Web Archiving & Preservation Task Force

National Library of Ireland, Dublin

First Re-convened Meeting

27 June 2018

In Attendance

- Carl Davies, BBC
  - Web and Twitter Archive
- Garret McMahon, Digital Repository Ireland
  - Preservation of data / dynamic web
- Natalie Milne, National Archives of Ireland
  - Web editor of NAI website
- Eilidh MacGlone, National Library of Scotland
  - Collecting alongside modern collections
- Els Breedstraet, EU Publications Office
  - All EU publications, including digital and web
- Garth Stewart, National Records of Scotland
  - Web continuity service
- Joanna Finnegan, National Library of Ireland
  - Interested in user service for web archives
- Maria Ryan, National Library of Ireland
  - Web archivists, looking to build a case for digital legal deposit to cover web
- Della Keating, National Library of Ireland
  - Stared Irish web archive 2012
- Leona Taylor, PRONI
  - On team responsible for digital preservation and web archiving
- Jason Webber, British Library
  - Engagement manager
- Kourosh Hassan Feissali, The National Archives UK
  - Interested in QA of UK web archive, improving processes

Virtual Attendees

- Tom Wilson, UNHCR
  - New start, Digital Preservation Team
  - Crawling UN web content
- Lori Donovan, Internet Archive
  - Programme manager for web archiving
- Jen Mitcham, Borthwick Institute for Archives, University of York
  - Digital Archivist, web archiving is now in institutional policy
- Laura Peaurt, University of Sheffield Library

Minutes

12.00 Lunch meet and greet

12.45 Welcome by Sandra Collins, NLI

13.00 Introduction by Sara Day Thomson

- introduced the work of DPC, her own research, and the reasoning for the re-launch of the Task Force – a very popular subject mentioned at DPC Member events.

13.15 Introduction to work at NLI by Maria Ryan

- growth in web archiving domain crawls in last 3 years
- Irish copyright law still does not extend to web content; still a challenge to making Irish web archives accessible

Revised Terms of Reference

- Defining some terms to a group which covers libraries, archives and other sorts of collecting bodies: web record creators or web content creators; defining publications and 'to publish' in a web content
- Legislative intervention
    - Reinforcing change and advocacy as a group
    - Aligning with DPC role in supporting and facilitating advocacy for digital preservation more generally
        - Letters to MPs
        - Public consultations
        - Other
- Public Consultations
    - influence national legislations
        - GDPR: exceptions for libraries and archives?
        - Copyright
- Monitoring
    - keeping an eye out for opportunities to respond to public consultations or other opportunities to intervene in legislative processes that affect web archiving and preservation
    - Laura, Internet Archive discusses process of reaching out to US legislators
        - letters to Congress
        - legal briefs
        - different avenues through courts (for interpretation of legislation) and through legislation (creation and amendments of laws)
        - IA work with Library of Congress (office of copyright is housed there)
- Request for an added clause for a Knowledge Exchange platform. Some possibilities:
    - forum (with alerts?) – on DPC website
        - benefit of an OS Forum is it will be easier to archive
    - Policy and tool exchange
    - Email thread
    - Slack

    - Blog
    - GitHub channel
    - Google Hangouts for chats between meetings
- Closed and Open Meetings (2.2 or 2.3 in Terms of Ref)
- Open Meetings for a subtopic groups to discuss a specific issues; e.g. Solutions – finding tools, event with service providers / vendors, social media
    - Laura from Sheffield – subtopic group about rights issues/how to comply with rights
    - Elements of resource discovery
    - Combine WAPTF meetings with with other events – DPC event? Others? – for those travelling
    - Invite experts on certain topics, eg procurement
    - These groups could organise and coordinate using Slack channel or other Forum functionalities
- Other possible function of WAPTF:
    - Keep a 'thread' or 'open conversation' about procurement and requirements
    - Survey of how people do web archiving
    - Facilitate support network for DIY
- Scoping issue brought up by Garret McMahon (DRI): Is data / underlying databases / dynamic web in scope for the Task Force
    - Consensus is yes
    - Best practice around long-term preservation
    - If good DP in place at creation of data interface / system
    - preserving data for re-usability - need for lobbying
    - Presenting data online
    - Maintaining utility
    - Funders/grant applications – become standard requirement for universities to provide URL
    - Open data portal
    - Datasets also should be delivered to central archive
- Lobbying also should be with funding bodies to implement web archiving requirements from beginning of research projects
    - UK HEI challenges
        - dead project websites
        - projects that lose funding
        - Contextual website around project at risk even when data goes to the repository – challenge is to keep them linked
    - Lori Donovan
        - Internet Archive (IA) relationships with funding organisations – as recipients and just in the same community
        - IMLS, private funders, govt funding
- Chatham House Rules
    - Member meetings: All DPC members to receive proceedings
    - Chatham House Rules to be decided on case-by-case basis for 'extra' meetings and subtopic meetings
- To remain a living document

Actions

- SDT to clarify language describing 'publication' and 'record/object'
- SDT to add requirement to section 2.1 for transparency and group consensus in order to represent Task Force

    - This agreement would depend on the motion being specifically related to the remit of the Task Force Group. "Keeping the web accessible" is perhaps the best catch-all mantra for directing the Group's activity.
- SDT to add funding bodies to lobbying objective in section 2.1
- SDT to add objective 2.4 to create option for sub-topic groups
- SDT to add clause 2.5 for DPC to create an online knowledge exchange and create proposal for what this will look like
    - KHF (TNA) to look at online options for Task Force communication e.g. forum, wiki, Trello etc.
- NLI, NRS, others happy to contribute to knowledge exchange of requirements building
- Task Force members to monitor for national/regional/local issues which would fall into the Group's focus area and feed back to Task Force at next meeting or remotely – to become a standing item on the agenda.
- SDT to draft 'common areas of concern' for the group and ask for volunteers to lead on these in designated sub-groups e.g. tools, website design, legal deposit etc.
- SDT to look at options at a vendor 'hot seat' event at which where various web archiving vendors would have the opportunity to showcase their services and products for Task Force members and perhaps hold clinics
- Further comments on Ts and Cs to be circulated remotely.

Show & Tell by NLI and Tour of Reading Rooms

Open Session

- Opportunities for collaboration and advocacy
    - Brexit as a hot topic that garners attention from the media
    - BL: EU referendum Twitter for Leave campaign garnered interest and raised awareness of UKWA, as did Grenfell Collections
- Making decisions about selection policies
    - BL archives web content for 'everyone', but really have to ensure support of academia
        - It's a risk to spend 4-5 years on a PhD when limited understanding by academic supervisors; so they encourage web archives as one source of data alongside other data
    - NLS: curator for Scottish Indy Ref collected from across all formats, incl. web
        - Collecting around a subject regardless of format
        - Less useful to separate by format
        - Trying to get historic index for what was happening at the time of capture
        - family tree of how web resources evolve – requires metadata
    - NLI: web archive collection policies mirror existing 'paper policies'
        - difficult to gain engagement from academics because of a culture of practice very attached to books
    - BL: difficulty finding method for citation to web resources that meets demands of academic researchers e.g. most material can only be viewed on site in the reading room on computers with many functions disabled, for example, users cannot copy and paste and cannot take photos of the screen, so have to transcribe URL by hand
        - 'golden handcuffs' of eLegal Deposit (but at least it is captured and maintained)

    - TNA: collect official gov't websites and social media accounts; gain permission from website manager for copyright, some exceptions for third party content used
        - Increasing number of requests from civil servants; they are starting to see value of web archives
    - BL: Citation issues with web archives - journal has publication date but websites only have a capture date, which doesn't mean much
    - Journalists also starting to starting to see usefulness of web archives, but academics less engaged
    - EU Pub Office: archive web publications by EU, but making decisions still difficult (eg social media?)
    - difficulty defining 'publication': ''if exposed to the public' is not a useful definition anymore
    - must be transparent about collecting practices so that context is evident
- NRS: condemned if you do, condemned if you don't
    - 150ish gov't agencies
    - 150ish private depositors
    - publish documentation about when sites are archived, how often crawled
- Showing demand for access (to demonstrate funding needs / make a business case)
    - Not about numbers but impact & risk
- Goal is to embed Web Archiving as Business as Usual
- Duplication better or worse?
    - NLI: uses too many resources
    - BL: more copies make the record more reliable
- EU Pub Office: would like to see standardization of metadata for web archiving
- BL: looking to implement Memento or something similar
- NLS: performs network analysis; Excel analysis to track sites you may not know about
    - Documenting QA
        - review QA from last harvest
        - excel
- TNA QA Checklist
    - Task Force review of checklist created by TNA for quality checking
    - similar list at EU Pub Office – also used to educate website owners, as a double check
- Contract with supplier covers QA?
    - EU Pub: no. only to website creators
    - quality v cost
    - NRS: basic QA from supplier then in-house check; different browsers
    - pick issues – can it be fixed or not
    - deciding what things to ask supplier to fix (20-30 captures per month)
    - EU Pub Office: only ask supplier to fix approaches or bugs that will improve future captures rather than 1-off patches
    - TNA: webrecorder patches – captured by TNA, patched in by supplier
- TNA URL QA
    - Xenu – crawl site for 404s and other errors and creates a list
        - checks if broken on live site
        - if working on live site, ask supplier to patch
        - under 1% not captured
    - NRS: impose time limit on QA
    - TNA: sometimes QA takes 3 mins, sometimes, 3 days
- York: is QA necessary?
    - NRS: cost-benefit; only if it's worth it
- TNA: trying to automate QA with MirrorWeb (supplier)
    - robot to do things like xenu 404 crawler

    - to eventually do large scale QA
- BL: using machine learning for QA
- QA for Twitter archive at TNA
    - because captured via API so QA simpler
    - check images to ensure capture
    - check if all tweets have been captured
    - 6 month embargo on Twitter archives
    - BBC Twitter playback
    - show & tell
    - searchable interface
    - currently backfilling from 300 BBC account owners
    - indexing outcomes
    - Twitter account for every radio show
    - Twitter archive acts as catalogue of content / guests on radio shows
- BBC YouTube archive
    - unique material uploaded
    - other archived at source
- BBC Facebook
    - downloading accounts (from owners / as owners) and trying to find ways of presenting the 'self-archiving' data
- EU Copyright article 13
    - what will be the impact on collecting?

Next Date for Web Archiving & Preservation Task Force

- 8 November 2018, NRS, Edinburgh

Close of Session

Minutes taken by Sara Day Thomson with additions from Garth Stewart (NRS)

Published 25/07/2018