Machine-readable Data

Application Programming Interfaces (APIs)

Terms and Conditions
Social Media Data

Institutional Accounts

Datasets

Data Archive

Our digital memory accessible tomorrow

Sara Day Thomson
sara.thomson@dpconline.org

www.dpconline.org
• Platform APIs
• Data Resellers
• Third-party Services
• Platform Self-Archiving Services
• Research Dataset Identifiers
Platform APIs
‘Backdoor into platform’

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

‘Walled garden’ of social media platforms
APIs

Rendered Tweet from Gnip 'Data Format'

JSON of tweet (lines 1-26 of 201) from Gnip 'Data Format'

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

<table>
<thead>
<tr>
<th>Standard Pricing Plan</th>
<th>Econom Pricing Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>$399/month</td>
<td>$199/month</td>
</tr>
<tr>
<td>3,000 new records/month</td>
<td>1,000 new records/month</td>
</tr>
<tr>
<td>Unlimited social accounts</td>
<td>10 social accounts</td>
</tr>
<tr>
<td>Unlimited data storage</td>
<td>Unlimited data storage</td>
</tr>
<tr>
<td>All archiving platform features</td>
<td>All archiving platform features</td>
</tr>
</tbody>
</table>

e.g. ArchiveSocial Standard

Pricing Plan

Start Archiving

Start Archiving

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

Facebook

- Download a copy of your Facebook data.

Twitter

- Manage the contacts imported to your address book.
- Download your entire Tweet history.
- Review the apps that you have given access to your Twitter account.
- Review the access you have given to third-party apps.

Google

- Your account, your data. Download a copy.
- Create an archive with your data from Google products.
- Manage archives

Sara Day Thomson
sara.thomson@dpconline.org
www.dpconline.org
Research Dataset Identifiers

- IDs for entries in a curated dataset
- e.g. in database, in spreadsheet
- Researchers can request dataset with same identifiers
- Not 100% accurate

Tweet (JSON)

```json
{"created_at": "Thu Apr 30 21:53:11 +0000 2015",
"id": 503895901623496700,
"id_str": "503895901623496700",
"text": "This is a test tweet @LoveforTestingI with"}
```
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 by Twitter in Oct 2014
  - Archived by Archive Team
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
Preserving Social Media

DPC Technology Watch Report

Public Release: Monday 15 February

dpconline.org/publications/technology-watch-reports
Look out for...

Preserving Transactional Data: a DPC Briefing Day
17 March 2016
London
A Joint Workshop with the DPC and the UK Data Service

For more info and to register:
bit.ly/1Qx0ibn

Coming soon...

Preserving Transactional Data
DPC Technology Watch Report
Summer 2016

Sara Day Thomson
sara.thomson@dpconline.org
www.dpconline.org
#thanks!