

Data Sharing

digital preservation and data sharing

dpconline.org

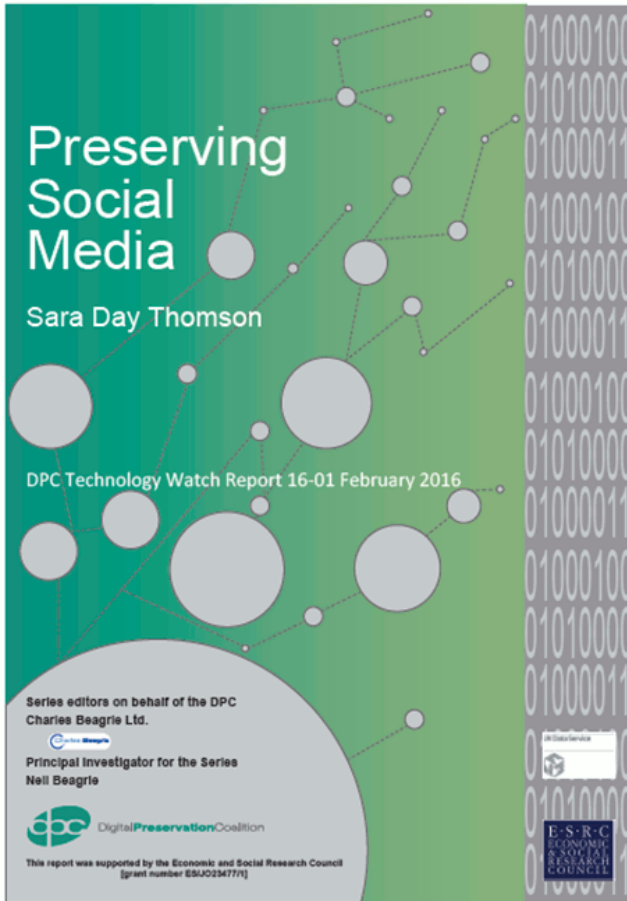


Digital **Preservation** Coalition

sara day thomson

tw: @sdaythomson

Technology Watch Report



Web Archiving & Preservation Task Force

A screenshot of the Digital Preservation Coalition website. The top navigation bar is green with white text for 'Digital Preservation Coalition' and 'Login'. Below it is a menu with 'ABOUT', 'NEWS', 'OUR WORK', 'KNOWLEDGE BASE', 'EVENTS', and 'BLOG'. A search bar with 'Google Custom Search' and a magnifying glass icon is visible. The main content area shows a breadcrumb trail: 'Home > Our Work > Working Groups and Task Forces > DPC > Our work > Web Archiving & Preservation Task Force'. The title 'Web Archiving & Preservation Task Force' is prominently displayed. A green callout box contains the text: 'New Terms of Reference for the reconvening of the Web Archiving & Preservation Task Force are now open for comment from members. See the bottom of this page for more information.' Below this, the section 'What is the Web Archiving & Preservation Task Force?' is followed by three paragraphs of text describing the task force's purpose and activities. A sidebar on the left contains the text 'ups ces', 'Bedern Group', 'Dedicated Support for Members', and 'Collaboration'.

<http://dx.doi.org/10.7202/twr16-01>



Digital Preservation Coalition

Some Key Challenges

- sharing rich, useable data for reproducible research
- scale & volume
- ethics & user awareness
- lack of standards or best practice
- vulnerability of content

Digital Preservation Approaches

- defining a Designated Community
- 're'-selection
- risk assessment & policy
- bit preservation & data security
- metadata & documentation
- advocacy

Sharing Twitter Datasets in a Data Repository



UK Data Service
Discover



Discover

Variable and question
bank

QualiBank

You are not logged in | [Login to Discover](#) | [Site Search](#) [FAQ](#) [Help](#) [Contact](#)

[About us](#) [Get data](#) [Use data](#) [Manage data](#) [Deposit data](#) [News and events](#)

Discover > Catalogue

Catalogue

SHARE

UK Data Service data catalogue record for:

After Woolwich twitter corpus

[Syntax](#)

[Download](#) | [DDI XML](#)

TITLE DETAILS

SN: 852078
Title: After Woolwich twitter corpus
Persistent identifier: [10.5255/UKDA-SN-852078](#)
Depositor: Martin Innes, Cardiff University
Principal investigator(s): Martin Innes, Cardiff University
Sponsor(s): Economic and Social Research Council
Grant number: ES/L008181/1
Other acknowledgements: Colin Roberts

SUBJECT CATEGORIES

Society and culture

ABSTRACT

Abstract copyright data collection owner.

After Woolwich Twitter Corpus represents social media data collected from Twitter to analyse social reactions to the murder of Drummer Lee Rigby in Woolwich on 22 May 2013. The dataset covers a roughly 12 month span from March 29th 2013 onwards. The data enabled the tracking of the evolution of public perceptions and sentiments in real-time as key events occur. The dataset comprises of a csv format file with Tweet IDs and Date for all collected tweets. All other relevant tweet data have been omitted to comply with the Twitter API Terms of use. In order to recreate the data, utilise the Twitter API to request each tweet by ID.

Project description:

The research will analyse social reactions to the murder of Drummer Lee Rigby in Woolwich on 22 May 2013 using social media data collected from Twitter, blogs and other sources. Such data uniquely enable the tracking of the evolution of public perceptions and sentiments in real-time as key events occur. They enable us to track the arc of social reactions from the crime scene through to the conclusion of the

Sharing Twitter Datasets in a Data Repository

The screenshot shows the Datorium website interface. At the top, there is a navigation bar with the 'gesis' logo and a language selector for German. Below this is the 'datorium' logo. A search bar and a 'View Item' button are visible. The main content area is divided into a left sidebar with navigation options (Search, Add data, About datorium) and a main panel. The main panel features the Leibniz-Gemeinschaft logo and a 'General Description' section. The description includes fields for Title, URI, Primary Researcher, Publication Year, Availability, Contributor, Subject Area, Topic Classification, Abstract, Geographical Area, Universe, Selection Method, Data Collection Mode, Survey Period, Licenses, and Notes. The dataset is titled 'Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness' and was published in 2016. It is available with restricted access and is contributed by Wolfgang Zenk-Möltgen at GESIS - Leibniz Institute for the Social Sciences. The abstract describes a collection of geotagged tweets from the US, including aggregated hashtag counts and shapefiles for geotagging.

This screenshot shows the login page of the Datorium website. The top navigation bar is identical to the previous screenshot. Below the navigation bar, there is a search bar and a 'Sign in' button. The main content area contains a message: 'The file you are attempting to access is a restricted file and requires credentials to view. Please login below to access the file.' Below this message is a login form with two input fields: 'E-Mail Address:' and 'Password:'. A 'Forgot your password?' link is located next to the password field. A 'Sign in' button is positioned below the input fields. At the bottom of the page, there is a registration prompt: 'Please register to document and publish your research data in datorium. Click here to register.'

Sharing Tweet IDs



Digital Preservation Coalition



```
1 {
2   "created_at": "Thu Apr 30 21:53:11 +0000 2015",
3   "id": 593895901623496700,
4   "id_str": "593895901623496704",
5   "text": "This is a #test tweet @LoveforTestingT with an image. http://t.co/ZvgHovKZq4",
6   "source": "<a href='\"http://twitter.com/\"' rel='\"nofollow/\">Twitter Web Client</a>",
7   "truncated": false,
8   "in_reply_to_status_id": null,
9   "in_reply_to_status_id_str": null,
10  "in_reply_to_user_id": null,
11  "in_reply_to_user_id_str": null,
12  "in_reply_to_screen_name": null,
13  "user": {
14    "id": 2993982541,
15    "id_str": "2993982541",
16    "name": "Test Demo",
17    "screen_name": "jondee_test",
18    "location": "Denver, CO",
19    "url": null,
20    "description": "this is a test account.",
21    "protected": false,
22    "verified": false,
23    "followers_count": 2,
24    "friends_count": 43,
25    "listed_count": 0,
26    "favourites_count": 0,
```

```
1 {
2   "created_at": "Thu Apr 30 21:53:11 +0000 2015",
3   "id": 593895901623496700,
4   "id_str": "593895901623496704",
5   "text": "This is a #test tweet @LoveforTestingT with
```



social media is vulnerable to loss

LinkedIn is not a storage service. You agree that we have no obligation to store, maintain or provide you a copy of any content or information that you or others provide.

LinkedIn User Agreement

The Twitter Entities shall not be liable for ... any data loss.

Twitter Terms of Service

Bit Preservation

Even if immediate sharing is not possible....

...preserving the bits, metadata, and documentation is still critical to making social media content accessible in future.



www.digitalbevaring.dk

Spectrum of Sharing Policies?

It doesn't have to be all or nothing...



Risk Assessment & Policy Development

Develop sharing
strategies based on the
dataset or classification
of datasets?



Public Awareness

... about linked administrative data

‘At the beginning, **low awareness** of the uses of social research drove **scepticism** about its value.’

‘Later in the dialogues, when participants had **learned more** about the aims and methods of social research, they tended to be **more positive** about its value.’

#thanks!

