

Gathering Social Media Data

Leon Derczynski

Functional utterances

Vowels



Velar closure: consonants

Speech



New modality: writing


Digital text

E-mail

Social media



**Increased
machine-
readable
information**

A large, solid black arrow pointing downwards, indicating a progression or flow from the top of the text block to the bottom.

Gartner "3V" definition:

1. Volume

2. Velocity

3. Variety

High volume & velocity of messages:

Twitter has ~20 000 000 users per month

They write ~500 000 000 messages per day

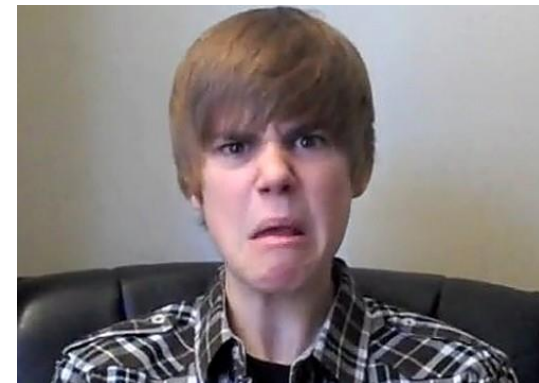
Massive variety:

Stock markets;

Earthquakes;

Social arrangements;

... Bieber



Twitter, LinkedIn, Facebook

Twitter has varied uptake per country:

- Low in China (often censored, local competitor – Weibo)
- Low in Denmark, Germany (Facebook is preferred)
- Medium in UK, though often complementary to Facebook
- High in USA

Networks have common themes:

- Individuals as nodes in a common graph
- Relations between people
- Sharing and privacy restrictions
- No curation of content
- Multimedia posting and re-posting

Other features: topics, closed groups, moderation, liking, media, groups, person discovery ..

Disclaimer: I Am Not A Legal Professional; caveat emptor!

1. Twitter

Opened in 2006 as a short message blogging service

Allows 'subscription' to interesting accounts

Anyone can post, most messages are public

Messages are <140 characters

Posts can come from PC, mobile, SMS, iPad etc

Specialised markup: #hashtags and @mentions

Has grown extremely popular

- 100 million active users; over 230 million tweets a day
<http://www.guardian.co.uk/technology/pda/2011/sep/08/twitter-active-users>

Public relations

Barack Obama

We just made history. All of this happened because you gave your time, talent and passion. All of this happened because of you. Thanks

Celebrity worship

Kidrauhl ♥

“One day you will forget me. You have a husband and be a mother. But I will never forget you, My Beliebers.” - Justin Bieber ε

Broadcasting & Activism

Ars Technica

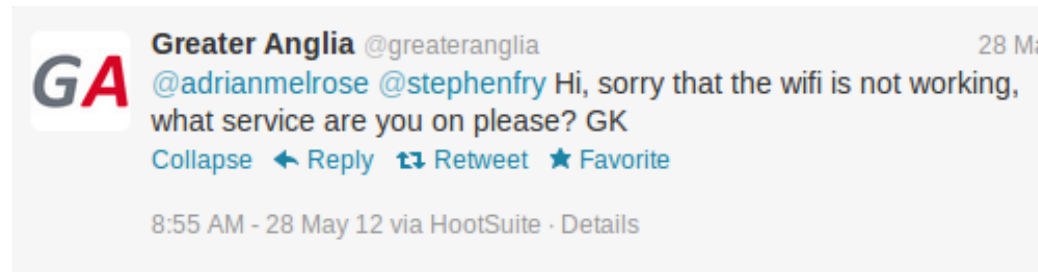
SOPA opponents unveil "Digital Bill of Rights" <http://arstechnica.com/tech-policy/20...> by [@nathanmattise](#)

Social uses

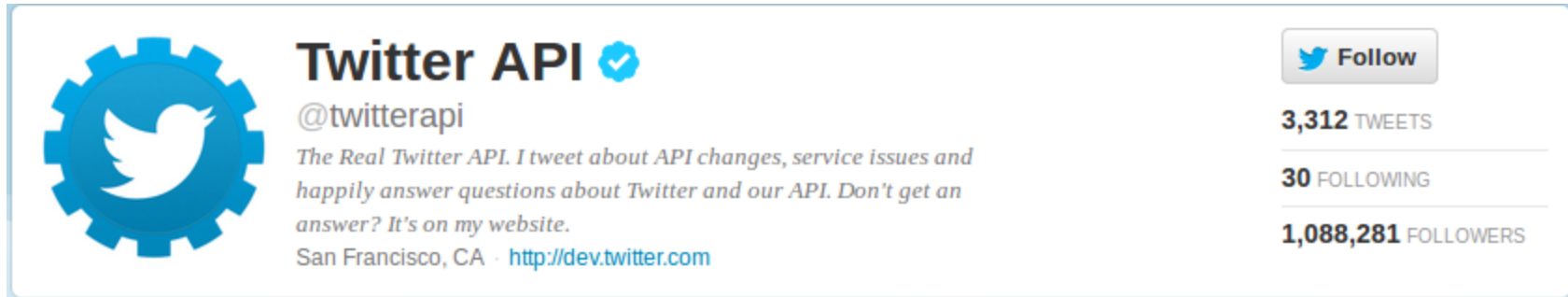
「ジャム」 **Jam Gregory**

[@RyanBibby](#): lots of people have been talking about it - need to make sure I watch it! Love [@ninaconti](#), got a signed DVD at [#EdFringe](#) :D


Conversations/Customer Support




Twitter User Profiles



A screenshot of a Twitter profile for 'Twitter API'. The profile features a blue gear icon with a white bird inside. The name 'Twitter API' is displayed in bold black text with a blue verified checkmark. The handle '@twitterapi' is shown below the name. The bio reads: 'The Real Twitter API. I tweet about API changes, service issues and happily answer questions about Twitter and our API. Don't get an answer? It's on my website.' The location is 'San Francisco, CA' and the website is 'http://dev.twitter.com'. On the right side, there is a 'Follow' button, '3,312 TWEETS', '30 FOLLOWING', and '1,088,281 FOLLOWERS'.



Twitter API 

@twitterapi

The Real Twitter API. I tweet about API changes, service issues and happily answer questions about Twitter and our API. Don't get an answer? It's on my website.

San Francisco, CA · <http://dev.twitter.com>

 Follow

3,312 TWEETS

30 FOLLOWING

1,088,281 FOLLOWERS

- Picture
- Name
- Location
- Website
- Bio (160 characters)

What is Twitter? (2)

Interest-graph social media

Following/follower relationship is typically not bi-directional

- 77.6% of user connections are not reciprocated (Kwak 2010)

A large graph in which mutual follower/following relationships comprise the edges

Twitterers can 'retweet' one another, so information propagates via the graph quickly

- RTs typically contain links to interesting content

Users can be organised in lists, which introduces groupings

Example Tweet metadata in JSON

```
{  "contributors":null,
  "text":"Automotive RDFa (a horribly researched SEO article on RDFa/Microformats):
http://ow.ly/5JSoS #somanerrorsitsfunny",
  "geo":null,
  "retweeted":false,
  "in_reply_to_screen_name":null,
  "truncated":false,
  "entities":{"urls":[{"expanded_url":null,"indices":[74,92],"url":"http://ow.ly/5JSoS"}],
    "hashtags":[{"text":"somanerrorsitsfunny","indices":[93,114]}],
    "user_mentions":[]},
  "in_reply_to_status_id_str":null,
  "id":94029193863639040,
  "source":"<a href=\"http://www.hootsuite.com\" rel=\"nofollow\">HootSuite</a>",
  "in_reply_to_user_id_str":null,
  "favorited":false,
  "in_reply_to_status_id":null,
  "retweet_count":0,
  "created_at":"Thu Jul 21 13:01:21 +0000 2011",
```

Example Tweet metadata in JSON (2)

```
"in_reply_to_user_id":null,  
"id_str":"94029193863639040",  
"place":{"id":"c799e2d3a79f810e",
```

```
  "bounding_box":{"type":"Polygon",  
    "coordinates":[[[6.6266397,35.4928765],  
                    [18.5203619,35.4928765],  
                    [18.5203619,47.0924397],  
                    [6.6266397,47.0924397]]]}
```

Type of place, e.g.
"city"

```
  "place_type":"country",  
  "name":"Italia",  
  "attributes":{,
```

```
    "country_code":"IT",  
    "url":"http://.../1/geo/id/c799e2d3a79f810e.json",  
    "full_name":"Italia",  
    "country":"Italia"
```

Country containing
the place of origin

```
  },
```

Example Tweet metadata in JSON (3)

```
"user":{"location":"Blacksburg, VA",  
  ...,  
  "statuses_count":2404,  
  "lang":"en",  
  "id":20446311,  
  ...,  
  "description":"Text from the user profile (max 160 chars)", ...,  
  "name":"User Name", ...,  
  "created_at":"Mon Feb 09 16:33:16 +0000 2009",  
  "followers_count":1239,  
  "geo_enabled":false, ...,  
  "url":"The author's URL (optional)",  
  "utc_offset":-21600,  
  "time_zone":"Central Time (US & Canada)", ..,  
  "friends_count":160, ...,  
  "screen_name":"twitter-user-name", ...,  
  "listed_count":189, ...  
}, ...
```

Embedded user information, can get out-of-sync, if the user changes it later

How to get tweets?

The REST API allows access timelines, tweeting, following, etc.

- REST/JSON based
- Requires registration, and developer / app keys
- Contains access to what was previously the Search API
- Core entities: tweets, users, entities, places
- Heavily rate-limited

The Streaming API streams tweets in real time

- Various strengths available, from 1% to 100% sample (~\$1M p.a.)
- May be filtered by language, location, user view, hashtag, search term

See <https://dev.twitter.com/docs>

2. LinkedIn

Opened in 2003 as a professional networking portal

Focus is on a CV-like profile

Allows connection to your contacts

Allows subscription and posting to forum-like groups

Event-focused rather than message focused

Posts can come from PC, mobile, SMS, iPad etc

260 million registered users



2. LinkedIn



Feed-based output; information on new relations

Focus on building networks: contact suggestions, contact history, people interested in you

A screenshot of a LinkedIn profile page. The top navigation bar includes the LinkedIn logo, a search bar, and various icons for notifications, settings, and profile management. The main content area is divided into several sections: a news feed with articles like "Europe's Tech Hubs: Let's Startup Somewhere Else" and "Bored at Work? Here's What To Do!"; a "Jobs you may be interested in" section featuring roles like "Lead Lawyer" at Siemens and "UK Financial Controller" at Insight UK; a section for new connections showing a recent connection with Rajat Malhotra; and a right-hand sidebar with sections for "You Recently Visited" (AcEmpire.co.uk), "Who's Viewed Your Profile" (9 people in the past 30 days), and "Who's Viewed Your Updates" (33 views). The bottom of the page shows a snippet of a post by Jose Maria Gomez Hidalgo about a conference.

2. LinkedIn



Data is available via API

No storage of data permitted: **“No LinkedIn data can be stored”**

- Except member ID
- User data can be stored only given explicit permission from that user
- Rationale: “LinkedIn users own their data. They need to have control over it. They might want to change it, change the visibility rules, or even delete it.”

Cross-referencing data is not permitted (via e.g. other networks)

- Creates problems for storing and communicating graph information
- Analysis must be live, but processing is not instantaneous – so no snapshots

API access is query driven: entities, items in streams

- Entities: people, stream, groups, mail, companies, job positions
- API is rate limited at application, user and developer level
- Limits quite high: e.g. 100k user profile queries per application per day

3. Facebook

Opened in 2004 as a university student directory

Communication is based on personal pages, to which messages are posted

Allows connection to your contacts

Allows subscription and posting to forum-like groups

Message focused, with comments and voting systems (unidirectional)

Posts can come from PC, mobile, SMS, iPad etc

1 200 million registered users

Extensive privacy options for users

3. Facebook



News items, with comments and likes

Access network connections, events and private messaging

A screenshot of the Facebook news feed interface. The top navigation bar is blue with the Facebook logo, a search bar, and links for Home, Profile, and Account. The left sidebar contains a user profile for Mark Robinson and a list of navigation options: News Feed, Messages (90), Events (3), Photos, Friends, Applications, Games, Groups (1), Marketplace, Friend Hug, and More. The main content area is titled "News Feed" and shows a "What's on your mind?" prompt. Below this are several news items: a comment by Fiona Baikie, Jackie James, and Nick Procter on Nick Procter's status; a post by Amy Simmonds about swimming; a post by Joe Mordey; a post by Danielle Eaton; a post by Clare Shewring; and a comment by Sarah McDermott on Maddy McDermott's photo. The right sidebar contains sections for Requests (3 event invitations, 1 group invitation, 1 Page suggestion), Suggestions (Myriam Stobart, Carol Power Gilhawley), Sponsored (Best value offers online, Super Free Sky+HD box), and Events (PRAYER VIGIL FOR JEREMY, Simone Mellor's birthday, Steve Blacker's birthday, Peter Francis's birthday, Prabal Ray's birthday). The bottom of the page has a "Connect with friends" button and a "Chat (Offline)" button.

3. Facebook



Main APIs for facebook data access: Graph, Public Feed (also others for web hosting, ads)

REST and JSON-based

- GET graph.facebook.com /{node-id}
- GET graph.facebook.com /{node-id}/{edge-name}
- Also POST, DELETE

Example response; fields vary depending on entity type

```
{
  "id": "4",
  "link": "https://www.facebook.com/zuck",
  "gender": "male",
  "username": "zuck",
  "picture": {
    "data": {
      "url": "https://fbcdn-profile-a.akamaihd.net/hprofile-ak-prn2/202896_4_1782288297_q.jpg",
      "is_silhouette": false
    }
  }
}
```

Many different entity types (messages, links, photos, events, posts, payments, videos..)

Optional FQL access – Facebook Query Language

One extra API: Keyword Insights

- Access to demographic information given keywords, locations

Storing social media data

What would help us do our science?

- NLP and network analysis tools often data-driven, preferring “as much data as possible”
- Not only do the messages change over time – meta-information also
- A minimum: something that helps others reproduce your work
- Abstract annotations over the raw data != the raw data

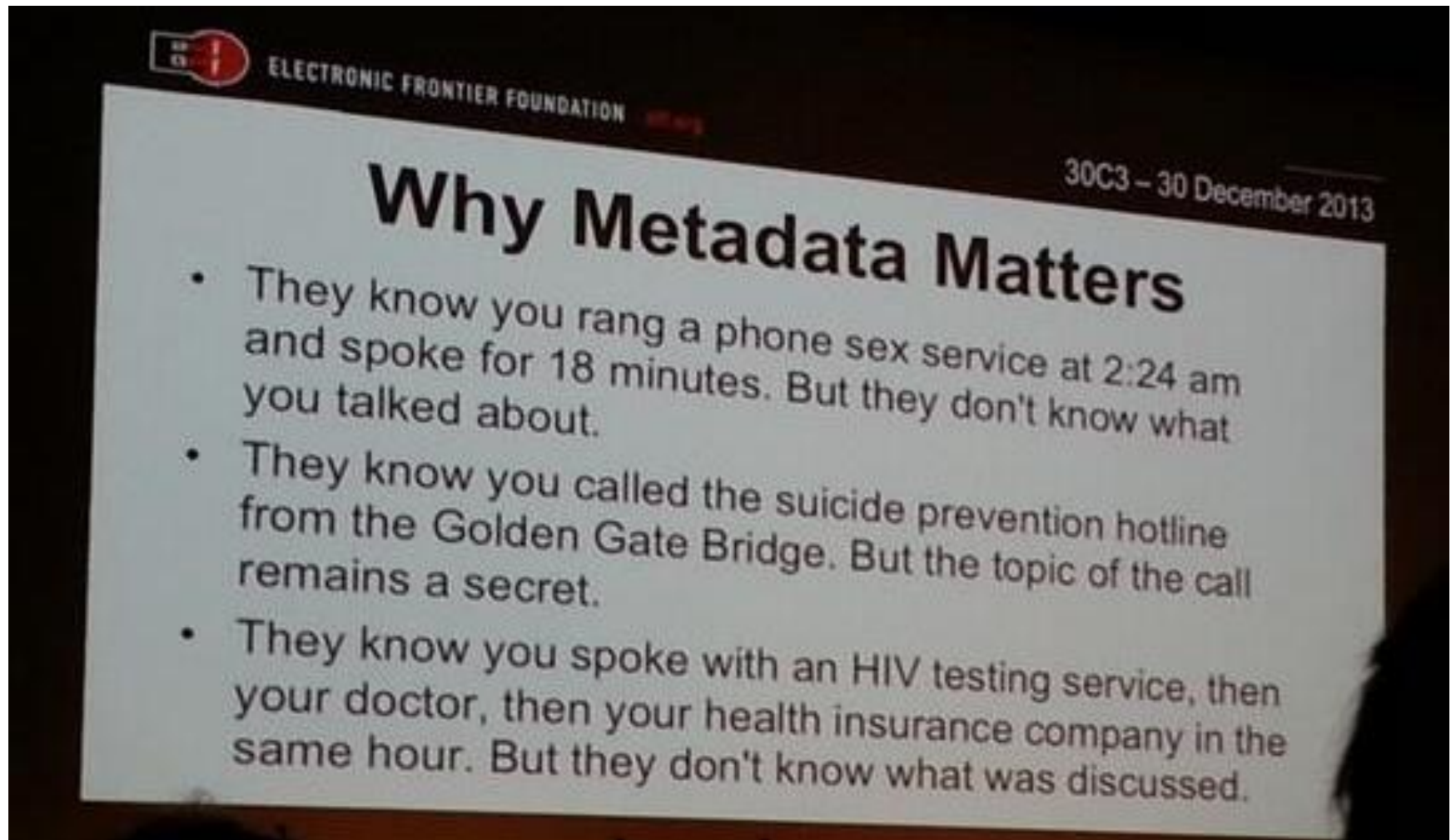
What native data can we safely store?

- LinkedIn: Object IDs only
- Twitter: IDs and the freshest seen API call result
- Facebook: Anything that the user has given us access to

Ethical considerations

- We all have something to hide (e.g. from identity thieves)
- Important that personal data cannot proliferate once its owner removes / changes it
- How long to retain for? NSA's minimum 15-year seems excessive
- **Metadata just as powerful as text data**
- **Text data weaker without metadata**

Storing social media data



(from Kurt Opshal's slides at the Chaos Communication Congress, photo by Marion Marschalek)

Social media corpora

Distribution concerns

- Social media corpora are difficult to distribute
- E.g. Twitter does not allow you to give other researchers/companies/anyone tweets you have collected and annotated
- Instead, distribute the tweet IDs and stand-off markup for the linguistic gold data
- The recipient re-collects all tweets himself, based on the IDs
- Necessary so user-deleted tweets are not propagated – privacy
- LinkedIn has even more stringent data sharing policy
- Facebook more relaxed, but data recipient must also have express permission from user

Corpus completeness

- However, in some cases (e.g. misinformation, smear tweets) messages can be deleted
- Makes re-creating the corpus is problematic
- Two classes of deletion:
 - Rapid deletions, usually within first few minutes (e.g. of spam, for editing the text)
 - Slower deletions (Petrovic et al. 2013)

Increased topic and entity drift: broader range of entities (Eisenstein 2013)

- Corpora age rapidly, and become less useful for some purposes (e.g. NEL)



Bontcheva, Derczynski, Funk, Greenwood, Maynard, Aswani 2013. TwitIE: An open-source information extraction pipeline for microblog text. RANLP

Eisenstein 2013. What to do about bad language on the internet. NAACL

Kwak, Lee, Park, Moon 2010. What is Twitter, a social network or a news media? WWW

Petrovic, Osborne, Lavrenko 2013. I Wish I Didn't Say That! Analyzing and Predicting Deleted Messages in Twitter. arXiv cs.SI 1305.3107