Web Archiving & Preservation Working Group

National Records of Scotland (NRS), Edinburgh

Meeting 004 (Sub-topic Meeting Social Media & Complex Content, DPC Members, Supporters, and Public)

16 January 2020

In Attendance

- Frances Bell, York Explore and The National Archives
- Caroline Brown, University of Dundee (Scottish Council on Archives)
- Lynn Bruce, NRS
- Alyson Campbell, DPC
- Paul Choi, Falkirk Community Trust (Scottish Council on Archives)
- Lynda Clark, University of Dundee
- Philippa Currie, University of the Highlands and Islands
- Carl Davies, BBC
- Alice Doyle, University of Stirling
- David Elliott, British National Yachting Archive
- Jessica Evershed, NRS
- Barbara Fuentes, NRS
- Sophie Ham, Dutch Digital Heritage Network
- Grainne Loughran, PRONI
- Teresa Maley
- Sarah Mason, Artefactual
- Valerie McCutcheon, University of Glasgow
- John McMillan, DPC
- Ray Moore, University of York
- Rosita Murchan, PRONI
- Claire Newing, The National Archives (UK)
- Marjolein Platjee, Bodleian Libraries
- Rachel Proudfoot, University of Leeds
- Marcel Ras, Dutch Digital Heritage Network
- Sean Rippington, University of St Andrews
- Giulia Carla Rossi, British Library

- Silvia Sevilla, EU Publications Office
- Luke Sloan, University of Cardiff (Social Data Science Lab)
- Hania Smerecka, Lloyds Banking Group
- Kate Tasker, University of California San Francisco
- Helen Taylor, Heriot-Watt University
- Sara Day Thomson, DPC (Chair)
- Pauline Ward, University of Edinburgh
- Dorothy Waugh, University of York
- Karin de Wild, University of Leicester
- Eve Wright, NRS
- Emma Yan, University of Glasgow

Apologies

- Maria Ryan, National Library of Ireland
- Jerry Jenkins, British Library
- Helena Byrne, British Library

Minutes

10.30 Meeting Opens & Introduction by Chair, Sara Day Thomson (DPC)

- First public open meeting
- Sold out with a waiting list
- Hashtag is #WAPWG20
- Forum to share experiences, establish common goals, inform policy development
- Group does NOT build tools or tell people what to do
- Today is a special edition focussing on social media and complex content – stuff that doesn't fit in to established web archiving programmes.

10.40 Overview of Complex Web & Social Media Preservation by Sara Day Thomson

- Complex means different things for different organisations. Depends on your capacity and expertise.
- Are you interested in official accounts, multimedia posts, photos, or the actual interactions between users?
- Common complex web objects include dynamic and interactive sites, embedded or streaming components, sites with external dependencies, or device-specific (e.g. iphone specific) apps.
- Often become obsolete very quickly.
- May be difficult to tell if something has been preserved 'correctly' if you don't know what it was meant to look like in the first place.
- Challenges include – dealing with outlying cases, tools are not evolving as quickly as the technology they are trying to capture, rights issues – copyright, personal data, restricted information, right to be forgotten, lack of in-house expertise.

- Opportunities include – may be able to capture experience of a site rather than the underlying data, heading off expensive preservation efforts at a later date, opportunity to trial and engage with new technologies and shape them to your needs, developing streamlined or automated capture processes.
- Social media – typically trying to preserve individual accounts or posts using API, screenshots, or Heritrix type crawl)
    - Datasets may be created based on certain parameters (e.g. time, users, geographical area).
    - Typically includes photos or other kinds of embedded content.
    - There will also be network information (e.g. user behaviour information)
    - Twitter API can produce json data
    - Can present API Twitter data in a spreadsheet (e.g. google doc) using tools that make use of the API.
- Challenges include T&Cs around sharing and reuse, which change frequently. Each platform has different functionality and export options. Difficult for scholars to produce reproducible research. Difficult to carry out traditional archival function on interactive, fluid datasets. Difficult to plan storage requirements.
- Opportunities include - legal compliance, source of 'naturally occurring data' for analysis, improve public services and better governance, online open-source investigations (e.g. BelingCat), opportunity to engage with new tools and service development (e.g. social feed manager, documenting the now, COSMOS, Wasim Ahmed's blog)

11.05 Overview of Collecting Approach to Complex Publications by Giulia Rossi (Curator of Digital Publications, British Library)

- British Library is one of 6 legal deposit libraries in the UK and have begun to collect 'emerging formats' under the Non-Print Legal Deposit Regulations. Set up a project to identify and explore needs in this area from 2017-2019.
- Identified issues e.g. born digital publications with no physical counterpart, device dependency, multiple media, use of non-standard format and media types, at risk of obsolescence, typically not part of existing collections.
- Prioritised web-based interactive narratives, books as mobile apps, and structured data.
- Since regulations change in April 2013 they are compelled to preserve a variety of born-digital publications includes eBooks, eJournals, websites, geospatial data, digital sheet music
- Opportunity to mitigate risk of a 'digital black hole', capture the nature of UK digital culture
- Benefits to researchers: ensures preservation of their publications, support access to emerging formats
- Benefits publishers by making their work accessible over the long term, provide examples of innovation
- How do they do this - determine user expectations and needs + understand technical complexities and dependencies, with the goals of maximizing long term reusability and provide a meaningful access experience.
- This has to be done selectively due to high costs, and low level of standardisation, following the terms of the NLPD regulations. Following prioritisation agreed by all Legal Deposit Libraries.
- Collection methods include file transfer, download via access code/password, web harvesting tools, description as a collection method (e.g. contextual information when full capture is not possible). Different approaches will be needed for different content types.
- Challenges - hard to 'horizon scan', geographical scope, copyright definition, continual change in digital publishing. DRM, version control, partial capture, lack of standards.
- Technical difficulties - strong hardware and software dependencies. How to provide access - technical and legal limitations. Restrictions around re-use.
- These challenges are not unique to legal deposit libraries, but to anyone dealing with new media.

- Collections include - interactive narratives collection on the UK Web Archive e.g. 80 Days and Breathe
- Next steps - keep adding to the collections. Further research into discovery and description. Better and different forms of access. Identify new formats to consider e.g. social media, chatbots, XR, AI, fanfiction.

11.25 Case Study: Web Collections in Museums by Karin de Wild (Digital Fellow, One by One: building digitally confident museums, University of Leicester, National Museums of Scotland)

- Museums as 'servants of memory'. How will museums preserve contemporary culture e.g. the web.
- Museums are not collecting vast sections of the web - tend to focus on specific sites created by artists.
- Challenge - what is the boundary of the art object - what is part of what must be preserved?
- Project with MOMA - Agent Ruby - case study for preserving internet art.
- Ruby is a chat bot from 1999, made with then state of the art technology. The interface is the primary focus of preservation - however Ruby is created in Java which is no longer supported by many browsers. Agent Ruby could have been migrated to a new interface by rewriting the code.
- 'Code resituation' avoids the replacement of the original artists' code to restore functionality by reusing as much of the original code as possible.
- Agent Ruby contains several links that no longer work but redirect to Internet Archive where possible.
- Agent Ruby is now a software-based installation in a museum rather than a website.
- MOMA did present her online in 'espace' for a while, but this is no longer active. Museums need to consider their own digital legacy
- Ruby could be downloaded to palm computers, but this technology is largely obsolete now. This can now be emulated on modern phones to provide a similar experience.
- Two important datasets - (i) Agent Ruby's answers dataset, which links to her facial expressions, and (ii) a record of every user and conversation she has ever had (including user data). Some of these transcripts have been published as part of the museum installation. Should/can this data be made publicly accessible? The chat logs were not created by the artists and are not perceived to be part of the artwork by some conservators.
- Agent Ruby was iteratively designed, with many versions. Some are preserved, some are not.
- Later version called 'Dina' and 'Peter Weibel' bot.
- Rhizome - online archive for preserving online art. How to describe provenance of internet art?
- Using linked data to describe the provenance of online art, recording links between different versions and different collections and actors/agents.
- Used the Prov-DM data model.
- There is not a standard way of describing provenance. Prov-DM could act as a translator between standards.
- [Slides show examples of how the ontology has been used]
- Challenges/Risks - how to preserve the interface, links, hardware, data, iterations/variants.

11.45 Q&A / Discussion

- What about art that is created with the understanding that it will not be permanent?
- Bodleian library web archive use Archive-IT for trying to capture social media archives but have some problems. Capturing a few Instagram sites using Heritrix
- TNA capture some social media as part of their web archiving programme. Use MirrorWeb. Some limitations on what can be captured via APIs so are looking at other technologies including WebRecorder. Looking at LinkedIn - some government departments use this quite a lot. GitHub is used by a lot of government departments (including TNA).

- BBC produces lots of apps which are hardware and software dependent. E.g. BBC sounds, news apps. So they record the user experience via screencasting rather than preserving the underlying code. Are devices such as iPhones being preserved?
- Saving the stuff is relatively easy but providing meaningful access is difficult (access is what we are preserving!)

12.00 Lunch

12.45 Case Study: Defining and Capturing Web-based Interactive Fiction by Lynda Clark (Postdoctoral Research Fellow, Innovation in Games and Media Enterprise, University of Dundee)

- There aren't many archives of videogames or interactive fiction. Those that exist often rely on videos of people playing games
- Project brief to create a collection and create something that relates to existing collections
- Process: identify objects, determine UK authorship, categorise, collect, QA, analyse what people are making and create new works
- Uses the definition of 'Interactive Narrative' as used by the British Library: interactive, narrative (can be non-linear or anti-story), non-standard, and web-based.
- Why is interactive narrative at risk? Normal reasons e.g. Dropbox changing their terms for sharing files. Can sometimes use comments and reviews to guess at what missing games were like.
- Identified 294 works by 114 creators. Hypertext, parser-based, choice-based (sometimes paywalled), multi-modal, video-game style, and uncategorisable
- There's a lack of standardisation, no centralised storage, no collection method for commercial (paywalled) works.
- Collected using the UKWA tools (W3Act) and Webrecorder.
- Hypertexts were captured well as standard websites, unless they heavily relied on images.
- Parser-based were captured well by Webrecorder, the UKWA didn't work so well for these.
- Choice-based - success varied from item to item.
- Multi-modal - used both UKWA tools and Webrecorder, each capturing different elements
- Avatar-based games - could be crawled if the creator had enabled auto-load on itch.io
- Some objects could not be collected satisfactorily - the limitations were recorded and as much contextual information as possible
- Webrecorder good for video and audio, though its time-consuming and largely manual
- ACT (tool) - is better for large scale automated captured but is more limited in what it can capture.
- Analysis - genre-mixed works were very common, as were 'slice of life' works.
- Public transport, tea, mental health, cats and metanarratives were common
- Twine is open source so code could be preserved easily.

12.45 Lunch

13.15 Case Study: Twitter Data for Social Science Research by Luke Sloan (Deputy Director, Social Data Science Lab, University of Cardiff)

- What is Twitter? Is it useful for Social Science research? It depends!
- Social Data Science Lab (SDSL) has funding to explore linking longitudinal data and Twitter data. There will be ethical concerns, preservation concerns, access concerns
- SDSL has used Twitter data to predict crime and election outcomes
- SDSL developed a tool called COSMOS to collect Twitter data using the API. There is now a web version. Intention is to democratise social media data collection.

- Cannot use Twitter API / COSMOS to collect retrospectively.
- COSMOS DEMO - using the downloadable version. Should be easier for non-technical experts to collect tweets. Can collect a random 1% sample of tweets produced, or filter by specific terms. Will return all tweets produced up to 1% of the global feed (this is capped by Twitter itself)
- Can then filter the dataset to narrow down more specifically to what you want.
- Can analyse using a range of tools - in this case a word cloud
- Original version of COSMOS was able to work out where Twitter users lived by studying their geodata - this is now being turned off.
- Have to be careful when it comes to drawing conclusions from any dataset - sampling must be done carefully.
- Do we understand what Twitter data really is - the software may be obscuring this a bit.
- Do users understand how the API works?
- Do users understand what JSON is and what they can learn from it? A single tweet can have over 150 associated attributes!
- Need to consider the tweet, the user and geography
- Confidentiality can be introduced by only saving the tweet ID rather than the user name. Should users be able to delete their tweets, including offensive or illegal tweets? Tweets from public officials? Teens in particular tweet to an 'imagined audience' of peers rather than researchers.
- Provide a secure environment for analysis that prevent individuals from being identified by combinations with other data.
- Many tweet attributes can be combined to identify an individual.

Exercise: Analysis of Disclosure Risk by Twitter JSON Attributes: see Worksheet

13:45 Q&A / Discussion

Q: Using COSMOS, are search criteria saved alongside search results? Need to know how the dataset was created.
A: No, but they are hoping to develop this.

Q: Can historic tweet datasets not created using COSMOS be ingested into COSMOS to be analysed?
A: Yes, this can be done. The dataset may need to be separated into smaller chunks for processing.

Q: Can COSMOS deal with other sources of social media data?
A: Only Twitter so far as this is the easiest site to harvest from. May expand to include Instagram.

14.00 Ethical Approaches to Archiving Social Media: an Intro by Sara Day Thomson

- Ethics are not the same as the law, or terms and conditions. Difficult to enforce in principle. Should inform policies and values statements.
- Different social media platforms have different terms and conditions.
- User awareness of how their social media presence can be used is relatively low. Not clear that they really understand consent forms, user agreements etc. Not clear that consent is informed.
- Who does the collection belong to, who is entitled to curate and preserve it, interpret it? See Michelle Caswell's Feminist Standpoint Appraisal.
- Examples include Documenting the Now - lack of user consent, potential for fraudulent use and manipulation. Potential for harm to tweeters.
- Difficult in applying traditional processing techniques due to scale/volume of archive.
- There are 'data haves' and 'data have nots' - big disparities in power in creation, management and ownership of data, and who has access to data and who does not.

- Ethical decisions should be considering the platform, user awareness and consent, ownership and authorship, economics of access. Try to empower people to make their own archival decisions where possible.

14.15 Small Group Activity and Feedback to Big Group

Ethical Deliberation: To Archive Twitter or Not to Archive Twitter

Proposition 1: 'The legal regulations and Twitter T&Cs are enough justification for preserving Twitter data as part of collections.'

- Twitter's terms and conditions are freely available and transparent. Everyone had to sign up for them.
- The legal basis is out of our hands and its common knowledge that tweets can get out of hand. It's for the common good that we collect Twitter.

Proposition 2: 'The legal regulations and Twitter T&Cs are enough justification for preserving Twitter data as part of collections, but we will only keep Tweet IDs (unique identifiers for each tweet) and dispose of all other tweet content and metadata.'

- Web archives cannot see the future of the information they hold. Data trends mean that it is very easy to identify people and there is a risk of identifying individuals in areas such as health and crime.

Proposition 3: 'The legal regulations and Twitter T&Cs don't go far enough to protect individuals represented in the data, so we have to seek permission from every user before collecting.'

Group 1

- The group looked at examples of people who have posted on Twitter without realising what the consequences are. Terms and Conditions of Twitter have changed since 2010. People who signed up in 2010 won't have understood the consequences. The 'Rules' of Twitter don't tell people what others do with the data. Therefore, people aren't informed about how their data is used. Twitter admitted they misused their own conditions by selling data to advertisers.
- We shouldn't stop archiving, but we should accession in the same way as oral history. We shouldn't collect data without knowing we can re-use it in the future – especially given the large size of the data and we may not be able to re-use in future if we don't ask now.
- Terms and Conditions, including Twitter rules, do not clearly state how others may use your data and are more focussed on what you the user may do. Some evidence from academic studies that users do not fully understand how their data can be used. Also anecdotal evidence that younger users believe they are tweeting to their peers only and do not fully understand how academics (and others) might use their tweets. Twitter has already admitted that some of the use settings have failed to work in the past - data has been shared even when users did not give permission.
- It not fair or ethical to expect that users understand that their tweets contain 150 different attributes that be used to personally identify them.
- It's also quite standard for social scientists gathering data to use consent form and there's no reason why we should not do this for twitter - we already have processes and forms in place that we can use that will have been created by our legal teams and ethics committees, with things like GDPR in mind.

- Doesn't make sense to invest resources in gathering data that we may not be able to use in the future - just gather consent to make it clear.

Group 2 added

- GDPR and right to be informed about collection and use of data means people should be informed at time of collecting that their data is being processed.

Proposition 4: 'The legal regulations and Twitter T&Cs don't go far enough to protect individuals represented in the data, so we will notify every user (through Twitter) and offer to remove any tweet or content upon request.'

- People don't understand the consequences of Twitter. We need to open up researcher and make users aware. To future proof collection we need an audit trail of consent. This would comply with the right to be forgotten. They would use automated replies to do this.

Proposition 5: 'The legal regulations and Twitter T&Cs don't go far enough to protect individuals represented in the data, so we will collect and preserve Twitter data, but store it in a secure dark archive.'

- Also had issues with Twitter's T&Cs. The dark archive would enable them to collect but to leave untouched. This would bring it into line with other digital deposits.

- They also noted that T&C's can change over time and these will probably be time-based.

Final Thoughts

'Archiving Twitter presents many challenges, but we have the ethical duty to do it.'

15.45 Remaining Items, Actions, & Next Steps

Q: What to do with sites that require an age verification - doesn't work with web crawlers. Webrecorder seems like an option but it's manual and time-consuming.
A: attempted this with interactive fiction but Heritrix didn't work

Q: Should we think more about our moral and ethical duty to preserve social media data, rather than ethical risks? Need to make sure we document people, have the data to be critical of social media companies, and to liberate data from them.

16.00 Close of session

Minutes were contributed by Sean Rippington, University of St Andrews, and Lynn Brice, National Records of Scotland with thanks from the DPC and the rest of the WAPWG.

Published 03/02/2020