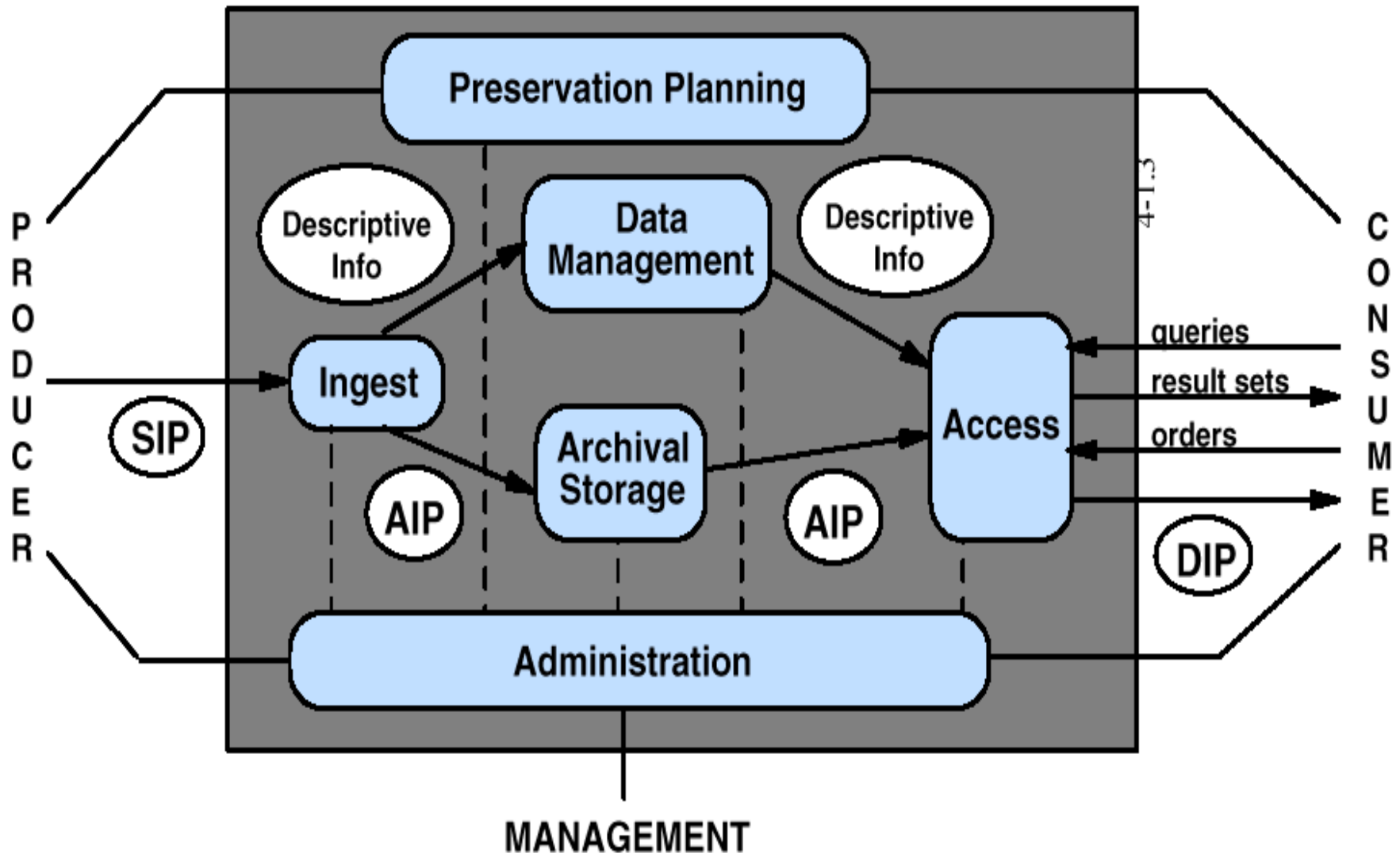
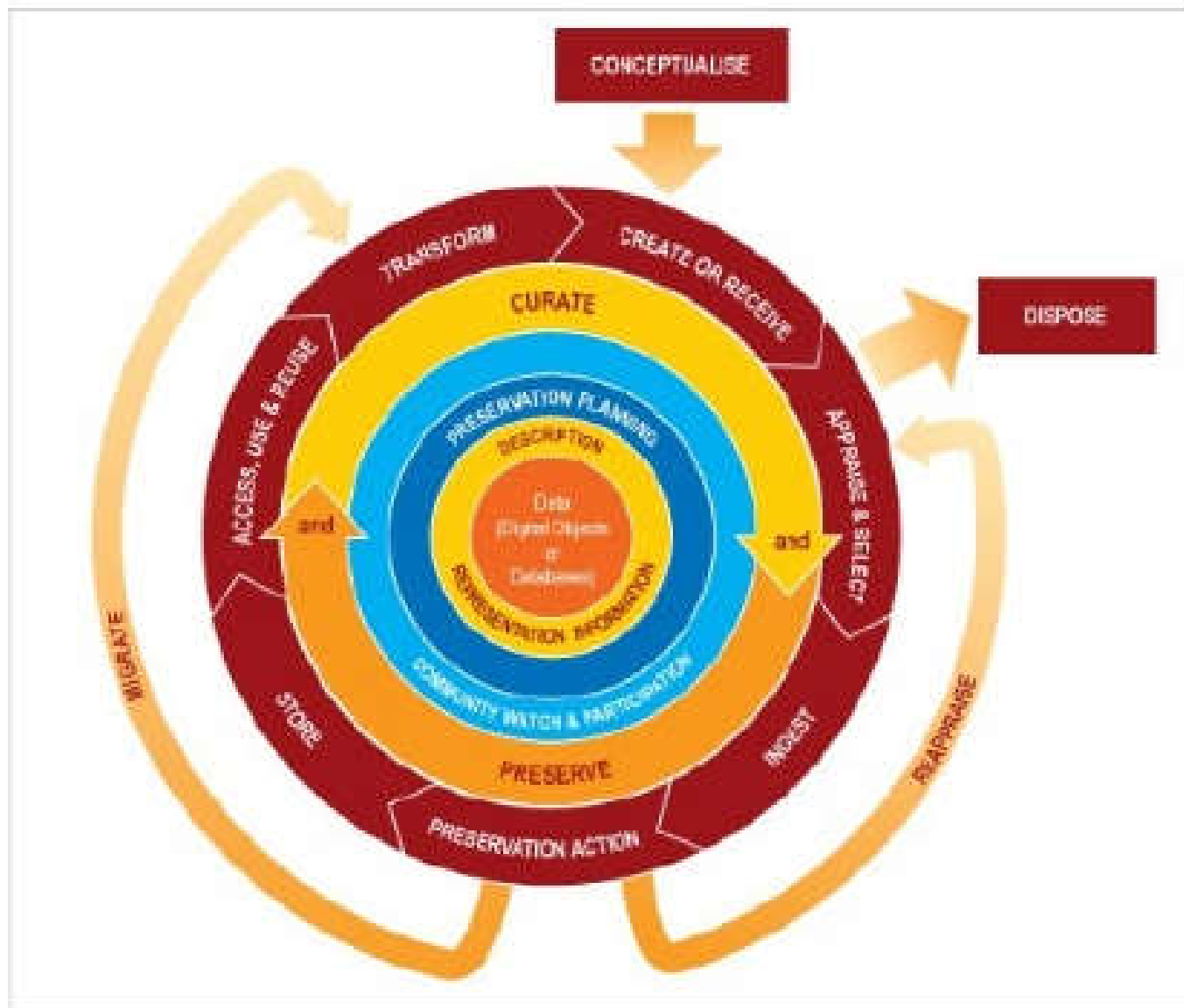


Image from -- Reference Model for an *Open Archival Information System* (OAIS) – CCSDS,2002, <http://www.ccsds.org/documents/650x0b1.pdf>



Digital Curation Centre: <http://www.dcc.ac.uk/resources/curation-lifecycle-model>

What is WEB Archiving, IIPC (International Internet Preservation Consortium) defines web archiving as:

Web archiving is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use.

Why Archive the Web

- ❑ Compliance and Juridical Evidence
- ❑ Research Potential – design studies, history, art history, language change and evolution, Innovation

<https://netpreserve.org/web-archiving/>

INTERNET ARCHIVE Explore more than 698 billion web pages saved over time



BROWSE HISTORY

Find the Wayback Machine useful? [DONATE](#)



Tools

- [Wayback Machine Availability API](#)
Build your own tools.
- [WordPress Broken Link Checker](#)
Banish broken links from your blog.
- [404 Handler for Webmasters](#)
Help users get where they were going.

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. [Visit Archive-It to build and browse the collections.](#)

Save Page Now

SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.
Only available for sites that allow crawlers.

<https://archive.org/web/>



Explore more than 698 billion web pages saved over time

ERPANET

Results: 50 100 500

Collection search: Main

<http://erpanet.org/>
<http://www.erpanet.org>
📅 2,191 📄 868 📁 59 📁 0
10,813 capture(s) from 2001 to 2016 | Site stats

<http://daedalus.lib.gla.ac.uk/>
erpanet eprints service
📅 501 📄 13 📁 0 📁 0
909 capture(s) from 2003 to 2014 | Site stats

<http://eprints.erpanet.org/>
<http://eprints.erpanet.org>
📅 336 📄 1 📁 0 📁 0
2,117 capture(s) from 2004 to 2015 | Site stats

https://web.archive.org/web/*/ERPANET

https://web.archive.org/web/20020523003521/http://www.erpanet.org/

Go APR MAY JUL 23 2002 2003 About this capture

ERPANET

Electronic Resource
Preservation and Access NETWORK

ABOUT
DOCUMENTATION
PROJECT PARTNERS

Mission Statement


The European Commission funded ERPANET Project will establish an expandable European Consortium, which will make viable and visible information, best practice and skills development in the area of digital preservation of cultural heritage and scientific objects. ERPANET will bring together memory organisations (museums, libraries and archives), ICT and software industry, research institutions, government organisations (including local ones), entertainment and creative industries, and commercial sectors (including for example pharmaceuticals, petro-chemical, and financial). The dominant feature of ERPANET will be the provision of a virtual clearinghouse and knowledge-base on state-of-the-art developments in digital preservation and the transfer of that expertise among individuals and institutions.

[ERPANET Brochure in printable PDF format.](#)

INTERNET ARCHIVE WaybackMachine http://www.erpanet.org/ 285 captures 23 May 2002 - 1 Jun 2002

Go APR JUN AUG 23 2002 2003 2004 About this capture

ELECTRONIC RESOURCE PRESERVATION and ACCESS NETWORK




ENGLISH FRANÇAIS DEUTSCH ITALIANO ESPAÑOL

© erpanet 2002/2004

Internet Archive Wayback Machine <http://www.erpanet.org/> 285 captures 23 May 2002 - 1 Jun 2022

https://web.archive.org/web/20050630083331/http://www.erpanet.org/ APR JUN AUG 30 2004 2005 2006 About this capture

english français deutsch italiano

ELECTRONIC RESOURCE PRESERVATION AND ACCESS NETWORK 

Home About Site Map Search Contact Us Users Private

news

AVAILABLE: Bern Public Lectures
Posted on 4th November 2004

Lectures by Ken Thibodeau (NARA) and Seamus Ross (HATII, Univ. Glasgow) are now available to consult. Ken Thibodeau outlines the ERA's plans of electronic archiving, and Seamus Ross introduces the newly launched Digital Curation Centre.


AVAILABLE: Ingest Strategies Guidance Document
Posted on 7th October 2004

This guidance document is intended to introduce ingest and its role in the development of a digital repository system. The appendix contains a companion guide and checklist when defining and/or selecting an ingest strategy, presenting a survey of the factors required for consideration.

ANNOUNCED: Workflow in Digital Preservation
Posted on 26th August 2004

ERPANET is pleased to announce its workshop on workflow in preservation. The three-day workshop is co-hosted by the Open Society Archives at Central Europe, Hungary, in October 13-15.


» Topic of the Month
» erpaAdvisory
» erpaAssessments
» erpaDirectory
» erpaEprints
» erpaEvents
» erpaGuidance
» erpaStudies
» Charter (v4.1)
» erpaDiscussion
» erpaDocumentation



Internet Archive Wayback Machine <http://www.erpanet.org/> 285 captures 23 May 2002 - 1 Jun 2022

Go DEC FEB 15 2007 2008 20

english français deutsch italiano

ELECTRONIC RESOURCE PRESERVATION AND ACCESS NETWORK 

Home About Site Map Search Contact Us Users Private

news

Papers from ERPANET Berne Conference Published
Posted on 21st July 2007


In October 2004 a group of leading international scholars met at the Swiss Federal Archives under the auspices of ERPANET to examine the topic of 'Managing and archiving records in the digital era: the same discipline or a difficult partnership of two different professions'. The debate was very lively. The papers from the conference have now been revised, edited, and published as 'Managing and Archiving Records in the Digital Era: Changing Professional Orientations', Niklaus Bütikofer, Hans Hofman, and Seamus Ross (eds), 2006, Baden: Hier+Jetzt. (ISBN 10: 3-03919-019-9) Copies can be ordered from the publisher at: <http://www.hierundjetzt.ch/>

ANNOUNCED: DCC Workshop on the Long-term Curation of Medical Databases
Posted on 11th October 2005

A joint DCC and Electronic Resource Preservation and Access Network (ERPANET) workshop on the Long-term Curation of Medical Databases will be held in the Gulbenkian Institute, Lisbon, Portugal, on the 13-14 October 2005. For more information about the programme, course fees and registration, see <http://www.dcc.ac.uk/training/mdb-2005>.

AVAILABLE: Bern Public Lectures

» Topic of the Month
» erpaAdvisory
» erpaAssessments
» erpaDirectory
» erpaEprints
» erpaEvents
» erpaGuidance
» erpaStudies
» Charter (v4.1)
» erpaDiscussion
» erpaDocumentation



Wayback Machine 134 captures 20 Aug 2008 - 23 Apr 2022

https://web.archive.org/web/20060820070336/http://www.digitalpreservationeurope.eu/about/

digital preservation Europe

Home & News | Contact Us | Site Map | RSS | Search | Users | Private

about DPE

[Mission Statement](#) | [Objectives](#) | [Project Partners](#) | [Staff Directory](#)

Mission Statement

Electronic resources are a central part of our cultural and intellectual heritage, but this material is at risk. Digital memory needs constant management, using new techniques and processes, to contain such risks as technological obsolescence. Risk begins before the digital record is created and continues for as long as the digital object needs to be retained. Digital preservation is too big an issue for individual institutions or even sectors to address independently. Concerted action at both national and international level is required. DigitalPreservationEurope, building on the earlier successful work of ERPANET, facilitates pooling of the complementary expertise that exists across the academic research, cultural, public administration and industry sectors in Europe.

DigitalPreservationEurope (DPE) fosters collaboration and synergies between many existing national initiatives across the European Research Area. DPE addresses the need to improve coordination, cooperation and consistency in current activities to secure effective preservation of digital materials. DPE's project partners lead work to:

- raise the profile of digital preservation;
- promote the ability of Member States acting together to add value to digital preservation activities across Europe;
- use cross-sectoral cooperation to avoid redundancy and duplication of effort;
- ensu

Information Society Technologies

» About DPE
 » Events
 » DPE Publications
 » Associate Partner Forum
 » ErpaEprints
 » Resources
 » Adding Information

Wayback Machine 212 captures 17 Jun 2008 - 12 May 2022

https://web.archive.org/web/20080212012820/http://www.digitalpreservationeurope.eu/

digital preservation Europe

home & news | contact us | site map | rss | search | staff

welcome

CeBIT
4-9 MARCH, HANNOVER GERMANY

DPE at CeBIT 2008
Preserving the future -
Stand B14, Hall 9, Future Parc

DigitalPreservationEurope(DPE), building on the earlier successful work of ERPANET, facilitates pooling of the complementary expertise that exists across the academic research, cultural, public administration and industry sectors in Europe.

DPE fosters collaboration and synergies between many existing national and international initiatives across the European Research Area. DPE addresses the need to improve coordination, cooperation and consistency in current activities to secure effective preservation of digital materials. DPE's success will help to secure a shared knowledge base of the processes, synergy of activity, systems and techniques needed for the long-term management of digital material.

Latest News:

Preserving the future - cutting edge digital preservation techniques showcased at CeBIT 4-9 March 2008 Stand B14, Hall 9, Future Parc
 Posted on 6th February 2008

The latest digital preservation research and technology developments and their relevance to the industry will be showcased at the stand 'We preserve

- about DPE
- news and events
- exchange programme
- DPE challenge
- DPE user community
- DPE registries
- DPE publications
- DRAMBORA
- ErpaEprints

DONATE

WayBackMachine

Results: 50 100 500

[Calendar](#) · [Collections](#) · [Changes](#) · [Summary](#) · [Site Map](#) · [URLs](#)Saved **212 times** between [June 17, 2006](#) and [May 12, 2022](#).

JAN					FEB					MAR					APR													
	1	2	3	4	5						1	2				1			1	2	3	4	5					
6	7	8	9	10	11	12	3	4	5	6	7	8	9	2	3	4	5	6	7	8	6	7	8	9	10	11	12	
13	14	15	16	17	18	19	10	11	12	13	14	15	16	9	10	11	12	13	14	15	13	14	15	16	17	18	19	
20	21	22	23	24	25	26	17	18	19	20	21	22	23	16	17	18	19	20	21	22	20	21	22	23	24	25	26	
27	28	29	30	31	24	25	26	27	28	29	23	24	25	26	27	28	29	27	28	29	30							
													30	31														
MAY					JUN					JUL					AUG													
			1	2	3	1	2	3	4	5	6	7			1	2	3	4	5					1	2			
4	5	6	7	8	9	10	8	9	10	11	12	13	14	6	7	8	9	10	11	12	3	4	5	6	7	8	9	
11	12	13	14	15	16	17	15	16	17	18	19	20	21	13	14	15	16	17	18	19	10	11	12	13	14	15	16	
18	19	20	21	22	23	24	22	23	24	25	26	27	28	20	21	22	23	24	25	26	17	18	19	20	21	22	23	
25	26	27	28	29	30	31	29	30	27	28	29	30	31	24	25	26	27	28	29	30	31							
SEP					OCT					NOV					DEC													
			1	2	3	4	5	6	1	2	3	4					1	1	2	3	4	5	6					
7	8	9	10	11	12	13	5	6	7	8	9	10	11	2	3	4	5	6	7	8	7	8	9	10	11	12	13	
14	15	16	17	18	19	20	12	13	14	15	16	17	18	9	10	11	12	13	14	15	14	15	16	17	18	19	20	

welcome

Digital Preservation Europe (DPE), building on the earlier successful work of ERPNANET, facilitates pooling of the complementary expertise that exists across the academic research, cultural, public administration and industry sectors in Europe.

DPE fosters collaboration and synergies between many existing national and international initiatives across the European Research Area.

DPE addresses the need to improve coordination, cooperation and consistency in current activities to secure effective preservation of digital materials. DPE's success will help to secure a shared knowledge base of the processes, synergy of activity, systems and techniques needed for the long-term management of digital material.

Latest News:

New translation of the DPE Briefing papers.

Posted on 24th February 2009

Cristina Faria provided Portuguese translation of the paper on [Automating semantic metadata extraction](#) by Di Yunhyong Kim from HATII University in Glasgow.

Maria Teresa De Gregori translated the paper on [Database Preservation](#) into Italian. [See list of all briefing paper translations.](#)

DRAMBORA Auditor's Training Course in Rome (March 23rd 2009 - 25th 2009)

Posted on 17th February 2009

Based on practical research and developed jointly by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) provides a methodology for self-assessment of digital preservation repositories. The toolkit (<http://www.repositoryaudit.eu>), has been evaluated and applied across a diverse range of organisations, such as national libraries, scientific data centres and archives. DPE is organising a successful series of training courses to train new DRAMBORA auditors. The third of these will be held at the Ministry for Cultural Heritage, Rome, Italy, in March 2009. [More Info](#)

New Briefing Paper about 'UMID – Unique Material Identifier' is available

Posted on 13th February 2009

A Unique Material Identifier (UMID) is a special code that is used to identify audiovisual (AV) materials. Text by Nadja Wallaszkovits, Christian Liebl, Phonogrammarchiv - Austrian Academy of Sciences. [Read it here.](#)

- ~ [about DPE](#)
- ~ [what is digital preservation](#)
- ~ [news and events](#)
- ~ [exchange programme](#)
- ~ [DPE challenge](#)
- ~ [DPE user community](#)
- ~ [DPE registries](#)
- ~ [DPE publications](#)
- ~ [DRAMBORA](#)
- ~ [PLATTER](#)
- ~ [ErpaEprints](#)
- ~ [unique identifier service](#)
- ~ [DPE video training](#)



BOOKMARK



Find us on

INTERNET ARCHIVE <http://digitalpreservationeurope.eu/> Go **MAR** **MAY** **JUL**
WayBackMachine [212 captures](#) 17 Jun 2006 - 12 May 2022 **17** **2016** **2017** **2018** About this capture

digitalpreservationeurope.eu

BUY THIS DOMAIN
This domain is FOR SALE - Diese Domain steht ZUM VERKAUF

INTERNET ARCHIVE WayBackMachine Explore more than 698 billion web pages saved over time

Search results for 'krys-corporus' showing 50 results. Includes a thumbnail of the website and a snippet: 'http://krys-corporus.eu/ hatii and dcc release krys i corpus to aid research'.

Wayback Machine interface showing a timeline of captures for 'http://www.krys-corporus.eu/'.

Home Register/Login Information Terms of Use Contact

Welcome to the website of the KRY S I Corpus.

The KRY S I corpus is a collection of over 6300 documents labelled with their genre classes. It was constructed as part of a research initiative to automate document genre classification driven by the Digital Curation Centre. It was carried out at the Humanities Advanced Technology and Information Institute (HATII), University of Glasgow between 2005 and 2008.

The notion of genre is deeply embedded in the way humans organise information. Identifying the genre of a document helps to characterise the physical and conceptual structure of the text, helping to capture the style and location of further information within the text. There have been very few genre-labelled corpora available to the research community. Our corpus is made available here to fill this gap and serve as a valuable resource for researchers in:

- metadata extraction,
- digital curation,
- text classification,
- text mining,
- computational linguistics,
- and, pattern recognition.

To access the Corpus, please register first by going to the page [Registration/Login](#). By registering to access the Corpus, you are agreeing to the specified [Terms of Use](#). The KRY S I Corpus is owned by the University of Glasgow. However, the copyrights to the documents within the Corpus are retained by the original copyright owners, and their permission might be required before you copy, use, or distribute any of the content. Please note that documents will be removed upon the request of the copyright owners without prior notice. Also, note that access to the corpus could be withdrawn should any misuse of the Corpus be detected.

All documents within the KRY S I Corpus have been collected from the Internet and are, thus, publicly available. While the authors tried to assure that no copyright law was violated, not all document owners could be contacted. Should you find that your document is in use unrighteously within this collection, please contact the KRY S I Corpus Manager at manager@krys-corporus.eu, quoting the document ID number within the collection, and the document will be removed immediately. Please be assured, that no content within the document has been altered except when this has been explicitly asked for by the copyright owner.

There is more information about the construction method and composition of the corpus on the [Information](#) page. Automated experiments we have conducted using the corpus are reported in the published papers listed on [this page](#).

If you would like to further help this corpus building initiative, please go to our

WayBack Machine interface showing a search for 'http://www.krys-corporus.eu/Register.html' with 'Latest' and 'Show All' buttons.

Hrm.

The Wayback Machine has not archived that URL.

This page is not available on the web because page does not exist

Click here to search for all archived pages under <http://www.krys-corporus.eu/>.

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.

Your use of the Wayback Machine is subject to the Internet Archive's [Terms of Use](#).

https://core.ac.uk/display/91457

Mail - Seamus Ross - O... Capture Calendar | Mus... restaurants Santa Maria... Official ticketing page... Näsäjo helsingborg esc... Wayback Machine ERPANET - Electronic Resou... 404 Not Found

https://www.krys-corpus.eu/info.html

Not Found

The requested URL /info.html was not found on this server.

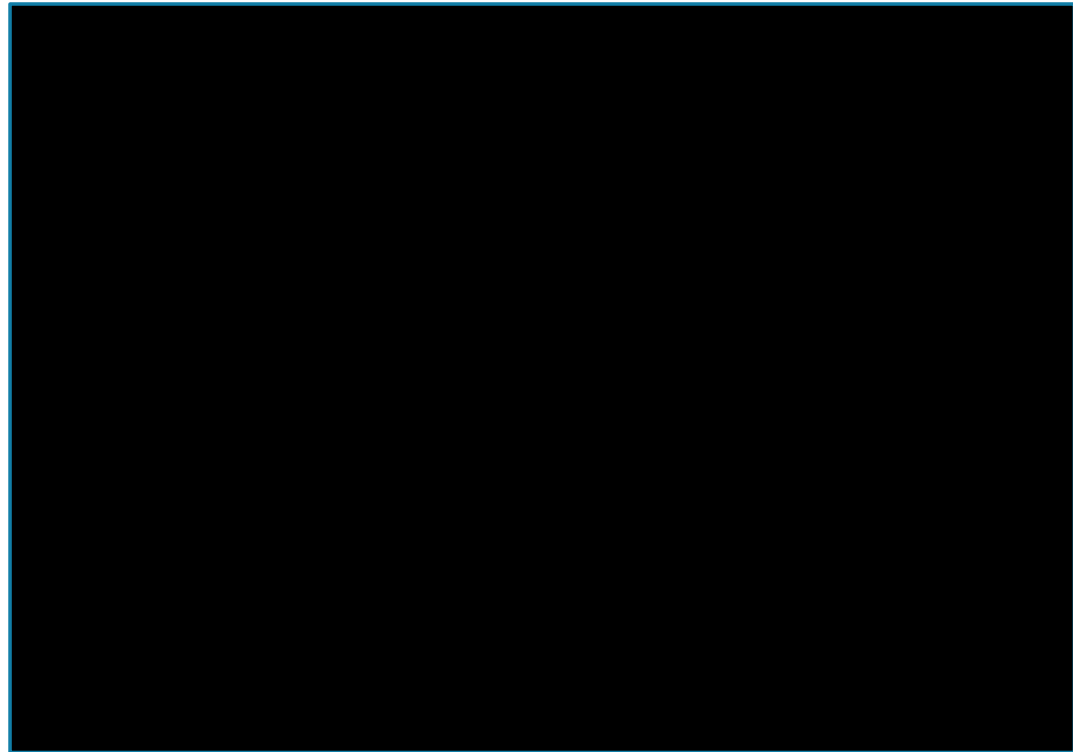
RESEARCH

Building a Document Genre Corpus: a Profile of the KRY5 I Corpus

Vera Berninger, Yunhyong Kim, Seamus Ross • 17 October 2008

Abstract

This paper describes the KRY5 I corpus (<http://www.krys-corpus.eu/info.html>), consisting of documents classified into 70 genre classes. It has been constructed as part of an effort to automate document genre classification as distinct from topic detection. Previously there has been very little work on building corpora of texts which have been classified using a non-topical genre palette. The reason for this is partly due to the fact that genre as a concept, is rooted in philosophy, rhetoric and literature, and highly complex and domain dependent in its interpretation ([11]). The usefulness of genre in everyday information search is only now starting to be recognised and there is no genre classification schema that has been consolidated to have applicable value in this direction. By presenting here our experiences in constructing the KRY5 I corpus, we hope to shed light on the information gathering and seeking behaviour and the role of genre in these activities, as well as a way forward for creating a better corpus for testing automated genre classification tasks and the application of these tasks to other domains



<https://www.krys-corpus.eu/> now points to a Swedish Site of dubious nature

The evolution of web archiving

Miguel Costa¹ · Daniel Gomes² · Mário J. Silva³

Received: 1 May 2015 / Revised: 12 April 2016 / Accepted: 12 April 2016 / Published online: 9 May 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract Web archives preserve information published on the web or digitized from printed publications. Much of this information is unique and historically valuable. However, the lack of knowledge about the global status of web archiving initiatives hamper their improvement and collaboration. To overcome this problem, we conducted two surveys, in 2010 and 2014, which provide a comprehensive characterization on web archiving initiatives and their evolution. We identified several patterns and trends that highlight challenges and opportunities. We discuss these patterns and trends that enable to define strategies, estimate resources and provide guidelines for research and development of better technology. Our results show that during the last years there was a significant growth in initiatives and countries hosting these initiatives, volume of data and number of contents preserved. While this indicates that the web archiving community is dedicating a growing effort on preserving digital information, other results presented throughout the paper raise concerns such as the small amount of archived data in comparison with the amount of data that is being published online.

Keywords Web archiving · Digital preservation · Survey

✉ Miguel Costa
migcosta@gmail.com

¹ Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

² Foundation for National Scientific Computing, Lisbon, Portugal

³ INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

1 Introduction

The world wide web has a democratic nature, where everyone can publish all kinds of information using different types of media. News, blogs, wikis, encyclopedias, photos, interviews and public opinions are just a few examples of this vast list. Part of this information is unique and historically valuable. For instance, the speech of a president after winning an election or the announcement of an imminent invasion of a foreign country, might become as valuable in the future as ancient manuscripts are today. However, since the web is so dynamic, a large amount of information is lost everyday. Several studies quantify this loss: 80 % of web pages are not available in their original form after 1 year [1]; 13 % of web references in scholarly articles disappear after 27 months [2]; 11 % of social media resources, such as the ones posted on Twitter, are lost after 1 year [3]. All this information will likely vanish in a few years, creating a knowledge gap about the present for future generations. We are already experiencing unsatisfied information needs due to missing pages or old formats of documents that are not readable by the latest software version.¹ Pioneers of the Internet, such as Vint Cerf, recently warned about the danger of future generations who will have little or no record of the twenty-first century.² International organizations are also concerned with the web ephemerality problem. The UNESCO recognized the importance of digital preservation in 2003, by stating that the disappearance of digital information constitutes an impoverishment of the heritage of all nations [4]. In 2010, the UNESCO endorsed the Universal Declaration on Archives, which states that archives play an essential role in the development of societies by safeguard-

¹ http://en.wikipedia.org/wiki/Digital_obsolescence.

² <http://www.bbc.com/news/science-environment-31450389>.

Miguel Costa, Daniel Gomes, & Mário J Silva, 2017, “The evolution of web archiving,” *International Journal of Digital Libraries* 18, pp., 191–205. <https://doi.org/10.1007/s00799-016-0171-9>

Observing Web Archives

The Case for an Ethnographic Study of Web Archiving

Jessica Ogden
University of Southampton
Southampton, UK
jessica.ogden@soton.ac.uk

Susan Halford
University of Southampton
Southampton, UK
susan.halford@soton.ac.uk

Leslie Carr
University of Southampton
Southampton, UK
lac@ecs.soton.ac.uk

ABSTRACT

This paper makes the case for studying the work of web archivists, in an effort to explore the ways in which practitioners shape the preservation and maintenance of the archived Web in its various forms. An ethnographic approach is taken through the use of observation, interviews and documentary sources over the course of several weeks in collaboration with web archivists, engineers and managers at the Internet Archive - a private, non-profit digital library that has been archiving the Web since 1996. The concept of *web archival labour* is proposed to encompass and highlight the ways in which web archivists (as both networked human and non-human agents) shape and maintain the preserved Web through work that is often embedded in and obscured by the complex technical arrangements of collection and access. As a result, this engagement positions web archives as places of knowledge and cultural production in their own right, revealing new insights into the performative nature of web archiving that have implications for how these data are used and understood.¹

KEYWORDS

web archiving, knowledge production, STS, materiality, information labour

ACM Reference format:

Jessica Ogden, Susan Halford, and Leslie Carr. 2017. Observing Web Archives. In *Proceedings of WebSci '17, Troy, NY, USA, June 25-28, 2017*, 10 pages. <https://doi.org/10.1145/3091478.3091506>

1 INTRODUCTION

The World Wide Web has emerged as the preeminent mechanism for global communication, political, economic and cultural exchange and more. Yet, at the same time, the Web is ephemeral. For a medium that has become pre-eminent, its dynamism and transience has become increasingly worrisome. These concerns have been illustrated in various longitudinal studies of link rot [55] and investigations which found that during a period between 2009 and 2012, on average 11% of online resources shared on social media failed to resolve one year later [60]. In this context, it is increasingly claimed that

¹This paper is based on data collected and fieldwork undertaken by the first author as part of their PhD research.



This work is licensed under a Creative Commons Attribution International 4.0 License.

WebSci '17, June 25-28, 2017, Troy, NY, USA.
© 2017 Copyright held by the owner/authors.
ACM ISBN 978-1-4501-4094-4/17/06.
<https://doi.org/10.1145/3091478.3091506>

the ephemerality of the Web demands intervention to preserve web content - in web archives - that reconstruct sites and the 'web experience' for posterity [2]. However, there has been rather less attention to the nature of this intervention: how it is done and why this matters. This paper explores the critical decisions being made now that will shape future generations' ability to understand the history of the Web.

1.1 Web Archiving

Web archiving has roots in a wider digital preservation movement which emerged in the 1980s-1990s, led by memory institutions to develop strategies for addressing the rise of personal computing and the impact of digital artefacts on their abilities to capture and preserve 'records of social phenomena' [61]. This was particularly fuelled by fears over the so-called 'digital dark ages', a term first used by Kuny [41] to describe a scenario where the development pace of technologies (used to produce digital objects) outweighs that of the investment in technologies, infrastructures and policies to preserve them long-term. As the world's information and communication platforms are increasingly born-digital and online, a diverse community of practitioners have positioned web archives as key to capturing and preserving digital cultural heritage, ensuring stability and access to pre-existing web resources and facilitating new knowledge via scholarly research. Web archives in their various forms - including social media archives - have thus become a sort of 'prosthesis' for the Web and a necessary pre-condition for any research into the Web(s) of the past and near-present.²

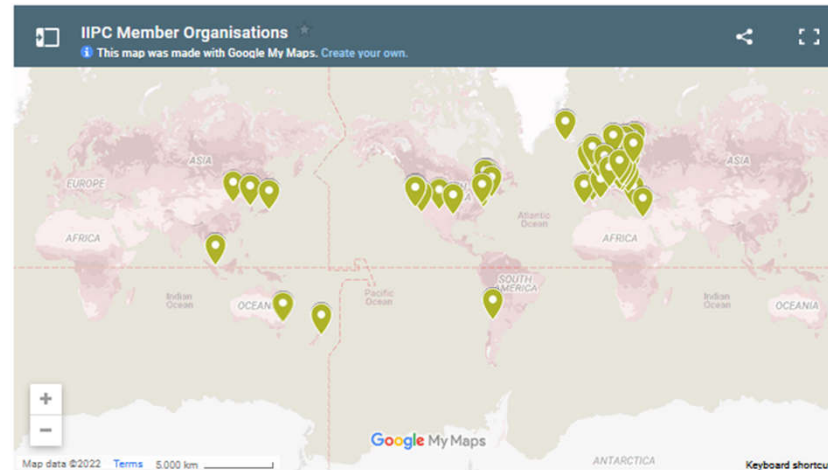
The history of web archiving has been documented to varying degrees in existing overviews [9, 13, 77] which chart the emergence of a field of practice around web archiving. Each have used a series of factors to characterise the domain over time, including: the tools and technologies used, the frequency and scale of selection/collection methods (e.g. broad versus targeted) and the various motivations behind the creation of web archives. These motivations may reinforce and represent, at least in part, a continuation of classical interpretations and analogue conceptions of the value and role of libraries and archives as institutions that provide access to cultural heritage, information and knowledge resources; facilitate evidence-based accountability and promote community memory and identity, amongst others [18, 26].

Web archiving projects have spanned from the large-scale collection of web resources by organisations such as the Internet Archive

²Inspiration for this analogy is taken from Derrida's [28] treatment of 'technological devices for archiving' as prostheses for memory formation and storage.

Jessica Ogden, Susan Halford, Susan and Leslie Carr, 2017, Observing web archives: The case for an ethnographic study of web archiving, In *Proceedings of WebSci' 17, Troy, NY, USA., June 25-28, 2017*. ACM. 10 pp.

IIPC members



ARCHIEFWEB.EU

ARQUIVO.PT

PORTUGUESE WEB ARCHIVE AT FCCN-FCT

BIBLIOTECA DE CATALUNYA

LIBRARY OF CATALONIA

BIBLIOTECA NACIONAL DE CHILE

NATIONAL LIBRARY OF CHILE

BIBLIOTECA NACIONAL DE ESPAÑA

NATIONAL LIBRARY OF SPAIN

BIBLIOTEKA NARODOWA

LOS ALAMOS NATIONAL LABORATORY RESEARCH LIBRARY

NACZELNA DYREKCJA ARCHIWÓW PAŃSTWOWYCH
POLISH NATIONAL ARCHIVE

NACIONALNA I SVEUČILIŠNA KNJIŽNICA U ZAGREBU
NATIONAL AND UNIVERSITY LIBRARY OF CROATIA

NARODNA BIBLIOTEKA SRBIJE
NATIONAL LIBRARY OF SERBIA

NARODNA IN UNIVERZITETNA KNJIŽNICA
NATIONAL AND UNIVERSITY LIBRARY OF SLOVENIA

<https://netpreserve.org/about-us/members/>

Search the history of over 698 billion web pages on the Internet.

Wayback Machine

Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.

Archive News
 Save our Safe Harbor, continued: Internet Archive Supports Libraries and Nonprofits in Submission to the Copyright Office
 June Book Talk: The Catalogue of Shipwrecked Books
 GITCOIN Grants: Donate a Few Tokens, Defend a Public Treasure

Top Collections at the Archive

- American Libraries: 3,561,460 items
- Audio Books & Poetry: 108,279 items
- LibriVox: 16,905 items
- Additional Collections: 20,919,434 items
- Live Music Archive: 247,147 items

INTERNET ARCHIVE | WEB | BOOKS | VIDEO | AUDIO | SOFTWARE | IMAGES

ABOUT | BLOG | PROJECTS | HELP | DONATE | CONTACT | JOBS | VOLUNTEER | PEOPLE

HOME | EXPLORE | LEARN MORE | CONTACT US

The leading web archiving service for collecting and accessing cultural heritage on the web. Built at the Internet Archive.

Welcome to Archive-It! Contact Us

Explore Collections [Show All Collections](#)

- North Africa & the Middle East 2011**
By Internet Archive Global Events
A collection of websites, news coverage, and commentary includes the most recent events in Libya, Egypt and Sudan. Our partners at Library of Congress,...
- Fulbright Scholar and Alumni Archive**
By Fulbright Academy of Science & Technology
The Fulbright Academy of Science and Technology promotes the work of 300,000+ alumni of the Fulbright
- Japan Earthquake 2011**
By Virginia Tech: Crisis, Tragedy, and Recovery Network
This collection depicts the events after the Earthquake and Tsunami in Japan in March 2011. Our partners at Virginia Tech: Crisis, Tragedy, and Recovery...

Explore Collecting Organizations [Show All Organizations](#)

- University of Texas Southwestern Medical Center**
The UT Southwestern Digital Library & Learning Center supports the information
- Florida International University Libraries**
Everglades Explorer is a library, archive and research service with customized
- Temple University**
The Temple University Archives, located within the Libraries' Special Collections Research Center, is the principal repository for and steward of...

SIGN UP | LOG IN | UPLOAD

INTERNET ARCHIVE Explore more than 698 billion web pages saved over time

Wayback Machine

Results: 50 100 500



- Tools**
- Wayback Machine Availability API
 - Chrome Extension
 - Firefox Add-on
 - Safari Extension
 - MS Edge Add-on
 - iOS app
 - Android app

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit [Archive-It](#) to build and browse the collections.

Collection Search

Enter any keyword PDFs

This service is based on indexes of specific data from selected Collections.

Save Page Now

Capture a web page as it appears now for use as a trusted citation in the future.

[FAQ](#) | [Contact Us](#) | [Terms of Service](#) (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.

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

<https://archive.org/> ; <https://web.archive.org/> ; <https://archive-it.org/>



HOME | EXPLORE | LEARN MORE
CONTACT US

The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive



Page Not Found - 404

Sorry, the page you requested was not found.
You may want to try:

- Checking the address for a typo.
- Starting with the navigation links on this page.
- Returning to the previous page.



Archive-It
Built at the **Internet Archive**

2014 Archive-It
The leading web archiving service
for collecting and accessing
cultural heritage on the web

Home

Learn More

Contact Us

About Archive-It
News/Press
Meet the Team
Publications



BRITISH LIBRARY

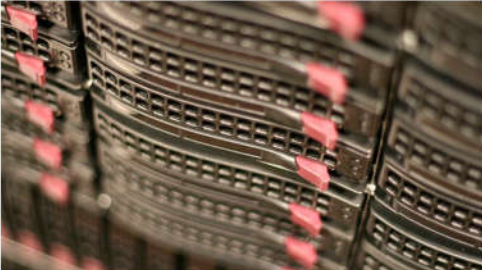
Search our website

Our website Main Catalogue

Catalogues & Collections Discover & Learn What's On Visit Business Support Shop Join

Collection guides

UK Web Archive



Subjects

Digital scholarship
Undertake innovative research with our digital collections and data

Contemporary Britain
Collections reflecting contemporary British society and culture

Content published to the web changes rapidly and is at a high risk of loss. Web Archiving collects, preserves and makes available web resources from the UK domain.

About the collection

The British Library has been collecting websites since 2005, initially on a selective basis and since

DET KGL. BIBLIOTEK

COLLECT AND RETURN NEW USER ABOUT US OPENING HOURS PÅ DANSK LOGIN

FIND MATERIALS INSPIRATION EVENTS SERVICES VISIT US SEARCH

Home / Find materials / Collections / Netarkivet

Collections

Netarkivet

Research access
Collaboration on web archives
About collecting internet material

Netarkivet

We are responsible for collecting and preserving the Danish part of the Internet as part of the Danish Legal Deposit Act. The goal is to ensure the material can be used for future research purposes.

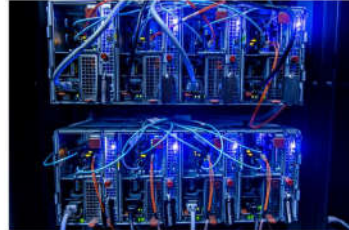


Photo: Thomas Søndergaard

Netarkivet Smurf - N-gram visualisation

Search for words in html pages in Netarkivet for each year. The number of results found is compared with the total number of html pages from that year.


Since 2005, we have collected material from the Danish part of the Internet and preserved it in our web archive. More precisely, this means material published on the Internet in Danish, by Danes or addressed to Danes. The material is part of Denmark's cultural heritage, which the library must preserve for posterity

Web Archive Luxembourg

WHAT WE DO HOW IT WORKS WHAT WE HAVE FAQ CONTACT

Luxembourg Web Archive

Preserving the Luxembourg web for future generations



LIBRARY LIBRARY OF CONGRESS

Web Archives Search our site

Library of Congress - Web Archives - Collections with Web Archives

Library of Congress

Web Archive

Search Web Archives Collections with Web Archives

Featured Content

THE BOSTON ANNIE BARNES AOSTA CORSE The Library of Congress THE MOVEMENT BOOK

Collections with Web Archives

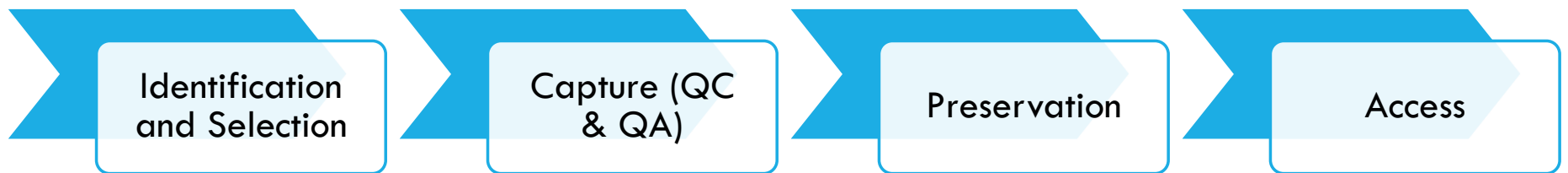
Collection: Afghanistan, Iran, Pakistan, and Tajikistan Election Web Archive

Collection: Afghanistan, Iran, Pakistan and Tajikistan Government Web Archive

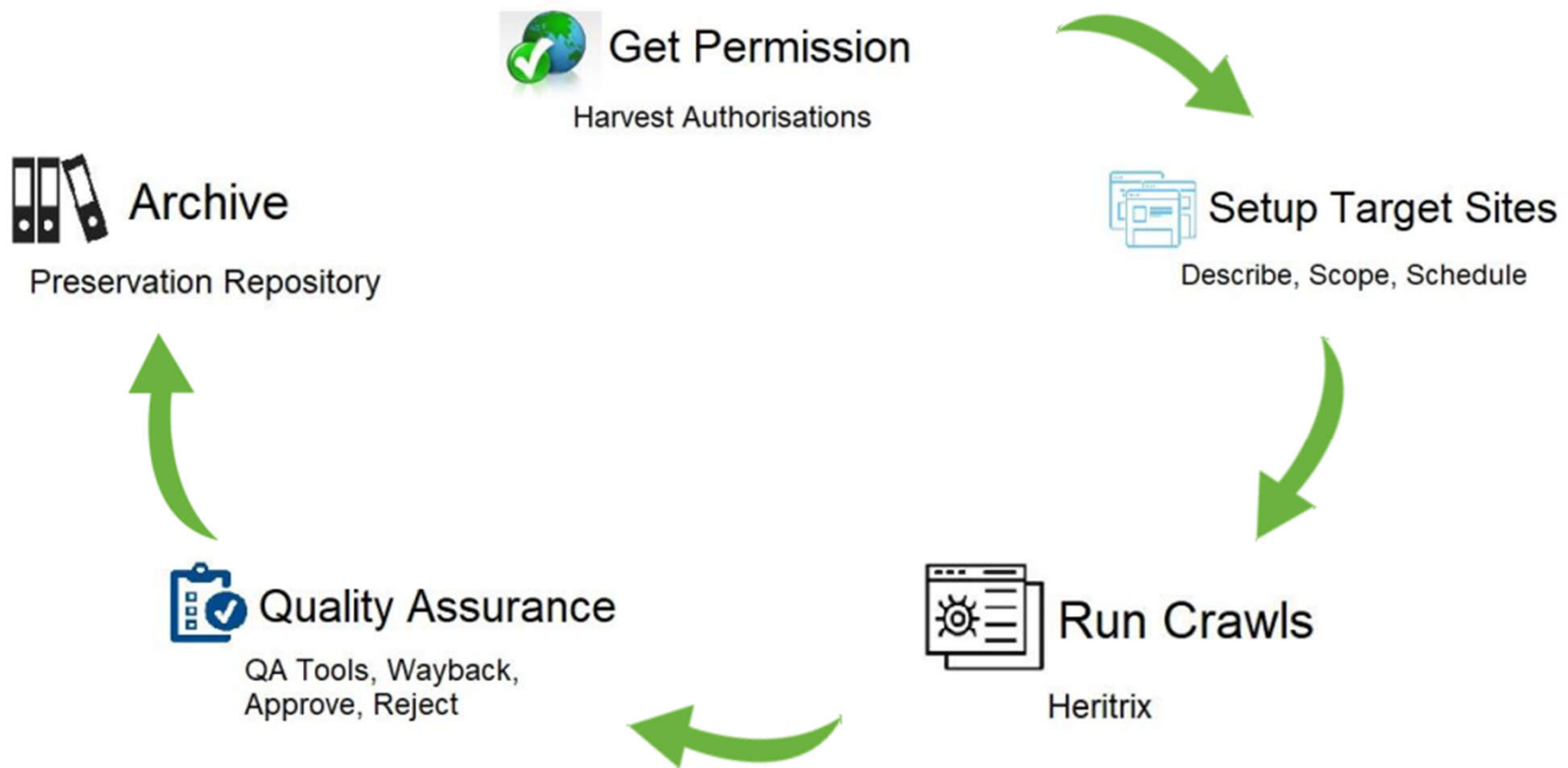
Collection: Afghanistan Web Archive

Collection: African Government Web Archive

- <https://www.webarchive.lu/>
- <https://www.kb.dk/en/find-materials/collections/netarkivet>
- <https://www.bl.uk/collection-guides/uk-web-archive>
- <https://www.loc.gov/web-archives/collections/?st=gallery>



What is the WCT?



Ben O'Brien and Hanna Koppelaar, 2018, Web Curator Tool (WCT) Tutorial, IIPC Web Archiving Conference 2018, Wellington NZ, http://netpreserve.org/ga2018/wp-content/uploads/2018/11/IIPC_WAC2018-Ben_O%E2%80%99Brien_Hanna_Koppelaar-Web_Curator_Tool_Tutorial.pdf

← → ↻ 🏠 <https://github.com/internetarchive/heritrix3/wiki> 80% ☆

🐙 Product Team Enterprise Explore Marketplace Pricing

Search / Sign in Sign up

📄 internetarchive / heritrix3 Public 🔔 Notifications

<> Code Issues 49 Pull requests 7 Actions Projects Wiki Security Insights

Home

AdrthegameDEV edited this page 2 days ago · 4 revisions

This is the public wiki for the Heritrix archival crawler project.

Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project.

Heritrix (sometimes spelled heretrix, or misspelled or mis-said as heratrix/heritix/ heretix/heratix) is an archaic word for heiress (woman who inherits). Since our crawler seeks to collect and preserve the digital artifacts of our culture for the benefit of future researchers and generations, this name seemed apt.

All topical contributions to this wiki (corrections, proposals for new features, new FAQ items, etc.) are welcome! Register using the link near the top-right corner of this page.

Heritrix is designed to respect the <http://www.robotstxt.org/robotstxt.html> and <http://www.robotstxt.org/meta.html>, and collect material at a measured, adaptive pace unlikely to disrupt normal website activity.

If you notice our crawler behaving poorly – The Internet Archive uses archive.org_bot as User Agent when crawling – please send us email at archive-crawler-agent@lists.sourceforge.net.

(If you see a different User-Agent in your logs that still says 'heritrix', it may be someone else using this open-source software. In such a case, even if we can't directly change how your site is crawled, we are happy to help you interpret your logs and identify, contact, or block the source of any troublesome crawling.)

`⁺` The newer wildcard extension to robots.txt is not yet supported (<https://github.com/internetarchive/heritrix3/issues/250>).

The most up to date release packages are the Heritrix 3.x <https://github.com/internetarchive/heritrix3/releases>. These are also available on <https://search.maven.org/search?q=g:org.archive.heritrix>.

Pages 162



Structured Guides:

- [Getting Started with Heritrix](#)
- [Operating Heritrix](#)
- [Configuring Crawl Jobs](#)
- [REST API](#)
- [Glossary](#)

[Wiki index](#)

[FAQs](#)

User Guide

- [Introduction](#)
- [New Features in 3.0 and 3.1](#)
- [Your First Crawl](#)
- [Checkpointing](#)
- [Main Console Page](#)
 - [Main Console Data Elements and Operations](#)
- [Profiles](#)
- [Heritrix Output](#)

<https://github.com/internetarchive/heritrix3/wiki>

APPROACHES TO WEB TARGETTING

Broad Crawls

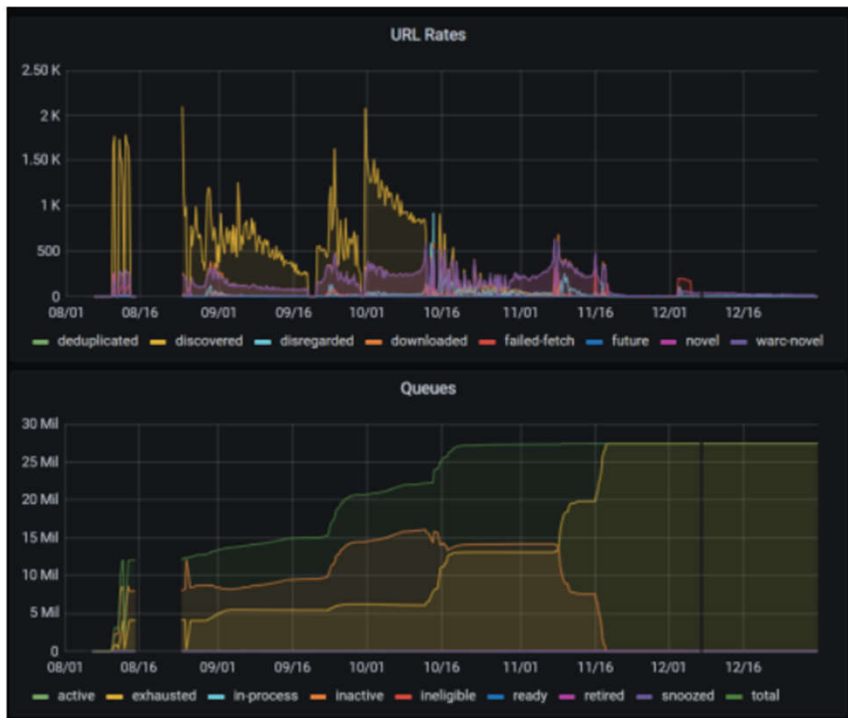
- National domain (top-level domain) crawls
- Websites and website classes that align with aspects of the larger frame

Selective Crawls

- Permission target-based crawls
- Thematic collection/crawls
- Events
- Special Interests
- Urgent and Emerging situations

2021 Domain Crawl


As in 2020, the 2021 Domain Crawl was run on the Amazon Web Services cloud. This time, following improvements to Heritrix and building on prior experience, the crawl ran more smoothly and efficiently than in 2020, using less memory and disk space for the crawl frontier. The crawler was started up early in August for penetration testing, and then taken down while the security concerns were addressed. The actual crawl began on the 24th of August, starting with 10 million seed URLs, and the vast majority of the crawl had completed by mid-November. Most of the 27 million hosts we visited were crawled completely, but ~57,200 hosts did hit the 500MB size cap. However, some of these were content distribution networks (CDNs), i.e. services hosting resources for other sites, so some caps were lifted manually and the crawl was allowed to continue.



<https://blogs.bl.uk/webarchive/2022/01/ukwa-2021-technical-update.html>

The Federal Council > FDHA > NL

Corona Contact Homepage Site map DE FR IT EN

 Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss National Library NL

Search:

Your Research Collections Services Exhibitions and events Publications and research projects Information for professionals About us

Swiss National Library > Information for professionals > e-Helvetica > Websites

[← e-Helvetica](#)

Websites – Web Archive Switzerland

[Websites](#)

[FAQs on web archiving](#)

- Submission form
- Background document, information sheets and glossary
- The film
- Web Archive Switzerland: training
- Information about importing data

The Swiss National Library (NL) has set for itself the objective to establish a collection of websites of patrimonial importance and by doing so will be able to build a long-term repository to preserve and guarantee future access to this significant source of Swiss culture. This collection called Web Archive Switzerland is selective in its approach and can by no means aim for exhaustiveness. The websites are stored in the long-term digital repository of the NL. The task of building this long-term digital repository is accomplished thanks to a partnership between the Swiss National Library and Swiss cantonal libraries and other interested institutions. The cantonal libraries are responsible for the identification and pre-listing of websites while the NL is responsible for the collecting, cataloguing in Helveticat, long-term archiving and providing access to the websites. Cantonal libraries submit their websites to the NL via a submission web form. The NL then harvests these websites and the corresponding bibliographic records are included in Helveticat, the

[top of page](#)


Information about importing data

After submitting the website using the online form and once the snapshot has been successfully archived, a bibliographic record will be generated in Helveticat.

The following documents list the MARC21 fields that appear in the bibliographic records for websites in Helveticat

1. Content taken from the submission web form (first version)
2. Content added automatically (first version)
3. Content that has been adapted over time

 [Web Archive Switzerland: MARC21 fields extracted from the submission web form, following RDA \(in German\) \(PDF, 449 kB, 03.02.2022\)](#)

 [Web Archive Switzerland: MARC21 fields extracted from the submission web form, following AACR2 \(in German\) \(PDF, 458 kB, 03.02.2022\)](#)

More on this topic

[Helveticat](#) 

FAQs on web archiving

Here you find answers to questions concerning the collection, harvesting, archiving and the use of the web archive.

Last modification 03.02.2022

[^ Top of page](#)





WEBSITE

Le site de l'Université populaire du Canton de Genève : UPCGe = Web der Volkshochschule (UP) des Kantons Genf = Sito internet dell'Università Popolare del Canton Ginevra = Website of the Université Populaire of Geneva Canton

Genève : Université populaire du canton de Genève

TOP

DETAILS

LINKS

SEND TO

Details

Title	Le site de l'Université populaire du Canton de Genève : UPCGe = <u>Web</u> der Volkshochschule (UP) des Kantons Genf = Sito internet dell'Università Popolare del Canton Ginevra = Website of the Université Populaire of Geneva Canton
Publication	Genève : Université populaire du canton de Genève
Language	French German Italian English
Other title	Titre espagnol <2012-2013>: Universidad popular del Cantón de Ginebra > Titre portugais <2012-2013>: Universidade popular do Cantão de Genebra > Université populaire du Canton de Genève <2012-2013> >
Contributor	Université populaire du canton de Genève >
Note	Archivé par la Bibliothèque nationale suisse Description faite à partir du site <u>Web</u> (visionné le 23.10.2015); Description n'est pas mise à jour; Anciennes URL: http://www.upcge.ch
Other Note	Texte en: Français, Allemand, Italien, Anglais <2015->
URI	http://permalink.snl.ch/bib/sz991017981710003976 http://permalink.snl.ch/bib/sz001701386

https://nb-helvetica.primo.exlibrisgroup.com/discovery/fulldisplay?docid=alma991017981710003976&context=L&vid=41SNL_51_INST:helvetica&lang=en&search_scope=MyInstitution&adaptor=Local%20Search%20Engine&tab=LibraryCatalog&query=any,contains,web%20archive&offset=0

14,650 results in 10 groups

Sort by Date descending Availability All

Tweak your results

Units

- > 2021
- > 2020
- > 2019
- > 2018
- > 2017
- > 2016
- > 2015
- > 2014
- > 2013
- > 2012

- 1820
- 1825
- 1715
- 1718
- 1704
- 2533
- 2684
- 364
- 146
- 141

Language

Domain

Content type

Web Site restricted access	Web Site restricted access	Web Site restricted access	Web Site restricted access	Web Site restricted access	Web Site restricted access	Web Site restricted access	Web Site restricted access	Web Site restricted access
Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...	Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...	Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...	Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...	Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...	Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...	Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...	Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...	Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...
Publication year 2021	Publication year 2020	Publication year 2019	Publication year 2018	Publication year 2017	Publication year 2016	Publication year 2015	Publication year 2014	
Unit 2021-10-23	Unit 2020-10-23	Unit 2019-10-23	Unit 2018-10-23	Unit 2017-10-23	Unit 2016-12-28	Unit 2015-10-23	Unit 2014-10-23	
Domain https://www.upcge.ch	Domain https://www.upcge.ch	Domain http://www.upcge.ch	Domain http://www.upcge.ch	Domain http://www.upcge.ch	Domain http://www.upcge.ch	Domain http://www.upcge.ch	Domain http://www.upcge.ch	Domain http://www.upcge.ch
Call Number bel-2004887	Call Number bel-1748722	Call Number bel-1422087	Call Number bel-1224138	Call Number bel-1033499	Call Number bel-785244	Call Number bel-508362	Call Number bel-430960	Call Number bel-430960
URN urn:nbn:ch:bel-2004887	URN urn:nbn:ch:bel-1748722	URN urn:nbn:ch:bel-1422087	URN urn:nbn:ch:bel-1224138	URN urn:nbn:ch:bel-1033499	URN urn:nbn:ch:bel-785244	URN urn:nbn:ch:bel-508362	URN urn:nbn:ch:bel-430960	URN urn:nbn:ch:bel-430960
https://upcge.ch/fr/	https://upcge.ch/fr/	http://www.upcge.ch/	http://www.upcge.ch/	http://www.upcge.ch/fr/component1.php/cagallery/5-galerie/detail/180-galerie1.php?url=fr&nbw=1&mb=communeant	http://www.upcge.ch/de/component1.php/cagallery/5-galerie/detail/192-galerie	http://www.upcge.ch/fr/component1.php/cagallery/5-galerie/detail/190-galerie1.php?url=fr&nbw=1&mb=communeant	http://www.upcge.ch/	http://www.upcge.ch/
Wayback access restricted	Wayback access restricted	Wayback access restricted	Wayback access restricted	Wayback access restricted	Wayback access restricted	Wayback access restricted	Wayback access restricted	Wayback access restricted
Open snapshot	Open snapshot	Open snapshot	Open snapshot	Open snapshot	Open snapshot	Open snapshot	Open snapshot	Open snapshot
Web Site restricted access	Web Site restricted access							
Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...	Le site de l'Université populaire du Canton de Genève : UPCGe = Web de...							
Publication year 2013	Publication year 2012							
Unit 2013-10-23	Unit 2012-10-21							
Domain http://www.upcge.ch	Domain http://www.upcge.ch							
Call Number bel-340164	Call Number bel-339983							

https://www.e-helvetica.nb.admin.ch/search?group=bel-272634&sort=ehs_publication_date%20desc

BRITISH LIBRARY

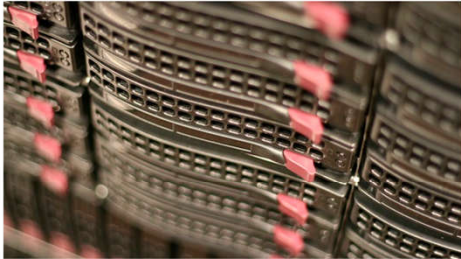
Search our website
Our website Main Catalogue

Catalogues & Collections | Discover & Learn | What's On | Visit | Business Support | Shop

Join

Collection guides

UK Web Archive



Subjects

Digital scholarship
Undertake innovative research with our digital collections and data

Contemporary Britain
Collections reflecting contemporary British society and culture

Content published to the web changes rapidly and is at a high risk of loss. Web Archiving collects, preserves and makes available web resources from the UK domain.

← → ↻ 🏠 🔒 <https://www.webarchive.org.uk> 30% ☆ 🗑️ 📄 📌 ⌵ ☰

UKWA UK Web Archive Home Topics and Themes [Look at a website](#) [About Us](#) [Contact Us](#) Help at the bottom right... 2025/06/22 12:02:07 GMT+0 Language

Search the UK Web Archive

Enter a specific website URL, e.g. www.bbc.com or any word or phrase.


What we do

The UK Web Archive (UKWA) collects millions of websites each year, preserving them for future generations. Use this site to discover old or obsolete versions of UK websites, search the text of the websites and browse websites organised by different topics and themes.


The UKWA is a partnership of the [six UK Legal Deposit Libraries](#).

Topics and Themes


Topics and Themes are groups of websites brought together on a particular theme by librarians, curators and other specialists, often working in collaboration with key organisations in the field.



British Stand-up Comedy Archive
Collection named and administered by Stuart Miller.



French in London
This collection of websites has been selected by Louise Hour Hepton.



News Sites
158 sites are included in this collection.

<https://www.bl.uk/collection-guides/uk-web-archive#> and <https://www.webarchive.org.uk/>

← → ↻ 🏠 https://www.webarchive.org.uk/en/ukwa/search?text=Brexit&search_location=full_text&reset_filters=false&content_type=Web+Page 30% ☆

UKWA
UK WEB ARCHIVE

Home Topics and Themes Save a UK website About Us Contact Us

Clear all filters

Access: Viewable online
Document type (include)

Accessing Content

- Viewable Online (1,178,320)
- At Libraries (6,138,885)

Domain

- politics.co.uk (1,451,661)
- twitter.com (1,125,348)
- sky.com (1,211,448)
-

Document Type

- Web Page (11,736,520)
-

Suffix

- .com (4,888,611)
- .uk (2,079,781)
- .org (1,195,251)
-

Date Collected

From:

To:

Topics and Themes

- IT Collection (1,798,151)
- The Queen's Official Birthday 2016 (1,243,002)
- EU Referendum (1,091,201)
-

Brexit

Enter a specific website URL (e.g. www.ukia) or any word or phrase...

Tips/Notes for using the UK Web Archive

Search results: 11,736,520 results for "Brexit"

Sort by:

1 2 3 4 Next >

No to Brexit | Progress | News and debate from the progressive co...

<http://www.progressivemove.org.uk/2013/10/19/no-to-brexit-from-labour-keep-it-in-europe/>
Progress Magazine Search: No to Brexit Dennis MacShane MEP | Posted on 19 October 2012 | Comments: 3 Web
Date collected: 2013-05-14

Point of View: BREXIT talk is already hurting the UK economy - new

http://www.britishtaxauthority.org/point_of_view_brexit_talk_is_already_hurting_the_uk_economy
European Reform Blog About Home » Blog » Point of View: BREXIT talk is already hurting the UK economy Point
Date collected: 2014-11-26

Progress | News and debate from the progressive community | Tag Arc...

<http://www.progressivemove.org.uk/tag/progress/>
The Progress 150 Club Subscribe to Progress magazine Search: Topics No to Brexit Dennis MacShane MEP
Date collected: 2013-05-11

Cameron plays with fire in risking a 'Brexit' (Leicester City Liber...

<http://leicesterliberals.org.uk/news/articles/2013/05/23/cameron-plays-with-fire-in-risking-a-brexit/?display=Mobile>
in riskin - Brexit January 23, 2013 12:37 PM By Bill Newton Dunn MEP Originally published by East
Date collected: 2013-05-12

Cameron plays with fire in risking a 'Brexit' (Leicester City Liber...

<http://leicesterliberals.org.uk/news/articles/2013/05/23/cameron-plays-with-fire-in-risking-a-brexit/?display=Accessible>
Search: Cameron plays with fire in riskin - Brexit January 23, 2013 12:37 PM By Bill Newton Dunn MEP
Date collected: 2013-05-12

Cameron plays with fire in risking a 'Brexit' (Rutland & Melton Lib...

<http://rutlandliberals.org.uk/news/articles/2013/05/23/cameron-plays-with-fire-in-risking-a-brexit/?display=Mobile>
European Group Links Keep in Touch: Your Views Join The Party Cameron plays with fire in riskin - Brexit
Date collected: 2013-05-02

Cameron plays with fire in risking a 'Brexit' (Rutland & Melton Lib...

<http://meltonliberals.org.uk/news/articles/2013/05/23/cameron-plays-with-fire-in-risking-a-brexit/?display=Accessible>
Brexit January 23, 2013 12:37 PM By Bill Newton Dunn MEP Originally published by Law Mollamb Liberal
Date collected: 2013-05-02

The Freedom Association has been shortlisted for the Brexit Prize |...

<http://www.faf.net/2014/03/26/the-freedom-association-has-been-shortlisted-for-the-brexit-prize/>
Free is not in the interests of freedom: The Freedom Association has been shortlisted for the Brexit
Date collected: 2014-11-26

What is available in our Reading Rooms?

Millions of websites collected under Legal Deposit. There are two British Library locations (London and Boston Spa in Yorkshire), six others in the UK and one in Ireland that have access to the UK Web Archive collection. You can view these resources at the six UK Legal Deposit Libraries (in 9 locations).

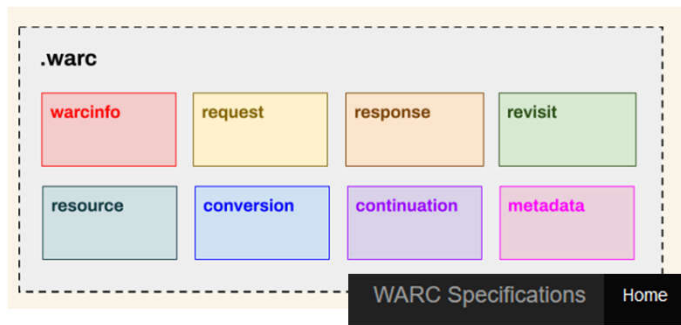
Locations: The British Library (St. Pancras and Boston Spa), National Library of Wales (Aberystwyth and Cardiff), National Library of Scotland (Edinburgh and Glasgow), Bodleian Library (Oxford), Cambridge University Library, and the Library of Trinity College, Dublin.

Curated collections: There are over 100 curated collections available in the 'Topics and Themes' section of the UK Web Archive website.

You will need to use a Library PC to access the web archive collection.

THIS IS A DOMAIN OF EXPERIMENTATION AND LEARNING

- What is the process of collecting websites/webpages?
- How are the results stored?
- How do we preserve collected web materials in the long term?
- What are the options and ways in which web access be provided?
- What kinds of tools do Users of web archives need?



About

Title
 — The WARC Format 1.0 standard

Latest version
 — [See version 1.1](#)

Previous version
 — None

Issues
 — [View issues on GitHub](#)

Contents

- 1 Scope
- 2 Normative references
- 3 Terms, definitions and acronyms
 - 3.1 Terms and definitions
 - 3.1.1 WARC record
 - 3.1.2 WARC record content block
 - 3.1.3 WARC record payload
 - 3.1.4 WARC record header
 - 3.1.5 WARC named fields
 - 3.1.6 WARC logical record
 - 3.2 Acronyms
- 4 File and record model
- 5 Named fields
 - 5.1 General
 - 5.2 WARC-Record-ID (mandatory)
 - 5.3 Content-Length (mandatory)
 - 5.4 WARC-Date (mandatory)
 - 5.5 WARC-Type (mandatory)
 - 5.6 Content-Type

The WARC Format 1.0

Web sites and web pages emerge and disappear from the world wide web every day. For the past ten years, memory organizations have tried to find the most appropriate ways to collect and keep track of this vast quantity of important material using web-scale tools such as web crawlers. A web crawler is a program that browses the web in an automated manner according to a set of policies; starting with a list of URLs, it saves each page identified by a URL, finds all the hyperlinks in the page (e. g. links to other pages, images, videos, scripting or style instructions, etc.), and adds them to the list of URLs to visit recursively. Storing and managing the billions of saved web page objects itself presents a challenge.

At the same time, those same organizations have a rising need to archive large numbers of digital files not necessarily captured from the web (e.g., entire series of electronic journals, or data generated by environmental sensing equipment). A general requirement that appears to be emerging is for a container format that permits one file simply and safely to carry a very large number of constituent data objects for the purpose of storage, management, and exchange. Those data objects (or resources) must be of unrestricted type (including many binary types for audio, CAD, compressed files, etc.), but fortunately the container needs only minimal knowledge of the nature of the objects.

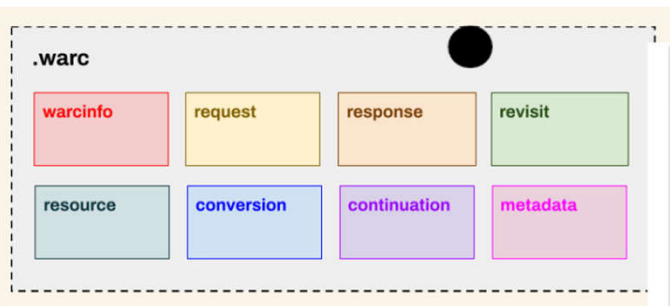
The WARC (Web ARChive) file format offers a convention for concatenating multiple resource records (data objects), each consisting of a set of simple text headers and an arbitrary data block into one long file. The WARC format is an extension of the ARC File Format [ARC] that has traditionally been used to store “web crawls” as sequences of content blocks harvested from the World Wide Web. Each capture in an ARC file is preceded by a one-line header that very briefly describes the harvested content and its length. This is directly followed by the retrieval protocol response messages and content. The original ARC format file is used by the Internet Archive (IA) since 1996 for managing billions of objects, and by several national libraries.

The motivation to extend the ARC format arose from the discussion and experiences of the International Internet Preservation Consortium (IIPC), whose members include the national libraries of Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, Sweden, The British Library (UK), The Library of Congress (USA), and the Internet Archive (IA). The California Digital Library and the Los Alamos National Laboratory also provided input on extending and generalizing the format.

The WARC format is expected to be a standard way to structure, manage and store billions of resources collected from the web and elsewhere. It will be used to build applications for harvesting (such as the open source Heritrix web crawler), managing, accessing, and exchanging content. The way WARC files will be created and resources will be stored and rendered will depend on software and applications implementations.

Besides the primary content recorded in ARCs, the extended WARC format accommodates related secondary

<https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/>



187 lines (163 sloc) | 5.5 KB

```

1  WARC/1.0
2  WARC-Type: warcinfo
3  WARC-Record-ID: <urn:uuid:fb6cf0a-6160-4550-b343-12188dc05234>
4  WARC-Date: 2014-01-03T03:03:22Z
5  Content-Length: 196
6  Content-Type: application/warc-fields
7  WARC-Filename: live-20140103030321-wwwb-app5.us.archive.org.warc.gz
8
9  software: LiveWeb Warc Writer 1.0
10 host: wwwb-app5.us.archive.org
11 isPartOf: liveweb
12 format: WARC file version 1.0
13 conformsTo: http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf
14
15
16 WARC/1.0
17 WARC-Type: response
18 WARC-Record-ID: <urn:uuid:6d058047-ede2-4a13-be79-90c17c631dd4>
19 WARC-Date: 2014-01-03T03:03:21Z
20 Content-Length: 1610
21 Content-Type: application/http; msgtype=response
22 WARC-Payload-Digest: sha1:B2LTWWPUOYAH7UIPQ7ZUPQ4VMBSVC36A
23 WARC-Target-URI: http://example.com?example=1
24 WARC-Warcinfo-ID: <urn:uuid:fb6cf0a-6160-4550-b343-12188dc05234>
25
26 HTTP/1.1 200 OK
27 Accept-Ranges: bytes
28 Cache-Control: max-age=604800
29 Content-Type: text/html
30 Date: Fri, 03 Jan 2014 03:03:21 GMT
31 Etag: "359670651"

```

From example at:

https://github.com/webrecorder/pywb/blob/main/sample_archive/warcs/example.warc



Web Archiving

Background

- Web archiving is conducted under the Library and Archives of Canada Act, section 8 (2) (sampling from the Internet for digital preservation purposes). Library and Archives Canada's (LAC) latest policy instruments recognize web-based resources as unique, born-digital documentary heritage. Collecting and preserving web resources ensures future access and research use.
- The Web Archiving Program began at LAC in December 2005 and has been an ongoing operational activity since 2013.
- Web archiving is a digital preservation discipline and is practiced by over 50 international memory institutions, mostly national libraries. The field is advanced primarily by the International Internet Preservation Consortium (IIPC) of which LAC is a founding member. In 2019, Sylvain Bélanger is serving as the Treasurer and a member of the Steering Committee.
- LAC employs a robust methodology for collecting web resources and social media, which includes comprehensive crawls of the Government of Canada (GC) web presence; curating thematic research collections (e.g., Centenary of the First World War, Canada 150, Federal Elections, Olympic and Paralympic Games); documenting important events in Canadian history as they unfold (e.g., Humboldt Broncos junior hockey team bus accident, forest fires in western Canada); engaging in "rescue" or preservation archiving of resources at known risk (e.g., the website of the National Inquiry into Missing and Murdered Indigenous Women and Girls); and supplementing other library and archival collections with web holdings, in collaboration with other internal and external experts (e.g., Truth and Reconciliation Web Archive).

Considerations

- Currently, no public access is available for LAC's non-federal web holdings, which comprise 50% of the total collections (30 terabytes). Funding to develop additional services and a comprehensive access portal is being proposed for the Central Agency Funding Request for Digital Optimization.

Key Public Messages

- LAC's web archiving methodology includes five main activities: 1. Domain crawls of the GC 2. Curation of thematic web and social media collections 3. Event-based crawling 4. Preservation archiving of resources at known risk and 5. Supplementing library collections or archival fonds with web holdings.
- The collection currently comprises nearly 1.5 billion digital objects and 60 terabytes of data. As of 2016, web archival holdings accrue at a minimum rate of 13 terabytes per fiscal year.

<https://www.bac-lac.gc.ca/eng/transparency/briefing/2019-transition-material/Pages/digital-web-archiving.aspx>

Library and Archives Canada (LAC/BAC)

- Introduction: <https://www.bac-lac.gc.ca/eng/about-us/about-collection/Pages/web-social-media-archiving.aspx>
- Web Archive Service:
 - The Government of Canada Web Archive is currently April 2022 offline: <https://www.bac-lac.gc.ca/eng/discover/archives-web-government/Pages/web-archives.aspx>
 - There are also thematic collections, such as the COVID-19 material: <https://www.bac-lac.gc.ca/eng/about-us/about-collection/Pages/documenting-2020-covid-19-pandemic.aspx>

The screenshot shows the Library and Archives Canada website. At the top, there is a navigation bar with the Government of Canada logo and the text "Government of Canada" and "Gouvernement du Canada". To the right, there are links for "Canada.ca", "Services", "Departments", and "Français". Below this, the main header features the "Library and Archives Canada" logo and a large red maple leaf. A search bar is visible with the text "Search BAC-LAC.gc.ca" and a "Search" button. Below the search bar, there are four menu items: "Discover the Collection", "Search the Collection", "Services for the Public", and "Services and programs". A breadcrumb trail reads "Home → Discover the Collection → Government of Canada Web Archive". A light blue banner at the top of the main content area contains the text "Renewing our web presence" and a close button. Below this, the main heading is "Government of Canada Web Archive". A yellow attention box with a warning icon contains the text: "Attention Please note that the Government of Canada Web Archive is currently not available. We apologize for the inconvenience that this may cause." At the bottom left, there is a "Share this page" button. At the bottom right, the text "Date modified: 2020-04-14" is displayed.

Library and Archives Canada (LAC/BAC)

- There are also thematic collections, such as the COVID-19 material: <https://www.bac-lac.gc.ca/eng/about-us/about-collection/Pages/documenting-2020-covid-19-pandemic.aspx>



The screenshot shows the Library and Archives Canada website. At the top, there is a navigation bar with the Government of Canada logo and the text "Government of Canada" and "Gouvernement du Canada". To the right, there are links for "Canada.ca", "Services", "Departments", and "Français". Below this, the main header features the "Library and Archives Canada" logo and a search bar with the text "Search BAC-LAC.gc.ca". A navigation menu includes "Discover the Collection", "Search the Collection", "Services for the Public", and "Services and programs". The breadcrumb trail reads "Home → About Us → About the Collection → Documenting the COVID-19 Pandemic". A light blue banner at the top of the main content area says "Renewing our web presence" with a close button. The main heading is "Documenting the COVID-19 Pandemic". Below this is the "Web and Social Media Preservation Program" section, followed by the "Coronavirus/COVID-19 Collection (Feb 2020–)" section. The text in this section states: "More than ever before, web archiving since 2020 has emerged internationally as a rapid-response means of documenting a crisis. The COVID-19 pandemic demonstrated that web archiving is one of the few immediate actions that information professionals and digital librarians and archivists can take to preserve a historical timeline and the primary resources about an extended crisis." It continues: "From the beginning of the COVID-19 pandemic in early 2020, Library and Archives Canada's (LAC) Web Archiving and Social Media Program team was fully engaged in documenting the evolution of the situation and its effects on Canadian society. The team curated a diverse collection that includes websites from government and non-government sources, as well as social media relating to the pandemic's impact on life in Canada." The "COVID-19 collection scope, priorities and highlights:" section lists the following items:

- French and English news media (daily newspaper crawls and targeted content)
- Public health information from all levels of government (federal, provincial and territorial government resources with a focus on public health communications)
- Impact on business and the economy (for example, corporate sites for affected industries)
- Health, science and medicine (for example, information about research efforts)
- Sites focused on social and cultural aspects, including religion, artistic and cultural expression, and impacts on families, children and education

Archives de l'internet

La BnF assure le dépôt légal de l'internet français. Sa collection de sites archivés, qui est parmi les plus anciennes et les plus riches dans le monde, est ouverte à toute personne justifiant d'une recherche.

DÉCOUVRIR

Les archives de l'internet conservées à la BnF représentent à ce jour plus d'1 pétaoctet de données. Les toutes premières collections, constituées à titre expérimental et par l'apport d'Internet Archive, remontent à 1996.

L'archivage du web s'inscrit depuis 2006 dans le cadre de la mission de dépôt légal de la BnF. Il porte sur le domaine français, c'est-à-dire les sites enregistrés en .fr, sous une extension liée au territoire national (.re, ou .bzh par exemple), ou sous extension générique (.com ou .org par exemple) à la condition qu'ils soient produits en France ou que leur auteur y soit domicilié.

Les collectes sont réalisées à l'aide d'un robot-logiciel qui explore les sites comme le ferait un internaute, en copiant à mesure de sa progression tous les éléments constitutifs des pages: textes, images, fichiers audio et vidéo, animations, feuille de style et liens. La collecte ne prétend pas à l'exhaustivité mais repose sur un principe de représentativité. La BnF conjugue à cet effet deux modes de collecte.

LA COLLECTE « LARGE »

Réalisée une fois par an, l'objectif de cette collecte est d'avoir un échantillon du plus grand nombre de sites possibles. La liste de ces sites lui est communiquée par des bureaux d'enregistrement partenaires, tels que l'Association française pour le nommage de l'internet en coopération (Afnic) et



- Découvrir
- Explorer
- Contribuer
- Contact
- Ressources

<https://www.bnf.fr/fr/archives-de-linternet>



Welcome to Archive-It!

Contact Us

Explore Collections

[Show All Collections](#)



Maryland State Document Collection

By University of Maryland

This collection contains material created by the State of Maryland related to state planning.



IT History Society

By IT History Society

The IT History Society has created this comprehensive archive of IT websites which is a valuable resource for historians, archivists and the general...



International Whistleblower Archive

By International Whistleblower Archive

The purpose of this collection is to provide documented information about whistleblowers and the act of whistleblowing in the United States and...

Explore Collecting Organizations

[Show All Organizations](#)



IT Historical Resource Sites

Collected by: [IT History Society](#)

Archived since: Feb, 2010

Description: The IT History Society (ITHS) is a world-wide group of over 500 members working together to assist in and promote the documentation, preservation, cataloging, and researching of Information Technology (IT) history. We offer a place where individuals, academicians, corporate archivists, curators of public institutions, and hobbyists alike can gather and share information and resources. This catalog of resource sites concerning IT history is the only one of its kind and is a valuable resource for IT historians and archivists alike.

Subject: [Computers & Technology](#)

Enter a search term on the right to search the text within the archived pages. Or for more search options, use the Advanced Search options below.

Advanced Search

Contains **all** of:

Exact phrase:

Not containing:

From the Host:

Results per host:

File format:

Capture date range:

From:

To:

[Advanced Search](#)

[Help with Search](#)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

[Search](#)

[Clear](#)

The following results were found for the term(s): **digital curation**

- No metadata results for **digital curation**, but there are up to 4586 matches within the page text.

Search Page Text

Page 1 of 230 (4,586 Total Results)

[Next Page](#)

Sort By: **Best Match**

Digital Curation | Smithsonian Institution Archives

URL: <https://siarchives.si.edu/what-we-do/digital-curation>

This text was captured on Oct 03, 2017 [Show All Captures](#)

Digital Preservation Often confused with the broader scope of **digital curation**, **digital** preservation is focused on the task of ensuring that **digital** collections are accessible to the public in the future

Content: text/html Size: 57 KB

[More Results from siarchives.si.edu](#)

DBLP: Digital Curation Curriculum Conference

URL: <http://dblp.uni-trier.de/db/conf/digcurv/>

This text was captured on Jan 03, 2016 [Show All Captures](#)

Maurizio Lunghi, Vittore Casarosa: Proceedings of the Framing the **Digital Curation** Curriculum Conference, Florence, Italy, 6-7 May, 2013.

Content: text/html Size: 4.0 KB

[More Results from dblp.uni-trier.de](#)

dblp: Digital Curation Curriculum Conference 2013

<https://archive-it.org/>

WEB ARCHIVING



Image: RaHul Rodriguez
<https://creativecommons.org/licenses/by-sa/2.0/>

The NYARC web resources program archives, preserves, and provides online public access to curated collections of websites in areas that correspond to the scope and strengths of the print collections at each research library, as well as to NYARC project websites and the institutional websites of the three museums. The initial phase of the NYARC web resources program was made possible through funding provided by The Andrew W. Mellon Foundation. For more information or to nominate websites for inclusion in NYARC's web archive collections, please email: webarchive@frick.org

SEARCH NYARC'S WEB ARCHIVE

Is your site archived by the NYARC Web Archiving Program? If so, let your audience know and put the ["Archived by NYARC" logo](#) on your homepage!

To learn more about our program, please see [Frequently Asked Questions: Web Archiving](#).

<https://nyarc.org/initiatives/web-archiving>



New York Art Resources Consortium (NYARC)

Archive-It Partner Since: Oct, 2010
Organization Type: [Museums & Art Libraries](#)
Organization URL: <http://www.nyarc.org>

Description: The New York Art Resources Consortium (NYARC) consists of the research libraries of three leading art museums in New York City: The Brooklyn Museum, The Frick Collection, and The Museum of Modern Art. The NYARC web resources program archives curated collections of websites in areas which correspond to the scope and strengths of the print collections at each research library, as well as NYARC project websites and the institutional websites of the three museums. To nominate websites for inclusion in NYARC's collections, please email: [webarchive\[at\]frick\[dot\]org](mailto:webarchive[at]frick[dot]org)

Narrow Your Results

Subject Sort By: Count | (A-Z)

- Arts & Humanities (10)
- Society & Culture (6)
- Universities & Libraries (1)

Sites and collections from this organization are listed below. Narrow your results at left, or enter a search query below to find a collection, site, specific URL or to search the text of archived webpages.

- Collections
- Sites
- Search Page Text

Page 1 of 1 (10 Total Results)

Sort By: [Collection Name \(A-Z\)](#) | [Collection Name \(Z-A\)](#)

Art Resources

Archived since: Mar, 2014

Description: Art-rich websites of significance to the study of art and art history, especially those at risk of disappearance from the live web.

Subject: [Arts & Humanities](#)

Artists' Websites

Archived since: Apr, 2014

Description: Websites of artists significant to the collections of the Brooklyn Museum, The Museum of Modern Art, and The Frick Collection.

Subject: [Arts & Humanities](#), [Society & Culture](#)

Auction Houses

Archived since: Oct, 2010

Description: Auction houses specializing in sales of art. Includes catalogs and price results.

Subject: [Arts & Humanities](#), [Society & Culture](#)

Brooklyn Museum

Archived since: Feb, 2014

Description: Brooklyn Museum's website features an extensive exhibition archive and rich video content with artist interviews, public lectures, and discussions with curators and scholars.

Subject: [Arts & Humanities](#), [Society & Culture](#)

Catalogues Raisonnés

Archived since: May, 2014

Description: Born-digital catalogues raisonnés (scholarly compilations of an artist's work)

Subject: [Arts & Humanities](#)

Museum of Modern Art

Archived since: Feb, 2014

Description: Archived versions of MoMA's websites (including MoMA collection records, exhibition sites, MoMA Magazine, and POST blog)

Subject: [Arts & Humanities](#)

New York Art Resources Consortium (NYARC)

Archived since: Mar, 2014

Description: The New York Art Resources Consortium (NYARC) consists of the research libraries of three leading art museums in New York City: The Brooklyn Museum, The Frick Collection, and The Museum of Modern Art. With funding from The Andrew W. Mellon Foundation, NYARC was formed in 2006 to facilitate collaboration that results in enhanced resources to research communities.

Subject: [Arts & Humanities](#)

New York City Galleries

Archived since: Sep, 2014

Description: Galleries and art dealers based in New York City

Subject: [Arts & Humanities](#), [Society & Culture](#)

Restitution of Lost or Looted Art

Archived since: Oct, 2014

Description: Sites dedicated to restitution efforts and provenance of artworks that may be lost, stolen or looted.

Subject: [Arts & Humanities](#), [Society & Culture](#)

The Frick Collection

Archived since: Jan, 2014

Description: The Frick Collection is an internationally recognized museum and research center, known for Old Master paintings, European sculpture and decorative arts.

Subject: [Arts & Humanities](#), [Society & Culture](#), [Universities & Libraries](#)

Page 1 of 1 (10 Total Results)

<https://www.archive-it.org/organizations/484>

Enter a search term on the right to search the text within the archived pages. Or for more search options, use the Advanced Search options below.

Advanced Search

Contains all of:

Exact phrase:

Not containing:

From the Host:
ex. www.archive-it.org

Results per host:
1 (default)

File format:
All formats

Capture date range:
From: To:

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Mohammad Ali Taraghjiah

The following results were found for the term(s): **Mohammad Ali Taraghjiah**

• No metadata results for **Mohammad Ali Taraghjiah**, but there are up to 1 matches within the page text.

Search Page Text

Page 1 of 1 (1 Total Results)

Sort By: [Best Match](#)

Whole Sale Catalogue
URL: <http://www.bonhams.com/cgi-bin/public.sh/pubweb/publicSite.r?screen=WholeCatalogue&SaleNo=16393>
This text was captured on Oct 08, 2010. [Show All Captures](#)
Ali Taraghjiah (Iran, b. 1943) Unlimited \$15,000 to 20,000 94 Abbas Kiarostami (Iran, b. 1940) Show White Series \$25,000 to 35,000 95 Rokni Haerizadeh (Iran, b. 1978) Spring Burst \$18,000 to 24,000 96
Consent: [text.html Size: 29 KB](#)
[More Results from www.bonhams.com](#)

Page 1 of 1 (1 Total Results)

You are viewing an archived web page, collected at the request of [New York Art Resources Consortium \(NYARC\)](#) using [Archive-It](#). This page was captured on 06:59:35 Oct 08, 2010, and is part of the [Auction Houses](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. Found 0 archived media items out of 0 total on this page.

[Print Sale Lots](#)
[Close Screen](#)

Bonhams¹⁷⁹³

Sale 16393 - Modern & Contemporary Arab, Iranian, Indian & Pakistani Art
Royal Mirage, Dubai

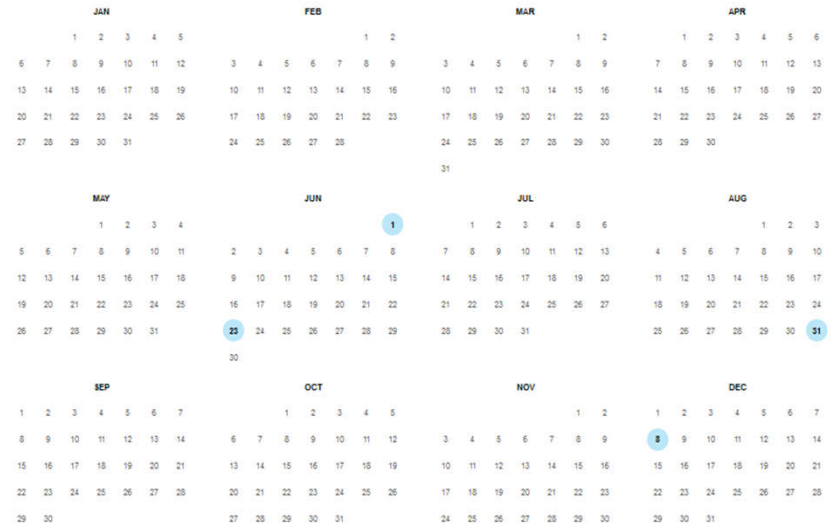
Lot	Description	Estimate
1	Youssef Kamel (Egypt, 1890-1971) Cairo street scene	\$4,000 to 6,000
2	Saliba Douailhy (Lebanon, 1912-1994) Landscape	\$4,000 to 6,000
3	Nasser Chaura (Syria, 1920-1992) Landscape	\$12,000 to 15,000
4	Moustapha Farroukh (Lebanon, 1902-57) Street Scene	\$3,000 to 5,000
5	Cesar Gemayel (Lebanon, 1898-1958)	\$15,000 to 25,000
6	Sohrab Sepehri (Iran, 1928-1980) Kashan	\$60,000 to 80,000
7	Iran Darroudi (Iran, b. 1936) Eshgh Khamoush Shodeh	\$30,000 to 50,000
8	Abdur Rahman Chughtai (Pakistan, 1897-1975) Maiden contemplating moths at a flame	\$45,000 to 55,000
9	Ustad Allah Bux (Pakistan, 1895-1978) Sohni and Mahinwal	\$8,000 to 12,000
10	Nazem Al Jaafari (Syria, b. 1918) From Jabal Al-Arab	\$25,000 to 28,000

- Group: Exhibition History
- Title: Best Film Not Playing at a Theater Near You: 2009
URL: <http://www.moma.org/calendar/exhibitions/1011/>
Loading Wayback Capture Info...
Group: Exhibition History
- Title: MoMA Starts: An 80th Anniversary Exhibition
URL: <http://www.moma.org/calendar/exhibitions/1012/>
Loading Wayback Capture Info...
Group: Exhibition History
- Title: The New Typography
URL: <http://www.moma.org/calendar/exhibitions/1013/>
Loading Wayback Capture Info...
Group: Exhibition History
- Title: Performance 7: Mirage by Joan Jonas
URL: <http://www.moma.org/calendar/exhibitions/1014/>
Loading Wayback Capture Info...
Group: Exhibition History
- Title: Tim Burton and the Lurid Beauty of Monsters
URL: <http://www.moma.org/calendar/exhibitions/1015/>
Loading Wayback Capture Info...
Group: Exhibition History
- Title: Cool Men in a Golden Age: Alfred Leslie and F
URL: <http://www.moma.org/calendar/exhibitions/1016/>
Loading Wayback Capture Info...
Group: Exhibition History
- Title: Lithuanian Cinema: 1990–2009
URL: <http://www.moma.org/calendar/exhibitions/1017/>
Loading Wayback Capture Info...
Group: Exhibition History

https://wayback.archive-it.org/4387*/http://www.moma.org/calendar/exhibitions/1013/ 67%

Museum of Modern Art, archived by New York Art Resources Consortium (NYARC)

Searched for <http://www.moma.org/calendar/exhibitions/1013/> RSS CDX
Saved 11 times between January 12, 2017 and December 8, 2019.



This calendar view maps the number of captures of <http://www.moma.org/calendar/exhibitions/1013/>, not how many times the site was actually updated.

COVID-19 — We are open under the Orange COVID-19 traffic light setting, with safety measures in place. [Find out more](#)

Reading room hours changed — Reading rooms remain closed on Mondays. This is a temporary measure to help us manage the impact of COVID-19. [Find out more](#)

Events About

Home > Collections > A-Z of our collections > New Zealand Web Archive

New Zealand Web Archive

The New Zealand Web Archive is a collection of archived New Zealand and Pacific websites. Use the web archive to see a visual history of how websites have changed over time. The web archive is part of the Alexander Turnbull Library collections.

On this page

- What's the New Zealand Web Archive?
- Access items in the New Zealand Web Archive
- Strengths of the web archives collection
- Nominate a site
- Related resources

Caring for your collections +

A-Z of our collections -

- Alexander Turnbull Library Collections
- Archive of New Zealand Music
- Arthur Nelson Field Collection
- AtoJs Online
- Carling Collection
- Cartographic Collection
- Corelli Collection
- Danish Collection
- Dorothy Neal White Collection
- Drawings, Paintings, and Prints
- Earp Collection

<https://natlib.govt.nz/collections/a-z/new-zealand-web-archive/>

Toggle Aotearoa

New search Journal Search Collection Discovery Help

COLLECTIONS /

Canterbury Earthquake = Te Rū o Waitaha

Websites included are those that focus on the September 4, 2010 and February 22, 2011 earthquakes and their immediate aftermath. Includes personal accounts and government advice on emergency housing, insurance claims, legal advice and support packages. Ko ngā paetukutuku e aro ana ki ngā rū i pā mai i te 4 o Hepetema, 2010 me te 22 o Pepuere, 2011 me ngā āhuatanga i muri tonu mai. Kei roto ko ngā kōrero ake a te tangata me ngā tohutohu a te kāwanatanga mō ngā whare ohotata, ngā kerēme inihua, ngā tohutohu ture me ngā kaupapa tautoko.

Sort items by Relevance Search

Items in this collection (88)

WEBSITE Four paws and whiskers .	WEBSITE Sounds for the South : [web site].	WEBSITE 2010 Canterbury earthquake : from Wikipedia, the free encyclopedia.	WEBSITE 12:51pm : Christchurch, one year after February 22, 2011.	WEBSITE Lyttelton Harbour Information Centre : providing quality local & visitor information.
WEBSITE Landcheck.org.nz .	WEBSITE EQ-IQ : Earthquake Commission.	WEBSITE Emergency housing help .	WEBSITE Christchurch says thanks : share	WEBSITE New Zealand Law Society .

<https://natlib-primo.hosted.exlibrisgroup.com/primo-explore/collectionDiscovery?vid=NLNZ&collectionId=81279506680002836>

New search Journal Search Collection Discovery Help

Canterbury Earthquake = Te Rū o Waitaha

Websites included are those that focus on the September 4, 2010 and February 22, 2011 earthquakes and their immediate aftermath. Includes personal accounts and government advice on emergency housing, insurance claims, legal advice and support packages. Ko nga paetukutuku e aro ana ki nga ru i pa mai i te 4 o Hепetema, 2010 me te 22 o Pepuere, 2011 me nga ahuatanga i muri tonu mai. Kei roto ko nga korero ake a te tangata me nga tohutohu a te kawanatanga mo nga whare ohotata, nga kereme inihua, nga tohutohu ture me nga kaupapa tautoko.

Sort items by: Relevance Search

Items in this collection (88)

- WEBSITE Four paws and whiskers .
- WEBSITE Seeds for the South : Jamb Stoj.
- WEBSITE 2010 Canterbury earthquake : from Wikijerida, the free encyclopedia.
- WEBSITE 12:23pm - Christchurch, one year after February 22, 2011.
- WEBSITE Lyttelton Harbour Information Centre: providing quality local & visitor information.
- WEBSITE Landcheck.org.nz.
- WEBSITE EQ IQ: Earthquake Commission.
- WEBSITE Emergency housing help .
- WEBSITE Christchurch says thanks : share
- WEBSITE New Zealand Law Society .

WEBSITE

Four paws and whiskers .

Fi
Christchurch, N.Z. : Fi, [2008]-

OPEN ACCESS

Online access >

Top

Online access


Send to

Details

Links

Online access

Sign in for request options Sign in



Four paws and whiskers
Fi.
Fi,[2008]-

Year
ALL v

Volume
ALL v

2011 03 06 available at: [National Digital Heritage Archive](#)
Open Access

https://natlib-primo.hosted.exlibrisgroup.com/primo-explore/fulldisplay?docid=NLNZ_ALMA11269970330002836&context=L&vid=NLNZ&lang=en_US&adaptor=Local%20Search%20Engine

← → ↻ 🏠 <https://ndhadeliver.natlib.govt.nz/webarchive/20110306153112/http://fourpawsandwhiskers.blogspot.com/> ☆ ☰

Te Puna Māhanga o Aotearoa
NATIONAL LIBRARY
of New Zealand

Four Paws and Whiskers
Sun, 06 Mar 2011 02:41:11 GMT

🔍 Share Report Abuse Next Blog»



four paws and whiskers



Home


LOADED WEB
See blogs and businesses for New Zealand

LOADED WEB
LOADED WEB

FOLLOWERS


MARCH 6, 2011

Cordon reduced in CBD



MY PICTURES FROM THIS BLOG - JUST CLICK ON ONE TO SEE THEM ALL AT PICASA

FI, FROM FOUR PAWS AND WHISKERS



Christchurch, New Zealand

[View my complete profile](#)

SUBSCRIBE TO

📧 Posts ▾

📧 All Comments ▾

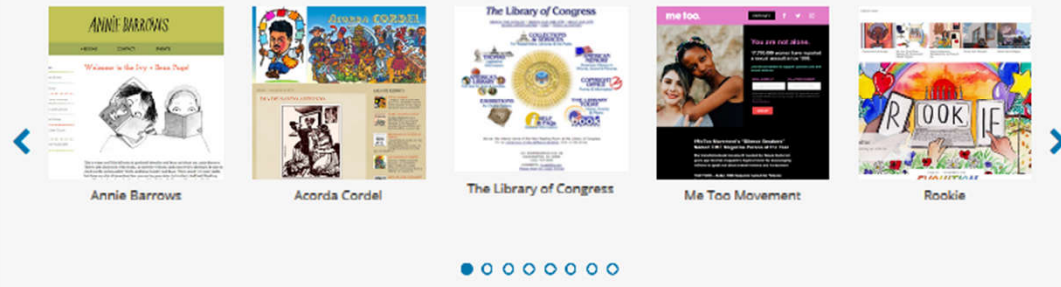
SUBSCRIBE VIA EMAIL

<https://ndhadeliver.natlib.govt.nz/webarchive/20110306153112/http://fourpawsandwhiskers.blogspot.com/>

FORMAT
Web Archive





Search Web Archives Collections with Web Archives

Featured Content



Collections with Web Archives

View Gallery Go Sort By Select Go

 <p>COLLECTION Afghanistan, Iran, Pakistan, and Tajikistan Elections Web Archive</p> <p>The Afghanistan, Iran, Pakistan and Tajikistan Elections Web Archive includes campaign sites archived during these countries' election seasons since 2014. This archive collection includes websites for government bodies, political parties, presidential candidates,...</p> <p>Collection Items: View 159 Items</p>	 <p>COLLECTION Afghanistan, Iran, Pakistan and Tajikistan Government Web Archive</p> <p>This collection consists of official government websites for Afghanistan, Iran, Pakistan and Tajikistan and provides an overview of the political, economic, administrative and social situation in these four countries. Several of these...</p> <p>Collection Items: View 389 Items</p>	 <p>COLLECTION Afghanistan Web Archive</p> <p>The Afghanistan Web Archive is comprised primarily of websites produced by the Afghan government, specifically the executive branch. Also included are a selection of sites from statistical reporting agencies, banking institutions, universities,...</p> <p>Collection Items: View 21 Items</p>	 <p>COLLECTION African Government Web Archive</p> <p>The African Government Web Archive provides links to information from key African government ministries, institutions and organizations for the 51 countries in Africa south of the Sahara. This will ensure that the...</p> <p>Collection Items: View 90 Items</p>
--	---	--	--

<https://www.loc.gov/web-archives/collections/>

LIBRARY OF CONGRESS

Library of Congress » Digital Collections » United States Supreme Court Nominations Web Archive » About this Collection

COLLECTION
United States Supreme Court Nominations Web Archive

About this Collection [Collection Items](#)

Featured Content

- The Alto Opinions : A Report of the Alto Project at the Yale ...
- Bush Nominates Judge John Roberts to Supreme Court - US ...
- Sonia Sotomayor Biography - University of Michigan Law School ...
- Elena Kagan - Wikipedia, the free encyclopedia
- American Bar Association Standing Committee on the Federal ...

About this Collection

[Listen to this page](#)

Rights & Access

Expert Resources

- [Research Guide: U.S. Supreme Court Nominations](#)
- [United States Senate](#)
- [Law Library of Congress](#)
- [Web Archiving Program Information](#)
- [Tips on Searching the Web Archive](#)

This collection consists of blogs, academic articles, Congressional press releases, and media articles related to the nominations of John Roberts, Harriet Miers, Samuel Alito, Sonia Sotomayor, and Elena Kagan for the United States Supreme Court. This content covers the years 2005, 2006, 2009, and 2010.

Collection Period: September 2005 to November 2010.

Frequency of Collection: Most sites in the collection were targeted for capture weekly, with some targeted for capture once or monthly.

Languages: Collection material in English, with Arabic and Spanish.

LIBRARY OF CONGRESS

United States Supreme Court Nominations Web Archive

ne Court Nominations Web Archive

Collection Items

Part of: [United States Supreme Court Nominations Web Archive](#) Available Online

View: List | **Go** Sort By: Select | **Go**

Web Archive	319
Online Format	
Web Page	315
PDF	3
Date	
2000 to 2099	318
1900 to 1999	5
Location	
United States	319
Israel	3
Canada	1
Darfur	1
Indiana	1
Minnesota	1
New York	1
Palestine	1
Sudan	1
Washington	1
More Locations >	
Part of	
United States Supreme Court Nominations Web Archive	319
United States Supreme Court Nominations Web Archive	318
Law Library of Congress	318
Researcher and Reference Services Division	125
Public Policy Topics Web Archive	121
Legal Blogs Web Archive	39
Guest Collections	22

WEB ARCHIVE
Opinio Juris
A weblog dedicated to reports, commentary, and debate on current developments and scholarship in the fields of international law and politics. Website, electronic | Electronic (Form).
Date: 2007-02-14

WEB ARCHIVE
How Appealing
Website, electronic | Electronic (Form).
Contributor: United States - Bashman, Howard - Supreme Court
Date: 2006-06-06

WEB ARCHIVE
The University of Chicago Law School Faculty Blog
Website, electronic | Electronic (Form).
Contributor: Law School - University of Chicago
Date: 2006-05-15

WEB ARCHIVE
The 10b-5 Daily
News and events related to securities class action litigation. Containing all facts, with particularity, and an occasional dose of commentary. Website, electronic | Electronic (Form).
Contributor: Roberts, Lyle
Date: 2007-03-01

Websites category

Contents

[Australian Web Archive](#)[Restricted content](#)[What can I do if I am concerned about an archived webpage?](#)[Disclaimer](#)

Explore archived websites from over 8 billion records stored on the Australian Web Archive. This includes material relevant to the cultural, social, political, research and commercial life and activities of Australia and Australians.

Check out **Related Pages** for more information.

Australian Web Archive

The Australian Web Archive contains:

- Australian websites selected for the PANDORA Web Archive (one of the world's first web archiving initiatives)
- Australian Government websites (formerly accessible through the Australian Government Web Archive)
- Websites with addresses that end in .au (collected annually)

These websites are saved as snapshots of how they appeared at the time that they were online, and then preserved on the archive as a copy.

Material will continue to be added to the Australian Web Archive and online material is collected in accordance with the Library's function as stated in the National Library Act 1960, the legal deposit provisions of the

Site search

Categories

[Newspapers and Gazettes](#)[Magazines and Newsletters](#)[Images, Maps and Artefacts](#)[Research and Reports](#)[Books and Libraries](#)[Diaries, Letters and Archives](#)[Music, Audio and Video](#)[People and Organisations](#)[Websites](#)[Searching](#) ▶[Navigating](#) ▶

Restricted content

Some archived webpages are restricted from public use for a variety of reasons. It may be because the organisation that published the webpage has asked for it to be restricted. Other reasons include:

Privacy (personal data):	Sensitive personal data that may make a person easily identifiable or locatable, and possibly subject to harm such as identify theft or acts of violence. This includes, but is not limited to, date of birth, address, medical data, LGBTQI status.
Privacy (other):	Content that may be considered an invasion of privacy but is not sensitive personal data.
Defamation:	Content that is subject to defamation proceedings.
Cultural protocols:	Content that may be in violation of recognised cultural protocols. This includes, but is not limited to, the content related to deceased persons from Indigenous communities.
Court order:	Content that cannot be published due to a court order. This includes, but is not limited to, content that is subject to a suppression order.
Criminal:	Content deemed criminal under legislation. This includes, but is not limited to, child pornography or child abuse images, content that advocates the committing of a terrorist act, and images falling under non-consensual image sharing laws.
Harmful:	Content that may be considered harmful by some but is not necessarily illegal. This includes, but is not limited to, pornographic images of consenting adults, hate speech and content that advocates unsafe behaviour.
Copyright:	Content that may be subject to copyright or licensing agreements. If the request relates to copyright, the applicant must supply evidence that they are the rights holder or their agent.
Commercial:	Content that may be of significant commercial advantage or sensitivity.
Protected government data:	Government data that is sensitive, official or classified and exempt from publication under the Freedom of Information Act 1982.
Other:	Any reason that falls outside the above listed categories.

Some are restricted permanently while others will be released after a period of time.

Don't worry - there are millions of more pages in the web archive that can help you with your research. Keep browsing.



From archive to analysis: accessing web archives at scale through a cloud-based interface

Nick Ruest¹ · Samantha Fritz² · Ryan Deschamps² · Jimmy Lin³ · Ian Milligan²

Received: 14 August 2020 / Accepted: 1 November 2020 / Published online: 6 January 2021
© The Author(s) 2021

Abstract

This paper introduces the Archives Unleashed Cloud, a web-based interface for working with web archives at scale. Current access paradigms, largely driven by the scope and scale of web archives, generally involve using the command line and writing code. This access gap means that subject-matter experts, as opposed to developers and programmers, have few options to directly work with web archives beyond the page-by-page paradigm of the Wayback Machine. Drawing on first-hand research and analysis of how scholars use web archives, we present the interface design and underpinning architecture of the Archives Unleashed Cloud. We also discuss the sustainability implications of providing a cloud-based service for researchers to analyze their collections at scale.

Keywords Web archives · Interface design · Digital humanities · Accessibility · Sustainability

1 Introduction

Nick Ruest, Samantha Fritz, Ryan Deschamps, Jimmy Lin, and Ian Milligan. 2021, “From Archive to Analysis: Accessing Web Archives at Scale through a Cloud-Based Interface.” *International Journal of Digital Humanities* 2, no. 1-3: 5–24. <https://doi.org/10.1007/s42803-020-00029-6>.

The DSA Toolkit Shines Light Into Dark and Stormy Archives

Themed web archive collections exist to make sense of archived web pages (mementos). Some collections contain hundreds of thousands of mementos. There are many collections about the same topic. Few collections on platforms like Archive-It include standardized metadata. Reviewing the documents in a single collection thus becomes an expensive proposition. Search engines help find individual documents but do not provide an overall understanding of each collection as a whole. Visitors need to be able to understand what individual collections contain so they can make decisions about individual collections and compare them to each other. The Dark and Stormy Archives (DSA) Project applies social media storytelling to a subset of a collection to facilitate collection understanding at a glance. As part of this work, we developed the DSA Toolkit, which helps archivists and visitors leverage this capability. As part of our recent International Internet Preservation Consortium (IIPC) grant, Los Alamos National Laboratory (LANL) and Old Dominion University (ODU) piloted the DSA toolkit with the National Library of Australia (NLA). Collectively we have made numerous improvements, from better handling of NLA mementos to native Linux installers to more approachable Web User Interfaces. Our goal is to make the DSA approachable for everyone so that end-users and archivists alike can apply social media storytelling to web archives.

by Shawn M. Jones, Himarsha R. Jayanetti, Alex Osborne, Paul Koerbin, Martin Klein, Michele C. Weigle, Michael L. Nelson

Editor's Note: This article makes use of Robust Links. Next to each hyperlink the reader will discover a menu that allows them to visit an archived version of the linked resource in case the current version has changed or is no longer available. Visit the Robust Links project for tools and more information on combating reference rot.

Web Archive Collections Are Too Large To Understand At A Glance

Web archives are invaluable for a variety of research studies. Historians have analyzed how humans interacted on extinct websites, like Geocities. Social scientists have used them to study the changes in social commerce over time. Journalists can use web archive evidence to bring attention to questionable medical practices and document changes in government policy.

Some archivists create themed web archive collections by selecting web pages for preservation that support a topic. Each web page, or original resource, can change over time. Archivists capture these original resources at specific points in time, turning each observation into a memento. The date and time of capture is that memento's memento-datetime. A TimeMap contains the set of mementos for an original resource. Archive-It is a popular platform for creating themed web archive collections. Themed collections also exist at the Library of Congress, Conifer, the Croatian Web Archive, the UK Web Archive, and the National Library of Australia's (NLA) PANDORA and Trove collections.

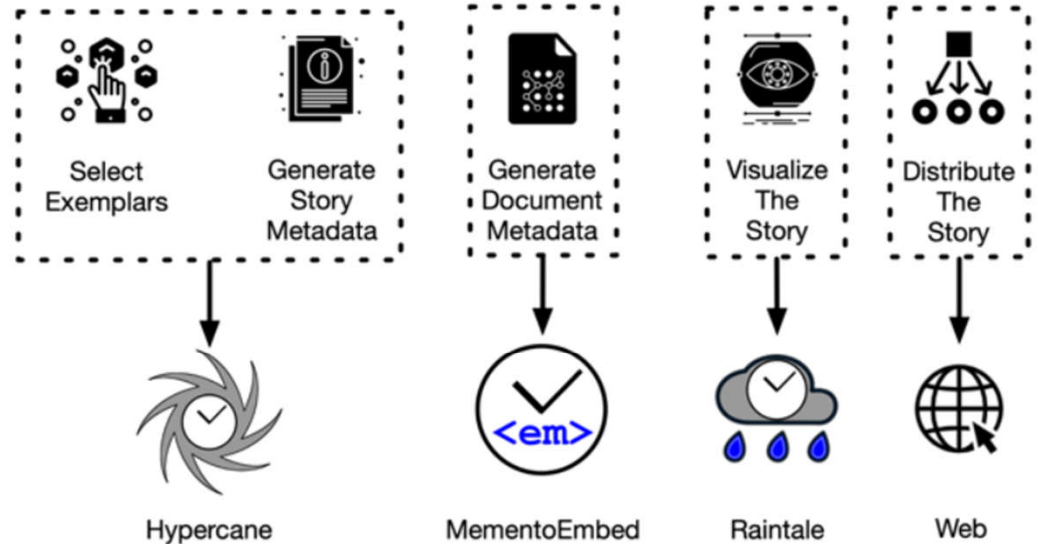


Figure 8. How the DSA Toolkit fits with the storytelling model shown in Figure 7.

S. M. Jones, H. Jayanetti, A. Osborne, P. Koerbin, M. Klein, M. C. Weigle, and M. L. Nelson, 2022, "The DSA Toolkit Shines Light Into Dark and Stormy Archives," Code4Lib Journal, Issue 53, 2022-05-09, <https://journal.code4lib.org/articles/16441>

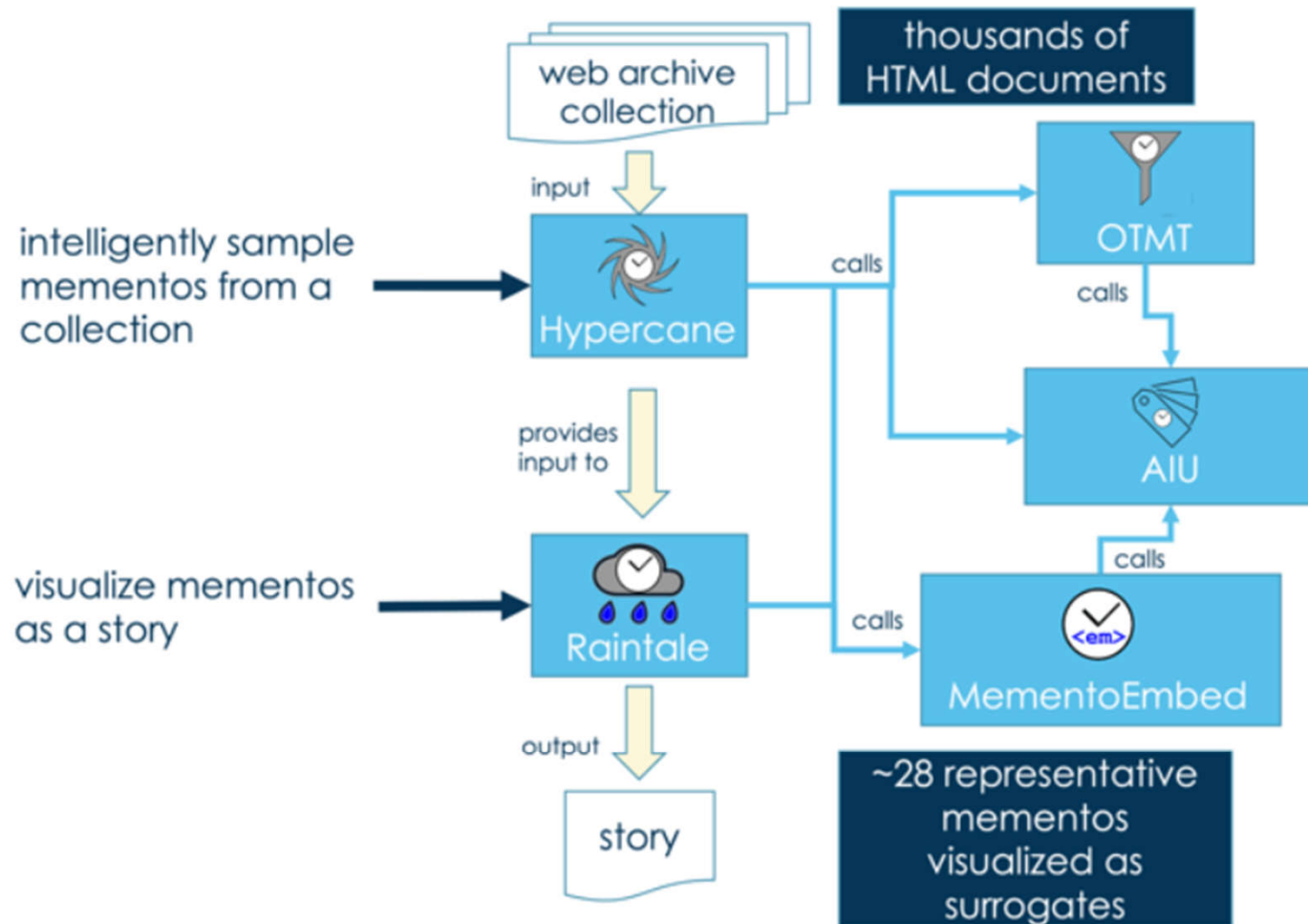


Figure 11. The DSA Toolkit workflow for producing a story relies heavily on Hypercane and Raintale.

S. M. Jones, H. Jayanetti, A. Osborne, P. Koerbin, M. Klein, M. C. Weigle, and M. L. Nelson, 2022, "The DSA Toolkit Shines Light Into Dark and Stormy Archives," Code4Lib Journal, Issue 53, 2022-05-09, <https://journal.code4lib.org/articles/16441>



Education: Open Data in Schools



How Open Data can benefit educational purposes

31/01/2018

Often, when Open Data is discussed, the focus lies most of the time on business opportunities and the various benefits that Open (Government) Data holds for the economy and government, in particular with regards to fostering business innovation and efficiency gains. Further benefits however have also been quantified on other dimensions, such as the environmental and societal ones. Concerning the latter, the focus lies here on the added value that Open Data brings for democracy by enhancing transparency of political processes as well as boosting participation of citizens in the decision-making at local, regional and/or national level. However, little attention is given on the use of Open Data in the field of education. More and more European countries have acknowledged this potential and have started some proof of concept projects to introduce Open Data in the school curriculum.

One example thereof is the **Open Data for Education competition in Northern Ireland**. In 2016, the government of Northern Ireland launched a competition for innovative ideas on how to use Open Data for Education (#ODNI4EDU). The objective was to use available [Open Data provided via the public sector portal to assist with teaching in primary and secondary schools](#). The best two ideas were awarded €20,000 towards developing their prototypes into classroom-ready applications. One of the winners was "Our Raging Planet", a platform that helps pupils learn about the possible impact of natural disasters. The application allows teachers to demonstrate what could happen if a disaster, which occurred at the other side of the planet, would strike at home. At the same time, it enables pupils to get acquainted with the geographic implications of global warming and how [data can help take informed decisions](#).

A further example comes from Germany where a **collaborative Open Data school project was launched in Moers**. In Moers, a town with approximately 100,000 inhabitants in Germany's federal state of North Rhine-Westphalia, the local administration and the Open Knowledge Foundation set out to explore the potential of Open Data in the classroom. The project – [DatenmachenSchule](#) – aims to help develop a series of software applications, together with the Adolfinum Gymnasium and the students of the Rhine-Waal University of Applied Sciences. Like the Northern Irish project, the idea here as well is to make school lessons more interactive by employing Open Data.

The project showed how data can be used in the school curriculum in a variety of subjects. In mathematics or computer science lectures, data was used in assignments. In subjects such as politics or social studies, pupils were able to learn more about their local government, its budget and spending, by using Open Data publicly available on the open budget platform [offenerhaushalt.de](#). Other school lessons included data on the demographic development of Moers' districts. For example, in the case of the demographic development data, seventh graders were asked to analyse the statistics and compare the insights to their personal experiences. In groups, pupils were then tasked to identify good locations for a nursing home, a kindergarten and a supply store for young families. They then reflected which data was useful and which data might be missing, before discussing further potential use cases of publicly available data. Furthermore, data on energy consumption in schools was used in environmental science classes. It provided pupils with insights into their own school's energy use and enabled them to see their school's performance compared to other schools in the area. Pupils were able to understand how they can help save energy and contribute to a sustainable use of resources. Such examples and many more use cases of Open Data in the classroom, can be found in the published [guidelines](#) of the DatenMachenSchule project.

A similar initiative was launched in **Switzerland as well, where steps have been taken to introduce Open Data in the school curriculum**. Data analysis is part of the [Lehrplan21](#) curriculum - a plan adopted in 2014 to harmonise the school curriculum in the 21 German-speaking and multi-lingual cantons. As part of the [Media and Informatics curriculum](#)

<https://data.europa.eu/en/datastories/education-open-data-schools>



Articles in this section

[Guide for new Archive-It users](#)

[Archive-It Video Curriculum](#)

[Known Web Archiving Challenges](#)

[Live chat support](#)

[Quickstart Guide for Trials](#)

[What is web archiving?](#)

[Support Ticket Submission](#)

[Set up and administer your account](#)

[Assign user access levels](#)

[Monitor your data budget](#)

[See more](#)

Archive-It Video Curriculum



Sylvie Rollason-Cass
Updated 1 month ago

[Follow](#)

On this page:

- [Getting Started](#)
 - [Navigating Archive-It](#)
 - [Administrative Functions](#)
 - [Pre-crawl Scoping](#)
 - [Test Crawls](#)
 - [PDF Only Crawls](#)
- [Post Crawl Analysis](#)
 - [Getting the most from your post crawl reports](#)
 - [Understanding your Hosts Report](#)
 - [Quality Assurance](#)
- [Advanced Training Webinars](#)
 - [Advanced Scoping](#)
 - [Archiving Video Content](#)
 - [Archiving Social Media](#)
 - [Advanced Quality Assurance](#)
 - [Access to Archive-It Collections](#)
 - [Under the Hood](#)
 - [Describing Web Archives](#)
 - [Intro to Brozzler](#)
 - [WARC Tools for Management and Preservation](#)

<https://support.archive-it.org/hc/en-us/articles/216489103-Archive-It-Video-Curriculum->



Digital sources and digital archives: historical evidence in the digital age

Trevor Owens¹  · Thomas Padilla² 

Received: 6 July 2019 / Accepted: 22 April 2020 / Published online: 4 May 2020
© Springer Nature Switzerland AG 2020

Abstract

As the cultural record becomes increasingly digital the evidentiary basis of history expands and shifts. How must historical scholarship change when the evidentiary basis shifts toward the digital? Through explorations of a series of born digital and digitized sources, we identify and discuss key issues relating to humanities scholars ability to develop claims and arguments grounded in digital sources and digital archives. In exploring these issues in digital source criticism, we work to provide practical guidance for scholars on key issues and questions to consider when working with born digital and digitized primary sources.

Keywords Digital history · Historiography · Research methods · Collections as data · Source criticism · Digitization · Archives

The world is full of potential primary sources. Almost anything can be a source. The rings of a tree testify to weather conditions and changes in climate (Cronon 1983). Probate records document the material goods individuals held at the end of their lives (Bushman 1992). Court proceedings offer insight into the experiences of the oppressed (Pagan 2003). Just as any kind of physical object might serve as a source, so does a digital source. As societies increasingly express themselves using digital means, the

Trevor Owens and Thomas Padilla, 2021, "Digital sources and digital archives: historical evidence in the digital age," *Int J Digit Humanities* 1, 325–341. <https://doi.org/10.1007/s42803-020-00028-7>

Full-Text and URL Search Over Web Archives



Miguel Costa

Abstract Web archives are a historically valuable source of information. In some respects, web archives are the only record of the evolution of human society in the last two decades. They preserve a mix of personal and collective memories, the importance of which tends to grow as they age. However, the value of web archives depends on their users being able to search and access the information they require in efficient and effective ways. Without the possibility of exploring and exploiting the archived contents, web archives are useless. Web archive access functionalities range from basic browsing to advanced search and analytical services, accessed through user-friendly interfaces. Full-text and URL search have become the predominant and preferred forms of information discovery in web archives, fulfilling user needs and supporting search APIs that feed third-party applications. Both full-text and URL search are based on the technology developed for modern web search engines. However, while web search engines enable searching over the most recent web snapshot, web archives enable searching over multiple snapshots from the past. This means that web archives have to deal with a temporal dimension that is the cause of new challenges and opportunities, discussed throughout this chapter.

1 Introduction

The World Wide Web has a democratic character, and everyone can publish all kinds of information using different types of media. News, blogs, wikis, encyclopaedias, photos, interviews and public opinion pieces are just a few examples. Some of this information is unique and historically valuable. For instance, online newspapers reporting the speech of a president after winning an election or announcing an imminent invasion of a foreign country might become as valuable in the future as

M. Costa (✉)
Vodafone, Lisbon, Portugal
e-mail: miguel.costa2@vodafone.com

Miguel Costa, 2021, “Full-Text and URL Search Over Web Archives,” In *The Past Web*, 71–84. Cham: Springer International Publishing.
https://doi.org/10.1007/978-3-030-63291-5_7.



When expectations meet reality: common misconceptions about web archives and challenges for scholars

Brenda Reyes Ayala¹

Received: 16 August 2020 / Accepted: 29 March 2021 / Published online: 12 June 2021
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

Abstract

As the study of digital history, politics, and culture emerges as an academic discipline, web archives will play a valuable role as sources of information. Those wishing to engage with web archives will need both specific technical skills and a high-level understanding of how the web works. This paper examines the nature and type of misconceptions that web archivists form when they create and utilise web archives. In order to carry out this research, the author qualitatively analyzed support tickets submitted by web archivists using the Internet Archive's Archive-It (AIT), the most popular web archiving service. The tickets comprised 2544 interactions between web archivists and AIT support specialists. This paper describes the expectations AIT users bring to web archives, and the differences between their expectations and the realities of the web archiving process. It identifies the most prominent misconceptions AIT users have about both web archives and the web itself, analyses the challenges these misconceptions can pose for researchers, and recommends ways in which these can be addressed.

Keywords Web archiving · Mental models · Digital humanities

1 Introduction

Brenda Reyes Ayala, 2021, "When Expectations Meet Reality: Common Misconceptions About Web Archives and Challenges for Scholars," International Journal of Digital Humanities 2, no. 1-3: 89–106. <https://doi.org/10.1007/s42803-021-00034-3>.

Developing Web Archiving Metadata Best Practices to Meet User Needs

Jackie M. Dooley
Karen Stoll Farrell
Tammi Kim
Jessica Venlet

ABSTRACT

The OCLC Research Library Partnership Web Archiving Metadata Working Group was established to meet a widely recognized need for best practices for descriptive metadata for archived websites. The Working Group recognizes that development of successful best practices intended to ensure discoverability requires an understanding of user needs and behavior. We have therefore conducted an extensive literature review to build our knowledge and will issue a white paper summarizing what we have learned. We are also studying existing and emerging approaches to descriptive metadata in this realm and will publish a second report recommending best practices. We will seek broad community input prior to publication.

Two recent surveys of users and managers of archived websites have shown that lack of a common approach to creating metadata is the most widely shared challenge for this community.^{1,2} In response, OCLC Research established a Web Archiving Metadata Working Group (WAM) to develop descriptive metadata best practices.³ At the group's first meeting in January 2016, we recognized that it would be inadvisable to develop best practices for descriptive metadata without first gaining a clear understanding of user needs and behavior in this context. We are taking this into account throughout the project.

1. Ricky Erway, "Thoughts from Partner Staff about Web Archiving," hangingtogether.org, October 29, 2015, <http://hangingtogether.org/?p=5450> (accessed January 18, 2017).
2. A research team led by Matthew Weber at Rutgers University surveyed users of web archives in the winter of 2016. They expect to publish their data late in 2016.
3. "Web Archiving Metadata Working Group," OCLC Research, last modified March 10, 2016, <http://oclc/wam> (accessed January 18, 2017).

Jackie M Dooley, Karen Stoll Farrell, Tammi Kim, and Jessica Venlet, 2017 "Developing Web Archiving Metadata Best Practices to Meet User Needs," *Journal of Western Archives*, Vol. 8 : Iss. 2 , Article 5,
DOI: <https://doi.org/10.26077/cffd-294a>
Available at: <https://digitalcommons.usu.edu/westernarchives/vol8/iss2/5>



The values of web archives

Valérie Schafer¹ · Jane Winters²

Received: 13 September 2020 / Accepted: 18 April 2021 / Published online: 10 June 2021
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

Abstract

This article considers how the development, promotion and adoption of a set of core values for web archives, linked to principles of “good governance”, will help them to tackle the challenges of sustainability, accountability and inclusiveness that are central to their long-term societal and cultural worth. It outlines the work that has already been done to address these questions, as web archiving begins to move out of its establishment phase, and then discusses seven key principles of good governance that might be adapted by and embedded within web archives: participation, consensus, accountability, transparency, effectiveness and efficiency, inclusivity and legality. The article concludes with a call to action for researchers and archivists to co-create the core values for web archives that will be required if they are to remain a vital part of our cultural heritage infrastructure.

Keywords Web archives · Good governance · Sustainability · Inclusiveness · FAIR data · Openness

In gathering, preserving, curating, publishing and/or analysing an intangible and massive born-digital heritage, key stakeholders, whether they are libraries, private

Valérie Schafer and Jane Winters, 2021, “The Values of Web Archives,” *International Journal of Digital Humanities* 2, no. 1-3: 129–44.
<https://doi.org/10.1007/s42803-021-00037-0>.

scientific **data**

OPEN

DATA DESCRIPTOR

DUKweb, diachronic word representations from the UK Web Archive corpus



Adam Tsakalidis^{1,2}, Pierpaolo Basile³, Marya Bazzi^{1,4,5}, Mihai Cucuringu^{1,5} & Barbara McGillivray^{1,6}  

Lexical semantic change (detecting shifts in the meaning and usage of words) is an important task for social and cultural studies as well as for Natural Language Processing applications. Diachronic word embeddings (time-sensitive vector representations of words that preserve their meaning) have become the standard resource for this task. However, given the significant computational resources needed for their generation, very few resources exist that make diachronic word embeddings available to the scientific community. In this paper we present DUKweb, a set of large-scale resources designed for the diachronic analysis of contemporary English. DUKweb was created from the JISC UK Web Domain Dataset (1996–2013), a very large archive which collects resources from the Internet Archive that were hosted on domains ending in '.uk'. DUKweb consists of a series word co-occurrence matrices and two types of word embeddings for each year in the JISC UK Web Domain dataset. We show the reuse potential of DUKweb and its quality standards via a case study on word meaning change detection.

Background & Summary

Word embeddings, dense low-dimensional representations of words as real-number vectors¹, are widely used in many Natural Language Processing (NLP) applications, such as part-of-speech tagging, information retrieval, question answering, sentiment analysis, and are employed in other research areas, including biomedical sciences² and scientometrics³. One of the reasons for this success is that such representations allow us to perform vector calculations in geometric spaces which can be interpreted in semantic terms (i.e. in terms of the similarity in the meaning of words). This follows the so-called distributional hypothesis⁴, according to which words occurring in a given word's context contribute to some aspects of its meaning, and semantically similar words share similar contexts. In Firth's words this is summarized by the quote "You shall know a word by the company it keeps"⁵.

Vector representations of words can take various forms, including count vectors, random vectors, and word embeddings. The latter are nowadays most commonly used in NLP research and are based on neural networks which transform text data into vectors of typically 50–300 dimensions. One of the most popular approaches for generating word embeddings is word2vec¹. A common feature of such word representations is that they are labour-intensive and time-consuming to build and train. Therefore, rather than training embeddings from scratch, in NLP it is common practice to use existing pre-trained embeddings which have been made available to the community. These embeddings have typically been trained on very large web resources, for example Twitter, Common Crawl, Gigaword, and Wikipedia^{6,7}.

Over the past few years NLP research has witnessed a surge in the number of studies on diachronic word embeddings^{8,9}. One notable example of this emerging line of research is¹⁰, where the authors proposed a method for detecting semantic change using word embeddings trained on the Google Ngram corpus¹¹ covering 8.5 hun-

A. Tsakalidis, P. Basile, M. Bazzi, et al. 2021, "DUKweb, diachronic word representations from the UK Web Archive corpus," *Sci Data* 8, 269. <https://doi.org/10.1038/s41597-021-01047-x>



Web-archiving and social media: an exploratory analysis

Call for papers digital humanities and web archives – A special issue
of international journal of digital humanities

Eveline Vlassenroot¹  · Sally Chambers² · Sven Lieber³ · Alejandra Michel⁴ ·
Friedel Geeraert⁵ · Jessica Pranger⁵ · Julie Birkholz⁶ · Peter Mechant¹

Received: 5 October 2020 / Accepted: 18 April 2021 / Published online: 22 June 2021
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

Abstract

The archived web provides an important footprint of the past, documenting online social behaviour through social media, and news through media outlets websites and government sites. Consequently, web archiving is increasingly gaining attention of heritage institutions, academics and policy makers. The importance of web archives as data resources for (digital) scholars has been acknowledged for investigating the past. Still, heritage institutions and academics struggle to ‘keep up to pace’ with the fast evolving changes of the World Wide Web and with the changing habits and practices of internet users. While a number of national institutions have set up a national framework to archive ‘regular’ web pages, social media archiving (SMA) is still in its infancy with various countries starting up pilot archiving projects. SMA is not without challenges; the sheer volume of social media content, the lack of technical standards for capturing or storing social media data and social media’s ephemeral character can be impeding factors. The goal of this article is three-fold. First, we aim to extend the most recent descriptive state-of-the-art of national web archiving, published in the first issue of International Journal of Digital Humanities (March 2019) with

Eveline Vlassenroot, Sally Chambers, Sven Lieber, Alejandra Michel, Friedel Geeraert, Jessica Pranger, Julie Birkholz, and Peter Mechant, 2021, "Web-archiving and social media: an exploratory analysis." International Journal of Digital Humanities 2, no. 1: 107-128.



Social media data archives in an API-driven world

Amelia Acker¹ · Adam Kreisberg²

Published online: 24 September 2019
© The Author(s) 2019

Abstract

In this article, we explore the long-term preservation implications of application programming interfaces (APIs) which govern access to data extracted from social media platforms. We begin by introducing the preservation problems that arise when APIs are the primary way to extract data from platforms, and how tensions fit with existing models of archives and digital repository development. We then define a range of possible types of API users motivated to access social media data from platforms and consider how these users relate to principles of digital preservation. We discuss how platforms' policies and terms of service govern the set of possibilities for access using these APIs and how the current access regime permits persistent problems for archivists who seek to provide access to collections of social media data. We conclude by surveying emerging models for access to social media data archives found in the USA, including community driven not-for-profit community archives, university research repositories, and early industry-academic partnerships with platforms. Given the important role these platforms occupy in capturing and reflecting our digital culture, we argue that archivists and memory workers should apply a platform perspective when confronting the rich problem space that social platforms and their APIs present for the possibilities of social media data archives, asserting their role as “developer stewards” in preserving culturally significant data from social media platforms.

Keywords APIs · Developer stewards · Platform perspective · Social media data archives

✉ Amelia Acker
aacker@ischool.utexas.edu
Adam Kreisberg
adam.kreisberg@simmons.edu

¹ School of Information, University of Texas at Austin, 1616 Guadalupe St, Suite 5.202, Austin, TX 78701, USA

² School of Library of Information Science, Simmons University, 300 The Fenway, Boston, MA 02115, USA

Amelia Acker and Adam Kreisberg, 2020, “Social Media Data Archives in an API-Driven World,” *Archival Science* 20(2), pp., 105–23. <https://doi.org/10.1007/s10502-019-09325-9>