Web Archiving & Preservation Working Group
Social Media & Complex Content

16 January 2020

National Records of Scotland
Edinburgh

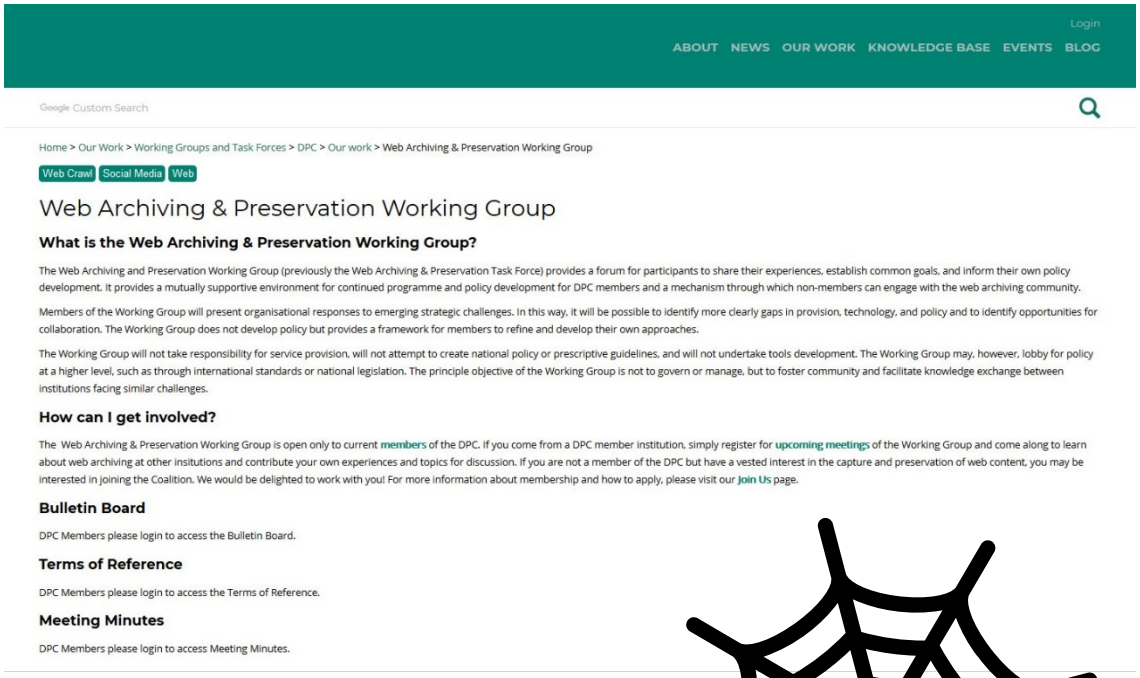
#WAPWG20

DigitalPreservationCoalition

dpconline.org

Welcome!

- Web Archiving & Preservation Working Group
- Forum for participants:
  - share their experiences
  - establish common goals
  - inform policy development



Login

ABOUT NEWS OUR WORK KNOWLEDGE BASE EVENTS BLOG

Google Custom Search

Home > Our Work > Working Groups and Task Forces > DPC > Our work > Web Archiving & Preservation Working Group

Web Crawl Social Media Web

Web Archiving & Preservation Working Group

What is the Web Archiving & Preservation Working Group?

The Web Archiving and Preservation Working Group (previously the Web Archiving & Preservation Task Force) provides a forum for participants to share their experiences, establish common goals, and inform their own policy development. It provides a mutually supportive environment for continued programme and policy development for DPC members and a mechanism through which non-members can engage with the web archiving community.

Members of the Working Group will present organisational responses to emerging strategic challenges. In this way, it will be possible to identify more clearly gaps in provision, technology, and policy and to identify opportunities for collaboration. The Working Group does not develop policy but provides a framework for members to refine and develop their own approaches.

The Working Group will not take responsibility for service provision, will not attempt to create national policy or prescriptive guidelines, and will not undertake tools development. The Working Group may, however, lobby for policy at a higher level, such as through international standards or national legislation. The principle objective of the Working Group is not to govern or manage, but to foster community and facilitate knowledge exchange between institutions facing similar challenges.

How can I get involved?

The Web Archiving & Preservation Working Group is open only to current members of the DPC. If you come from a DPC member institution, simply register for upcoming meetings of the Working Group and come along to learn about web archiving at other insitutions and contribute your own experiences and topics for discussion. If you are not a member of the DPC but have a vested interest in the capture and preservation of web content, you may be interested in joining the Coalition. We would be delighted to work with you! For more information about membership and how to apply, please visit our Join Us page.

Bulletin Board

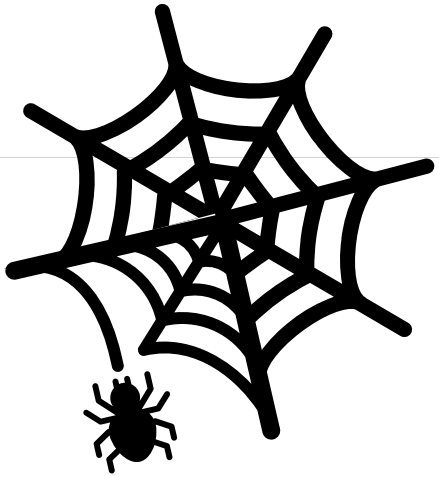
DPC Members please login to access the Bulletin Board.

Terms of Reference

DPC Members please login to access the Terms of Reference.

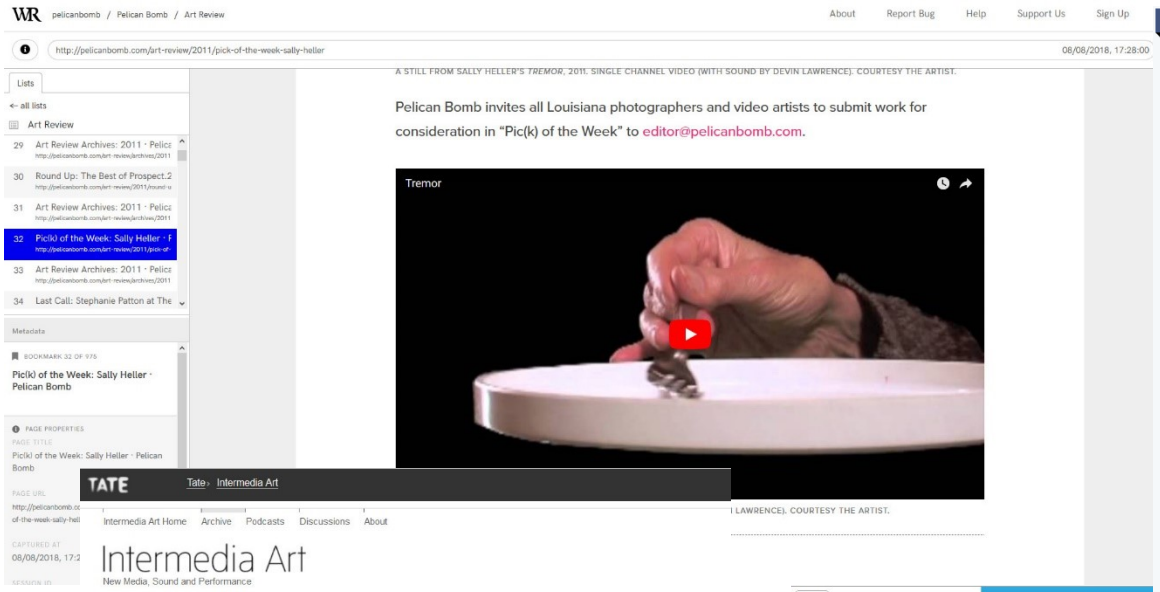
Meeting Minutes

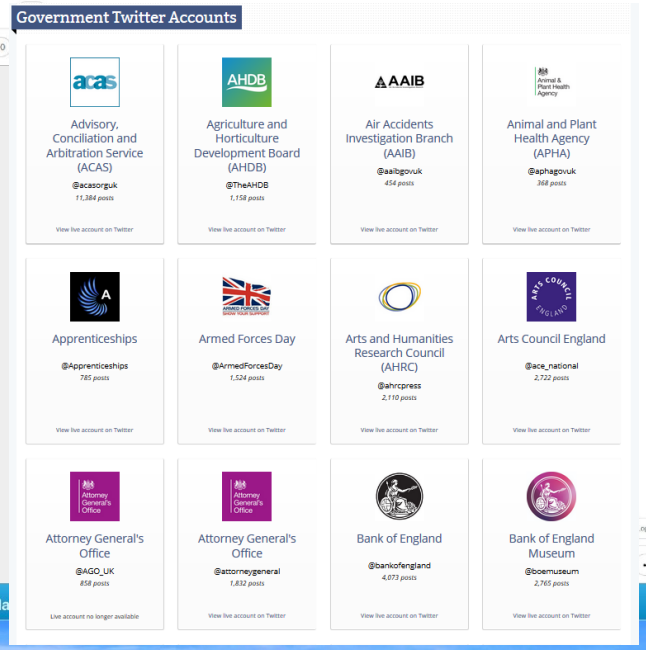
DPC Members please login to access Meeting Minutes.

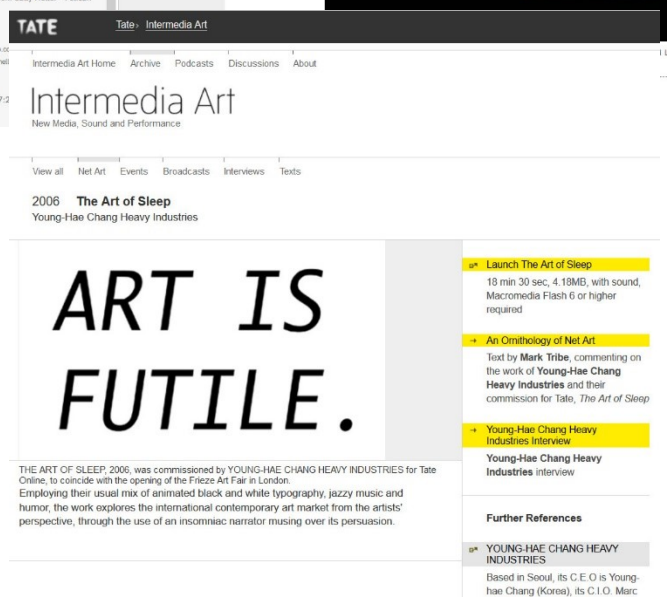


Special Edition:
Social Media & Complex Content

DigitalPreservationCoalition







Digital Preservation Coalition

… exists to secure our digital legacy

- Membership Organisation
- 90+ Members Worldwide
- 7 Supporters
- Events, Training, Research, Consultancy, Guidance, Publications, Awards, and more!

dpconline.org

JOIN US

Upcoming Events

Aye, Preserve!
(free, open to all)
12 March, Glasgow

DPC Briefing Day:
managing sensitive data
2 April 2020, Manchester

DPC Briefing Day:
preserving 3D data
30 April, York

Upcoming Events (DPC Members)

Web Archiving Webinar

Ep 1: Approaches & Tools - Large-scale and Collaborative

30 Jan 2020, 4pm (UK)

Web Archiving Webinar

Ep 2: Approaches & Tools - DIY & Microservices

6 Feb 2020, 3pm (UK)

Connecting the Bits (& more!)

22-24 June 2020, London

Upcoming Events

DigitalPreservationCoalition

WAPWG goes to the Netherlands!

October 20

More to come

Today's Programme

10.00 Registration, Tea & Coffee

10.30 Welcome & Introductions

10.45 Overview of Complex Web & Social Media Preservation

11.05 Overview of Collecting Approach to Complex Publications by Giulia Rossi

11.25 Case Study: Web Collections in Museums by Karin de Wild

11.45 Q&A

12.00 Lunch

12.45 Case Study: Defining and Capturing Web-based Interactive Fiction by Lynda Clark

13.15 Case Study: Twitter Data for Social Science Research by Luke Sloan

13.45 Q&A

14.00 Ethical Approaches to Archiving Social Media

14.15 Ethical Deliberation

15.00 Feedback to Big Group

15.45 Remaining Items & Next Steps

16.00 Close

Overview of Complex Web & Social Media Preservation

DigitalPreservationCoalition

How is it different?

Examples of Complex Content

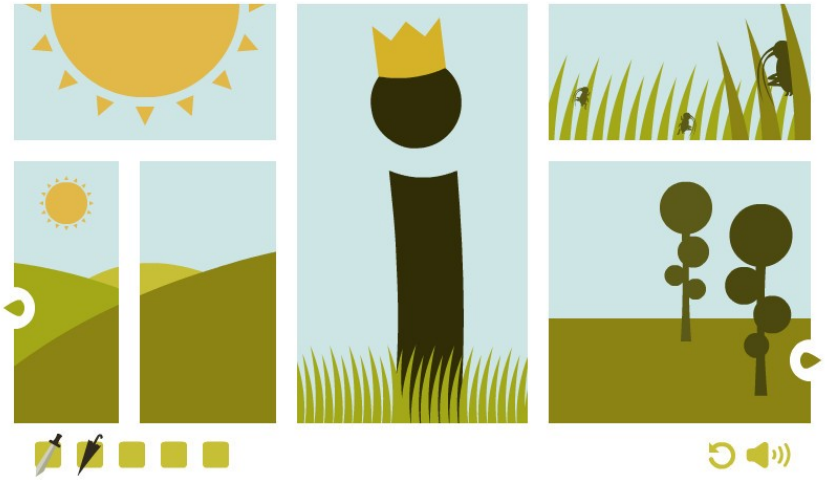
- Webpage with streaming video or audio
- Web-based storybook designed for a tablet or iPad
- Comic that requires user interaction
- Web-based video games

Examples of Social Media

- A corporate or government Twitter or Facebook account
- Event recordings uploaded to YouTube
- Photos shared on Flickr, Facebook, or Instagram
- Data about interactions between users

Complex Web Objects

- Dynamic & interactive
- Embedded or streaming components
- External Dependencies:
  - Databases
  - Other websites
  - Local data (e.g. location services)
- Device specific (ex. optimized for tablet or mobile)



The Empty Kingdom, Daniel Merlin Goodbrey

Breathe – a digital ghost story, Kate Pullinger

Challenges

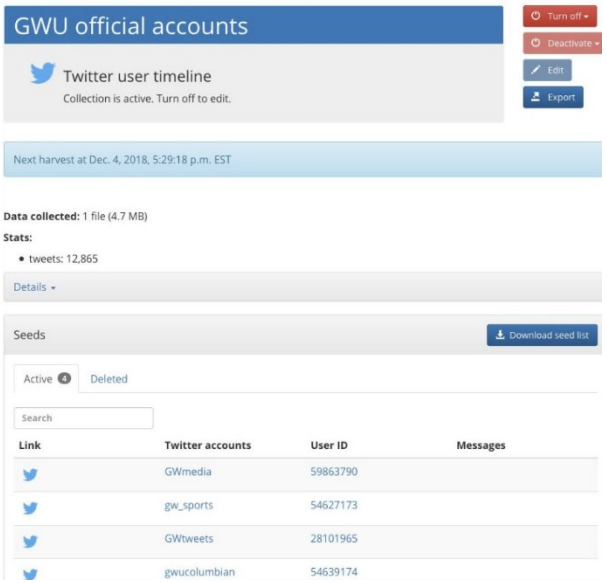
- Doesn’t fit into established workflows
- Difficult or impossible to capture with available tools
- Consists of components with a rapid obsolescence rate (esp. mobile apps)
- Complicated rights issues due to multiple contributors or developers
  - copyright
  - personal data
  - restricted information
- Lack of in-house expertise

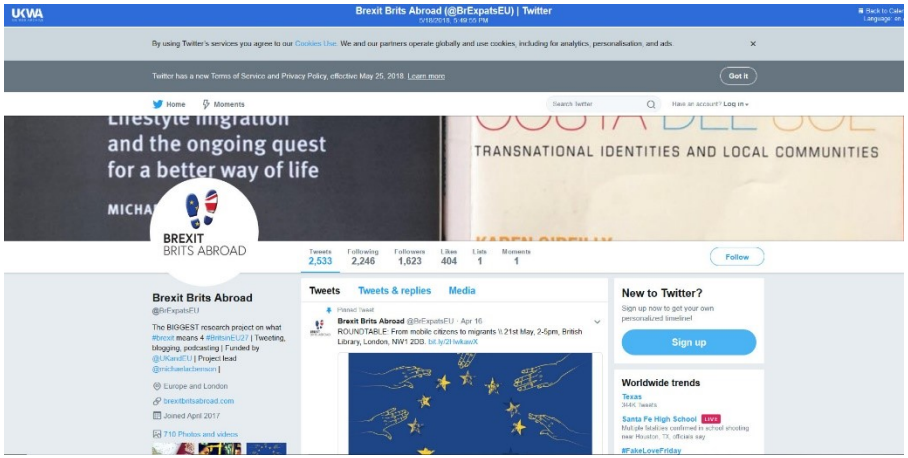
Opportunities

- Capturing a web resource that better reflects the original experience
- Preventing the loss of a content type vulnerable to rapid obsolescence
- Heading off an expensive process further down the line
- Trialling and providing feedback for new tool and service development
- Developing capacity for web archiving that can be used to capture an increasing number of types of objects and records

Social Media

- Individual posts or accounts (ex. screenshot, Heritrix type crawl, API tool)
- Datasets generated based on set parameters
- Images or embedded content
- Network information (ex. number of users, how users are connected, location of users, flow of information)





Social Media … as Data

DigitalPreservationCoalition

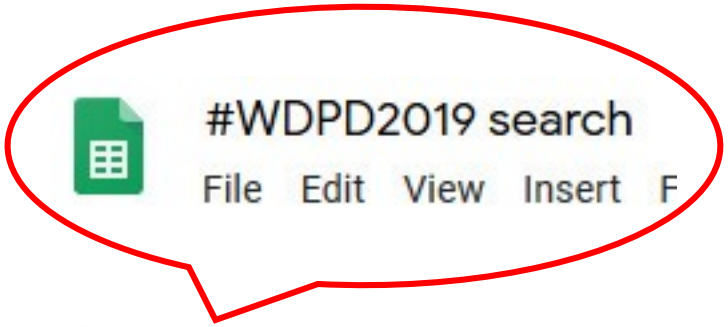
JSON

```
{
    "created_at": "Thu Apr 30 21:53:11 +0000 2015",
    "id": 593895901623496700,
    "id_str": "593895901623496704",
    "text": "This is a #test tweet @LoveforTestingT with an image. http://t.co/ZvgHovKZq4",
    "source": "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>",
    "truncated": false,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "in_reply_to_screen_name": null,
    "user": {
        "id": 2993982541,
        "id_str": "2993982541",
        "name": "Test Demo",
        "screen_name": "jondee_test",
        "location": "Denver, CO",
        "url": null,
        "description": "this is a test account.",
        "protected": false,
        "verified": false,
        "followers_count": 2,
        "friends_count": 43,
        "listed_count": 0,
        "favourites_count": 0,
```


Twitter Interface

Social Media … as Data

DigitalPreservationCoalition



#WDPD2019 search

File Edit View Insert Format Data Tools Add-ons Help TAGS Last edit was made seconds ago by sday t

Share

id_str

	A	B	C	D	E	J	K	L	M	N	O	P	Q	R
1	id_str	RT @SDArchives: It's	text	created_at	time	from_user_id_str	in_reply_to_status_ic	source	profile_image_url	user_followers_count	user_friends_count	user_location	status_url	entities_str
2	119251	___Sarka	RT @ObsoleteMediaUK: Happy World Digital Preservation Day! #WDPD2019. Explore hundreds of physical media formats and help identify those most at risk. https://t.co/wh8zBh3uIM https://t.co/CfHKdzsTgI	Thu Nov 07 18:52:2	07/11/2019 18:52:23	740450442		<a href="http://twitter	http://pbs.twimg.com	81	563	Czech Republic :'(	http://twitter.com/___	{"hashtags":[{"text":"V
3	119245	__anjanette	oh my god oh my god oh my god. Library of Queensland 🙌 #WDPD2019 🌈 https://t.co/Pyakj1eekt	Thu Nov 07 14:50:4	07/11/2019 14:50:48	549412718		<a href="http://twitter	http://pbs.twimg.com	1758	774	Minneapolis' largest	http://twitter.com/__a	{"hashtags":[{"text":"V
4	119243	__anjanette	RT @ObsoleteMediaUK: Happy World Digital Preservation Day! #WDPD2019. Explore hundreds of physical media formats and help identify those most at risk. https://t.co/wh8zBh3uIM https://t.co/CfHKdzsTgI	Thu Nov 07 13:51:2	07/11/2019 13:51:26	549412718		<a href="http://twitter	http://pbs.twimg.com	1758	774	Minneapolis' largest	http://twitter.com/__a	{"hashtags":[{"text":"V
5	119228	_allieq	RT @ArchivesSarah: Follow us on Twitter! Our team is going to be introducing ourselves over the course of #WDPD2019 (which for me is still tomorrow). https://t.co/vWFI7pExnj	Thu Nov 07 03:45:4	07/11/2019 03:45:49	16836678		<a href="http://twitter	http://pbs.twimg.com	211	385	Edmonton, AB	http://twitter.com/_all	{"hashtags":[{"text":"V
6	119256	_BAnQ	Au total, BAnQ met près de 18 millions de fichiers de documents patrimoniaux numériques à votre disposition en ligne! 🤩 #WDPD2019 https://t.co/7WZb3YfNDs	Thu Nov 07 21:51:5	07/11/2019 21:51:58	25117576		<a href="https://mobi	http://pbs.twimg.com	35175	1771		http://twitter.com/_BA	{"hashtags":[{"text":"V
7	119252	_Ben_Franklin1	RT @LibSkrat: Hey, it's #WorldDigitalPreservationDay! #WDPD2019 My wish for all of us is that this work be taken seriously and resourced appropriately. There's a lot of nonsense out there about this. It harms our collective ability to do it. https://t.co/uyI7kk5SYs	Thu Nov 07 19:47:1	07/11/2019 19:47:15	865677961919422464		<a href="http://twitter	http://pbs.twimg.com	4548	1894	Philadelphia, Londo	http://twitter.com/_Be	{"hashtags":[{"text":"V
8	119256	_BobbieWilliams	RT @heritage_adrian: Happy World Digital Preservation Day everyone. My #WDPD2019 blog is up … more on this topic to come from @RIBA @dpc_chat and friends in the future https://t.co/dmddI7EoQz	Thu Nov 07 22:18:3	07/11/2019 22:18:30	537353883		<a href="http://twitter	http://pbs.twimg.com	338	224	London	http://twitter.com/_Bo	{"hashtags":[{"text":"V
9	119249	_Brett_Johnston	RT @CounteringCrime: This World Digital Preservation Day #WDPD2019 we want to recognize the work of ACCO members who are fighting to capture evidence of trafficking in antiquities and human remains on social media - vital evidence that must be preserved. @AnthroPaulloy @DamienHuffer @electricarchaeo https://t.co/0IA5Q80XMA	Thu Nov 07 17:39:4	07/11/2019 17:39:46	1036545223155798017		<a href="http://twitter	http://pbs.twimg.com	198	2272	Albany, Western Aust	http://twitter.com/_Br	{"hashtags":[{"text":"V
10	119242	_BRICMICS	RT @ARAArchiTech: Digital archives need to be managed actively. Show them some love! #WDPD2019 #LoveArchives https://t.co/bpR2bOyKgx	Thu Nov 07 13:13:3	07/11/2019 13:13:32	864056181194723330		<a href="http://twitter	http://pbs.twimg.com	424	786	Around GB and Irela	http://twitter.com/_BF	{"hashtags":[{"text":"V
11	119242	_BRICMICS	RT @Postcard_Jim: Important https://t.co/mGuhWXwEtV by @MickyLindlar Many points for us all to engage with: Thanks @dpc_chat for sharing https://t.co/mGuhWXwEtV #WDPD2019 https://t.co/d3whAbyAPc	Thu Nov 07 13:13:0	07/11/2019 13:13:08	864056181194723330		<a href="http://twitter	http://pbs.twimg.com	424	786	Around GB and Irela	http://twitter.com/_BF	{"hashtags":[],"symbo
12	119238	_BRICMICS	RT @Postcard_Jim: @_BRICMICS @bodleianlibs @BDLSS @dpc_chat I look forward to represent @ARAUK_IE @ARAArchiTech at today's @_BRICMICS meeting. The #Archives & #RecordsManagement sector focuses on #digipres today #WDPD2019 & every day. We are learning across sectors, making @_BRICMICS an important partner, along with @dpc_chat @ADS_Update	Thu Nov 07 10:27:4	07/11/2019 10:27:48	864056181194723330		<a href="http://twitter	http://pbs.twimg.com	424	786	Around GB and Irela	http://twitter.com/_BF	{"hashtags":[],"symbo
13			On World Digital Preservation Day #WDPD2019 it is apt that @BRICMICS meets to consider #digipres & all other aspects of #map curation.											

Challenges

- Platform Terms & Conditions prevent sharing and re-use
- T&Cs change frequently, which may result in the suspension of other services or restricted access
- Each platform has different functionality and export options
- New platforms & updates to interface and functionality
- For scholarship: difficult to create reproducible research
- For archiving: difficult to select, describe, and build collections due to interactive and fluid conversations
- Rapid growth can make storage planning difficult
- Rights: personal data, copyright, right to be forgotten
- Ethics (more later): user awareness, consent for re-use, representation
- Expensive, if not captured quickly after creation

Opportunities

- Important source of cultural heritage
- Compliance with business, legal, and other regulations
- Source of 'naturally occurring' data for analysis
- Improve public services and support better governance
- Online, open-source investigation (such as Bellingcat, WITNESS)
- Trialling and providing feedback for new tool and service development
  - Social Feed Manager
  - Documenting the Now (twarc, Hydrator)
  - COSMOS
  - Wasim Ahmed Blog (LSE Blog, Twitter as data)