preserving social media
applying principles of digital preservation
to social media archiving

dpconline.org

digital preservation coalition

sara day thomson
tw: @sdaythomson
Some Key Challenges

- capturing 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
- metadata & documentation
- advocacy
- New, updated version released summer 2016
- Written and reviewed by over 40 specialists in digital preservation
- Available on dpconline.org with additional embedded resources, such as short training videos
Who Is the Future?

Designated Community (OAIS Reference Model)

• define the main user group
• anticipate future user needs
• extract specialist knowledge to create metadata
• develop user requirements
• useful in lieu of established professional standards and best practice
UK Data Service

Catalogue

UK Data Service data catalogue record for:

After Woolwich twitter corpus

SN: 852978
Title: After Woolwich twitter corpus
Persistent identifier: 10.5255/UKDA-08-852978
Depositor: Martin Innes, Cardiff University
Principal investigator(s): Martin Innes, Cardiff University
Sponsor(s): Economic and Social Research Council
Grant number: ES/L009189/1
Other acknowledgements: Gavin 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 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 re-use the data, users must request access 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
Future-proofing

‘re’-selection

- limited data storage
- restrictions imposed by platform APIs
- personal and sensitive information
- implications of longer time scale: 10 years? 20? 50?

Risk Assessment

Policy Development

Mitigate Problems

Reassure Users

Embed into business as usual
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
Future-telling: Bit Preservation

Even if immediate access is not possible....

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

*Dialogue on Data*
Ipsos MORI Social Research Institute
Advocacy Conundrum

Robust, varied research to demonstrate value of social media as research data

Better, more open platform policies for sharing social media research data

Research has been possible *despite* restrictive platform policies
Alignment with National Narratives

Social movements reflected in social media

Transparent government & open data initiatives
#thanks!