PRESERVING SOCIAL MEDIA: A BRIEFING DAY

27 October 2015
Wellcome Collection
London
Machine-readable Data

Application Programming Interfaces (APIs)

Terms and Conditions

Our digital memory accessible tomorrow

Sara Day Thomson
sara.thomson@dpconline.org

www.dpconline.org
Social Media Data

Institutional Accounts

Datasets

Data Archive
Our digital memory accessible tomorrow

Sara Day Thomson
sara.thomson@dpconline.org

www.dpconline.org

Strategies

• Platform APIs
• Data Resellers
• Third-party Services
• Platform Self-Archiving Services
• Research Dataset Identifiers
• Research Documentation & Metadata
Platform APIs
‘Backdoor into platform’

Common Characteristics
• Access Tokens
• Rate Limits
• Samples Sizes
• JSON data
• XML data
• Metadata (geolocation, user ID, gender, etc.)

‘Walled garden’ of social media platforms
Our digital memory accessible tomorrow

Sara Day Thomson
sara.thomson@dpconline.org

www.dpconline.org
Data Resellers

- Commercial service
- Provide data on behalf of platforms
- May provide other services, e.g. analytics
- Cater to corporations
- Built for consumer analysis

Reseller: Gnip
Service: PowerTrack*
Delivers: 100% of tweets
Pricing: $US 2000 per month plus $US 0.10 per 1000 tweets delivered

Twitter’s official reseller: Gnip
Facebook Topic Data reseller: Datasift

Third-party Services

Common Characteristics:

• Commercial service or partnership
• Deliver archiving services
• Customised Solutions
• Solution for highly specialised technical skills

e.g. ArchiveSocial Standard Pricing Plan

<table>
<thead>
<tr>
<th>Plan</th>
<th>Service Description</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>3,000 new records/month</td>
<td>$399/month</td>
</tr>
<tr>
<td>Unlimited</td>
<td>Social accounts</td>
<td></td>
</tr>
<tr>
<td>Unlimited</td>
<td>Data storage</td>
<td></td>
</tr>
<tr>
<td>Econom</td>
<td>Compliance at any time</td>
<td>$199/month</td>
</tr>
<tr>
<td></td>
<td>Social accounts</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data storage</td>
<td></td>
</tr>
</tbody>
</table>

Start Archiving

Sara Day Thomson
sara.thomson@dpconline.org

www.dpconline.org
Platform Self-Archiving Services

- Service for account owner
- Structured and unstructured data
- Some provide JSON

Google

Your account, your data.
Download a copy.

Create an archive with your data from Google products.

Manage archives

Facebook

Download a copy of your Facebook data.

Manage the contacts imported address book.

Twitter

Download your entire Tweet history.

Review the apps that you have given your Twitter account.

Review the account.

Sara Day Thomson
sara.thomson@dpconline.org

Our digital memory accessible tomorrow
Challenges

• Platform Terms & Conditions
• Brokering Agreements with Platforms
• Ethics of User Privacy and Awareness
• Selection
• Long-term Preservation
Platform Terms and Conditions

Common Policies
• Terms of Service
• User Agreement
• Privacy Policy
• Terms of Use
• Developer Agreement or Developer Policy

Developer Agreement or Policy: Common Characteristics
• Controls use of API
• Forbids sharing of data (incl. cloud storage)
• Forbids preservation of deleted data
Brokering Agreements with Platforms: Twitter

Transfer of Data: Twitter Research Access at the Library of Congress

• Agreement 2010
• Entire archive of tweets from 2006 – 2010
• All on-going tweets
• Authenticated research access only
• Not yet available
Ethics of User Privacy and Awareness

• Accidental disclosure
  – Anonymization
  – De-identification

• User Awareness
  – Commercial Use (profit)
  – Research Use (public good)

• Unequal access
  – Technical Skills
  – Resources

‘... combined, these [conditions] provide a more powerful picture, and thus a more manipulable population.’

– Ralph Schroeder ‘Big Data and the brave new world of social media research’ in Big Data & Society (2014)
Selection

• The ‘Conversation’
  – across users
  – across social media services

• Embedded Media
  – images
  – audio
  – moving images
  – URLs

• Uniformity of language, tags, or keywords?
• Spelling errors
• Obscenity or libel
• Shortened links (e.g. bitly, tinyurl)
• Relevance
Long-term Preservation

- Vulnerable to loss
- No legal or regulatory requirement for platform to preserve
- Changes in platform policy and ownership
- Historical data less commercially valuable
- Expensive to curate and store

- Image hosting service
- Shut down by Twitter in Sept 2014
- Data taken over by Twitter in Oct 2014
- Archived by Archive Team
Conclusions & Recommendations

• Integrating Web Archives
• Collaboration
• Tools & Services
• Further Reading
Integrate Social Media and Web Archives

• Temporal consistency
• Wider context for social media
• Re-direct URLs to archived webpages
• e.g. UK Government Twitter archive
• e.g. Arcomem combined crawler with integrated social media API tool
Collaboration

Between Collecting Institutions
- Share resources
- Share access
- Share methods and knowledge

Collecting Institutions with Researchers and End Users
- Understand end user needs
- Outreach
- Promote collections

Sara Day Thomson
sara.thomson@dpconline.org
www.dpconline.org
Tools & Resources

**ARCOMEM** (open source)
About: [http://www.arcomem.eu/open-source](http://www.arcomem.eu/open-source)

**Social Feed Manager** (open source)

**SODATO** (proprietary)
About: [http://link.springer.com/chapter/10.1007/978-3-319-06701-8_27](http://link.springer.com/chapter/10.1007/978-3-319-06701-8_27)
To Create an Account: [http://cssl.cbs.dk/software/sodato](http://cssl.cbs.dk/software/sodato)

**COSMOS** (open source)
To download: [http://www.cs.cf.ac.uk/cosmos/download-cosmos](http://www.cs.cf.ac.uk/cosmos/download-cosmos)

**NCSU Tools** (open source)
Lentil: [https://www.lib.ncsu.edu/social-media-archives-toolkit/collecting/lentil-user-guide](https://www.lib.ncsu.edu/social-media-archives-toolkit/collecting/lentil-user-guide)

Sara Day Thomson
sara.thomson@dpconline.org
www.dpconline.org
Tools & Services

**NCSU Tools** (open source)
Social Media Combine:
https://www.lib.ncsu.edu/social-media-archives-toolkit/collecting/combine-user-documentation

**NCSU Social Media Archives Toolkit** (resource)
https://www.lib.ncsu.edu/social-media-archives-toolkit

**TWARC** (open source tool)
To Download:
https://github.com/edsu/twarc

Overview of Research Tools for Twitter and other platforms
'Using Twitter as a data source: An overview of current social media research tools' by Wasim Ahmed
http://blogs.lse.ac.uk/impactofsocialscience/2015/07/10/social-media-research-tools-overview
Further Reading


#thanks!

Our digital memory accessible tomorrow

Sara Day Thomson
sara.thomson@dpconline.org

www.dpconline.org