

Web Archiving & Preservation Working Group (previously Task Force)

Tate Britain, London

Meeting 003 (General Meeting, DPC Members-only)

18 July 2019

In Attendance

- Helena Byrne, British Library
- Fionnuala Cavanaugh, Tate
- Sarah Colborne, University of Nottingham
- Carl Davies, BBC Archive
- Barry Dunkley, Tate
- Tom Ensom, Tate
- Patricia Falcao, Tate
- Jennifer Febles, HSBC
- Barbara Fuentes, National Records of Scotland
- Andrew Gent, Tate
- Sarah Haylett, Tate
- Hannah Jordan, University of Oxford
- Christopher King, Tate
- Louise Lawson, Tate
- Grainne Loughran, PRONI
- Claire Newing, The National Archives UK
- Michael Popham, University of Oxford
- Maria Ryan, National Library of Ireland
- Tony Steer, BBC Archive
- Garth Stewart, National Records of Scotland
- Sara Day Thomson, DPC
- Geoff Walker, BBC Archive
- Jason Webber, British Library
- Kyrie Whitsett, Internet Archive

In Virtual Attendance

- Eilidh MacGlone, NLS

- Silvia Sevilla, EU Publications Office
- Carlos Rarugal, BL

Minutes

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

- Welcome to new and returning members
- Change of designation from Task Force to Working Group. A Task Force concludes with the completion of a particular task, but the WAPTF has on-going and evolving needs. From now on the group will be referred to as the Web Archiving and Preservation Working Group.
- Quick round of delegate introductions, including name, institution, and role in web archiving.

10.40 Tate Welcome

- Welcome to Tate and context for relevant work at Tate by Deborah Potter, Head of Conservation

10.50 Review of Previous Meeting, Chair

- Quick review of minutes from 8 November 2019, Edinburgh

11.00 Open Floor to Agenda Items

- No new items

11.15 Open Discussion: Small Group Breakout

- 3 smalls groups based on the groups from the November 2018 meeting, with the aim of furthering the discussions of the previous meeting
- Advocacy and Legislative Consultation: draft a letter to an MP or other influential figure to advocate for supportive legislation
- Social Media and Complex Web Content: discuss if there is enough common ground among institutions to build a standardized model for documenting and archiving social media (by platform? In general?)
- Getting the most of third-party suppliers and writing requirements, including an exit strategy

12.15 Feedback to Big Group

1. Advocacy and Legislative Consultation

-- start of notes from Jason Webber

Letter to MP

Goals

- Ireland: web archiving legislation - including open access (Ireland has digital deposit - e journals, ebooks)
- UK: Open access; Data mining; Secondary data collection and usage (Ireland)

Other considerations

- Challenges in current environment - reading room access offers several barriers
- ATI and other digital researchers

Types of people wanting more access

- Journalists
- Computer scientists
- Digital humanity students
- Genealogists
- Interactive fiction authors
- Artists
- Web designers

Motivations

- National libraries should collect and protect the nations memory
- The web is lost or changes regularly
- Many people don't know how much they will miss their content when it is gone
- The web is the history of our time. Not preserving and giving access is like burning a library
- Preservation without access is pointless, web archiving is preserving the access
- Public available at publish, why not make it

Dear MP/TD

I am very concerned about the preservation and access to the nations historic websites. These public documents may only be on the live web for a short period of time but represent this nations history.

Currently the NLI is only legally able to collect these important resources with the explicit permission of the website owner. Not collecting the nations web space is denying future generations this important history.

In addition to collecting the nations websites it is important that this resource is given the wisest possible access. Open access to this resource should be the default with the opportunity for content owners to request that their website be reading room only. The benefits of open access are that academic researchers, journalists and members of the public can easily connect with the digital past.

-- end of notes from Jason Webber

2. Social Media and Complex Web-based Content

- Platforms too diverse and with such different functionality that it's not possible to create a single set of significant properties or requirements to apply to all social media
- Even difficult to create model for single platform, such as Twitter, because institutional needs are so diverse. For example, one institution might want official institutional social media account while another might want to capture high-quality captures of a single-complex object, such as an Instagram artwork. Some end users may want to view individual posts, others might want to data mine large quantities of content.
- Perhaps more useful to create example workflows or approaches dependent on scenarios and requirements to share as guidance

3. Third Party suppliers and Requirements, esp. Exit Strategy

-- start of notes from Garth Stewart

Discussion: Why members are interested in web archiving contracts

Discussion started by discussing whether it would make sense to capture and archive the access logs to your web archive. TNA are doing so, and aim to present these to users in due course.

Carl from BBC advised it was crucial to know exactly what you need before embarking a procurement exercise. The extremely complex nature of the BBC web estate (15m page on the BBC website alone) has led them to conclude that they require third party assistance in attempting to archive this. It is tricky to predict the market appetite at the moment for our types of service, so to help with this BBC are currently issuing a request for information to potential bidders. It was felt that certain requirements could be delivered contractually, but could be combined with existing in-house knowledge and processes. BBC are also interested in rolling out web continuity redirection for their sites, given the public record value of the Beeb's web content.

Tate carry out some web archiving in-house, but are considering a contractual service as they require dedicated technical assistance in areas such as emulation and replay of old content.

HSBC use a third party supplier for capturing some content from their corporate sites. However their working context means that they are unable to consider using any open source options to improve on this.

BBC: risk of in-house system is that staff turnover can impede business continuity.

Discussion: Contract Implementation and Maintenance

TNA: try and build a lot of time into the beginning of your contract to test processes and set clear expectations. TNA also have monthly calls with their supplier for review purposes.

BBC/NRS: try and influence your suppliers' roadmap as much as you can, but be wary of cost potential.

Don't forget the significance and responsiveness of the service's takedown function. This was particularly crucial to the BBC whenever they needed to takedown historic content in light of the Jimmy Saville sexual abuse scandal.

Discussion: what happens when companies go bust/exit strategies

The importance of building a clear exit strategy into your contract was discussed at large. HSBC highlighted that they had found it very beneficial to test the exit strategy of their digital preservation system.

TNA advised that they hold copies of WARC files of their entire archive in their digital repository. If they were suddenly to lose the services of their web archiving supplier, access would go down for a while

TNA: there's no doubt that building up your subject knowledge and market experience puts you in a much better position for contractual change.

Discussion: what's the minimum tenure for a web archiving in order to ensure proof-of-concept and be useful?

NRS: period of three years worked for us. Gave us subject knowledge, staff development, integration with peers network, and awareness of the market place.

What should any web archiving contract include as a minimum?

1. Capture functionality
2. QA functionality
3. Access Functionality
4. Takedown Functionality
5. Search
6. Exit Strategy
7. Security of data

-- end of Garth Stewart notes

12.45 Lunch

13.30 Tour of Tate Library and Archives

- Cancelled

14.30 Tate Show & Tell: Intermedia Art by Patricia Falcao and Sarah Haylett

- Current stages of evaluating intermedia art website and strategies for preserving it
- Slides available on the DPC Working Group page:

15.00 Intermedia Art Website Decommissioning Strategies, Fionnuala Cavanaugh (Tate)

- Current stages of evaluating intermedia art website and strategies for preserving it
- Slides available on the DPC Working Group page:

15.15 Big Group Open Session (remaining items)

- BBC Archive give a demo of the new web interface for the BBC Archive
- Live URL: <https://www.bbc.co.uk/archive/>
- Internet Archive capture from 18 July 2019:
<https://web.archive.org/web/20190718182811/https://www.bbc.co.uk/archive/>

15.45 Actions & Next Steps

- Next meetings will include General Meetings for DPC Members only as well as open, sub-topic groups focusing on 1) social media & complex web-based content; 2) Advocacy, Outreach, and Education for Web Archives; and 3) Web Archiving 'Trade Show' – Vendor Demos and Q&A
- Volunteer requested to help plan and host these meetings, follow-up email to request volunteers

16.00 Close of session

Minutes were contributed by Jason Webber, British Library and Garth Stewart, National Records of Scotland with thanks from the DPC and the rest of the WAPWG

Published 16/08/2019