404 Page Not Found OOPS

BOOK OF ABSTRACTS
#EWAVirtual 2020

Engaging with Web Archives
‘Opportunities, Challenges and Potentialities’
(#EWAVirtual)

21-22 September 2020
Maynooth University, Arts and Humanities Institute
Co. Kildare, Ireland

Maynooth University
National University of Ireland Maynooth
Arts and Humanities Institute

Published by:

#EWAVirtual Conference Organisers, Maynooth University Arts and Humanities Institute

Edited by

Sharon Healy, Michael Kurzmeier, Helena La Pina and Patricia Duffe

DOI: http://doi.org/10.5281/zenodo.4058013

Table of Contents

Introduction

Engaging with Web Archives

'Opportunities, Challenges and Potentialities', (#EWAVirtual), 21-22 September 2020, Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland.

Maynooth University Arts and Humanities Institute are delighted to be hosting the first international EWA conference which aims to:

- Raise awareness for the use of web archives and the archived web for research and education across a broad range of disciplines and professions in the Arts, Humanities, Social Sciences, Political Science, Media Studies, Information Science, Computer Science and more;
- Foster collaborations between web archiving initiatives, researchers, educators and IT professionals.
- Highlight how the development of the internet and the web is intricately linked to the history of the 1990s.

What is Web Archiving?

Pioneered by the efforts of the Internet Archive in 1996, national libraries and cultural heritage organisations quickly realised the need to preserve information and content that was born on the web. It was this awareness that gave rise to technologies, specifically web crawler programmes, used for web archiving. According to the International Internet Preservation Consortium, '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.' Due to serious concerns about the loss of web-born heritage, there has been a continuous growth of web archiving initiatives across the globe.

Why should we care?

For example, in Ireland — The first connection to the Internet as we know it (via TCP/IP), went live in Trinity College Dublin in June 1991. The first web server and website in Ireland can be traced back to 1991/92 in University College Cork (CURIA project); and other websites followed in 1993 from IONA Technologies, TCD Maths, IEunet, and University of Limerick. The growth of Irish websites was slow at first, but this changed by the end of 1995 due to international developments in browser technology, and the growth of internet service providers in Ireland (see TechArchives, How the internet came to Ireland; David Malone, Early Irish Web Stuff).

THERE ARE SIMILAR SCENARIOS AROUND THE WORLD

As researchers begin to negotiate and write the history of their countries for the 1990s, whether it is social, cultural, political or even economic, it seems inevitable that they will also need to consider their histories of IT – in terms of how the introduction of the internet and the WWW began to infiltrate the fabric of life, work and play.

The archived web is now an object of study in many countries, and there has been a lot of work done already to build research infrastructures and networks. But more needs to be done to promote awareness of the availability of web archives, and how they can be utilised as resources for research going into the future. And certainly, much more needs to be done in the realms of how web archives can be incorporated as resources in education, and how the use of web archives can be taught.

International literature using web archives for research and historical inquiry is growing; yet the question of how to effectively use the archived web for qualitative and quantitative research still remains open; and how to integrate the use of web archives into teaching is a path yet to be explored. Furthermore, existing web archiving efforts find it hard to exchange knowledge and take on larger projects, partially due to the lack of opportunities for exchange between the disciplines and educators.

The EWA organisers would also like to extend their sincerest thanks and appreciation to the following organisations and institutions for their kind support and efforts to make this conference event possible:

- Maynooth University Arts and Humanities Institute
- Maynooth University, Department of Sociology
- Maynooth University, Department of Media Studies
- Maynooth University, Department of Computer Science
- Maynooth University, Department of History
- National Library of Ireland, Web Archive
- TechArchives, Ireland
- University College Cork, Digital Arts & Humanities
- University College Dublin, School of History
- AGREXIS AG

Maynooth University National University of Ireland Maynooth | Arts and Humanities Institute

Maynooth University National University of Ireland Maynooth

UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh | Digital Arts & Humanities Ealaíona Digiteacha agus Daonnachtaí

nli Leabharlann Náisiúnta na hÉireann National Library of Ireland | < web archive > < cartlann ghréasáin >

agrexis

TECHARCHIVES

If you require more information or have any questions please feel free to email us: ewaconference@gmail.com

Follow us on Twitter:

- @EWAConf
- @MU_AHI
- #EWAVirtual

Welcome from Sharon Healy and Michael Kurzmeier

#EWAVirtual 2020 Conference Co-Chairs

On behalf of the organising committee of the first international Engaging with Web Archives conference, we would like to welcome all delegates to Maynooth University Arts and Humanities Institute for what we hope will be a stimulating event within the realms of engaging with web archives and web archiving activities. We are proud to announce that this is the first web archive conference of its nature ever to be held in Ireland; and, the first virtual conference to be held in Maynooth University for 2020. The programme contains, 35 paper presentations, and 2 distinguished Keynote speakers. We are delighted to extend a warm welcome to the two keynotes speakers: Prof. Niels Brügger of Aarhus University, Denmark; and Prof. Jane Winters of School of Advanced Study, University of London. UK.

#EWAVirtual brings together speakers who are historians, digital humanists, media scholars, social scientists, information and IT professionals, computer scientists, data consultants, librarians and archivists from Ireland, the United Kingdom, Europe, Canada, and the United States. To all the speakers, we appreciate your kindness, support and patience when the initial conference, scheduled in the Spring of 2020 was postponed, and your continued enthusiasm, cooperation and collaboration when we announced it would become a virtual event. We are also indebted to the Chairs of each session. Each one volunteered their services enthusiastically to assure the smooth running of the conference.

Our gratitude is extended to the tireless efforts of the organising committee. Its dedication, from the reviewing of papers, to the logistical components of organising the first physical conference. Then to find the motivation, and spirit to reorganise the event as a virtual conference, is greatly appreciated.

To all at Maynooth University and the band of volunteers, we appreciate your time, talent, and storyboard of ideas. Without your support and dedication, this conference would not be possible. A special shoutout goes to Professor Thomas O'Connor and Ann Donoghue from Maynooth University Arts and Humanities Institute. Their unfailing support, advice and kind assistance was invaluable throughout the entire processes of planning both EWA conferences (from the physical to the virtual).

Also, to all our sponsors and supporters, we appreciate all your encouragement, sound advice and uplifting messages. Particularly, we are grateful to the year-long encouragement and support by the committed staff at the National Library of Ireland.

To all the speakers, guests, volunteers, chairs and attendees, we thank you.

Together we have all played a part in the transformation of #EWA20 to #EWAVirtual.

All the Best

Sharon & Michael

Professor Niels Brügger

The variety of European web archives - potential effects for future humanities research

The aim of this keynote is to open up a discussion of how the great variety of European web archives may affect future humanities research based on the archived web as a source. The keynote is divided in two main sections. First, the different web archiving forms in Europe are briefly mapped with a focus on which countries do have a web archive, archiving strategies, and access conditions. Second, it is discussed how this state of affairs may affect transnational research projects, spanning more web archives. The case of the national Danish web domain is used as a stepping stone to evaluate to what an extent such a study can be replicated in other European countries, thus enabling transnational comparisons.

Niels Brügger is a Professor in Media Studies, Head of NetLab, part of the Danish Digital Humanities Lab, and head of the Centre for Internet Studies at Aarhus University in Denmark. He is a Coordinator of the European network RESAW, a Research Infrastructure for the Study of Archived Web Materials, and the managing editor of the international journal *Internet histories: Digital technology, culture and society.*

Professor Brügger has initiated the research projects "Probing a Nation's Web Domain — the Historical Development of the Danish Web" (2014-) and "the history of dr.dk, 1996-2006" (2007-), and co-initiated the research infrastructure project NetLab (2012-17) within the Digital Humanities Lab. His research interests are the history of the Internet as a means of communication, and Digital Humanities, including archiving the Internet as well as the use of digital research tools. Other interests include media theory, the Internet, and the relation between the two with a view to (re)evaluating the status and relevance of existing media theories and methods.
Recent publications include:

- *The Historical Web and Digital Humanities*, eds. N. Brügger, D. Laursen (Routledge, 2019)
- *The SAGE Handbook of Web History* eds. N. Brügger, I. Milligan (SAGE, 2019),
- *The Archived Web: Doing History in the Digital Age* (MIT Press, 2018).
- *Web 25: Histories from the first 25 years of the World Wide Web* ed. Niels Brügger (New York: Peter Lang, 2017)

Professor Jane Winters

Web archives as sites of collaboration

Openness to collaboration has been one of the defining characteristics of web archiving and web archive studies from the outset. The challenges posed by the archiving and preservation of born-digital data, including web archives, are simply too great to be solved by individuals or single organisations. This keynote will present some of the partnerships which have moved the field forward in the past decade, suggest some new avenues for collaboration in the future, and consider how the required knowledge and skills can be developed within universities and the cultural heritage sector to ensure that current web archiving initiatives are sustainable.

Jane Winters is a Professor of Digital Humanities and Pro-Dean for Libraries in the School of Advanced Study at the University of London. She is responsible for developing digital humanities and has led or co-directed a range of digital projects, including most recently Big UK Domain Data for the Arts and Humanities; Digging into Linked Parliamentary Metadata; Traces through Time: Prosopography in Practice across Big Data; the Thesaurus of British and Irish History as SKOS; and Born Digital Big Data and Approaches for History and the Humanities.

Professor Winters is a Fellow and Councillor of the Royal Historical Society, and a member of RESAW (Research Infrastructure for the Study of the Archived Web), the Academic Steering & Advocacy Committee of the Open Library of Humanities, the Advisory Board of the European Holocaust Research Infrastructure, the Advisory Board of Cambridge Digital Humanities, and the UK UNESCO Memory of the World Committee. Jane's research interests include digital history, born-digital archives (particularly the archived web), big data for humanities research, peer review in the digital environment, text editing and open access publishing.

Recent publications include:

- 'Giving with one hand, taking with the other: e-legal deposit, web archives and researcher access', in *Electronic Legal Deposit: Shaping the Library Collections of the Future*, ed. Paul Gooding and Melissa Terras (London: Facet Publishing, 2019);
- 'Negotiating the born digital: a problem of search', *Archives and Manuscripts*, 47:4 2019;
- 'Negotiating the archives of UK web space', in *The Historical Web and Digital Humanities: the Case of National Web Domains*, ed. Niels Brügger and Ditte Laursen (London: Routledge, 2019);

- ‘Web archives and (digital) history: a troubled past and a promising future?’ in *The SAGE Handbook of Web History*, ed. Niels Brügger and Ian Milligan (SAGE Publications Ltd., 2019)

DAY ONE: 21 September 2020

9.45 (IRE) / 10.45 (CEST)

WELCOME

Professor Tom O'Connor, Director of Maynooth University Arts and Humanities Institute

Michael Kurzmeier, #EWAVirtual Co-Chair, (Maynooth University)

10.00 (IRE) / 11.00 (CEST)

KEYNOTE

Chair: Joanna Finegan (National Library of Ireland)

Professor Niels Brügger, Aarhus University:
The variety of European web archives — potential effects for future humanities research

11.00 (IRE) / 12.00 (CEST)

Session 1: Archiving Initiatives

Chair: Jason Webber (UK Web Archive, British Library)

- Maria Ryan (National Library of Ireland): *The National Library of Ireland's Web Archive*: *preserving Ireland's online life for tomorrow*
- Sara Day Thomson (University of Edinburgh) *Developing a Web Archiving Strategy for the Covid-19 Collecting Initiative at the University of Edinburgh*
- Dr. Kees Teszelszky (KB – National Library of the Netherlands): *Internet for everyone: the selection and harvest of the homepages of the oldest Dutch provider XS4ALL (1993-2001)*

12.00 (IRE) / 13.00 (CEST)

Session 2: Collaborations

Chair: Patricia Duffe (Maynooth University)

- Dr. Brendan Power (The Library of Trinity College Dublin): *Leveraging the UK Web Archive in an Irish context: Challenges and Opportunities*

- Sarah Haylett & Patricia Falcao (Tate): *Creating a web archive at Tate: an opportunity for ongoing collaboration*

12.40 (IRE) / 13.40 (CEST)

Session 3: Archiving Initiatives (lightning round)

Chair: Rebecca O'Neill (Maynooth University)

- Rosita Murchan (Public Record Office of Northern Ireland): *PRONI Web Archive: A Collaborative Approach*

- Inge Rudomino & Marta Matijević (Croatian Web Archive, National and University Library in Zagreb – NSK): *An overview of 15 years of experience in archiving the Croatian web*

- Robert McNicol (Kenneth Ritchie Wimbledon Library): *The UK Web Archive and Wimbledon: A Winning Combination*

14.00 (IRE) / 15.00 (CEST)

Session 4: Research Engagement & Access

Chair: Chris Beausang (Maynooth University)

- Dr. Peter Mechant; Sally Chambers; Eveline Vlassenroot (Ghent University); Friedel Geeraert (KBR – Royal Library and the State Archives of Belgium): *Piloting access to the Belgian web-archive for scientific research: a methodological exploration*

- Sharon Healy (Maynooth University): *Awareness and Engagement with Web Archives in Irish Academic Institutions*

14.40 (IRE) / 15.40 (CEST) / 09:40 (EDT)

Session 5: Archiving Initiatives

Chair: Sara Day Thomson (University of Edinburgh)

- Anisa Hawes (Independent Curatorial Researcher): *Archiving 1418-Now using Rhizome's Webrecorder: observations and reflections*

- Nicole Greenhouse (New York University Libraries): *Managing the Lifecycle of Web Archiving at a Large Private University*

15.30 (IRE) / 16.30 (CEST)

Session 6: Social Science & Politics

Chair: Dr. Claire McGinn (Institute of Art, Design and Technology, Dún Laoghaire)

- Benedikt Adelmann MSc & Dr. Lina Franken (University of Hamburg): *Thematic web crawling and scraping as a way to form focussed web archives*

- Andrea Prokopová (Webarchiv, National Library of the Czech Republic): *Metadata for social science research*

- Dr. Derek Greene (University College Dublin): *Exploring Web Archive Networks: The Case of the 2018 Irish Presidential Election*

16.10 (IRE) / 17.10 (CEST) / 11:10 (EDT)

Session 7: Collaborations & Teaching

Chair: Dr. Joseph Timoney (Maynooth University)

- Olga Holownia (International Internet Preservation Consortium): *IIPC: training, research, and outreach activities*

- Dr. Juan-José Boté (Universitat de Barcelona): *Using web archives to teach and opportunities on the information science field*

16.50 (IRE) / 17.50 (CEST) / 10:50 (CST)

Session 8: Research of Web Archives

Chair: Sally Chambers (Ghent Centre for Digital Humanities, Ghent University)

- Bartłomiej Konopa (State Archives in Bydgoszcz; Nicolaus Copernicus University): *Web archiving – professionals and amateurs*

- Prof. Lynne M. Rudasill & Dr. Steven W. Witt (University of Illinois at Urbana-Champaign): *Opportunities for Use, Challenges for Collections: Exploring Archive-It for Sites and Synergies*

DAY TWO: 22 September 2020

9.45 (IRE) / 10.45 (CEST)

WELCOME

Michael Kurzmeier, EWA Co-Chair (Maynooth University)

10.00 (IRE) / 11.00 (CEST)

KEYNOTE

Chair: Maria Ryan (National Library of Ireland)

Professor Jane Winters, School of Advanced Study, University of London:
Web archives as sites of collaboration

11.00 (IRE) / 12.00 (CEST)

Session 9: Research Approaches
Chair: Jason Webber (UK Web Archive, British Library)

- Dr. Peter Webster (Independent Scholar, Historian and Consultant): *Digital archaeology in the web of links: reconstructing a late-90s web sphere*

- Michael Kurzmeier (Maynooth University): *Web defacements and takeovers and their role in web archiving*

11.40 (IRE) / 12.40 (CEST)

Session 10: Culture & Sport

Chair: Gavin Mac Allister (Irish Military War Museum)

- Dr. Philipp Budka (University of Vienna; Free University Berlin): *MyKnet.org: Traces of Digital Decoloniality in an Indigenous Web-Based Environment*

- Helena Byrne (British Library): *From the sidelines to the archived web: What are the most annoying football phrases in the UK?*

12.30 (IRE) / 13.30 (CEST)

Session 11: Research (lightning round)

Chair: Dr Julie Brooks (School of History, University College Dublin)

- Caio de Castro Mello Santos & Daniela Cotta de Azevedo Major (School of Advanced Study, University of London): *Tracking and Analysing Media Events through Web Archives*

- Dr. Eamonn Bell (Trinity College Dublin): *Reanimating the CDLink platform: A challenge for the preservation of mid-1990s Web-based interactive media and net.art*

- Hannah Connell (King's College London; British Library): *Curating culturally themed collections online: The Russia in the UK Special Collection, UK Web Archive*

14.00 (IRE) / 15.00 (CEST) / 9.00 (EST)

Session 12: Youth & Family

Chair: Dr. Lina Franken (University of Hamburg)

- Katie Mackinnon (University of Toronto): *DELETE MY ACCOUNT: Ethical Approaches to Researching Youth Cultures in Historical Web Archives*

- Dr. Susan Aasman (University of Groningen): *Changing platforms of ritualized memory practices. Assessing the value of family websites*

14.40 (IRE) / 15.40 (CEST)

Session 13: Source code and app histories

Chair: Prof. David Malone (Hamilton Institute, Maynooth University)

- Dr. Anne Helmond (University of Amsterdam) & Fernando van der Vlist (Utrecht University): *Platform and app histories: Assessing source availability in web archives and app repositories*

- Dr. Janne Nielsen (Aarhus University) *Exploring archived source code: computational approaches to historical studies of web tracking*

15.30 (IRE) / 16.30 (CEST) / 10.30 (EST)

Session 14: AI and Infrastructures

Chair: Dr. Juan-José Boté (Universitat de Barcelona)

- Mark Bell; Tom Storrar; Dr. Eirini Goudarouli; Pip Willcox (The National Archives, UK); David Beavan; Dr. Barbara McGillivray; Dr. Federico Nanni (The Alan Turing Institute): *Cross-sector interdisciplinary collaboration to discover topics and trends in the UK Government Web Archive: a reflection on process*

- Dr. Jessica Ogden (University of Southampton) & Emily Maemura (University of Toronto): *A tale of two web archives: Challenges of engaging web archival infrastructures for research*

16.10 (IRE) / 17.10 (CEST)

Session 15: WARC and OAIS

Chair: Kieran O'Leary (National Library of Ireland)

- Consultative Committee for Space Data Systems (CCSDS), Data Archive Interoperability (DAI) Working Group; Michael W. Kearney III; David Giaretta; John Garrett; Steve Hughes: *What's missing from WARC*? (Abstract/Bio)

16.45 (IRE) / 17.45 (CEST) / 08:45 (PDT)

Session 16: Web Archives as Scholarly Dataset

Chair: Michael Kurzmeier (Maynooth University)

- Dr. Helge Holzmann & Mr. Jefferson Bailey (Internet Archive): *Web Archives as Scholarly Dataset to Study the Web*

17.15 (IRE) / 18.15 (CEST)

An Irish Tale / Scéal Éireannach

17.45 (IRE) / 18.45 (CEST)

The Future of EWA

Sharon Healy & Michael Kurzmeier (Maynooth University)

Session 1: Archiving Initiatives

The National Library of Ireland's Web Archive: preserving Ireland's online life for tomorrow

Maria Ryan
(National Library of Ireland)

Keywords: Collection development, national domains, web archives, research, datasets

ABSTRACT

The National Library of Ireland (NLI) was founded in 1877 and its mission remains the same today; to collect, protect and make available the memory of Ireland. The library cares for a collection of over ten million physical items, with collections including manuscripts, photographs, prints and drawings and an extensive ephemera collection. In the 21st century, the NLI is working towards meeting the challenges of the digital world; collecting, preserving and providing access to a born digital record of Irish life. This presentation aims to examine the NLI web archive and highlight its importance to the documentation of Irish society and culture.

In 2011, the general and presidential election provided the catalyst for a pilot web-archiving project. Following the success of this project, the NLI focused on establishing the web-archiving programme by archiving political, cultural and social websites, capturing a record of elections, budgets, the decade of commemorations and historic events such as the 2015 marriage referendum. In 2016, the NLI received its first full time web archivist and launched a significant promotional drive around the 2016 commemorative project ‘Remembering 1916, Recording 2016’.

In 2017, The NLI also undertook a domain crawl of the Irish web, allowing for the capture of a wider range of websites and greater amounts of data, when compared with the selective web archive. The 2017 crawl encompassed all of the Irish top-level domain and other relevant websites that could be recognised as being hosted in Ireland but outside the .ie domain. It also used language detection software to identify Irish language websites outside the national domain. The crawl

amounted in almost 40 TB of unique data, which is preserved in the NLI. However, due to legislative restrictions, this data cannot be made available to researchers.

In the past nine years, the NLI web archive has grown and developed into what is now an established collecting strand in the NLI. Workflow development and a comprehensive collecting strategy has seen the web archive grow and mature. The NLI has embarked up to new opportunities for collaboration and research. Collaboration is at the heart of the values of the NLI and it has helped us broaden our collections and provide datasets to new researchers.

The future of research lies largely in born digital archives. The social, political and historical researchers of the future will require a record of the 21st century in Ireland. In other words, they will need web archives. This presentation will explore how the NLI is dedicated to building an Irish web archive that will document Irish life for decades to come.

Biography:

Maria Ryan is an assistant keeper and web archivist at the National Library of Ireland. A qualified archivist, she is co-chair of the IIPC training working group and a member of the NLI's diversity and inclusion committee.

Developing a Web Archiving Strategy for the Covid-19 Collecting Initiative at the University of Edinburgh

Sara Day Thomson
(Digital Archivist, Centre for Research Collections, University of Edinburgh)

Keywords: Covid-19, web archiving strategy, challenges, opportunities for collaboration; web archive collections

ABSTRACT

In this talk, the Digital Archivist at the University of Edinburgh will discuss the process (so far) for developing a strategy for capturing and preserving web-based submissions to their Collecting Covid-19 Initiative. She will also present plans for using this process as a springboard to develop a wider institutional programme(s) of web archiving.

In April, the Centre for Research Collections (CRC) put out an open call for members of the university community to submit materials that document their experiences of the Covid-19 pandemic and lockdown [1]. Depositors are invited to submit their digital records using a web form embedded on the university website [2].

At the time of the open call, the CRC did not have an established web archiving programme. Therefore, a new strategy had to be developed in response to the influx of web-based submissions (and other relevant web pages identified by the collecting team). This strategy, further, had to address the identified concerns of the Initiative: namely speedy deployment, but also handling sensitive material, understanding potential research uses, and balancing metadata requirements with low-barrier submission requirements.

The project team is now in the early stages of a partnership with the UK Web Archive through the National Library of Scotland. The CRC team will curate a special collection for the Collecting Covid-19 Initiative using the UKWA's infrastructure and guidance. Recognising some of the limitations of this approach, the Digital Archivist will supplement the Collecting Covid-19 collection with manual captures using OS tools, such as Conifer / Webrecorder Desktop and TAGS.

In order to make the most use of this strategy, the Digital Archivist has invited the project team to view these steps as a pilot study for wider web archiving programmes. This pilot will include an evaluation of methods for:

- gathering and analysing user needs and requirements
- choosing an approach, either collaboration with the UKWA or OS tools
- training, both staff and researchers, to capture web content as part of their work
- outreach to the wider university community to raise awareness of web archiving and of available archived web resources

Currently, the focus is finding a robust and reliable way to capture, curate, and preserve web-based submissions to the Covid-19 Collecting Initiative. However, in the coming months, the Digital Archivist hopes to lay the groundwork for next steps. First and foremost, she aims to host a series of focus groups (potentially virtually) with key researchers in collaboration with the Research Data Support team to better gather information about research needs and to raise the profile of available archived web content.

References:

[1] University of Edinburgh, Staff News, ‘Covid-19 experiences to be documented’ https://www.ed.ac.uk/news/students/2020/covid-19-experiences-to-be-documented

[2] University of Edinburgh, Collecting Covid-19 Initiative, https://www.ed.ac.uk/information-services/library-museum-gallery/crc/collecting-covid-19-initiative

Biography:

Sara Day Thomson is Digital Archivist at the University of Edinburgh where she looks after the management and preservation of digital materials across collections. She joined the University from the Digital Preservation Coalition where she was Research Officer, supporting the development of new methods and technologies to ensure long-term access to digital data. She reconvened the DPC’s Web Archiving and Preservation Working Group, a forum for organisations to share experiences in archiving web content. She also contributed to the development of IIPC & DPC Beginner Web Archiving Training materials and is the author of Preserving Social Media, a DPC Technology Watch Report.

Internet for everyone: the selection and harvest of the homepages of the oldest Dutch provider XS4ALL (1993-2001)

Dr. Kees Teszelszky
(Koninklijke Bibliotheek - National Library of the Netherlands)

Keywords: web archiving, web archaeology, web incunables, homepages, early web

ABSTRACT

“Web incunables” can be defined as those websites which were published in the first stage of the world wide web between 1990 and 1998. The early sites of the nineties were made at the start of publishing texts on the web and mark the frontier between analogue prints on paper and digital publications on the web. The first Dutch homepage and web incunable was put online in 1993: the same year one of the oldest Dutch internet provider XS4ALL (“Access for All”) started to offer its services to customers for the first time. This provider was founded by hackers and techno-anarchists in this year. It attracted a large group of creative Dutch internet pioneers after the start in May 1993 who have built at least 10,000 homepages between 1993 and 2001, of which a large part is still online in some form.

We can consider the remaining homepages as the most interesting born digital Dutch heritage collection still online and waiting to be studied. As XS4ALL was promoting and facilitating the building of these sites, the early web designers, artists, activists, writers and scientists were eagerly experimenting with the possibilities of the new medium in content, design and functionality. As XS4ALL was not so much seen as a company, but more as a society, many customers remained faithful to this provider till now. Due to this, a large amount of homepages of the early Dutch web can still be found at this provider. This heritage is however in danger. Dutch telephone company KPN took XS4ALL over in 1998 and announced in January 2019 to end this brand in near future. This is the reason why Koninklijke Bibliotheek - National Library of the Netherlands (KB-NL) started a web archiving project the same year to identify and rescue as much web incunables and early homepages as possible which are still hosted by this provider. This project was generously sponsored by SIDN-fonds and Stichting Internet4ALL.

This paper describes the method and first results of the ongoing pilot research project on internet archaeology and web incunables of KB-NL. It is about web archiving a selection of web incunables published on the Dutch web before 2001 which mirror the development of Dutch online culture on the web. I will describe the methods and sum up the experiences with selecting and harvesting homepages and mapping the Dutch digital culture online by link analysis of this collection. I will discuss also the characteristics of web materials and archived web materials, among others the first Dutch interactive 3D house, a virtual metro line for the digital city of Amsterdam, the "Stone Age Computer" and the first Dutch online literature magazine. I will also explain the use of these various materials (harvested websites, metadata link clouds, context information) for future research on the history of the Dutch web.

Biography:

Kees Teszelszky (1972) is a historian and curator of the digital collections at the Koninklijke Bibliotheek - National Library of The Netherlands. He graduated at the University of Leiden (Political Science, 1999) and at the University of Amsterdam (East European Studies, 1998) and obtained his PhD at the University of Groningen (Cultural History, 2006). He has been involved in research on web archiving and born digital sources since 2012. His present research field covers the selection, harvest and presentation of born digital sources at the KB. He is currently involved in projects on internet archaeology in the Netherlands, mapping the Frisian and Dutch national web domain, online news and the historic sources of our Post-truth era.

Leveraging the UK Web Archive in an Irish context: Challenges and Opportunities

Dr Brendan Power
(The Library of Trinity College Dublin)

Keywords: web archives, collaboration, legal deposit, 1916 Easter Rising

ABSTRACT

This paper will discuss a project to curate an archive of websites undertaken by The Library of Trinity College Dublin. The context for these projects was the UK legal deposit environment in which the six Legal Deposit Libraries (LDL's) work together to help preserve the UK's knowledge and memory. In 2013 the legal deposit remit was extended to include non-print, electronically published material, which means the LDL's may now capture and archive any freely available websites that are published or hosted in the UK. This happens in the Legal Deposit UK Web Archive, with the British Library providing the technical and curatorial infrastructure, and all LDL's contributing at both the strategic and planning level, and through curating themed collections. In this paper I will present a case study which demonstrates how The Library of Trinity College Dublin has explored the challenges and opportunities of utilising the research potential of this vast new resource.

The 1916 Easter Rising collection was a collaborative project in 2015/2016 between The Library of Trinity College Dublin (University of Dublin), the Bodleian Libraries (University of Oxford), and the British Library. The project aimed to identify, collect, and preserve websites that contribute to an understanding of the 1916 Easter Rising, with the aim of enabling critical reflection on both the Rising itself, and how it was commemorated in 2016. The project was a test case for effective collaboration between libraries in multiple jurisdictions helping to explore how themed, curated web archive collections can promote the potential of web archives to a wider audience. The presentation will review the project and outline the challenges and opportunities that emerged as it progressed. In particular, it will highlight the challenges that arose from working across multiple jurisdictions, and the implications of different legislative frameworks for archive curation and collection building.

Biography:

Brendan Power is Digital Preservation Librarian at The Library of Trinity College Dublin. He holds a BA from Dublin City University, an MPhil and PhD in History from Trinity College, the University of Dublin, and an MLIS from University College Dublin. A former Postdoctoral Research Fellow at Trinity College Dublin, he acted as the Web Archive Project Officer on the 1916 Easter Rising Web Archive and has previously published on this project.

Creating a web archive at Tate: an opportunity for ongoing collaboration

Sarah Haylett
(Tate)

Patricia Falcao
(Tate)

Keywords: web archives, net art, digital preservation, web-based art, archives

ABSTRACT

In the year 2000, Tate commissioned the first of fifteen net artworks for the then newly launched Tate website, Tate Online, which was devised as the fifth gallery. The commissioned artworks were meant to attract and challenge visitors to this still new online space. Initially these works were closely entwined with the main website, they were highlighted on the front page of the site, but as the number of works grew and Tate Online changed focus, these works were grouped together under the Intermedia Art microsite alongside contextualising texts, a programme of events and podcasts. The Intermedia website still exists online, but it has not been updated since 2012 and sits on a server that is now outdated and will eventually have to be decommissioned.

Tate does not archive its website, as a public body this is carried out by The National Archives UK Government Web Archive. It has a significant number of captures for the Intermedia website, but it is not consistent in capturing its interactive content - which was a key feature of several of the commissioned artworks. Therefore, due to these gaps and missing contextual information, there is not a representative or effective archived version of the Intermedia website, or the artworks available.

As part of the Andrew W. Mellon Foundation funded project Reshaping the Collectible: When Artworks Live in the Museum, a team of interdisciplinary researchers are looking at the history of the Net Art commissioning programme, the strategies to preserve the artworks and website as well

as looking to build Tate's capacity to collect internet art. The project is also an opportunity to go beyond the artwork collection and consider the same set of issues from the perspective of institutional records and the Tate Archive.

The developments in digital preservation, web archiving and more specifically in small scale web recording and emulation, means that this was the perfect moment to undertake extensive captures and documentation of the Intermedia Art website and individual artworks as they exist now. This has included extensive discussion with the artists who continue to host the works on their own servers.

This paper will present the different but complementary perspectives of both Tate's archive and Time-Based Media Conservation as they have worked together to understand the intricacies of documenting, conserving and maintaining the integrity and accessibility of web-based art and its online records in the contemporary art museum. It will discuss the tools and methodology used to archive the website and the plans to make it available as Tate's first website archived as a public record.

Biographies:

Patricia Falcao is a Time-based Media Conservator with a broad interest in the preservation of the digital components of contemporary artworks. She has worked at Tate since 2008, and currently works in the acquisition of media-based media artworks into the Collection. She currently collaborates with Tate's Research Department in the Reshaping the Collectible project, looking at the preservation of websites in Tate's context, as well as working with Tate's Technology team to continue to develop Tate's strategy for the preservation of high value digital assets. Patricia completed her MA at the University of the Arts in Bern with a thesis on risk assessment for software-based artworks. She continues to develop research in this field in her role as a Doctoral Researcher in the AHRC funded Collaborative Doctoral Program, between Tate Research and the Computing Department at Goldsmiths College, University of London. The subject of her research are the practices of software-based art preservation in collections, by artists and in the gaming industry.

Sarah Haylett is a professional Archivist; she received her MA in Archives and Records Management from UCL in 2014. She joined Tate in June 2018 having previously worked at Zaha Hadid Architects, The Photographers' Gallery and with private collectors. As part of the Reshaping the Collectible: When Artworks Live in the Museum project team, her research interests are rooted in the relationship between archival and curatorial theory and how, beyond a culture of compliance, Tate's record keeping can be more intuitive to research and collecting practice. She is very interested in sites of archival creation and intention, and how these are represented in artistic practice and the contemporary art museum.

PRONI Web Archive: A collaborative approach

Rosita Murchan
(Public Record Office of Northern Ireland - PRONI)

Keywords: Collaborations, challenges, resources, permissions, partnerships

ABSTRACT

The Public record of Northern Ireland web archive has been building its collection of websites for almost ten years, focusing initially on capturing the websites of our local councils and Government departments and those deemed historically or culturally important to Northern Ireland.

However, unlike the UK and Ireland, Northern Ireland do not have Legal deposit status and as a result we are sometimes limited as to what we can capture.

As the web archive has grown and evolved organically over the years with more and more requests for websites to be archived, PRONI has had to look at the issue of gaining permissions (and capturing sites without any legal deposit legislation) and on how we can continue to grow our collection with the limited resources we have available to us. One of the ways in which we are able to expand the scope of the collection is through collaborations not only with other institutes such as the British Library, that allow us to capture sites that would usually be outside our remit, but also by working in partnership with the other sections within our organisation.

The aim of this short presentation will be to look in more depth at PRONI's work with the web Archive, the strategies we have used to build it, our collaborative projects, and the challenges and obstacles we face as we continue to grow.

Biography:

Rosita Murchan has worked with the Public Record Office for two years and has been working solely on the web archive for one year.

An overview of 15 years of experience in archiving the Croatian web

Inge Rudomino
(Croatian Web Archive, National and University Library in Zagreb – NSK)

Marta Matijević
(Croatian Web Archive, National and University Library in Zagreb – NSK)

Keywords: legal deposit, Croatian Web Archive, web archiving, open access, online publication

ABSTRACT

National and University Library in Zagreb (NSK) began archiving Croatian web in 2004 , in collaboration with the University of Zagreb University Computing Centre (SRCE) when the Croatian Web Archive (HAW) was established. The basis for archiving web was the Law on libraries (1997) which subjected online publications to legal deposit. To harvest the web, HAW is using three different approaches: selective, .hr domain harvesting and thematic harvesting. In period from 2004 to 2010, HAW was based only on the concept of selective harvesting which implies that each resource is selected to be archived according to established Selection Criteria. Each title has a full level of bibliographic description and is retrievable in library online catalogue providing the end user with high quality archived copy. Special care is given to news portals which are archived daily. To each title and archived copy an URN:NBN identifier is assigned to ensure permanent access that is of great importance for future citations.

Since 2010, HAW conducts .hr domain crawls annually and harvests websites related to topics and events of national importance periodically. HAW's primal task is to ensure that harvested resources are preserved in their entirety, original format and with all the accompanying functionalities. Majority of harvested content is in open access.

The poster will present a fifteen years' experience of the National and University Library in Zagreb (NSK) in managing web resources with the emphasis on selective, domain and thematic harvestings as well as new website design with new functionalities.

Biographies:

Inge Rudomino: Senior librarian at Croatian Web Archive, National and University Library in Zagreb (Croatia). Graduated at Information Sciences (Librarianship), Faculty of Philosophy, University of Zagreb. From 2001 to 2007 works as a cataloguer in Department for Cataloguing Foreign Publications in National and University Library in Zagreb. Since 2007 works at Croatian Web Archive on tasks which include identification, selection, cataloguing, archiving, maintaining

Croatian Web Archive, communications with publishers, and promotion. Publishes articles in Croatian and conference proceedings in the field of web archiving.

Marta Matijević: MA is a librarian at Croatian Web Archive, National and University Library in Zagreb. Graduated Library and Information Science at Faculty of Humanities and Social Sciences in Osijek in 2016. From 2016 to 2018 has worked in academic and school libraries. Since 2019 works at Croatian Web Archive on identification, selection, cataloguing, archiving, maintaining Archive, communication with publishers and promotion. Her interests are web archiving and information theories and has published papers in such fields.

The UK Web Archive and Wimbledon: A Winning Combination

Robert McNicol
(Kenneth Ritchie Wimbledon Library, Wimbledon Lawn Tennis Museum)

Keywords: Tennis, Sport, Collaboration, Heritage, Preservation

ABSTRACT

Since January 2019, the Kenneth Ritchie Wimbledon Library, the world's largest tennis library, has been collaborating with the British Library on a web archiving project. The Wimbledon Library is curating the Tennis subsection of the UK Web Archive Sports Collection. The UK Web Archive aims to collect every UK website at least once per year and they also work with subject specialists to curate collections of websites on specific subjects. The ultimate aim is for the Tennis collection to contain all UK-based tennis-related websites. This will include websites relating to tournaments, clubs, players and governing bodies. It will also include social media feeds of individuals or organisations involved with tennis in the UK. Already we have collected the twitter feeds of all male and female British players with a world ranking.

We have also archived Wimbledon's own digital presence, including the award-winning Wimbledon.com, which celebrates its 25th anniversary in 2020. In addition to this we have archived Wimbledon's social media accounts, including those belonging to the Museum and the Wimbledon Foundation and its international digital presence in the form of the Wimbledon page on Weibo, a Chinese social media site. This falls within the scope of the project as, although the site is not an English language one, it is based in the UK.

The collaboration is mutually beneficial. For a small, specialist library such as ours, there are many advantages to having a partnership with the British Library. Equally, the UK Web Archive benefits from our specialist expertise in curating their Tennis collection.

In many ways, a project like this one is perfect for Wimbledon. Although our history and heritage are at the heart of everything we do, we're always innovating and striving to improve as well. That's why this project, which involves using the latest technology to preserve tennis history, is so exciting for us. This presentation will give an overview of why the Kenneth Ritchie Wimbledon Library wanted to get involved in web archiving, how the collaboration with the UK Web Archive came about and give an overview what has been collected so far.

Biography:

Since March 2016 I have worked as the Librarian of the Kenneth Ritchie Wimbledon Library, which is part of the Wimbledon Lawn Tennis Museum. Prior to this, I had a long career as a media librarian, mostly working in sport. From 2008 to 2016 I was Sport Media Manager at BBC Scotland in Glasgow. Before that, I also worked for the BBC in London and Aberdeen and I also worked briefly for ITV Sport and Sky Sports. I studied History at the University of Glasgow and Information and Library Studies at the University of Strathclyde.

Piloting access to the Belgian web-archive for scientific research: a methodological exploration

Dr. Peter Mechant
(Ghent University)

Sally Chambers
(Ghent University)

Eveline Vlassenroot
(Ghent University)

Friedel Geeraert
(KBR - Royal Library and
the State Archives of Belgium)

Keywords: research use of web archives, web-archiving, digital humanities, born-digital collections, digital research labs

ABSTRACT

The web is fraught with contradiction. On the one hand, the web has become a central means of information in everyday life and therefore holds the primary sources of our history created by a large variety of people (Milligan, 2016; Winters, 2017). Yet, much less importance is attached to its preservation, meaning that potentially interesting sources for future (humanities) research are lost. Web archiving therefore is a direct result of the computational turn and has a role to play in knowledge production and dissemination as demonstrated by a number of publications (e.g. Brügger & Schroeder, 2017) and research initiatives related to the research use of web archives (e.g. https://resaw.eu/).

However, conducting research, and answering research questions based on web archives - in short; ‘using web archives as a data resource for digital scholars’ (Vlassenroot et al., 2019) - demonstrates that this so-called ‘computational turn’ in humanities and social sciences (i.e. the increased incorporation of advanced computational research methods and large datasets into disciplines which have traditionally dealt with considerably more limited collections of evidence), indeed requires new skills and new software.

In December 2016, a pilot web-archiving project called PROMISE (PReserving Online Multiple Information: towards a Belgian StratEgy) was funded. The aim of the project was to (i) identify current best practices in web-archiving and apply them to the Belgian context, (ii) pilot Belgian

web-archiving, (iii) pilot access (and use) of the pilot Belgian web archive for scientific research, and (iv) make recommendations for a sustainable web-archiving service for Belgium. Now the project is moving towards its final stages, the project team is focusing on the third objective of the project, namely how pilot access to the Belgian web archive for scientific research. The aim of this presentation is to discuss how the PROMISE team approached piloting access to the Belgian web-archive for scientific research, including: a) reviewing how existing web-archives provide access to their collections for research, b) assessing the needs of researchers based on a range of initiatives focussing on research-use of web-archives (e.g. RESAW, BUDDAH, WARCnet, IIPC Research Working Group, etc. and c) exploring how the five persona's created as part of the French National Library's Corpus project (Moiraghi, 2018) could help us to explore how different types of academic researchers that might use web archives in their research. Finally, we will introduce the emerging Digital Research Lab at the Royal Library of Belgium (KBR) as part of a long-term collaboration with the Ghent Centre for Digital Humanities (GhentCDH) which aims to facilitate data-level access to KBR's digitised and born-digital collections and could potentially provide the solution for offering research access to the Belgian web-archive.

Bibliography

Brügger, N. & Schroeder, R. (Eds.). (2017). *The web as history: Using web archives to understand the past and present*. London: UCL Press.

Milligan, I. (2016). Lost in the infinite archive: the promise and pitfalls of web archives. International Journal of Humanities and Arts Computing, 10(1), 78-94. Doi: 10.3366/ijhac.2016.0161.

Moiraghi, E. (2018). Le projet Corpus et ses publics potentiels: Une étude prospective sur les besoins et les attentes des futurs usagers. [Rapport de recherche] Bibliothèque nationale de France. 2018. ⟨hal-01739730⟩

Winters, J. (2017). Breaking into the mainstream: demonstrating the value of internet (and web) histories. *Internet Histories*, 1(1-2), 173-179. https://doi.org/10.1080/24701475.2017.1305713.

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1(1), 85-111. https://doi.org/10.1007/s42803-019-00007-7

Biographies:

Dr Peter Mechant holds a PhD in Communication Sciences from Ghent University (2012). After joining research group mict (www.mict.be), Peter has been mainly working on research projects related to e-gov (open and linked data), smart cities, online communities and web archiving. As senior researcher, he is currently involved in managing projects and project proposals at a European, national as well as regional level.

Sally Chambers is Digital Humanities Research Coordinator at the Ghent Centre for Digital Humanities, Ghent University, Belgium and National Coordinator for DARIAH in Belgium. She is one of the instigators of an emerging Digital Research Lab at KBR, Royal Library of Belgium as part of a long-term collaboration with the Ghent Centre for Digital Humanities. This lab will facilitate data-level access to KBR's digitised and born-digital collections for digital humanities research. Her role in PROMISE relates to research access and use of Belgium's web-archive.

Eveline Vlassenroot holds a Bachelor Degree in Communication Sciences (Ghent University) and graduated in 2016 as a Master in Communication Sciences with a specialisation in New Media and Society (Ghent University). After completing additional courses in Information Management & Security at Thomas More Mechelen (KU Leuven), she joined imec-mict-Ghent University in September 2017. She participates in the PROMISE project (Preserving Online Multiple Information: towards a Belgian StratEgy), where she is researching international best-practices for preserving and archiving online information. She is also involved in several projects with the Flemish government regarding data standards, the governance of interoperability standards and linked open data.

Friedel Geeraert is a researcher at KBR (Royal Library) and the State Archives of Belgium, where she works on the PROMISE project that focuses on the development of a Belgian web archive at the federal level. Her role in the project includes comparing and analysing best practices regarding selection of and providing access to the information and data to be archived and making recommendations for the development of a long-term and sustainable web archiving service in Belgium.

Reimagining Web Archiving as a Realtime Global Open Research Platform: The GDELT Project

Dr. Kalev Hannes Leetaru

(The GDELT Project)

Keywords: GDELT Project; realtime; research-first web archive; news homepages

ABSTRACT

The GDELT Project (https://www.gdeltproject.org/) is a realization of the vision I laid out at the opening of the 2012 IIPC General Assembly for the transformation of web archives into open research platforms. Today GDELT is one of the world's largest global open research datasets for understanding human society, spanning 200 years in 152 languages across almost every country on earth. Its datasets span text, imagery, spoken word and video, enabling fundamentally new

kinds of multimodal analyses and reach deeply into local sources to reflect the richly diverse global landscape of events, narratives and emotions.

At its core, GDELT in the web era is essentially a realtime production research-centered web archive centered on global news (defined as sources used to inform societies, both professional and citizen-generated). It continually maps the global digital news landscape in realtime across countries, languages and narrative communities, acting both as archival facilitator (providing a live stream of every URL it discovers to organizations including the Internet Archive for permanent preservation) and research platform.

In contrast to the traditional post-analytic workflow most commonly associated with web archival research, in which archives are queried, sampled and analyzed after creation, GDELT focuses on realtime analysis, processing every single piece of content it encounters through an ever-growing array of standing datasets and APIs spanning rules-based, statistical and neural methodologies. Native analysis of 152 languages is supported, while machine translation is used to live translate everything it monitors in 65 languages, enabling language-independent search and analysis.

Twin global crawler and computational fleets are distributed across 24 data centers across 17 countries, leveraging Google Cloud's Compute Engine and Cloud Storage infrastructures, coupled with its ever-growing array of AI services and APIs, underpinning regional ElasticSearch and bespoke database and analytic clusters and all feeding into petascale analytic platforms like BigQuery and Inference API for at-scale analyses. This massive global-scale system must operate entirely autonomously, scale to support enormous sudden loads (such as during breaking disasters) and function within an environment in which both the structure (rendering and transport technologies) and semantics (evolving language use) are in a state of perpetual and rapid change.

Traditional web archives are not always well-aligned with the research questions of news analysis, which often require fixed time guarantees and a greater emphasis on areas like change detection and agenda setting. Thus, GDELT includes numerous specialized news-centric structural datasets including the Global Frontpage Graph that catalogs more than 50,000 major news homepages every hour on the hour, totaling nearly a quarter trillion links over the last two years to support agenda setting research. The Global Difference Graph recrawls every article after 24 hours and after one week with fixed time guarantees to generate a 152-language realtime news editing dataset cataloging stealth editing and silent deletions. Structural markup is examined and embedded social media posts cataloged as part of its Global Knowledge Graph. A vast distributed processing pipeline performs everything from entity extraction and emotional coding to SOTA language modeling and claims and relationship mapping. Images are extracted from each article and analyzed by Cloud Vision, enabling analysis of the visual landscape of the web. Datasets from quotations to geography to relationships to emotions to entailment and dependency extracts are all computed and output in realtime, operating on either native or translated content.

In essence, GDELT doesn't just crawl the open web, it processes everything it sees in realtime to create a vast archive of rich realtime research datasets. This firehose of data feeds into downloadable datasets and APIs to enable realtime interactive analyses, while BigQuery enables at-scale explorations of limitless complexity, including one-line terascale graph construction and geographic analysis and full integration with the latest neural modeling approaches.

Full integration with GCE, GCS and BigQuery couples realtime analysis of GDELT's rich standing annotations with the ability to interactively apply new analyses including arbitrarily complex neural modeling at scale. This means that GDELT is able to both provide a standing set of realtime annotations over everything it encounters and support traditional post-facto analysis at the effectively infinite scale of the public cloud.

From mapping global conflict and modeling global narratives to providing the data behind one of the earliest alerts of the COVID-19 pandemic, GDELT showcases what a research-first web archive is capable of and how to leverage the full power of the modern cloud in transforming web archives from cold storage into realtime open research platforms.

Biography

Dr. Kalev Hannes Leetaru - One of Foreign Policy Magazine's Top 100 Global Thinkers of 2013, Kalev founded the open data GDELT Project. From 2013-2014 he was the Yahoo! Fellow in Residence of International Values, Communications Technology & the Global Internet at Georgetown University's Edmund A. Walsh School of Foreign Service, where he was also an Adjunct Assistant Professor, as well as a Council Member of the World Economic Forum's Global Agenda Council on the Future of Government. His work has been profiled in the presses of more than 100 nations and in 2011 The Economist selected his Culturomics 2.0 study as one of just five science discoveries deemed the most significant developments of 2011. Kalev's work focuses on how innovative applications of the world's largest datasets, computing platforms, algorithms and mind-sets can reimagine the way we understand and interact with our global world. More on his latest projects can be found on his website at https://www.kalevleetaru.com/ or https://blog.gdeltproject.org.

Archiving 1418-Now using Rhizome's Webrecorder: observations and reflections

Anisa Hawes
(Independent Curatorial Researcher and Web Archivist)

Keywords: web archiving tools, social media, curation, process, Webrecorder

ABSTRACT

This paper explores the challenges of archiving https://www.1418now.org.uk/ and its associated social media profiles (Twitter, Instagram, and YouTube) using Rhizome's Webrecorder.

These web collections form an integral part of the Imperial War Museum's record of the 14-18Now WW1 Centenary Art Commissions programme and represent a recognition that essential facets of many of the Commissions would otherwise be absent from the archive.

Immediate public responses to Jeremy Deller's modern memorial event We're Here Because We're Here, for example, played out in the contemporary context of Web 2.0. Many people who encountered the memorial directly were moved to share their reflections on social media. Many others encountered the event indirectly: via messages, images, and videos which circulated on social networking platforms. In this way, the online sphere became an expanded site of public participation and experience. Meanwhile, imprinted engagement metrics and appended comments threads provided unprecedented curatorial insight into the artwork's impact and reach.

Webrecorder is a free, open-source web archiving tool developed by Rhizome. It enables high-fidelity capture of complex, interactive web pages, including social media sites. Written from the point of view of a curatorial researcher, this paper includes insights into the web archiving process and workflow. Combining work-in-progress screenshots and reflections extracted from my log notes, I'll explain how I have utilised Webrecorder's automation features and scripted behaviours alongside manual, action-by-action capture to build a rich collection, tackling the challenge of archiving both in-detail and at-scale.

Biography:

Anisa Hawes is an independent curatorial researcher and web archivist based in London, UK. As an embedded researcher at the Victoria and Albert Museum (2015-18) her work investigated how digital tools and software environments have altered design practice; and how the web and social media have produced new, participatory poster forms—such as memes which are appropriated as they circulate. Collaborating with Rhizome and British Library/UK Web Archive, she tested web archiving technologies to capture digital objects in the context of the platforms where they are created and encountered, whilst developing a framework of curatorial principles to support digital collecting.

Managing the Lifecycle of Web Archiving at a Large Private University

Nicole Greenhouse
(New York University Libraries)

Keywords: workflows, accessioning, description, quality assurance, context

ABSTRACT

New York University Libraries has been archiving websites since 2007. The collection, developed using the service Archive-It, consists of websites related to Labor and Left movements, the New York City downtown arts scene, contemporary composers, and university websites, totaling approximately 5000 websites and 13 terabytes of data. In 2018, I was hired as the first permanent structural archivist whose role is to solely manage the web archiving program. During this first year, it was important to the Archival Collections Management department in the NYU Libraries to incorporate web archiving in the greater workflows of the department as well as manage the day to day work that comes with web archiving, including capture, website submissions, quality assurance, and access and description. This presentation will discuss how we have developed a database to manage capture and quality assurance, as well as the ongoing project to accession recently added websites and create consistent description across all of the archived websites. The database allows us to track the lifecycle of each archived website and take advantage of the scoping and quality assurance tools provided by Archive-it but work around the service's limitations. The presentation will conclude with an overview of descriptive practices by creating accession records to track why curators and archivists add websites to the collection and update finding aids that provide a greater amount of contextual description that goes beyond Dublin Core and in line with

the department's descriptive policies to create transparent and standards compliant description in the context of the Special Collection's analog collections. By creating records that put the web archives in the context of the rest of the collections, NYU is able to promote the use of the archived websites.

Biography:

Nicole Greenhouse is the Web Archivist in the Archival Collections Management department at New York University Libraries. Nicole received her MA in Archives and Public History at NYU. She has previously worked at the Winthrop Group, the Center for Jewish History, and the Jewish Theological Seminary on a variety of analog and digital archives projects. She is currently the Communications Manager for the Web Archiving Section of the Society of American Archivists.

Thematic web crawling and scraping as a way to form focussed web archives

Benedikt Adelmann MSc
(University of Hamburg)

Dr. Lina Franken
(University of Hamburg)

Keywords: web crawling, scraping, thematic focussed web archives, discourse analysis

ABSTRACT

For humanities and social science research on the contemporary, the web and web archives are growing in their relevance. Not much is available when it comes to thematically based collections of websites. In order to find out about ongoing online discussions, a web crawling and scraping is needed as soon as a larger collection shall be generated as a corpus for further explorations.

Within the study presented here, we focus on the acceptance of telemedicine and its challenges. For the discourse analysis conducted (Keller 2005), the concept of telemedicine often is discussed within a broader field of digital health systems, while there are only few statements of relevance within single texts. Therefore, a large corpus is needed to identify relevant stakeholders and discourse positions and go into details of text passages – big data turns into small data and has to be filtered (see Koch/Franken 2019). Thematic web crawling and scraping (Barbaresi 2019: 30) is a mayor facilitator with these steps.

Web crawling has to start from a list of so-called seed URLs, which in our case refer to the main pages of web sites of organizations (e.g. health insurance companies, doctors' or patients' associations) known to be involved in the topic of interest. From these seed URLs, our crawl explores the network structure expressed by the (hyper)links between webpages in a breadth-first manner (see Barbaresi 2015: 120ff. for an overview of web crawling practices). It is able to handle content with MIME types text/html, application/pdf, application/x-pdf and text/plain. Content text is extracted and linguistically pre-processed: tokenization, part-of-speech tagging, lemmatization (reduction of word forms to their basic forms). If the lemmatized text contains at least one of some pre-defined keywords (see Adelmann et al. 2019 for this semantic-field based approach), the original content of the webpage (HTML, PDF etc.) is saved as well as the results of the linguistic

pre-processing. (Hyper)links from HTML pages are followed if they refer to (other) URLs of the same host. If the HTML page is a match, and only then, links are also followed if their host is different. We employ some heuristics to correct malformed URLs and avoid a variety of non-trivial equivalences since we are testing whether a URL has already been visited by the crawler. Of saved pages, the crawler records accessed URLs, date and time of access, and other metadata, including the matched keywords. URLs only visited (but not saved) are recorded without metadata; found links between them are as well. The script is published as hermA-Crawler (Adelmann 2019).

When using focussed web archives formed in this way, it is easy to use different approaches such as topic modelling (Blei 2012) or sentiment analysis (D'Andrea et al. 2015) on a larger base in order to support discourse analysis with digital humanities approaches.

References:

Adelmann, Benedikt; Andresen, Melanie; Begerow, Anke; Franken, Lina; Gius, Evelyn; Vauth, Mi-chael: Evaluation of a Semantic Field-Based Approach to Identifying Text Sections about Specific Topics. In: Book of Abstracts DH2019. https://dh2019.adho.org/wp-content/uploads/2019/04/Short-Papers_23032019.pdf.

Adelmann, Benedikt: hermA-Crawler. https://github.com/benadelm/hermA-Crawler.

Barbaresi, Adrien: Ad hoc and general-purpose corpus construction from web sources. Doctoral dissertation, Lyon, 2015.

Barbaresi, Adrien: The Vast and the Focused: On the need for thematic web and blog corpora. In: Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7), Cardiff, 2019. DOI: https://doi.org/10.14618/ids-pub-9025

Blei, David M.: Probabilistic topic models. Surveying a suite of algorithms that offer a solution to managing large document archives. In: Communications of the ACM 55 (2012), S. 77–84.

D'Andrea, Alessia; Ferri, Fernando; Grifoni, Patrizia; Guzzo, Tiziana: Approaches, Tools and Applications for Sentiment Analysis Implementation. In: International Journal of Computer Applications 125 (2015), S. 26–33. DOI: 10.5120/ijca2015905866.

Keller, Reiner: Analysing Discourse. An Approach from the Sociology of Knowledge. In: Forum: Qualitative Social Research Volume 6, No. 3, Art. 32 (2005). DOI: http://dx.doi.org/10.17169/fqs-6.3.19

Koch, Gertraud; Franken, Lina: Automatisierungspotenziale in der qualitativen Diskursanalyse. Das Prinzip des „Filterns“. In: Sahle, Patrick (ed.): 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd 2019). Digital Humanities: multimedial & multimodal.

Universitäten zu Mainz und Frankfurt, March 25 to 29, 2019. Book of Abstracts, pp. 89–91. DOI: 10.5281/zenodo.2596095

Biographies:

Benedikt Adelmann is a computer scientist at the University of Hamburg. Lina Franken is a cultural anthropologist at the University of Hamburg. Together, they are working within the collaborative research project "Automated modelling of hermeneutic processes – The use of annotation in social research and the humanities for analyses on health (hermA)". See https://www.herma.uni-hamburg.de/en.html.

Metadata for social science research

Andrea Prokopová
(Webarchiv, National Library of the Czech Republic)

Keywords: web archiving, metadata, big data, social sciences, data mining

ABSTRACT

The Czech web archive of National Library of the Czech Republic (Webarchiv) is one of the oldest in Europe (since 2000). It is therefore able to provide methodological support to new web archives and also has a large amount of harvested data. However, data cannot be provided due to copyright. At least there is the opportunity to use metadata of harvested web resources.

Two years ago, sociologists from the Academy of Sciences of the Czech Republic showed interest in the data for their research. This started their cooperation with the Czech web archive and also with the Technical University in Pilsen. These three institutions are currently working together to Development of the Centralized Interface for the Web content and Social Networks Data Mining.

The data sets that researchers prepare on their own using the interface can be used for various data analysis and interpretation of social trends and changes in the Internet environment. In the first phase of the project, a basic analysis of the content of the web archive took place. This revealed that the web archive contains nearly 9 and a half billion unique digital objects. These can be text, image, audio and video objects, or other digital objects (software, scripts, etc.). The analysis provided accurate information on how many objects are in the Webarchive with the current size.

The next phase was the programming work itself. There is already a prototype of the search engine that is in the process of internal testing.

Bibliography:

BRÜGGER, Niels, Niels Ole FINNEMANN, 2013. The Web and digital humanities: Theoretical and methodological concerns. Journal of Broadcasting & Electronic Media [online]. 2013, s. 66-80. ISSN 1550-6878. Dostupné z: http://thelecturn.com/wp-content/uploads/2013/07/The-web-and-digital-humanities-Theoretical-and-Methodological-Concerns.pdf

KVASNICA, Jaroslav, Marie HAŠKOVCOVÁ a Monika HOLOUBKOVÁ. Jak velký je Webarchiv? E-zpravodaj Národní knihovny ČR [online]. Praha: Národní knihovna ČR, 2018, 5(5), 6 - 7 [cit. 2020-01-22]. Dostupné z: http://text.nkp.cz/o-knihovne/zakladni-informace/vydane-publikace/soubory/ostatni/ez_2018_5.pdf

KVASNICA, Jaroslav, Andrea PROKOPOVÁ, Zdenko VOZÁR a Zuzana KVAŠOVÁ. Analýza českého webového archivu: Provenience, autenticita a technické parametry. ProInflow [online]. 2019, 11(1) [cit. 2020-01-22]. DOI: 10.5817/ProIn2019-1-2. ISSN 1804-2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/2019-1-2 Webarchiv: O Webarchivu [online].

Praha, 2015 [cit. 2020-01-22]. Dostupné z: https://www.webarchiv.cz/cs/o-webarchivu

Biography

I work as data analyst at Czech Webarchive and also in a project called *Centralized Interface for the Web content and Social Networks Data Mining*. Our goal is to provide datasets of metadata to scientists from humanities especially sociologists for their future research and data analýzy. Webarchiv is a part of NationaI Library of the Czech Republic. We harvest and archive all web sources with the Czech domain. I study Library studies and information science at Masaryk University, so I currently work in my field. I am a typical book worm with a creative soul and a passion for photography.

Exploring Web Archive Networks:
The Case of the 2018 Irish Presidential Election

Dr. Derek Greene
(University College Dublin)

Keywords: web archives, network analysis, data analysis, case study

ABSTRACT

The hyperlink structure of the Web can be used not only for search, but also to analyse the associations between websites. By representing large collections of web pages as a link network, researchers can apply existing methodologies from the field of network analysis. For web archives, we can use these methods to explore their content, potentially identifying meaningful historical trends.

In recent years the National Library of Ireland (NLI) has selectively archived web content covering a variety of political and cultural events of public interest. In this work, we analyse an archive of websites pertaining to the 2018 Irish Presidential Election.

The original archive consists of a total of 57,065 HTML pages retrieved in 2018. From this data we extracted all links appearing in these pages and mapped each link to a pair of domains. For our case study, we focus only on pairs of domains for which both the source and target are distinct, yielding 28,555 relevant domain pairs.

Next, we created a directed weighted network representation. In this network, each node is a unique domain. Each edge from node A to node B indicates that there are one or more links in the pages on domain A pointing to domain B. Each edge also has a weight, indicating the number of links between two domains. This yielded a network with 263 nodes and 284 weighted directed edges. Using network diagrams generated on this data, we can visualise the link structure around the sites used to promote each presidential candidate, and how they relate to one another.

This work highlights the potential insights which can be gained by using network analysis to explore web archives. These include the possible impact on collection development in the NLI selective web archive and the further study of the archived Irish web.

Biography:

Dr. Derek Greene is Assistant Professor at the UCD School of Computer Science and Research Investigator at the SFI Insight Centre for Data Analytics. He has over 15 years' experience in AI and machine learning, with a PhD in Computer Science from Trinity College Dublin. He is involved in a range of interdisciplinary projects which involve applying machine learning methods in fields such as digital humanities, smart agriculture, and political science.

IIPC: training, collecting, research, and outreach activities

Dr. Olga Holownia
(International Internet Preservation Consortium / British Library)

Keywords: web archiving, web archiving training, collaborative collections, Covid-19 web archive collections, web archiving resources

ABSTRACT

The basis of founding the International Internet Preservation Consortium (IIPC) in 2003 was the acknowledgement of "the importance of international collaboration for preserving Internet content for future generations". Over the years, the IIPC members have worked together on multiple technical, curatorial, and educational activities. They have developed standards and supported open source web archiving tools and software. The annual General Assembly (GA) and Web Archiving Conference (WAC) have provided a forum for exchanging knowledge and forging new collaborations not only within the IIPC but also within the wider web archiving community and beyond. This talk will give an update on the most recent activities, including the IIPC funded projects as well as initiatives led by the working groups: training, collecting, and research, all of which fall under membership engagement and outreach overseen by the IIPC Portfolios.

One of the key initiatives this year has been the "Novel Coronavirus (Covid-19) outbreak" transnational collection coordinated by the IIPC Content Development Group and organised in partnership with the Internet Archive. Over 9000 sites from over 140 countries and over 160 top level domains were made available through Archive-It seven months after the collection was launched in February 2020. We have also been publishing blog posts documenting the IIPC members' efforts at capturing and archiving web content related to the pandemic within the national domains.

This year also saw the publication of training materials designed and produced by the IIPC Training Working Group in partnership with the Digital Preservation Coalition. The first module comprising eight sessions, is aimed at curators, policy makers and managers or those who would like to learn about the basics of web archiving, including what web archives are, how they work,

and how web archive collections are curated. The programme helps acquire basic skills in capturing web archive content, but also how to plan and implement a web archiving programme.

In terms of research activities, alongside the repository of web archiving resources at the University of North Texas (UNT) Digital Library and enhancing the metadata in the Zotero bibliography, we have been promoting the outcomes of the IIPC funded projects through a series of webinars organised by the Research Working Group. Among the funded projects are a set of introductory Jupyter Notebooks developed by Tim Sherratt, the creator of the GLAM Workbench, and LinkGate, a tool for graph visualisation of web archives aided by an inventory of use cases. The former project was led by the UK Web Archive based at the British Library, in partnership with the Australian and the New Zealand web archives, the latter is a collaboration between Bibliotheca Alexandrina and the National Library of New Zealand.

References

About IIPC: https://netpreserve.org/about-us
IIPC Working Groups: https://netpreserve.org/about-us/working-groups
IIPC Projects: https://netpreserve.org/projects
IIPC General Assembly and Web Archiving Conference: https://netpreserve.org/general-assembly
IIPC collections in the UNT Digital Library: https://digital.library.unt.edu/explore/partners/IIPC
IIPC members' COVID-19 collections: https://netpreserveblog.wordpress.com/tag/covid-19-collection
"Novel Coronavirus (Covid-19) outbreak" collaborative collection: https://archive-it.org/collections/13529

Biography

Olga Holownia is Programme and Communications Officer based at the British Library. She manages the communications and provides support to the programmes of the International Internet Preservation Consortium (netpreserve.org). Her key projects include the organisation of the annual IIPC General Assembly and Web Archiving Conference as well as associated training and events. She is a co-chair of the IIPC Research Working Group.

Using Web Archives to Teach and Opportunities in the Information Science Field

Dr. Juan-José Boté
(Universitat de Barcelona)

Keywords: digital preservation, teaching, web archives, emulator, archiving software

ABSTRACT

Web archives are a useful tool for teaching different subjects to students, not only for history but also for teaching courses such as digital preservation, information architecture, or metadata structures.

The digital preservation of web archives offers a unique set of challenges when teaching students about information science. The first one is teaching about search strategies. Web archives have specific search tools and it is necessary to develop search strategies before beginning any search. For instance, one of the main challenges for students is in learning how to look for information through collections or looking for a precise website.

Secondly, in addition to search strategies, the students need to learn how to find and use old software to run images, videos, or other informational content. Part of the search process includes checking whether the archived software was commercial and whether it is possible to use for free with some limitations. Therefore, to run old software which can be downloaded from web archives, sometimes it is also necessary to use emulators to run the old software. Emulators are not always found in web archives and may not be available and students must add a further step in order to run old software.

In addition, when students set up archiving software, it is useful to know how it works. Testing the possibilities of archiving software is often kept to small scenarios because of the limitations of the course. Exposure to archiving software would permit students to learn the process of building small collections or creating new datasets of archived websites.

In this paper I explore different uses of the information science field when using web archives as a resource for teaching, which is especially helpful in a digital preservation course.

Biography:

Juan-José Boté is Assistant Professor at Universitat de Barcelona where he is also the coordinator of the Postgraduate Program on Social Media Content. His research is focused on digital preservation and cultural heritage.

Web archiving - professionals and amateurs

Bartłomiej Konopa
(State Archives in Bydgoszcz; Nicolaus Copernicus University)

Keywords: web archives, professional web archiving, amateurish web archiving, ArchiveTeam, comparative study

ABSTRACT

Web archiving can be defined as "any form of deliberate and purposive preserving of web material" (Brügger, 2011). That broad definition allows us to divide web archiving on numerous levels and distinguish many types of it. One of the possible distinctions is between professional and amateurish archiving. As professional archive one can treat big projects led mainly by national libraries, which employs experts and have strict regulations, like for example UK Web Archive and Danish Netarkivet. They are interested in national Webs and mainly preserve resources from one ccTLD in routine and repeatable crawls. Sometimes these archives build special collections, but very often they are predictable and related to "real world" events, for instance national elections. On the other side, as amateurish, one can recognize initiatives like ArchiveTeam, which are open for Internet users and does not have rigorous rules. They react to what is happening on the Web, observe endangered websites and services and try to preserve it. Their actions are spontaneous and disposable, but precisely aimed on the resources that would be lost. Both sides are trying to preserve web resources, because they consider them as digital heritage, which needs to be saved for the future generations. However, despite the mutual goal, professional and amateurish archives visibly differ in the way they function and materials they are interested as described above. The paper will search for these differences and analyse its influence on how and what will be archived, and then available for those, who want to experience and research the past Web. To reach this goal the author will compare UK Web Archive and Netarkivet with ArchiveTeam. Main source of information about these projects will be papers, news and their websites. The most important elements of these studies will be selection policy and criteria, scope, frequency and methods of archiving, and access rules. It will show differences in thinking about

Web, its border and ways of preserving and sharing this digital heritage. These factors will have also an impact on what resources will be available for later studies.

Biography:

Bartłomiej obtained his master's degree in archival science in 2007, currently he is a senior archivist at the State Archives in Bydgoszcz and a PhD student at the Nicolaus Copernicus University in Toruń (Poland). He is preparing a doctoral dissertation on Web archives, which are his main research interest. He collaborated with the web archiving lab "webArch", which is a pioneering project to popularize this issue in Poland.

Digital archaeology in the web of links: reconstructing a late-90s web sphere

Dr Peter Webster
(Independent Scholar, Historian and Consultant)

Keywords: web spheres, method, link graphs, link analysis, reconstruction

ABSTRACT

As interest in Web history has grown, so has the understanding of the archived Web as an object of study. But there is more to the Web than individual objects and sites. This paper is an exercise in understanding a particular ‘web sphere’.

Niels Brügger defines a web sphere as ‘web material … related to a topic, a theme, an event or a geographic area’ (Brügger 2018). I posit a distinction between ‘hard’ and ‘soft’ web spheres, defined in terms of the ease with which their boundaries may be drawn, and the rate at which those boundaries move over time. Examples of hard web spheres are organisations that have clear forms of membership or association: eg. the websites of the individual members of the European Parliament. The study of ‘soft’ web spheres tends to present additional difficulties, since the definition of topics or themes is more difficult if not expressed in institutional terms. The definition of ‘European politics’ may be contested in ways that ‘membership of the European Parliament’ may not. I present a method of reconstructing just such a soft web sphere, much of which is lost from the live web and exists only in the Internet Archive: the web estate of conservative Christian campaign groups in the UK in the 1990s and early 2000s.

The historian of the late 1990s has a problem. The vast bulk of content from the period is no longer on the live web; there are few, if any, indications of what has been lost – no inventory of the 1990s Web against which to check; of the content that was captured by the Internet Archive, only a superficial layer is exposed to full-text search, and the bulk may only be retrieved by a search for the URL. We do not know what was never archived, and in the archive it is difficult to find what we might want, since there is no means of knowing the URL of a lost resource.

We need, then, to understand the archived Web using only the technical data about itself that it can be made to disclose. This method of web sphere reconstruction is based not on page content but on the relationships between sites, i.e., the web of hyperlinks. The method is iterative, involving the computational interrogation of large datasets from the British Library and the close examination of individual archived pages, along with the use of printed and other non-digital sources. It builds upon recent studies which explore the available primary sources from outside the Web from which it may be reconstructed (Nanni 2017; Teszelszky 2019, Ben-David 2016; Ben-David 2019). It develops my earlier work in which the method was applied to smaller, less complex spheres (Webster 2017; Webster 2019).

References:

Ben-David, Anat. 2016. What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. New Media and Society 18, 1103-1119. https://doi.org/10.1177/1461444816643790

Ben-David, Anat. 2019. National web histories at the fringe of the Web: Palestine, Kosovo and the quest for online self-determination. In: *The Historical Web and Digital Humanities: the Case of National Web domains*, eds Niels Brügger & Ditte Laursen, 89-109. London: Routledge.

Brügger, Niels. 2018. *The archived Web: Doing history in the digital age*. Cambridge, MA: MIT Press.

Nanni, Federico. 2017. Reconstructing a website's lost past: methodological issues concerning the history of Unibo.it. *Digital Humanities Quarterly* 11. http://www.digitalhumanities.org/dhq/vol/11/2/000292/000292.html

Teszelszky, Kees. 2019. Web archaeology in The Netherlands: the selection and harvest of the Dutch web incunables of provider Euronet (1994–2000). *Internet Histories* 3, 180-194, DOI: 10.1080/24701475.2019.1603951

Webster, Peter. 2017. Religious discourse in the archived web: Rowan Williams, archbishop of Canterbury, and the sharia law controversy of 2008. In: *The Web as History,* eds Niels Brügger & Ralph Schroeder, 190-203. London: UCL Press.

Webster, Peter. 2019. Lessons from cross-border religion in the Northern Irish web sphere: understanding the limitations of the ccTLD as a proxy for the national web. In: *The Historical Web and Digital Humanities: the Case of National Web domains*, eds Niels Brügger & Ditte Laursen, 110-23. London: Routledge.

Biographies:

Dr Peter Webster is an independent scholar and consultant, and founder and managing director of Webster Research and Consulting (UK). He has published widely on the use of Web archives for contemporary history.

Web defacements and takeovers and their role in web archiving

Michael Kurzmeier
(Maynooth University)

Keywords: defaced websites; hacktivism; cybercrime archives; Geocities; web archives

ABSTRACT

This paper will provide insight into the archiving and utilization of defaced websites as ephemeral, non-traditional web resources. Web defacements as a form of hacktivism are rarely archived and thus mostly lost for systematic study. When they find their way into web archives, it is often more as a by-product of a larger web archiving effort than as the result of a targeted effort. Aside from large collections such as Geocities, which during a crawl might pick up a few hacked pages, there also exists a small scene of community-maintained cybercrime archives that archive hacked web sites, some of which are hacked in a hacktivist context. By examining sample cases of cybercrime archives, the paper will show the ephemerality of their content and introduce a framework for analysis.

As more and more of our daily communication happens digitally, marginalized and counter-public groups have often used the new media to overcome real-world limitations. This phenomenon can be traced back to the early days of the Web. This paper will provide an overview of defacements on the web and show the role web archives play in understanding these phenomena. Web defacements are ephemeral content and as such especially prone to link rot and deletion. They can provide not only information on the history of a single web page; they can also be seen as artifacts of a struggle for attention. Contextualized with metadata and the original page, defacements can add help restore such lost histories. The current state, however, is that only a number of collections are still online with only one collection still accepting new material and none being in a condition to be used for academic research. Finding relevant defacements in collections like the mentioned is a challenge, especially since there is little conformity in terms of content, language and layout between people hacking websites. The paper will introduce different approaches to methodology for identifying defacements and related pages.

Biography:

Michael Kurzmeier is a fourth-year PhD candidate in Digital Humanities and recipient of the Irish Research Council Postgraduate Scholarship. His research interest is the intersection between technology and society. His PhD thesis investigates the use of hacktivism as a tool of political expression. The research is grounded in an understanding of a contested materiality of communication, in which hacktivism is one method to occupy contested space. Michael is working with Kylie Jarrett (MU Media Studies) and Orla Murphy (UCC Digital Humanities). ORCID: https://orcid.org/0000-0003-4925-5197.

MyKnet.org: Traces of Digital Decoloniality in an Indigenous Web-Based Environment

Dr. Philipp Budka
(University of Vienna; Free University Berlin)

Keywords: MyKnet.org, indigenous web-based environment, digital decoloniality, internet history, anthropology

ABSTRACT

This paper discusses traces of digital decoloniality (e.g., Deem 2019) by exploring the history of the indigenous web-based environment MyKnet.org. By considering the cultural and techno-social contexts of First Nations' everyday life in Northwestern Ontario, Canada, and by drawing from ethnographic fieldwork (e.g., Budka 2015, 2019), it critically reviews theoretical accounts and conceptualizations of change and continuity that have been developed in an anthropology of media and technology (e.g., Postill 2017). In so doing, it examines how techno-social change and cultural continuity can be conceptualized in relation to each other and in the context of (historical) processes of digital decoloniality.

In 1994, the tribal council Keewaytinook Okimakanak (KO) established the Kuh-ke-nah Network (KO-KNET) to connect indigenous people in Northwestern Ontario' remote communities through and to the internet. At that time, a local telecommunication infrastructure was almost non-existent. KO-KNET started with a simple bulletin board system that developed into a community-controlled ICT infrastructure, which today includes landline and satellite broadband internet as well as internet-based mobile phone communication. Moreover, KO-KNET established services that became widely popular among the local indigenous communities such as the web-based environment MyKnet.org.

MyKnet.org was set up in 1998 exclusively for First Nations people to create and maintain personal homepages within a cost- and commercial-free space on the web. Particularly between 2004 and 2008, MyKnet.org used to be extremely popular mainly because of two reasons. First, MyKnet.org enabled people to establish and maintain social relationships across spatial distance in an

infrastructurally disadvantaged region. They communicated through homepage's communication boxes and they linked their homepages to the pages of family members and friends. Creating thus a "digital directory" of indigenous people in Northwestern Ontario. Second, MyKnet.org contributed to different forms of cultural representation and identity construction. Homepage producers utilized the service to represent and negotiate their everyday lives by displaying and sharing pictures, music, texts, website layouts, and artwork.

During fieldwork in Northwestern Ontario (2006-2008), many people told me stories about their first MyKnet.org websites in the early 2000s and how they evolved. People vividly described how their homepages were designed and structured and to which other websites they were linked. To deepen my interpretation and understanding of these stories, I used the Internet Archive's Wayback Machine to recover archived versions of these websites whenever possible. Thus, the Wayback Machine became an important methodological tool for my research into the decolonial history of MyKnet.org and related practices and processes of techno-social change and cultural continuity.

References:

Budka, P. (2019). Indigenous media technologies in "the digital age": Cultural articulation, digital practices, and sociopolitical concepts. In S. S. Yu & M. D. Matsaganis (Eds.), Ethnic media in the digital age (pp. 162-172). New York: Routledge.

Budka, P. (2015). From marginalization to self-determined participation: Indigenous digital infrastructures and technology appropriation in Northwestern Ontario's remote communities. Journal des Anthropologues, 142-143(3), 127–153.

Deem, A. (2019). Mediated intersections of environmental and decolonial politics in the No Dakota Access Pipeline movement. Theory, Culture & Society, 36(5), 113–131.

Postill, J. (2017). The diachronic ethnography of media: From social changing to actual social changes. Moment. Journal of Cultural Studies, 4(1), 19–43.

Biography:

Philipp Budka is a Lecturer in the Department of Social and Cultural Anthropology, University of Vienna, and the M.A. program Visual and Media Anthropology at the Free University Berlin. His research areas include digital anthropology and ethnography, the anthropology of media and technology as well as visual culture and communication. He is the co-editor of Ritualisierung – Mediatisierung – Performance (Vienna University Press, 2019) and Theorsising Media and Conflict (Berghahn Books, in press). His research has also been published in journals and books such as Journal des Anthropologues, Canadian Journal of Communication and Ethnic Media in the Digital Age (Routledge, 2019).

From the sidelines to the archived web: What are the most annoying football phrases in the UK?

Helena Byrne
(British Library)

Keywords: Football, Annoying Football Phrases, Shine, UK Web Archive, Web Archive Case Study

ABSTRACT

As the news and TV coverage of football has increased in recent years, there has been growing interest in the type of language and phrases used to describe the game. Online, there have been numerous news articles, blog posts and lists on public internet forums on what are the most annoying football clichés. However, all these lists focus on the men's game and finding a similar list on women's football online was very challenging. Only by posting a tweet with a survey to ask the public "What do you think are the most annoying phrases to describe women's football?" was I able to collate an appropriate sample to work through.

Consequently, the lack of any such list in a similar format highlights the issue of gender inequality online as this is a reflection of wider society. I filtered a sample of the phrases from men's and women's football to find the top five most annoying phrases. I then ran these phrases through the UK Web Archive Shine interface to determine their popularity on the archived web. The UK Web Archive Shine interface was first developed in 2015, as part of the Big UK Domain Data for the Arts and Humanities project. This presentation will assess how useful the Trends function on the Shine interface is to determine the popularity of a sample of selected football phrases from 1996 to 2013 on the UK web. The Shine interface searches across 3,520,628,647 distinct records from .uk domain, captured from January 1996 to the 6th April 2013.

This paper goes through the challenges of using the Shine interface to determine: what are the most annoying football phrases on the archived UK web. By using this example, it highlights how working with this resource differs from working with digitised publications and what strategies can be employed to gain meaningful answers to research questions. It is hoped that the findings

from this study will be of interest to the footballing world but more importantly, encourage further research in sports and linguistics using the UK Web Archive.

References:

Helena Byrne. (2018). What do you think are the most annoying phrases to describe women's football??https://footballcollective.org.uk/2018/05/18/what-do-you-think-are-the-most-annoying-phrases-to-describe-womens-football/ (Accessed August 26, 2018)

Andrew Jackson. (2016). Introducing SHINE 2.0 – A Historical Search Engine. Retrieved from: http://blogs.bl.uk/webarchive/2016/02/updating-our-historical-search-service.html (Accessed August 26, 2018)

Biography:

Helena Byrne is the Curator of Web Archives at the British Library. She was the Lead Curator on the IIPC CDG 2018 and 2016 Olympic and Paralympic collections. Helena completed her Master's in Library and Information Studies at UCD in 2015. Previously she worked as an English language teacher in Turkey, South Korea and Ireland.

Tracking and Analysing Media Events through Web Archives

Caio de Castro Mello Santos
(School of Advanced Study,
University of London)

Daniela Cotta de Azevedo Major
(School of Advanced Study,
University of London)

Keywords: Digital Humanities; Media Events; Web Archives; Discourse Analysis

ABSTRACT

Throughout the last two decades, media outlets have grown more reliant on online platforms to spread news and ideas. Web Archives are a valuable tool to analyse the recent past as well as the present social and political context.

However, the use of Web Archives to conduct research can be challenging due to the amount of data and its access limits. This project aims to develop mechanisms to extract, process and analyse data in order to provide scholars with a model to explore the impact of massive media events in the last couple decades. Two events have been taken as case studies: The London 2012 and Rio 2016 Olympics and the European Parliamentary Elections from 2004 to 2019. Regarding the Olympics, we aim to understand how online media have described the legacies of the London 2012 and Rio 2016 Olympics and how the choices made by the gatekeepers (news editors, journalists) influence the narrative about the consequences of both events. Whereas the study of the media coverage of the European elections can shed light on how political concepts such as nationalism and integration have an impact on the European public opinion and its attitudes towards European Institutions. Given the geographical and the temporal range of these projects, we will focus on different yet complementary Web Archives initiatives such as the Internet Archive, the UK Web Archive and Arquivo.pt.

This project is being developed as part of the Cleopatra Training Network under a PhD in Digital Humanities. Therefore, this research is combining traditional methods such as Discourse Analysis through a qualitative close reading with quantitative computational methods through distant reading. This approach aims to provide examples of how to apply this type of data to the interpretative methodologies of the Social Sciences.

Biographies:

Daniela Major: Early Stage Researcher at School of Advanced Study. Her doctoral project is on the Media coverage of the European Elections 2004-2019. She holds a master of letters in Intellectual History from the University of Saint Andrews and is a former research fellow at Arquivo.pt.

Caio Mello: Early Stage Researcher at the School of Advanced Study/University of London. Journalist with a master's in communication (UFPE – Brazil). Former research fellow at the Center for Advanced Internet Studies (CAIS - Germany).

Reanimating the CDLink platform: A challenge for the preservation of mid-1990s Web-based interactive media and net.art

Dr. Eamonn Bell
(Trinity College Dublin)

Keywords: compact disc, Web, preservation, music, interactive multimedia

ABSTRACT

The Voyager Company realised the creative and commercial potential of mixed-mode CD-ROMs as the platform par excellence for interactive multimedia. Starting in 1989, with the release of a HyperCard-based interactive listening guide for Beethoven's Symphony No. 9, Voyager tightly integrated rich multimedia, hyperlinked text, and high quality audiovisual recordings into over 50 software releases for Mac and PC well into the late 1990s. Consolidating their expertise in computer-controlled optical media with Laserdics, Voyager developed AudioStack: a set of extensions for the HyperCard environment that allowed fine-grained software control of high-fidelity audio stored on conventional optical media. AudioStack led to a cross-platform technology designed for use on the web called CDLink, comprising CD-ROM controller drivers, extensions for Macromedia Shockwave and the plain-text Voyager CDLink Control Language.

CDLink enabled and inspired commercial ventures and amateur productions alike, such as Sony Music's short lived ConnecteD experiment, the small but dedicated community of fan-sites that published time-synced lyric pages alongside hyperlinked commentaries for popular records, and even experimental sonic net.art in Mark Kolmar's Chaotic Entertainment (1996). As Volker Straebel (1997) has pointed out, Kolmar's work used CDLink files to probabilistically remix and

loop the contents of the user's own CD collection in code, evincing similar tactics of creation by contemporary experimental musicians and sound artists. Owing to the mostly obsolete hardware and software dependencies of the CDLink platform and the challenges posted by the fading born-digital traces of the mid-1990s Web, CDLink-dependent artifacts create difficulties for preservation and access. I summarise the above-mentioned developments that culminated in CDLink and describe the challenges of preserving Kolmar's artwork and making it available for future audiences, as well as those of the larger so-called "extended CD" ecosystem, which flourished during this decade.

Biography:

Eamonn Bell is a Research Fellow at the Department of Music, Trinity College Dublin. His current research focus is on the cultural history of the digital Audio CD format told from a viewpoint between musicology and media studies. In 2019, Eamonn was awarded a Government of Ireland Postdoctoral Fellowship in support of this two-year project, 'Opening the "Red Book"'. He holds a doctorate in music theory from Columbia University (2019), where he wrote a dissertation on the early history of computing in the analysis of musical scores. He also holds a bachelor's degree in music and mathematics from Trinity College Dublin (2013). His research engages the history of digital technology as it relates to musical production, consumption, and criticism in the twentieth century.

Curating culturally themed collections online: The 'Russia in the UK' Special Collection, UK Web Archive

Hannah Connell
(King's College London; British Library)

Keywords: Curatorship, diaspora, media, community, web archiving

ABSTRACT

The researcher-curated special collection, Russia in the UK, is part of the UK Web Archive, hosted by the British Library. This collection comprises a selection of websites created for and by the Russian-speaking population in the UK.

This paper will explore the challenges for creating and maintaining web archival collections. I will discuss difficulties in determining the parameters of this special collection. Alongside the impact

of the single-curatorial voice in shaping a collection, this paper will address the ways in which the legal and technical infrastructure underlying web archiving affects the shape of a collection. I will examine how the decision-making process behind curating and expanding this collection encourages reflection on the specific cultural context of Russian migration to the UK and complicates the notion of a culturally-themed diaspora collection.

The Russia in the UK special collection is public but still growing. This collection is valuable for researchers both as a resource for further research, and as a means of questioning research practices. The practice of creating and maintaining a special collection such as the Russia in the UK collection influences the shape of the collection and the online representation of the diasporic community it reflects. This paper will examine how the ongoing process of research and selection can be broadened to include new curators. I will discuss the ways in which a broader community can be involved in the curation process and the development of this special collection in the future.

Biography:

Hannah is undertaking an AHRC funded collaborative PhD studentship with the British Library and King’s College London exploring interwar migration from Russia through Russian-language émigré publishing. The selection of the content for the UKWA ‘Russia in the UK’ special collection forms part of this research, reflecting the ways in which diasporic communities continue to preserve and contribute to a shared identity though new forms of media today.

DELETE MY ACCOUNT: Ethical Approaches to Researching Youth Cultures in Historical Web Archives

Katie Mackinnon
(University of Toronto)

Keywords: web history, web archives, research ethics, youth cultures, 1990s web

ABSTRACT

Over the past 25 years the web has become an "unprecedentedly rich primary source…it is where we socialise, learn, campaign and shop. All human life, as it were, is vigorously there" (Winters, 2017). Web archives, as an increasingly important resource for writing social, cultural, political, economic, and legal histories, pose new challenges for historians who must learn how to "navigate this sea of digital material" (Milligan, 2012). Throughout these past few decades, young people have been a focus of digital cultures and participation (Turkle, 1995; Kearney, 2006; Scheidt, 2006; Ito et al., 2010; boyd, 2014; Vickery, 2017; Watkins et al., 2018). The early web communities of GeoCities that are available on the Internet Archive are a unique and incredibly fruitful resource for studying youth participation in the early web (Milligan, 2017) in a way that gives youth voices autonomy and agency. New challenges emerge when applying computational methodologies and tools to youth cultures in historical web archives at scale.

This paper considers the challenges in: 1) researching and writing about the phenomenon of young people divulging personal details about their lives without the possibility of informed consent; 2) accurately contextualizing web pages within wider online communities and; 3) engaging with socio-political climates that young people were experiencing and exploring the Web that focuses on the intersections of race, gender, sexuality, class, geography, and cultural and social pressures.

The EU's "Right to be Forgotten" (2014) and GDPR (2018) call into question the regularity with which young people become "data subjects" through their proximity to social networking sites, either through family, friends or themselves. Young people's data is subject to commodification, surveillance, and archiving without consent. Researchers engaging with historical web material have a responsibility to develop better practices of care. This paper further develops frameworks

to ethically research young people's historical web content in digital archives that accounts for the sensitive nature of web materials (Adair, 2018; Eichhorn, 2019), lack of consent protocols available to historical web researchers (Aoir IRE 3.0, 2019), and the ways in which computational methods and big data research attempts often fail to anonymize data (Brügger & Milligan, 2018). Web history research puts living human subjects at the forefront of historical research, which is something that historians are not particularly well-versed in. This paper surveys ethical approaches to internet and web archive research (Lomborg, 2018; Schäfer & Van Es, 2017; Whiteman, 2012; Weltevrede, 2016), identifies gaps in studying historical web youth cultures and suggests next steps.

Works Cited:

Adair, Cassius. (2019). "Delete Yr Account: Speculations on Trans Digital Lives and the Anti-Archival." Digital Research Ethics Collaboratory. http://www.drecollab.org/

Brugger, Niels and Ian Milligan. (2018). The SAGE Handbook of Web History. London: Sage.

Bruckman, Amy, Kurt Luther, and Casey Fiesler. 2015. "When Should We Use Real Names in Published Accounts of Internet Research?," in Eszter Hargittai and Christian Sandvig (eds) Digital research confidential: the secrets of studying behavior online. Cambridge, Mass: MIT Press.

DiMaggio, P., E. Hargittai, C. Celeste and S. Shafer. (2004). "Digital inequality: From unequal access to differentiated use." In Social Inequality, ed. K. Neckerman. Russel Sage Foundation.

Eichhorn, Kate (2019). The end of forgetting: growing up with social media. Cambridge, Mass: Harvard University Press.

franzke, a.s., Bechmann, A., Zimmer, M. & Ess, C.M. (2019) Internet Research: Ethical Guidelines 3.0, Association of Internet Researchers, www.aoir.org/ethics.

Ito et al. (2010). Hanging Out, Messing Around, and Geeking Out: Kids Living and Learning with New Media. MIT Press.

Jenkins, H., M. Ito, and d. boyd. (2016). Participatory Culture in a Networked Era: A Conversation on Youth, Learning, Commerce, and Politics. Polity.

Kearney, M. C. (2006). Girls Make Media. Routledge.

Kearney, M. C. (2007). "Productive spaces girls' bedrooms as sites of cultural production spaces." Journal of Children and Media, 1, 126-141.

Lincoln, S. (2013). "I've Stamped My Personality All Over It": The Meaning of Objects in Teenage Bedroom Space." Space and Culture, 17(3), 266–279.

Lomborg, Stine. (2018). “Ethical Considerations for Web Archives and Web History Research,” in SAGE Handbook of Web History, eds. Niels Brügger and Ian Milligan.

Milligan, Ian. (2017). “Pages by Kids, For Kids”: Unlocking Childhood and Youth History through Web Archived Big Data,” in The Web as History, eds. Niels Brügger and Ralph Schroeder, UCL Press.

Schäfer, Mirko Tobias, and Karin Van Es. (2017). The datafied society: studying culture through data. Amsterdam University Press.

Scheidt, L. A. (2006.) “Adolescent diary weblogs and the unseen audience,” in Digital Generations: Children, Young People, and New Media, ed. D. Buckingham and R. Willet. Erlbaum.

Skelton T. and Valentine G. (1998). Cool Places: Geographies of Youth Cultures. Routledge.
Turkle, Sherry. (1995). Life on the Screen: Identity in the Age of the Internet, Simon and Schuster.

van Dijck, José, Thomas Poell, and Martijn de Waal. (2018). The Platform Society; Public Values in a Connective World. New York: Oxford University Press.

Vickery, J. R. (2017). Worried about the wrong things: Youth, risk, and opportunity in the digital world. Cambridge, MA: MIT Press.

Watkins, S. C. et. al. (2018). The Digital Edge: How Black and Latino Youth Navigate Digital Inequality. NYU Press.

Weltevrede. Esther. (2016). Repurposing digital methods. The research affordances of platforms and engines. PhD Dissertation, University of Amsterdam

Whiteman, Natasha. (2012). “Ethical Stances in (Internet) Research,” in Undoing Ethics, by Natasha Whiteman, 1–23. Boston, MA: Springer US, 2012.

Winters, Jane. (2017) “Breaking in to the mainstream: demonstrating the value of internet (and web) histories,” Internet Histories, 1:1-2, 173-179.

Biography:

Katherine (Katie) Mackinnon is a Ph.D. candidate at the University of Toronto in the Faculty of Information. She researches web histories, including early uses of the internet by young people in the 1990s through a case study of the popular website, ‘GeoCities’. She is particularly interested in using web archives to conduct historical work, focusing on youth expressions of identity and community within their specific socio-political contexts.

Changing platforms of ritualized memory practices.
Assessing the value of family websites

Dr. Susan Aasman
(University of Groningen)

Keywords: web archives, vernacular culture, amateur media, web archaeology, technologies of memory

ABSTRACT

In this presentation I want to introduce research on current personal digital archival practices, as they have shifted from private spaces to more public platforms. I would especially like to discuss the value of concrete everyday practices of storing and sharing multimodal family records on late nineties/early 21st century family web sites. In addition, I will address the vulnerability of these archival practices, introducing a casus of a particular family web site hosted by the famous Dutch provider XS4all who will close its service permanently. Although the National Library of the Netherlands (KB) started to collect XS4all websites, when it comes to selecting and preserving online personal archives, there is still a need to raise awareness about these deeply meaningful memory practices. For one, these type of practices of memory staging do have a history that is much older that the history of the web suggests; they belong to a long durée history of technologies of memory production and distribution. At the same time, understanding these family oriented websites as designed in the nineties and early 200s gives us an excellent opportunity to understand the specificities of the shift from private to public, and from analogue to digital.

This research is part of larger agenda that addresses the urgent issue of long-term preservation of amateur media and how technological, political, social and cultural factors influence how we appraise and archive the often ephemeral nature of amateur media expressions. In particular, digital material poses multiple challenges, one of them the sustainability of many forms and formats of amateur media. The challenge is a shared task of public cultural heritage institutions, commercial, scholars and individuals alike. The archival strategies and the choices of what to keep and what to delete may resonate for decades to come. The presentation will argue that the complexities and contradictions that characterize present-day amateur media culture are mirrored by and reproduced in the complexities and contractions of archiving digital memories. There are no simple solutions and there are no simple guidelines, as amateur media archives – whether personal or collective or

whether they are analogue or digital - have been caught up in ethical, emotional, commercial, political contested areas and bear the burden of being technological, material, and personal.

Biography:

Dr. Susan Aasman is associate professor at the Centre for Media and Journalism Studies and Director of the Centre of Digital Humanities at the University of Groningen (NL). Her field of expertise is in media history, with a particular interest in amateur film and documentaries, digital cultures and digital archives, web history and digital history. She was a senior researcher in the research project 'Changing Platforms of Ritualised Memory Practices: The Cultural Dynamics of Home Movie Making'. Together with Annamaria Motrescu-Mayes, she is the co-author of Amateur Media and Participatory Culture: Film, Video and Digital Media (Routledge 2019). Recently she started working on web archival and web historical projects. She co-edited – together with Kees Teszelszky and Tjarda de Haan - a special issue on Web Archaeology for the journal TMG/Journal for Media History (https://www.tmgonline.nl/).

Platform and app histories: Assessing source availability in web archives and app repositories

Dr. Anne Helmond
(University of Amsterdam)

Fernando van der Vlist
(Utrecht University)

Keywords: platforms, apps, web historiography, web archiving, app archiving

ABSTRACT

In this presentation, we discuss the research opportunities for historical studies of apps and platforms by focusing on their distinctive characteristics and material traces. We demonstrate the value and explore the utility and breadth of web archives and software repositories for building corpora of archived platform and app sources. Platforms and apps notoriously resist archiving due to their ephemerality and continuous updates. As a result of rapid release cycles that enable developers to develop and deploy their code very quickly, large web platforms such as Facebook and YouTube change continuously, overwriting their material presence with each new deployment. Similarly, the pace of mobile app development and deployment is only growing, with each new software update overwriting the previous version. As a consequence, their histories are being overwritten with each update, rather than written and preserved. In this presentation, we consider how one might write the histories of these new digital objects, despite such challenges.

When thinking of how platforms and apps are archived today, we contend that we need to consider their specific materiality. With the term materiality, we refer to the material form of those digital objects themselves as well as the material circumstances of those objects that leave material traces behind, including developer resources and reference documentation, business tools and product pages, and help and support pages. We understand these contextual materials as important primary sources through which digital objects such as platforms and apps write their own histories with web archives and software repositories.

We present a method to assess the availability of these archived web materials for social media platforms and apps across the leading web archives and app repositories. Additionally, we conduct a comparative source set availability analysis to establish how, and how well, various source sets

are represented across web archives. Our preliminary results indicate that despite the challenges of social media and app archiving, many material traces of platforms and apps are in fact well preserved. The method is not just useful for building corpora of historical platform or app sources but also potentially valuable for determining significant omissions in web archives and for guiding future archiving practices. We showcase how researchers can use web archives and repositories to reconstruct platform and app histories, and narrate the drama of changes, updates, and versions.

Biographies:

Anne Helmond is an assistant professor of New Media and Digital Culture at the University of Amsterdam. Her research interests include software studies, platform studies, app studies, digital methods, and web history.

Fernando van der Vlist is a PhD candidate at Utrecht University and a research associate with the Collaborative Research Centre "Media of Cooperation" at the University of Siegen. His research interests include software studies, digital methods, social media and platform studies, app studies, and critical data studies.

Exploring archived source code: computational approaches to historical studies of web tracking

Dr. Janne Nielsen
(Aarhus University)

Keywords: archived source code: computational approaches; historical studies; web tracking

ABSTRACT

This paper presents different ways of examining archived source code to find traces of tracking technologies in web archives. Several studies have shown a prolific use of tracking technologies used to collect data about web users and their behavior on the web (e.g. Altaweel, Good & Hoofnagle, 2015; Roesner, Kohno & Wetherall, 2012; Ayenson, Wambach, Soltani, Good & Hoofnagle, 2011; see also the review of existing tracking methods in Bujlow, Carela-Espanol, Lee & Barlet-Ros, 2017). Tracking is used for a multitude of purposes from authorisation and personalisation over web analytics and optimisation to targeted advertising and social profiling. The extent of web tracking and the magnitude of data collected by powerful companies like

Facebook and Google have caused concerns about privacy and consent. To better understand the spread of tracking and the possible implications of the practices involved, it is important to study the development leading up to today. Most studies of web tracking study the current web but to study the historical development of tracking, we can turn to web archives.

The distinctive nature of archived web as "reborn digital" (Brügger, 2018) means that a study using archived web must always address the specific characteristics of this source and the associated methodological issues (Brügger, 2018; Masanès, 2006; Schneider & Foot, 2004) but a study of tracking technologies in the archived web poses additional, new methodological challenges. Tracking technologies are part of what could be called the environment of a website (cf. Helmond, 2017) but it is not part of what is usually considered the 'content', which the web archives aim to collect and preserve (Rogers, 2013). Tracking can also depend on technologies that are often difficult to archive (e.g. content based on JavaScript, Flash or similar). None the less, it is still possible to find traces of tracking technologies in web archives. One approach, inspired by the work of Helmond (2017), is to study the archived source code of websites. This paper presents a study of tracking technologies on the Danish web from 2006 to 2015 as it has been archived in the Danish national web archive Netarkivet. The study experiments with computational methods to map the development of different tracking technologies (e.g. http cookies and web beacons). The paper discusses the main methodological challenges of the study and shows how a profound knowledge of the specific archive and the changes in archiving strategies and settings over time is necessary for such a study.

References:

Altaweel, I., Good, N., & Hoofnagle, C. J. (2015). “Web Privacy Census”. Technology Science.

Ayenson, M. D., Wambach, D. J., Soltani, A., Good, N., & Hoofnagle, C. J. 2011. “Flash Cookies and Privacy II: Now with Html5 and Etag Respawning.” Ssrn.com. July 29.

Bujlow, T., Carela-Espanol, V., Lee, B.-R., & Barlet-Ros, P. 2017. “A Survey on Web Tracking: Mechanisms, Implications, and Defenses”. Proceedings of the IEEE, 105(8), 1476–1510.

Brügger, N. 2018. The Archived Web: Doing History in the Digital Age. Cambridge: MIT Press.

Helmond, A. 2017. Historical website ecology: Analyzing past states of the web using archived source code. In N. Brügger (Ed.), Web 25: histories from the first 25 years of the World Wide Web (pp. 139–155). New York: Peter Lang.

Masanès, J. 2006. Web Archiving: Issues and Methods. In J. Masanes (Ed.), Web Archiving (pp. 1–53). Springer.

Roesner, F., Kohno, T., & Wetherall, D. 2012. “Detecting and Defending Against Third-Party Tracking on the Web”. Presented at the 9th USENIX Symposium on Networked Systems Design.

Rogers, R. 2013. Digital methods. Cambridge: MIT Press.

Schneider, S. M. & Foot, K. A. 2004. “The Web as an Object of Study”. New Media & Society, 6(1), 114–122.

Biography:

Janne Nielsen is an Assistant Professor, PhD, in Media Studies and a board member of the Centre for Internet Studies at Aarhus University. She is part of DIGHUMLAB, where she is head of LARM.fm (a community and research infrastructure for the study of audio and visual materials) and part of NetLab (a community and research infrastructure for the study of internet materials). Her research interests include media history, cross media, web historiography, web archiving, web tracking, privacy and consent.

Cross-sector interdisciplinary collaboration to discover topics and trends in the UK Government Web Archive: a reflection on process

Mark Bell
(The National Archives, UK)

Tom Storrar
(The National Archives, UK)

David Beavan
(The Alan Turing Institute)

Dr. Eirini Goudarouli
(The National Archives, UK)

Dr. Barbara McGillivray
(The Alan Turing Institute)

Dr. Federico Nanni
(The Alan Turing Institute)

Pip Willcox
(The National Archives, UK)

Keywords: Discovery, Machine Learning, Collaboration, Machine Assisted Exploration, Scale

ABSTRACT

This paper proposes a discussion of a collaboration between The National Archives and The Alan Turing Institute to use artificial intelligence technologies to enable the navigation and comprehension of the UK Government Web Archive (UKGWA) at scale.

The National Archives are the official archive of UK government holding over 1000 years of history. Since 1996 The National Archives have been archiving UK government websites and social media output that are publicly accessible through the UKGWA. Users of the UKGWA can browse sites or use the very effective full text search service to find content in over 350 million documents (and counting). Search relies on keyword matching and is most effective when combined with domain knowledge, but most of our users don’t have this. There is currently no way to view the UKGWA as a whole or to group similar material together. Research into UKGWA users indicates they expect an “intuitive” search experience, allowing them to navigate this massive dataset, with search results surfacing relevant results. That type of search experience requires resource intensive data engineering and natural language processing methods that handle a high volume of queries, neither of which is currently available.

With The Alan Turing Institute, the national institute for data science and AI, we proposed a Data Study Group (DSG) to bring together experts from across and beyond academia to work on a data challenge for a week. Held in December 2019, the challenge focuses on discoverability of the UKGWA, applying advanced machine learning and natural language processing approaches to tasks such as creating a subject matter overview of the archive, machine assisted exploration, and identifying the emergence, growth, and decay of topics over time.

This talk will explain the challenges that we face when it comes to explore, understand, analyse and interpret the UKGWA; will focus on the collaboration between The National Archives and The Alan Turing Institute; and will present the work of selection and preparation of data prior to the challenge, as well as the process and outcomes of the challenge week itself – what went well, what didn't, what surprised us. We will also discuss next steps and how we will seek to implement the outcomes of this collaboration. This will include the challenges of turning a complex research prototype developed in a technical environment into something that can be practically integrated into the UKGWA interface to meet the needs of, and be understood by, our users.

We would welcome the thoughts of conference participants on this work to date, including on how it can be made useful to researchers, web archives, and their users.

Biographies:

Mark Bell is Senior Digital Researcher at The National Archives. He has worked as researcher on the AHRC funded project Traces Through Time on which he developed statistical methodologies for record linkage, and on the ESPRC funded ARCHANGEL which explored the use of Distributed Ledger Technology to provide trust in archived born-digital material. Mark's research interests cover a broad range of areas including Handwritten Text Recognition, Crowdsourcing, applications of Machine Learning to archival processes, and of course the challenges of working with large scale web archives.

Tom Storrar is the Head of Web Archiving at The National Archives. He has led the Web Archive team for over 10 years, transforming the way that web archiving is performed. Tom has spoken at a number of international conferences about the challenges of web archiving. As well as the day to day challenges of maintaining the archive, he has also defined collection policies around web pages, social media accounts, and even code repositories, as well as managing the migration to cloud based archiving.

David Beavan is Senior Research Software Engineer – Digital Humanities in the Research Engineering Group (also known as Hut 23) in The Alan Turing Institute. He has been working in the Digital Humanities (DH) for over 15 years, working collaboratively, applying cutting edge computational methods to explore new humanities challenges. He is Co-Investigator for two Arts and Humanities Research Council (AHRC) funded projects: Living with Machines and Chronotopic Cartographies, is Co-organiser of the Humanities and Data Science Turing Interest

Group and is Research Engineering's challenge lead for Data Science for Science (and also humanities) and Urban Analytics.

Eirini Goudarouli is a member of the Research Team at The National Archives. Her current research interests include digital humanities and digital archives. She is particularly interested in bringing together methods and theories from a range of disciplines that could essentially contribute to the rethinking of digital, archival and collection-based research. Eirini is the Co-Investigator of the International Research Collaboration Network in Computational Archival Science (IRCN-CAS), funded by the Arts and Humanities Research Council.

Barbara McGillivray is Turing Research Fellow at The Alan Turing Institute and the University of Cambridge. She has always been passionate about how Sciences and Humanities can meet. She completed a PhD in Computational Linguistics from the University of Pisa in 2010 after a degree in Mathematics and one in Classics from the University of Florence (Italy). Before joining the Turing, she was language technologist in the Dictionary division of Oxford University Press and data scientist in the Open Research Group of Springer Nature.

Federico Nanni is a Research Data Scientist at The Alan Turing Institute, working as part of the Research Engineering Group, and a visiting fellow at the School of Advanced Study, University of London. He completed a PhD in History of Technology and Digital Humanities at the University of Bologna focusing on the use of web archives in historical research and has been a post-doc in Computational Social Science at the Data and Web Science Group of the University of Mannheim. He also spent time as a visiting researcher at the Foundation Bruno Kessler and the University of New Hampshire, working on Natural Language Processing and Information Retrieval.

Pip Willcox is Head of Research at The National Archives. She has a background in digital editing and book history, focussing first on encoding medieval manuscripts and later on early modern printed books. More recently she has worked on projects linking collections and semantic web technologies, and social machines. She has developed a framework for an experimental humanities, using digital simulation to close-read and explicate interpretation of the archive. Her focus for the past several years has been on multidisciplinary engagement with collections, enabling digital research and innovation.

A tale of two web archives: Challenges of engaging web archival infrastructures for research

Jessica Ogden
(University of Southampton)

Emily Maemura
(University of Toronto)

Keywords: national web archives, researcher engagement, infrastructure studies

ABSTRACT

Web archives (WAs) are a key source for historical web research, and recent anthologies provide examples of their use by scholars from a range of disciplines (Brügger, 2017; Brügger 2018;

Brügger & Schroeder, 2017). Much of this work has drawn on large-scale collections, with a particular focus on the use of national web domain collections (Brügger & Laursen, 2019; Hockx-Yu, 2016). This previous work demonstrates how WAs afford new scholarship opportunities, yet little work has addressed how researcher engagement is impacted by the complexity of WA collection and curation. Further research has begun to address the impact of specific organizational settings where the technical constraints interact with policy frameworks and the limitations of resources and labour (Dougherty & Meyer, 2014; Hockx-Yu, 2014; Maemura et al. 2018; Ogden et al., 2017). Here, we extend this work to consider how these factors influence subsequent engagement, to investigate the very real barriers researchers face when using WAs as a source for research.

This paper explores the challenges of researcher engagement from the vantage point of two national WAs: the UK Web Archive at the British Library, and Netarkivet at the Royal Danish Library. We compare and contrast our experiences of undertaking WA research at these institutions. Our personal interactions with the collections are supplemented by observations of practice and interviews with staff, in an effort to investigate the circumstances that shape the ways that researchers use WAs. We compare these two national WAs along several dimensions, including: the legal mandates for collection; the ontological decisions that drive practices; the affordances of tools and technical standards; everyday infrastructural maintenance and labour; and the ways in which all of the above constructs the interfaces through which WAs are researched.

Our approach explores the materiality of WAs data across these two sites to acknowledge the generative capabilities of web archiving and reinforce an understanding that these data are not given or 'natural' (Gitelman, 2013). We highlight how the sociotechnical infrastructure of web archiving shapes researcher access, the types of questions asked, and the methods used. Here, access is conceived of not only in terms of 'open' versus 'closed' data, but rather as a spectrum of possibilities that orientates researchers to particular ways of working with data, whilst often decontextualising them from the circumstances of their creation. We question which kinds of digital research are afforded by national WAs, particularly when the scoping of collection boundaries on ccTLDs (top level domains) creates 'artificial geographic boundaries' (Winters, in press). Through this process we recognise and centre the assumptions about collection and use that are embedded in these research infrastructures, to facilitate a discussion of how they both enable and foreclose on particular forms of engagement with the Web's past.

Bibliography:

Brügger, N. (2018). The Archived Web: Doing History in the Digital Age. Cambridge, MA: MIT Press.

Brügger, N. (Ed.). (2017). Web 25: histories from the first 25 years of the World Wide Web. New York: Peter Lang.

Brügger, N., & Laursen, D. (Eds.). (2019). The historical web and digital humanities: The case of national web domains. Abingdon: Routledge.

Brügger, N., & Schroeder, R. (Eds.). (2017). The Web as History: Using Web Archives to Understand the Past and the Present. London: UCL Press. Retrieved from http://oapen.org/download?type=document&docid=625768

Dougherty, M., & Meyer, E. T. (2014). Community, tools, and practices in web archiving: The state-of-the-art in relation to social science and humanities research needs. Journal of the Association for Information Science and Technology, 65(11), 2195–2209. https://doi.org/10.1002/asi.23099

Gitelman, L. (Ed.). (2013). "Raw data" is an oxymoron. Cambridge, Massachusetts; London, England: The MIT Press.

Hockx-Yu, H. (2014). Access and Scholarly Use of Web Archives. Alexandria: The Journal of National and International Library and Information Issues, 25(1), 113–127. https://doi.org/10.7227/ALX.0023

Hockx-Yu, H. (2016). Web Archiving at National Libraries Findings of Stakeholders' Consultation by the Internet Archive. Internet Archive. Retrieved from https://archive.org/details/InternetArchiveStakeholdersConsultationFindingsPublic

Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If These Crawls Could Talk: Studying and Documenting Web Archives Provenance. Journal of the Association for Information Science and Technology, 69(10), 1223–1233. https://doi.org/10.1002/asi.24048

Ogden, J., Halford, S., & Carr, L. (2017). Observing Web Archives: The Case for an Ethnographic Study of Web Archiving. In Proceedings of the 2017 ACM on Web Science Conference (pp. 299–308). Troy, New York, USA: ACM Press. https://doi.org/10.1145/3091478.3091506

Winters, J. (in press, 2019). Giving with one hand, taking with the other: E-legal deposit, web archives and researcher access. In P. Gooding & M. Terras (Eds.), Electronic Legal Deposit: Shaping the library collections of the future. London: Facet Publishing.

Biography:

Jessica Ogden, University of Southampton; jessica.ogden@soton.ac.uk
Jessica Ogden is a PhD Candidate based in Sociology and the Web Science Centre for Doctoral Training at the University of Southampton. Jessica's research focuses on the politics of data, web archiving and digital data scholarship.

Emily Maemura, University of Toronto; e.maemura@mail.utoronto.ca

Emily Maemura is a PhD candidate at the University of Toronto's Faculty of Information (iSchool). Her research focus is on web archiving, including approaches and methods for working with web archives data and research collections, and capturing diverse perspectives of the internet as an object and/or site of study.

What's missing from WARC?

(Consultative Committee for Space Data Systems (CCSDS), Data Archive Interoperability (DAI) Working Group)

Mr. Michael W. Kearney III
Sponsored by Google, Huntsville,
Alabama, USA.

Mr. John Garrett
Garrett Software, Columbia,
Maryland USA

Mr. David Giaretta
PTAB Ltd, Dorset, UK.

Mr. Steve Hughes
Jet Propulsion Laboratory,
California Institute of Technology,
Pasadena, California, USA

Keywords: OAIS; WARC; CCSDS; HTML; MIME

ABSTRACT

This presentation will explain why the WARC format, by itself, is not adequate to preserve websites. As a brief justification of the claim, it is well known that a WARC file essentially captures the information sent from a website. However, by itself, this is not enough for long term preservation for the following reasons. Right now, there are suitable, readily available, Web browsers which can deal with current websites, supporting HTML standards, but often making guesses about how to display important but badly constructed web pages. In future these will not necessarily be available. More importantly websites not only display pages but also download files. The WARC file may show a MIME type of "application/vnd.ms-excel", which is a hint to the web browser to use MS Excel to show a spreadsheet. But what do the columns mean? For example, a column labelled "speed" may seem easy to understand but a speed of 10 mm/hour is very different from a speed of 10 miles/second. The WARC file does not provide enough information. The presentation will also explain what can be done to supplement WARC to fix these problems utilizing the long-term preservation practices of OAIS.

Biographies:

Mike Kearney is an engineering graduate of the University of Kentucky. He worked for NASA for 34 years in Systems Engineering and Technology positions; including chairmanship of the international standards body CCSDS, until retiring from NASA in 2015. He is now working with the non-profit Space Infrastructure Foundation and volunteers time for Google who sponsors attendance at Digital Preservation forums.

David Giaretta has led developments of standards in digital preservation (ISO 14721), in particular audit and certification of repositories (ISO 16363 and 16919) and developed practical and coherent solutions and services that will help repositories seeking ISO certification while adding value to their holdings.

Steve Hughes is a Principal Computer Scientist at the National Aeronautics and Space Administration (NASA) Jet Propulsion Laboratory. Three decades of experience with NASA's official archive for Solar System Exploration science data, the Planetary Data System. Chief architect for the archive's information architecture which is based on principles from the Open Archive Information System (OAIS) Reference Model (ISO-14721) and the ISO/IEC 11179 Metadata Registry (MDR) standard. Member of the Primary Trusted Digital Repository Accreditation Board (PTAB). Associate member of Jet Propulsion Laboratory's Center for Data Science and Technology, a virtual center for research, development and operations of data intensive and data-driven science systems. He was awarded the NASA Exceptional Public Service Medal for exceptional service to NASA science missions and data archives, architecting and implementing data intensive systems, information models, and ontologies for three decades.

John Garrett is an engineering graduate from Missouri University for Science and Technology and a Computer Science graduate of Johns Hopkins University. He spent 25 years working as a contractor for NASA's National Space Science Data Archive, including many years representing their needs and interests while developing digital preservation standards. He was instrumental in developing the OAIS Reference Model and continues to help lead the CCSDS DAI efforts developing OAIS related standards and standards for certifying Trustworthy Digital Repositories.

Background on the CCSDS DAI Working Group:

CCSDS is the Consultative Committee for Space Data Systems. It started in 1982 developing data and communications interoperability standards for data systems (flight and ground) that are used in space missions. While CCSDS is organized by space agencies, it is inclusive of other non-space organizations, industry and academia. CCSDS consists of about 22 working groups, one of which is the Data Archive Interoperability WG. The DAI WG is focused on long-term digital preservation archives. With extensive support from non-space-industry organizations (national archives and libraries from various countries, academia, other industry domains, etc.), the DAI WG developed the Reference Model for OAIS. Due to its wide applicability, OAIS became broadly adopted outside of the space industry. CCSDS and DAI standards are procedurally adopted by and published by ISO (as CCSDS functions as ISO TC20/SC13). The DAI has published many standards that support OAIS and that are applicable to some space-related archives as well as other "generic" preservation archives globally.

Web Archives as Scholarly Dataset to Study the Web

Dr. Helge Holzmann
(Internet Archive)

Jefferson Bailey
(Internet Archive)

Keywords: data processing, extraction, derivation, access, research

ABSTRACT

The Internet Archive (IA) has been archiving broad portions of the global web for over 20 years. This historical dataset, currently totaling over 20 petabytes of data, offers unparalleled insight into how the web has evolved over time. Part of this collecting effort has included the ability to support large-scale computational research efforts analyzing this collection. This presentation will update efforts within IA to support computational use of its web archive, approaching this topic through description of both program and technical development efforts.

Web archives give us the opportunity to process the web as if it was a dataset, which can be searched, analyzed and studied, temporally as well as retrospectively. However, web data features some very specific traits that raise new challenges to deal with when providing services based on the contained information. Our Web Data Engineering efforts are tackling these challenges in order to discover, identify, extract and transform archival web data into meaningful information for our users and partners, by hiding all the complexity and abstract away technical details.

Engineering has traditionally been the systematic application and combination of existing methods to build a desired system or thing. Data Engineering is different from this in that engineering here does not refer to creating something but transform the data in a way that it is more useful for what should be achieved. As part of this, new tools and processes are developed to accomplish this transformation more effectively as well as efficiently in terms of resources and time.

The talk will outline different computational research services for historical web archive data, along with technical challenges, novel developments and opportunities as well as considerations to make when working with this unique dataset, including:

- Researcher support scenarios
- Data limitations, affordances, and complexities
- Extraction, derivation, and access methods
- Infrastructure requirements
- Relevant tools and technologies
- Collection development and augmentation

In covering these topics through the lens of specific collaborations between IA and computational researchers performing large-scale analysis of web archives, this presentation will illuminate issues and approaches that can inform both the implementation of similar programs at other web archiving institutions and also help researchers interested in data mining web collections better understand the possibilities of studying web archives and the types of services they can expect to encounter when pursuing this work.

This overview is meant to showcase the latest achievements and upcoming data services from the Internet Archive's web archiving and data services group. Details about the way we and our systems work will be presented together with APIs and programming libraries that are ready to use as well as new features that are to be expected soon.

Biographies:

Helge Holzmann is Web Data Engineer at Internet Archive. Helge started working for the Archive in August 2018. Before, he earned his Master of Computer Science and worked as a researcher in Germany, striving for his PhD on efficient access methods for web archives, which resulted in publications at different conferences and journals, including TPDL, JCDL, BigData, SIGIR, WWW as well as the International Journal on Digital Libraries. He is passionate about big data, especially if there's a temporal aspect to it, and is glad to contribute to a non-profit organization that holds one of the biggest collections of free data in the world. In addition to creating innovative services by deriving new value from this unique dataset, Helge is happy to support libraries and institutions interested in accessing the data as a consultant located in Europe.

Jefferson Bailey is Director of Web Archiving & Data Services at Internet Archive. Jefferson joined Internet Archive in Summer 2014 and manages Internet Archive's web archiving services including Archive-It, used by over 650 institutions to preserve the web, as well as domain-scale and contract harvesting and indexing services. He works closely with partner institutions on collaborative technology development, computational research support, and data services. He is PI on multiple grants focused on systems interoperability, data-driven research use of web archives, and digital preservation initiatives. He was Chair of the Steering Committee of the International Internet Preservation Consortium (IIPC) until 2019.

Session 17: An Irish Tale / Scéal Éireannach

Born-digital displaced records: The disappearance of the GAA websites

Helena La Pina
(Maynooth University)

Keywords: Irish culture; GAA; archived websites; born-digital displaced records

ABSTRACT

This year, the author completed an MA in Historical Archives in Maynooth University, and produced a thesis titled: 'Displaced archives, and the core components in the debates surrounding repatriation'. The thesis utilises secondary literature in archival science, information/records management, and interdisciplinary scholarship to investigate the dilemmas associated with displaced archives. During the thesis research process, the author discovered that there was a limited amount of scholarship dealing with the displacement of electronic records, and a scarcity of scholarship regarding the displacement of born-digital records. This presentation aims to open a discussion on how archived websites, might also be understood as displaced born-digital records. In doing so, the author discusses a research study, which explores the presence of the Gaelic Athletic Association (GAA) web heritage in the Internet Archive's Wayback Machine.

Danielson (cited in Winn, 2015) offers an interpretation of displaced archives as 'archival materials that have been lost, seized, requisitioned, confiscated, purchased under duress, or otherwise gone astray'. Inkster (1983) proffers that a displaced or misplaced document comes under three definitions: the document is missing, the document is estray (which is the legal definition of a document not in possession of its owner), or the document is fugitive. The Society of Archivists (SAA) define fugitive as connoting 'materials that are not held by the designated archives or library charged with their preservation.' Displaced archives are also referred to as misplaced archives, expatriated archives, seized archives, archives in exile, and migrated archives (Inkster, 1983; Garaba, 2011; Winn, 2015). However, as Garaba argues, whatever term is used to describe displaced records and for whatever reason, the fundamental fact remains, they are not where they should be.

In this presentation, the author provides an analysis of the official GAA website, archived in the Wayback Machine within a certain timeframe. It also covers, on the periphery, other 'unofficial'

GAA archived websites. While chronicling the important role the GAA has played in Irish society, the author observes what dates were used for capturing and why the randomness of captures is not calibrated with end-of-season competitions like the All-Ireland final. The author discusses how the disappearance of GAA websites from the live web, fit the description of a missing cultural record. The author also highlights how the capture of GAA websites in the Wayback Machine, offers an interpretation of born-digital displaced record, in so far as the record is not where it should be.

References:

Garaba, Francis (2010) An investigation into the management of the records and archives of former liberation movements in east and southern Africa held by national and private archival institutions (PhD Dissertation, University of KwaZulu-Natal, South Africa, 2010) (https://researchspace.ukzn.ac.za/xmlui/handle/10413/1495)

Inkster, Carole M. (1983) Geographically misplaced archives and manuscripts: problems and arguments associated with their restitution, Archives and Manuscripts, 11(2), pp 113-124 (https://publications.archivists.org.au/index.php/asa/article/view/7559)

Winn, Samantha R. (2015) Ethics of access in displaced archives, Provenance, Journal of the Society of Georgia Archivists, 33(1), pp 6-13 (http://digitalcommons.kennesaw.edu/provenance/vol33/iss1/5)

Society of American Archivists, Dictionary of archival terminology, (https://dictionary.archivists.org/entry/fugitive.html).

Biography:

Helena La Pina recently completed an MA in Historical Archives at Maynooth University. Titled, ‘Displaced archives, and the core components in the debates surrounding repatriation’, her thesis investigates the dilemmas associated with displaced archives within the context of archival practices, and the justifications, rationales, and challenges for repatriation.

Recording Ireland's technology heritage: Lessons learned

John Sterne
(TechArchives project, Ireland)

Keywords: IT Histories; technology heritage

ABSTRACT

At its public launch in June 2016 the TechArchives project reached out to people with experience of past generations of information technology in Ireland and asked them to record personal testimonies. This work is continuing. As the project evolved, however, it became more concerned about the limited quantity and quality of historic material. It is therefore developing processes and methods to locate, catalogue and preserve digital evidence of significant actions and events.

Biography:

John Sterne is the founder of the TechArchives project. In the past he worked as a researcher, author, reporter and editor.