Document Distribution Note
Release to Members: Immediate
Release to Public: 08/06/2015

DigitalPreservationCoalition

Note: These are rough notes taken live during the event

Web Archives as Scholarly Sources

RESAW, University of Aarhus, Denmark, 8-10 June 2015

Organised by RESAW, Aarhus University, the State and University Library (Denmark), the Royal Library (Denmark), l'Institut des sciences de la communication du CNRS, Université de Lille 3, the Institute of Historical Research (University of London), the University of Amsterdam, the British Library, and Leibniz University Hannover/ALEXANDRIA.

See the conversation on Twitter through #resaw_eu

DAY 1

Keynote 1: Meghan Dougherty, Loyola University of Chicago

'Finding a Material Record of Info Culture'

- Methodology for analysing contemporary culture
- Archaeological methods could be applied to the history of the web
- Methods and policies for studying the web
- To study social change online is to study the historical web
- New communication technologies – how they influence society and culture – how we interact with them and are moulded by their limitations
- How we 'come to write those stories as history'
- Should focus on the *processes* we use to collect, use, and analyse those web artefacts
- How communication technology shapes our view of the world important to capture
- 'life lived with media' rather than discrete content
- How we adapt to new communication media important to capture
- Everyday life lived with media
- Reliance on new media ecological system – e.g. bots to prove 'you are human' (reddit)
- Examples of scenarios where the interactions provide more meaning than the content alone; e.g. When our digital tools are the extension of *others* into our space through the web via tech support
- Patterns in archiving are moving away from rich curation that looks at nuances of web experience, but rather emphasis on web as tool – digital as tool for research – extracting data rather than preserving entire experience of living the web
- Tools an active component of cultural networks – use us as much as we use them
- Effects on archives – shifts us towards access content and rather than dynamic info space
- Flow across internet determined by scripts and other structures
- Web not a discreet number of artefacts linked for preservation later

Note: These are rough notes taken live during the event

- We're excluding the characteristics that will be important for future researchers telling story of beginning of the web
- Current web archiving focuses on discrete pieces of content (pages, pdfs, etc) but does not lend itself to building tools that examine larger ecology – interactions, expectations, and how we move through web
- Lived experience of structure of web ecological system better approached through 'systems theory' – organism of parts
- Not an organisation of software but of processes and interactions
- Invisible dynamics – imperative for gaining meaning from archived web
- **Should explore more inclusive approach – things are not isolated, things are distributed, relationships are often invisible, bound in entangled relationships**
- Archive process itself introduces errors
- Traditional research and preservation methods don't account for much of what we experience – only work in stabilised social structures
- Web archives are: 1- incomplete, 2- unreliable, 3- difficult to search, 4- difficult to analyse
- In current web archives, I can tell you what I see, but not what it represents
- The Onion: Internet archaeologist – remains of Friendster (see: http://www.theonion.com/video/internet-archaeologists-find-ruins-of-friendster-c-14389)
- In traditional archaeology, they assume gaps and missing pieces, in web archiving, we don't know what we don't know – more difficult to create expectation for what we can and can't understand
- We don't know what methods produce 'good results'
- **Web archiving can be much more valuable than just capturing what the internet looked like yesterday.**
- If you are an archivist, reach out to researchers who have 'weird' questions and figure out how to help them answer their questions
- Research will need more collaborate effort because of different specialist knowledge and skills required
- Treating web archives like big data blinds us to the larger digital media ecology

3 Papers in Large Auditorium

1- Globalising web archiving efforts
An academic library perspective
Karen Farrell and Frank-Wilson

Web archiving in US Universities

- Responsible for their own content
- Some focus on their own and close university content
- Some exceptions – Columbia, Standford
- Coordinate with subject specialists
- Web archives complement greater library collections

Challenges:

Note: These are rough notes taken live during the event

- Without a web archiving unit?
- How to collect internationally?

1st effort: endangered languages web archives (subject-driven)

- 2 knowledge specialists
- Technical assistant (student)
- Archive-It

*Collection Development process much like the traditional processes: selection, preservation, access

Area studies librarianship

- Area expertise
- Language skills
- Collaboration
- Unique workflows

Web Archiving

- An integral element of collection development for area and international studies collections- workflows best practices etc integrated into existing collection policies

Ethical considerations

- Need to archive ephemeral websites for scholarly use (endangered archive project requests permission 3 times then archive without permissions; falls under 'fair use')
- Ethical – cultural imperialism, neo-colonialism (e.g. migrated archives from one region to another)
- Web archiving constinues this power blanace between 'the rich north' and the poor south : who decides what to archive
- 'who shapes the transmission of a country's history?'
- In absence of cyber laws, existing ethical conduct guidelines for conducting international research (e.g. African studies association ethical conduct guidelines)

Helen Hockx-Yu comments:

- Uneven distribution of content on the web
- Example for international collaboration – death of Nelson Mandela

2- The Unknown Aspects of Web Archives
Helen Hockx-yu, Head of Web Archiving at the British Library

- Resources rarely identical to the 'original' or 'live' websites
- Records of point-in-time HTTP transactions bn the web servers hosting content and crawlers requesting them

- Crawl date / time used to recognise web archives
- Misleading and easily mistaken as dates of publications

Note: These are rough notes taken live during the event

- Not explained to users

Euro Pent Office 2007 case
- Archived website from the internet Archive submitted as evidence
- Did not meet the standard of proof
- Difficulty in establishing date of availability
- '…whether and how it has been modified since the date it originally appeared on a web site'

When this lack of explanation for the architecture of web archives matters:
Error: 'Resource not in archive'

Common error mssg appears for different reasons

1. Intended boundary:
- No permissions for linked content
- Not allowed by robots.txt
- Edge of an archive (depth where crawler stops)
- Data volume limitation

2. Technical limitations, e.g. dynamic content the crawler could not collect

In UK Web archive: avoid dead end in navigation
-sear
-link to live web
-find archived copies elsewhere

Temporal inconsistency
- Single pages with same date point but actually only existed, say, 5-10 years apart

Does it matter?
- Motivation: avoid pages with 'holes' or 'gaps' – idiosyncratic?
- Some degree of temporal drift doesn't matter? – sometimes intellectual content doesn't actually change, only appearance (CSS files)
- Allow scholars to answer underlying conceptual questions and develop methodology

Maximum transparency remains the best remedy

Solutions:

Momento – reconstructing the web (shows how page is rendered over time)

What can we do?
- Framework for assessing temporal coherence
- The momento approach – timeline

Henriette Roued-Cunliffe comments:

Note: These are rough notes taken live during the event

- The problems came to light at crucial once researchers began using them as scholarly source
- How have you communicated these problems with archived web? Offer some guidance / best practice for researchers based on these problems

Question: in web archives maybe we should stop focusing on 'the original' – no definition of what his means in web archives

3- Archiving Online Do-It-Yourself Culture
Henriette Roued-Cunliffe
@henrietteroued
www.roued.com
research@roued.com
hdm329@hum.ku.dk

- More questions than answers at this stage
- How can we archive information shared by DIY culture and make it widely accessible
- By 'access', she means machine-readable data that can be obtained through web-based service
- Sharing is key –DIY barriers lowered due to internet
- Positive and negative views on new accessibility through internet
- Geographically dispersed members in DIY refashioning community
- Multimedia sources – videos, photos, text, etc comprise these communities
- More focus on dissemination of knowledge of 'how to' but recent new focus on maintaining the longevity of this information
- Sustainable platforms – trying to archive the information for the long-term – through blog posts – attempted organisation by tags – but users used them inconsistently – archivist attempt to read and tag each individual post

DAY 2

Keynote in Large Auditorium
Ditte Laursen, State and University Library
Per Møldrup-Dalum, State and University Library

Why is it important to know 'the story' of the web archive?

- to evaluate quality
- to evaluate reliability
- Etc.

Why is it difficult to tell the story of an archive?

- Moving target – web technology and content evolves really fast
- Tool required to look at archives also evolve really fast
- No benchmark for data so difficult to tell if analysis is accurate

Netarkivet – methods for evaluating web archive

Note: These are rough notes taken live during the event

- Informal interviews with staff
- Review of publications and newsletters
- Internal documentation

Data mining - challenges

- Size: 592TB – need something more manageable
- Reduced by looking at just metadata; down to 5 billion URLs
- No standard method
    - UNIX hackery
    - Java written for talk at hand
    - Statistical analysis and charting in R
    - others

2 Major 'Stories' in Netarkivet

1) Legal

- 2005 new legal deposit law to crawl Danish domain (.dk)
- 3 annual snapshots or triggered by major events
- 1997-2005: permissions-based crawls
- Growth in .dk domains
- 2006 'dynamic' materials appear
- 1998 harvesting starts
- Processing personal data – security measures to protect personal data, but data still made available to researchers
- Controlled data-mining: easier for archivists / curators to screen data for researchers

2) Technical Stories

HTTP Response codes
200: ☺
404: ☹
301, 302: :-/

- Rise in codes 301 (message to indicate webpage removed permanently) and 302 (webpage has been moved to X location)
- Responses since 2006 – is web becoming more dynamic and we are just not effective at harvesting it?
- But…
- 404 accounts for only 1.5% of returned codes (285 million pages) – is that significant enough to be a problem?

Solutions:

- Keep improving crawlers – move to Heritrix 3
- Use other crawlers: e.g. Umbra and CrawlJax which may be more equipped to keep up with how web evolves

Note: These are rough notes taken live during the event

Access:

- on premises print (1999 - 2005) using the 'Monk Machine'

More 'stories' of archive…

Media types:
2007: rise of video
2008: video exceeds capacity to harvest, also streaming begins to grow
2012: new tool developed for collecting YouTube videos

Limitation to .dk legal deposit

- Danish domains other than .dk
- 91 other Danish domains by 2012

Future curatorial practices:

- Customised crawler tools

Implications for research:

- Get to know your archive
- Re-frame your research question accordingly
- Be aware of your tools
- Double check results; try to validate results based on alternative sources

For web archives:

- Outreach! Proactively let the research community know about your resources
- Promote the collections
- Be aware of types of research questions being posed to web archives
- Evaluate pros/cons of tools used
- Quality control

PANEL: Systems, syntax, and snippets: accounting for software in web history

1- 'Shaping the social web: recovering the contributions of bulletin board system operators'
Kevin Driscoll, Microsoft Research

Why are bulletin board system operators important to the history of the web?

Survey – why is the internet important?

- Top responses:
- Job searches
- Finding romance

- But…

- Very few interviewed knew very little about how the internet works or where it comes from

Note: These are rough notes taken live during the event

- Misunderstanding of 'narrative' of history of web – some knowledge of use for military, some misunderstanding about the role of Steve Jobs at the time of his death

What other narratives for the web can be written, narratives about how web is actually 'lived'?

Social History of Web

- Users of bulletin boards were groups excluded from more traditional forms of communication (e.g. gay community, niche hobbies, etc.)
- Bulletin boards a whole new venue of communication, unique because users have no knowledge of who their audience really is

Bulletin Board community today

- Groups of enthusiasts
- e.g. Bo Zimmerman – private collector of BBS software
- decentralised network for exchange of BBS software and knowledge

How BBSs are used

- Sources of Friction: users who thought internet was boring, just 'highways' of information with no where to 'stop off'
- Is BBS on-ramp or off-ramp?

Stigma of commercialisation

- Academics wary of commercialisation, who did not participate in BBS community, were informing the press and government about them

BBS

- First generation of internet service providers

2- 'Perl and the web that was'
Michael Stevenson, University of Groningen
m.p.stevenson@rug.nl

Transitional period in web history, roughly 1997-2001
e.g. Slashdot (based on Perl)
rise of dynamic web ('web 1.5')

Background of Perl

- Developed in 1980s for system administration
- Good for text manipulation and for large documents
- e.g Usenet
- modules (CPAN) – become incorporated into web archive

Metaphors for relationship between Perl and the web

- 'Swiss army chainsaw'

Note: These are rough notes taken live during the event

- 'glue' keeping web together

Perl & Culture

- Open source focus

Media Historical Approach

- Perl as toolbox for building and imagining the web
- Perl helps people conceptualise the web
- CGI incorporated, etc. – Perl is a broken web for archives
- CGI – information retrieval vs. dynamic publishing created by CGI
- Now Perl also a data source
- Lowers distinction between amateur and professional

3- 'Digging into the software interfaces of the social web: APIs as tools to reconstruct missing social media content in archived websites'
Anne Helmond
University of Amsterdam
a.helmond@uva.nl
@silvertje

- Partially enabled by javascript which allows us to embed resources such as social plugins
- On a website and to upload external content and functionality
- And these social plug-ins, such as Facebook comments, known as 'data pour' – snippet of code on the website that creates a container for sending and receiving content from external databases
- Ex. Crawled page of Huffington Post article with Facebook comments plugin; [seen: code provided to web masters but comments not there]
- Technically functions as API call
- APIs considered the 'modern glue' of the internet – allow different social media networks to interact, as well as for social media platform to interact with external websites and with apps
- Allow platforms to extend their own features outside their own platform boundaries into external websites, e.g. the 'like' button

APIs glue the real-time web with the archived web

- Archived web pages do not have same functionality as live web – more problematic with rise of javascript and the dynamic web

PANEL: Between medium and archive: researching YouTube as a popular Archive

1- 'Structuring Mediated Memories: YouTube's sociotechnical practices and the Syrian War
Rik Smit
University of Groningen
@riksmit86

Document Distribution Note
Release to Members: Immediate
Release to Public: 08/06/2015

Note: These are rough notes taken live during the event

'YouTube IS an archive'

It is…
- 'Citizen witnessing' (Allan)
- 'Hybrid

Sociotechnical curating practices:
- Tagging
- Filtering
- Describing

Political because the archive is arranged -> YouTube anticipates search behaviour
- This guides interpretation
- Hence, structure the archive and partly determine whose voice is heard

Curating and Readying the algorithm
- Classifactory imagination (Beer)
- Visibility and invisibility (Bucher)
- Human and non-human actors – in process of curation
- Strategy and effectiveness
- Curators of storage and display (Gehl)

Which content?...
Who uploaded the videos
- Existing media companies - Represent 'citizen videos' sent to them
- Activist media

Which frames are dominant in YouTube archive?
e.g. in Syria
-questions
-accusations
-moral statements

'mediated prospective memory'

Influence of YouTube will impact the future

2- 'ECS 2014 Online – Winning through YouTube' – Case STudy
Henrik Smith-Sivertsen

- YouTube central to studies of the digital music revolution
- Platform for amateurs
- Usefulness of evaluating web archives in numbers? Sometimes qualitative is invaluable

Document Distribution Note
Release to Members: Immediate
Release to Public: 08/06/2015

Note: These are rough notes taken live during the event

Archiving Eurovision 2014 (Online) in Copenhagen

- Mostly manual selection
- Email alerts
- Prominent misses due to things not tagged with popular or official tags

What was in scope?

- Official outputs
- Fan videos
- ESC Top videos done by viewers

Documentation in Excel

3- 'YouTube as an Archive – archiving YouTube'
Susan Aasman

Ric Prelinger: YouTube is an 'ideal form of archive'

- Comprehensive
- Open to contributions from users

YouTube is a 'default' archive
-for students, for everyday users, and professional media researchers

<u>Changing practices</u>
Manuel Castells: 'Every cultural expression, from the worst to the best, from the most elitist to the most popular, comes together in this digital universe that links up in a giant, a historical supertext, past, present, and future manifestations of the communicative mind. By so doing, they construct a new symbolic environment. They make virtuality our reality'. (The Information Age - Economy, Society and Culture)

YouTube – from digital video repository to social platform to cultural archive

Wayback Machine possible tool for reconstructing deleted or private videos

John Hartley: YouTube is a 'probability' archive

<u>Argument for archiving Facebook</u>

- Catherine Marshall and Frank Shipman; IIPC 2015, Stanford, Should we archive Facebook? Why?
- As a Heterogeneous personal store
    - Survey of FB users: should FB be archived? Majority: No.
- How many 'dead bodies' do we need to reconstruct a picture of the past?
- Cooperative collections – Brewer Kahle

Longpaper: 'Challenges in Archiving Social Media Data for Research: the case of Twitter'
Katrin Weller

Document Distribution Note
Release to Members: Immediate
Release to Public: 08/06/2015

Note: These are rough notes taken live during the event

katrin.weller@gesis.org
@kweller

- SCOPUS searches for Twitter and other social media shows an increase in use of social media in research
- Twitter more popular for research than other platforms
- Twitter is used for studying political communication, crises communication, major events analysis, etc. – not necessarily historians
- Top disciplines who use Twitter data: social scientists, communications scholars, but computer scientists dominate (from SCOPUS database)

Current project:

- interviewing social media researchers in different fields
- in different career levels
- 42 qualitative interviews
- Twitter mostly, but also other platforms

What is so special about social media data?:

- researchers value the immediateness of reactions on social media
- value the fact that it's naturally occurring data
- value the structure of the data (timestamps, author is identifiable, identifiable networks, new format with structured metadata)
- challenges of social media research (outweigh the benefits atm)
- greatest challenges: data access and data sharing

Struggle to get data from Twitter and other platforms and once they obtained a dataset they wanted to share, they encountered problems from other sources

Cause problems for research quality, and for validation; one researcher: I can't share my data so my study cannot be replicated, it can't be tested for review, also, it means my data can't be made available for other researchers to discover and use for new, innovative studies – there is no open data

One source of restriction: platform companies, e.g. Twitter, inc. Terms of Service – you cannot share data collected through an API to third parties, and certainly not in machine-readable formats like JSON or XML. Twitter has requested institutions remove datasets from their websites

Challenges to individual studies:
One researcher obtains a dataset for a study; wants to archive that ONE dataset for future use – not interested in archiving all of Twitter

Archiving Twitter at large scale not a feasible option any time soon – based on progress at LoC; strategy should be to focus on small datasets

'Twitter data' – collected through API, hundreds to the millions of tweets, or user data, Twitter and elections (collected sizes of datasets, different types of APIs, different types of tools, many researchers do not specify in their written paper,

Document Distribution Note
Release to Members: Immediate
Release to Public: 08/06/2015

Note: These are rough notes taken live during the event

Attempts to archive datasets loose rich context, lose visual interface – reduced to spreadsheet or JSON file – only field that can be preserved or shared is the Tweet ID

Future researchers who want to use the same dataset will have to call from the API again, however, any deleted tweets will not be in the set due to User Rights

Archiving Twitter involves ethical questions: user intent not for research use

Challenges, in sum:

- Twitter Terms and Conditions
- Deleted content
- Changing nature of Twitter
- Lost content (e.g. retweet via button or copy/paste?)
- Lose the conversation, lose the 'stories', lose meaning
- Can only target datasets by hashtags if you know what they are, may miss data that doesn't include hashtag
- User names change
- May Lose URLs and images

Possible solutions:

- Identify groups who have collected corpora of Twitter data
- Build a set of standards around citing process of capturing dataset
- Qualitative studies to understand how people use social media data

Document Distribution Note
Release to Members: Immediate
Release to Public: 08/06/2015

Note: These are rough notes taken live during the event

Workshop: 'Elements, Graphs, and Entities: Analyzing Web Archive Datasets'
Vinay Goel, Senior Data Engineer, Internet Archive
Jefferson Bailey, Director of Web Archiving Programs, Internet Archive

Goals:

- Foster discussion on datasets, access, research
- Give familiarity with available types of datasets and demystify derivation
- Provide skills in working with data
- Mine that data!
- Use and understand visualization and presentation tools
- Help build community and contribute to training

Workshop Agenda

- Discuss emerging research methods and web archives
- Background on research service models
- Overview and walkthrough of web archive datasets
- Leverage IA infrastructure for large-scale processing to produce research datasets
- Increase use, visibility, and value of web archive collection

- Research Services

- Some background efforts
- Expand access models to web archives
- Enable new insights into collections
- Facilitate computational analysis use cases
- Leverage IA infrastructure to help smaller efforts

Wen Archive Datasets
WAT Datasets – web archive transformation, key metadata from every resource
LGA Datasets – Longitudinal Graph analysis, what links to what over time
WANE Datasets – we archive named entities, names of people, places, organisations

+CDX Dataset

Problem is researchers need their own clusters

What is CDX?

WAT Dataset
Extensible file format
WAT contains:

- key metadata extracted from (W)ARCs for every resources
- For each URL

    - provenance metadata (timestamps, server IP, response code, doc size, content type, etc)
- Also for each HTML
    - text data (doc title, meta-keywords

WAT Advantages

- Smaller (size is about 18% of a (w)ARC
- Reference: one-to-one mapping to (W)Arc
- Packaging: records in JSON

WAT Example

LGA Dataset
Contain:

- what url links to what URLS and when
- every single link in your entire collection

'Purely linking information'

LGA Advantages:

- Size about 1% of a WARC
- Reference: complete collection over time
- Packaging: extremely compact text files
- Enables: identify important websites; study how relationships change over time; interpret graphs in a specific time slice and with varying degrees of granularity

Example: Dynamic Graph Visualization using LGA data –Vinay

WANE Dataset

Contains:

- named entities extracted from WARCs for every text resource (using Stanford NER)
- entities are the names of people, places

Named entity disambiguation – some researchers working on

Classification tools allow in some noise

For more information: Webarchive.jira.com (Archive-it research services)

Workshop:

https://github.com/vinaygoel/ars-workshop

https://webarchive.jira.com/wiki/display/Iresearch/Web+Archive+Analysis+Workshop

Document Distribution Note
Release to Members: Immediate
Release to Public: 08/06/2015

Note: These are rough notes taken live during the event

DAY 3

Short Paper: 'Web History as Time Travel: Digital Nostalgia and Collaborative Filtering in Public Engagement with the Internet Archive's WayBack Machine'
Megan Sapnar Ankerson, University of Michigan

'Nostalgia for Web 1.0'
For quieter, safer more simple times
Glitchy 8-bit excitement
Garish, flashy, neon, blinking, clip-art filled wonder

Prelim Questions:

- How do everyday users engage with web archives in the service of digital nostalgia?
- How does the general public understand the Wayback machine?
- What images are selected to represent the old web?
- How does the discourse of ti'time travel' figure in ways archives are constructed as portals to the past

Seen as accurate metaphor, that WayBack actually takes you back to a perfectly preserved moment of the web

Discourse analysis – pilot study

- Google
- blogs with snapshots of Wayback Machine
- mainstream pop press

Metareial-semiotic analysis

- draws on feminist technoscience studies (Haraway 1997;Suchman 2008)

Thierry Bardini – history of how the computer

Computational metaphors are conceptual systems that shape the ways we imagine computing in different historical moments

Illustration: 'Peabody's Improbably History' – Rocky and Bullwinkle (Edgar Allen Poe episode) – going back to fix the unstable past

Paradoxes – collecting everything – literally a window to past, but also, past is unstable

'History' – terminology
-not another name for the past (Taylor)

Historiography, something we have to learn how to do

Collaborative Filtering

Note: These are rough notes taken live during the event

- technical term for recommendation systems
- How the Internet Archive first prioritized which sites should be archived
- Cultural term for the selective and situation process

Short paper: 'Multi-layered archives: how the past of the Internet becomes present again on the web'
Camille Paloque-Berges (conservatoire national des arts et metiers, Paris)

Geneology of the web

- Usenet as a Computer-Mediated-Communication (CMC)
- 1979 born on Unix – 1989 'electronic conferencing, as the most popular use of social use of computer netwrorks

Geneology of Usenet web archiving

- Multi-layer – metaphor for showing the sedimentation of different computer network experiences in web archiving

- Methodological proposal for describing and analysising
- History of Usenet 'no archive policy' to a Web archiving cultural economy
- A sedimentation both critical and methodological
- Usenet user communities' reception
- 'ego' searcher' and 'alter searches': referential and cultural memory

Evolution of the access and search interface:

- flattering the hierarchy
- confusion of results

Conclusion

- IA and Archive tm has made an effort to archive a part of Usenet
- Google still hve the most comprehensive Usenet archive
- We researchers ware willing to make an effort to build Usenet corpora, and even standardize it
- To libraries: please archive at least you country / linguistic regional branch of Usenet: it is a crucial part of Internet History (andits textfiles light and easy to handle)

Short Paper: 'A view on reduction: why web archiving needs to be focused to become common use'
Matthias Weber, European Central Bank

What are websites? Two opinions
Importance of webarchiving for commercial sector – as evidence for auditing

Do not have an interest in archiving unless they can make money, are under pressure from public, are under pressure from legislation.

Note: These are rough notes taken live during the event

More progress in US than European Union

Suggestions for administrations or companies and their records keeping

- Focus on internet and Intranet pages
- Collection profile in harmonisation with other digital and non-digital resources
- Mainly consideration of the core tasks and aims of an organisation
- Also consideration of legal, compliance and related aspects

Bring into harmonisation with all other resources in other formats in order to avoid redundancies

Integration into a long-term preservation programme (and thus☺
Reducation…
Of amount
Of formats
Of technical peculiarities

Panel: Research Explorations of the UK Domain Data Archive

1- The Experience of researching Eurosceptiism using the Big Data Domain
Richard Deswarte, Uni East Anglia

Big UK Domain Data for the Arts and Humanities

Using web archives as sources for traditional research questions – without knowledge of new methodologies for using new sources

Read the 10 ten reports

- Richard Deswarte a 'failure' at using web archives like traditional resources – suffered from hubris
- Frustration
- Too little then too much (312 hits - .5% to 1.1 million
- Serendipity
- Meaningful results
    - uncertain start
    - we presence, growth but is that due to web or movement
    - local sites but relationship remains unclear

Big data – big problem?

- researcher problem? Archive problem?
- Search focus

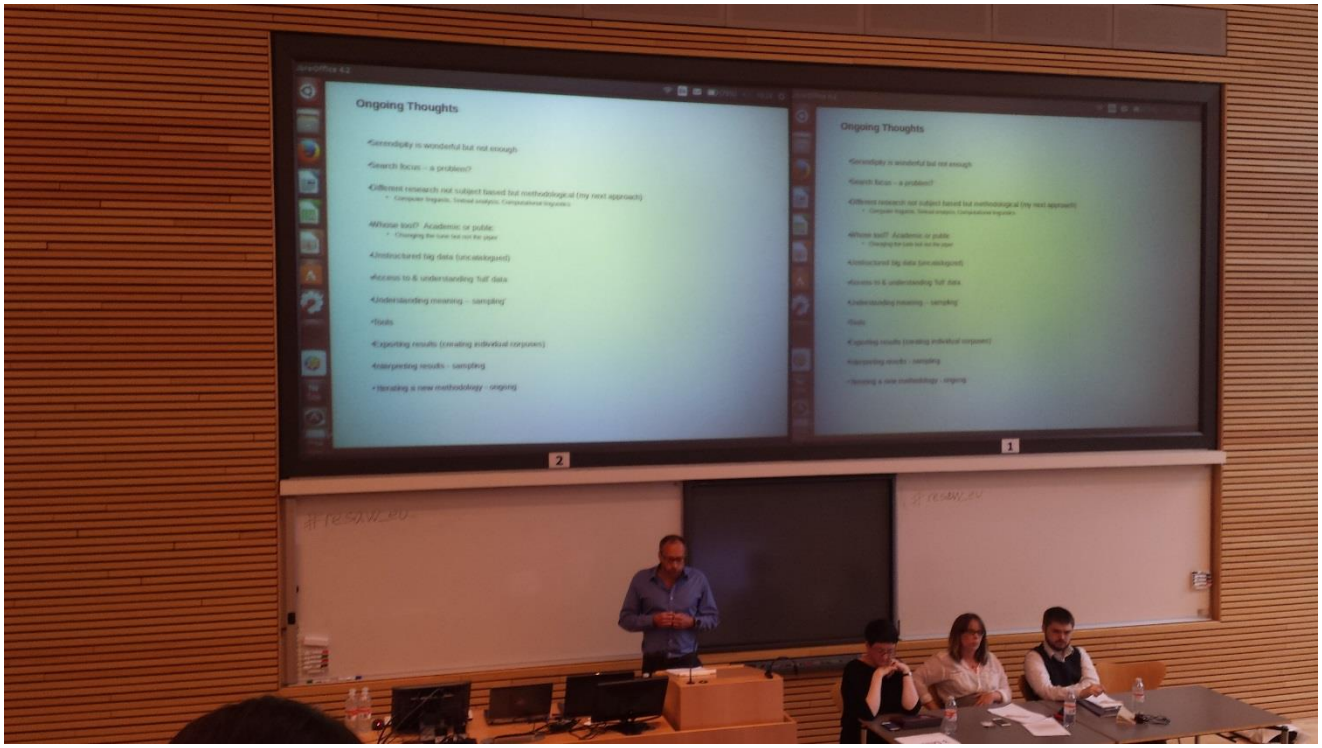
Full-text search doesn't give hierarchical results of other search engines

Sampling – way forward? Way back?

Note: These are rough notes taken live during the event

- But content is unstructured

Ongoing thoughts



2- 'Online reactions to institutional crises: BBC Online and the aftermath of Jimmy Savile'
Rowan Aust, Royal Holloway

- Savile we never fully be removed from the BBC, his long and wide-reaching career will permeate the archive
- His presence online is harder to restrain
- Used advanced search to find Savile on the website and to track where/ when he had been taken down

BBC is amending content mentioning Savile – lack of dates means unclear when exactly it was removed – only timestamps on harvests

3- 'A view on reducation: Why webarchiving needs to be fcused to become common use'
Gareth Millward

Lessons from failure

- First strategy to Build reliable corpora
- Not possible through UK Web Archive through initial methods (RNIB)
- Question needed adjustment, not results

Note: These are rough notes taken live during the event

- Though there were some quantitative results
- Went qualitative – looking at how organisations websites had changed over time – trends, fashions, fluctuating budgets

Next step would be link analysis – how did organisations use the web?

Conclusions: Internet history methodologies need to be taught to historians before letting them loose in the web archive

About this document

Version 1	Written at workshop	08/06/2015	SDT
Version 2	Distributed	08/06/2015	DPC members