5.
Tools and methods for data collection, secondary use of collected datasets

Universially useful skills and tools

- R
- Python
- (Excel)
- Tableau (http://www.tableau.com/)

Data Collection

Data access options

- APIs
- Official resellers
- „manual“ forms of collection
- Re-using published datasets
- Third party tools
- (Crowdsourcing)
- (via social media companies)

Inequality in data access possibilities

- Data haves and data have nots
  - Financial reasons
  - Connections to companies
  - Different skills
  - …

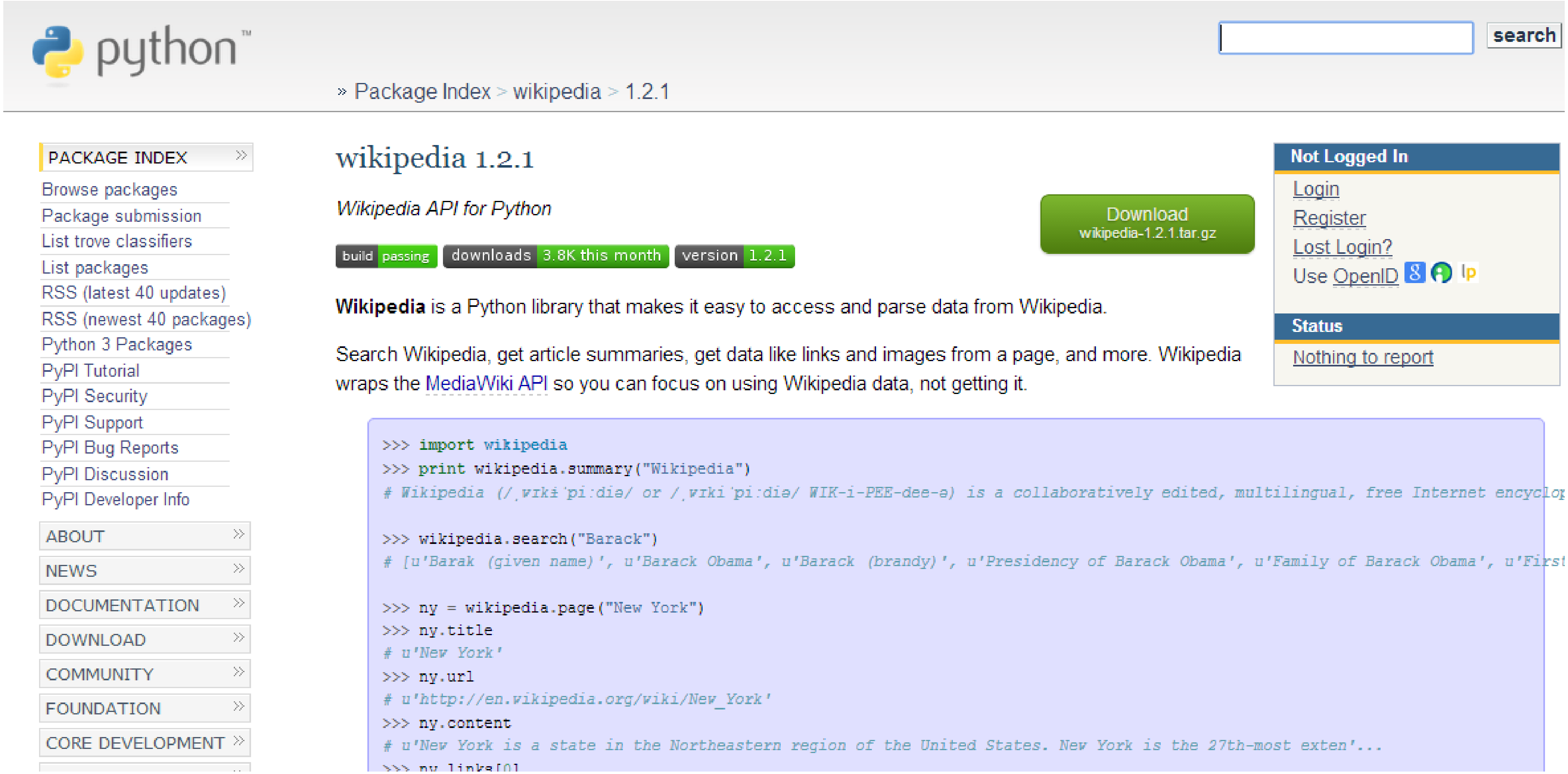
boyd, d., and Crawford, K. 2012. Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5):662–679. DOI: 10.1080/1369118X.2012.678878

Data access options

Via APIs (Application Programming Interface)

- Not developed for scholarly uses – but still frequently used.
- With several limitations
- May change their terms of services

Wikipedia API for Python



wikipedia 1.2.1

Wikipedia API for Python

Wikipedia is a Python library that makes it easy to access and parse data from Wikipedia.

Search Wikipedia, get article summaries, get data like links and images from a page, and more. Wikipedia wraps the MediaWiki API so you can focus on using Wikipedia data, not getting it.

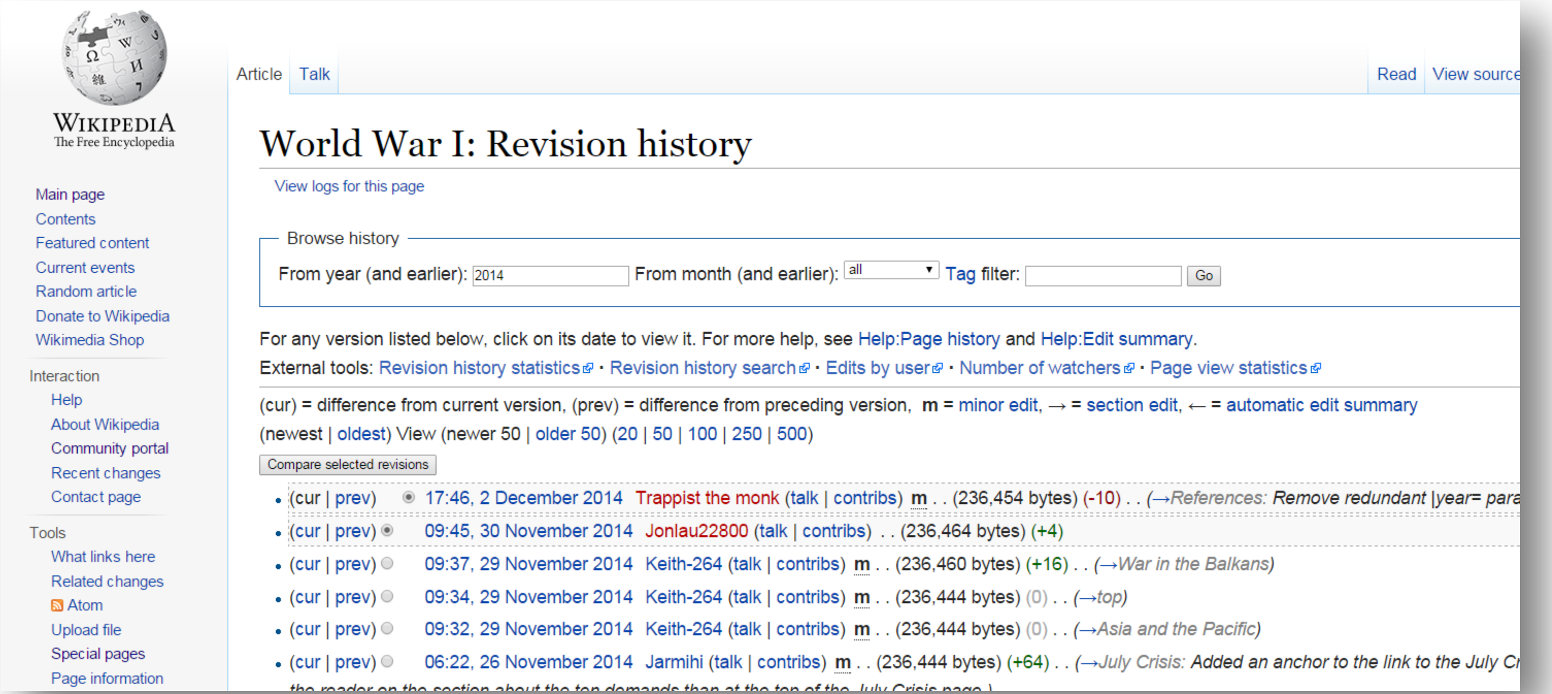
```
>>> import wikipedia
>>> print wikipedia.summary("Wikipedia")
# Wikipedia (/ˌwɪkɨˈpiːdiə/ or /ˌwɪkiˈpiːdiə/ WIK-i-PEE-dee-ə) is a collaboratively edited, multilingual, free Internet encyclop

>>> wikipedia.search("Barack")
# [u'Barak (given name)', u'Barack Obama', u'Barack (brandy)', u'Presidency of Barack Obama', u'Family of Barack Obama', u'First

>>> ny = wikipedia.page("New York")
>>> ny.title
# u'New York'
>>> ny.url
# u'http://en.wikipedia.org/wiki/New_York'
>>> ny.content
# u'New York is a state in the Northeastern region of the United States. New York is the 27th-most exten'...
>>> ny.links[0]
```

Wikipedia

- Editable articles about present and past events
- Collective intelligence?



World War I: Revision history

View logs for this page

Browse history

From year (and earlier): 2014 From month (and earlier): all Tag filter: Go

For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary.

External tools: Revision history statistics · Revision history search · Edits by user · Number of watchers · Page view statistics

(cur) = difference from current version, (prev) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary

(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur | prev) ● 17:46, 2 December 2014 Trappist the monk (talk | contribs) **m** . . (236,454 bytes) (-10) . . (*→References: Remove redundant |year= para*
- (cur | prev) ● 09:45, 30 November 2014 Jonlau22800 (talk | contribs) . . (236,464 bytes) (+4)
- (cur | prev) ○ 09:37, 29 November 2014 Keith-264 (talk | contribs) **m** . . (236,460 bytes) (+16) . . (*→War in the Balkans*)
- (cur | prev) ○ 09:34, 29 November 2014 Keith-264 (talk | contribs) **m** . . (236,444 bytes) (0) . . (*→top*)
- (cur | prev) ○ 09:32, 29 November 2014 Keith-264 (talk | contribs) **m** . . (236,444 bytes) (0) . . (*→Asia and the Pacific*)
- (cur | prev) ○ 06:22, 26 November 2014 Jarmihi (talk | contribs) **m** . . (236,444 bytes) (+64) . . (*→July Crisis: Added an anchor to the link to the July Cr*
the reader on the section about the ten demands than at the top of the July Crisis page.)

Wikipedia: Edit Wars

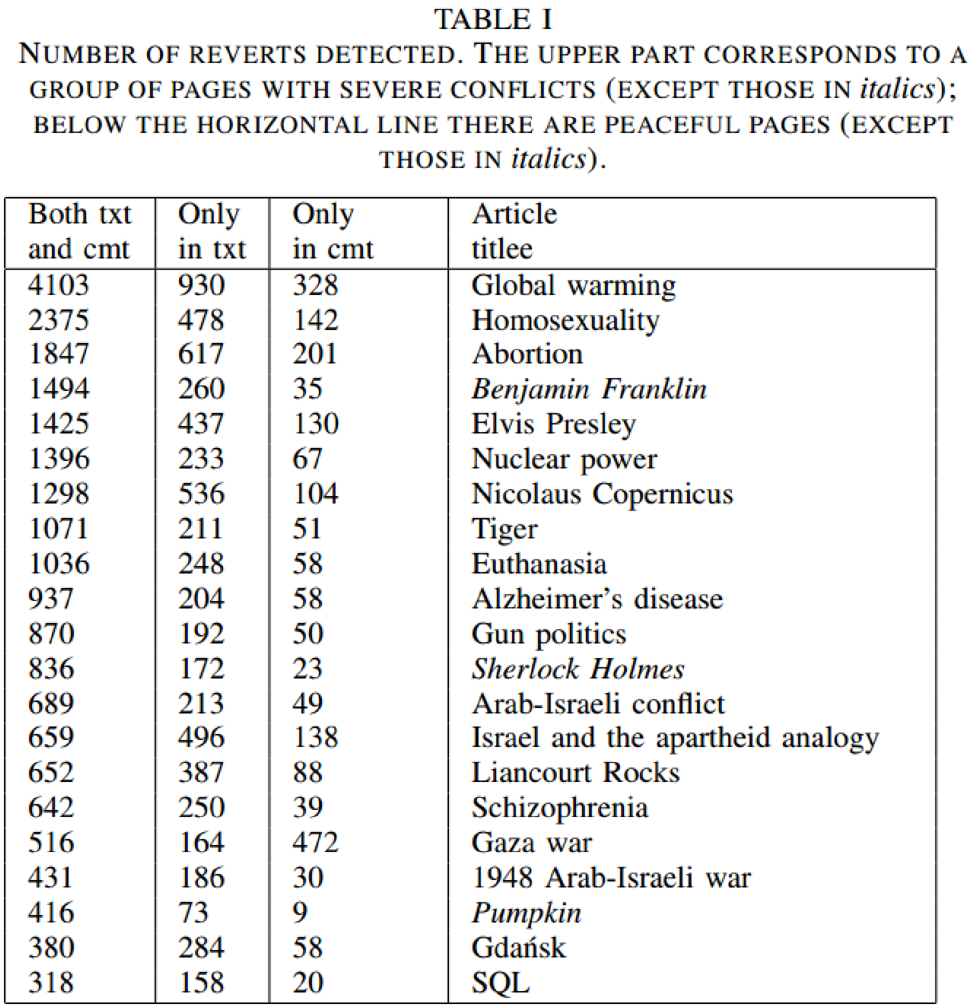


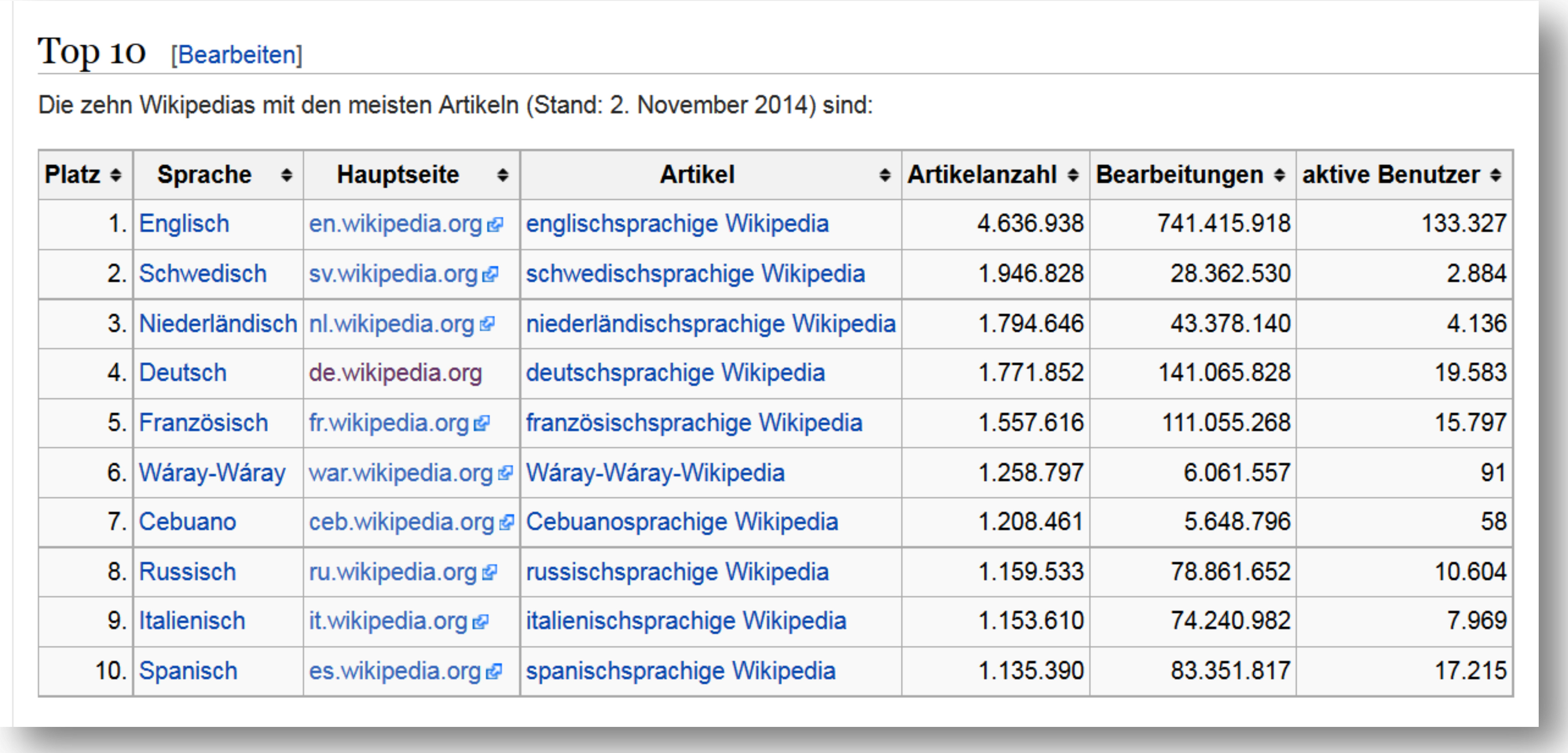
TABLE I

Number of reverts detected. The upper part corresponds to a group of pages with severe conflicts (except those in *italics*); below the horizontal line there are peaceful pages (except those in *italics*).

Both txt and cmt	Only in txt	Only in cmt	Article titlee
4103	930	328	Global warming
2375	478	142	Homosexuality
1847	617	201	Abortion
1494	260	35	*Benjamin Franklin*
1425	437	130	Elvis Presley
1396	233	67	Nuclear power
1298	536	104	Nicolaus Copernicus
1071	211	51	Tiger
1036	248	58	Euthanasia
937	204	58	Alzheimer's disease
870	192	50	Gun politics
836	172	23	*Sherlock Holmes*
689	213	49	Arab-Israeli conflict
659	496	138	Israel and the apartheid analogy
652	387	88	Liancourt Rocks
642	250	39	Schizophrenia
516	164	472	Gaza war
431	186	30	1948 Arab-Israeli war
416	73	9	*Pumpkin*
380	284	58	Gdańsk
318	158	20	SQL

Yasseri, T., Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of conflicts in Wikipedia. PLOS ONE 7(6): e38869.

Wikipedia: Cultural Disparities?



Top 10 [Bearbeiten]

Die zehn Wikipedias mit den meisten Artikeln (Stand: 2. November 2014) sind:

Platz	Sprache	Hauptseite	Artikel	Artikelanzahl	Bearbeitungen	aktive Benutzer
1.	Englisch	en.wikipedia.org	englischsprachige Wikipedia	4.636.938	741.415.918	133.327
2.	Schwedisch	sv.wikipedia.org	schwedischsprachige Wikipedia	1.946.828	28.362.530	2.884
3.	Niederländisch	nl.wikipedia.org	niederländischsprachige Wikipedia	1.794.646	43.378.140	4.136
4.	Deutsch	de.wikipedia.org	deutschsprachige Wikipedia	1.771.852	141.065.828	19.583
5.	Französisch	fr.wikipedia.org	französischsprachige Wikipedia	1.557.616	111.055.268	15.797
6.	Wáray-Wáray	war.wikipedia.org	Wáray-Wáray-Wikipedia	1.258.797	6.061.557	91
7.	Cebuano	ceb.wikipedia.org	Cebuanosprachige Wikipedia	1.208.461	5.648.796	58
8.	Russisch	ru.wikipedia.org	russischsprachige Wikipedia	1.159.533	78.861.652	10.604
9.	Italienisch	it.wikipedia.org	italienischsprachige Wikipedia	1.153.610	74.240.982	7.969
10.	Spanisch	es.wikipedia.org	spanischsprachige Wikipedia	1.135.390	83.351.817	17.215

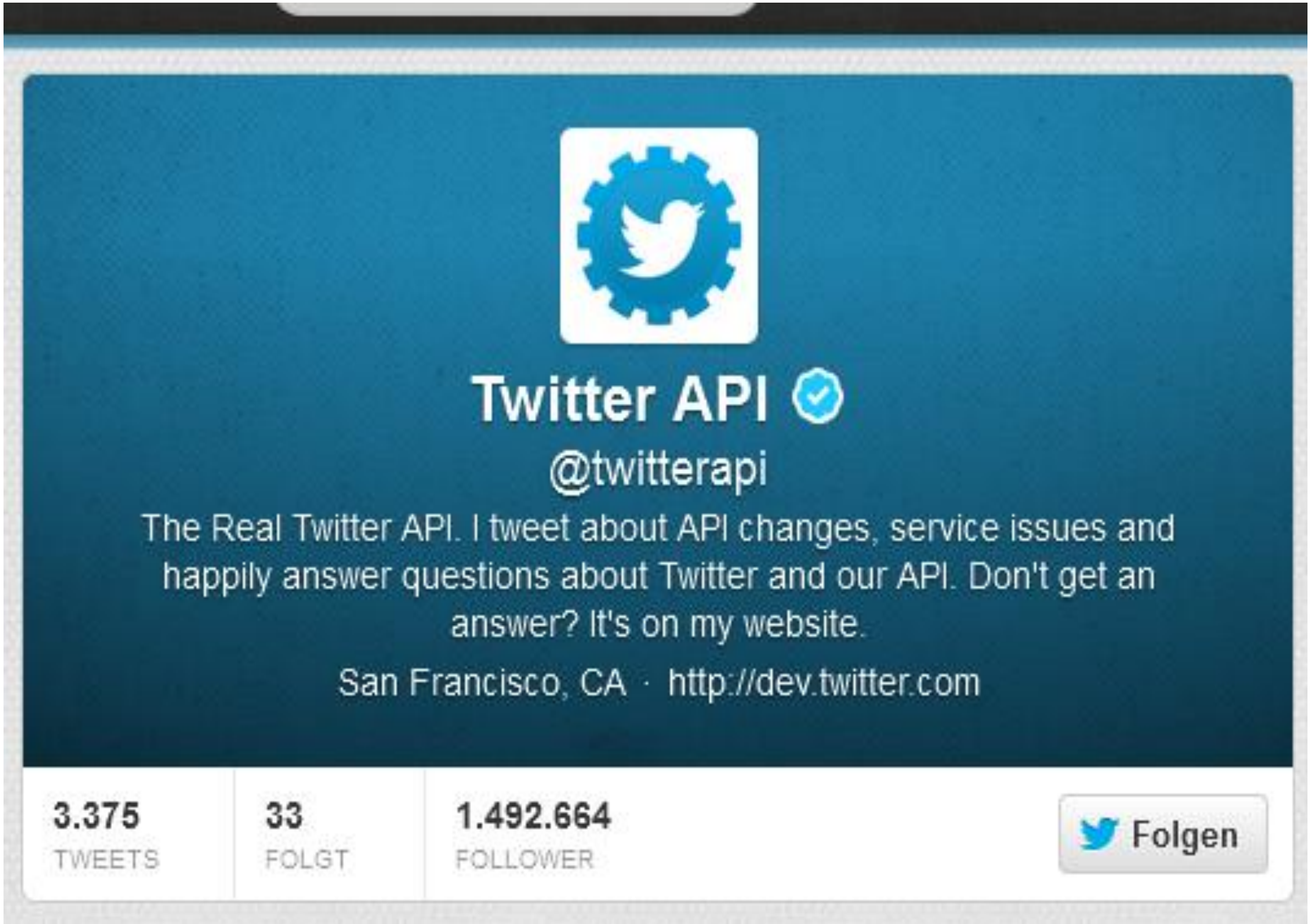
Wikipedia dumps

- Also available via the Internet Archive.
- Capture Wikipedia at one point in time.
- E.g. German Wikipedia is about 3 GB plus 15 GB for images.

https://dumps.wikimedia.org/
https://en.wikipedia.org/wiki/Wikipedia:Database_download

What can you do with Wikipedia data?

- Look at edit history.
- Use as text corpus.
- Compare text length, language use etc. for example across categories.
- Extract links.
- Compare language versions.
- …



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

3.375	33	1.492.664
TWEETS	FOLGT	FOLLOWER

https://dev.twitter.com/docs/api/1.1
https://dev.twitter.com/blog

Twitter APIs

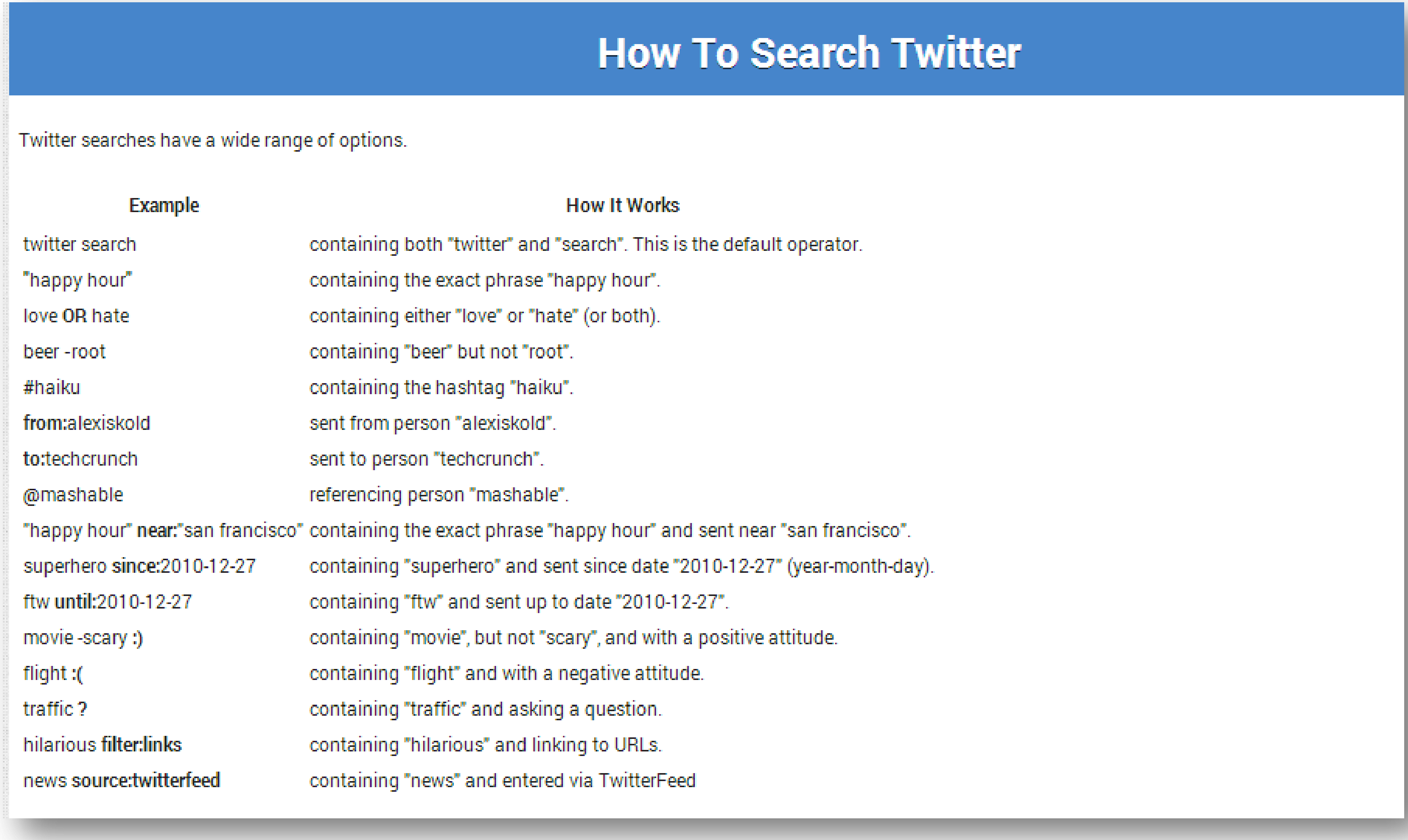
STREAMING API

- Push-based, Live-Stream
- Public stream vs. User stream
- Researchers need tools to keep connected to the stream.

REST API

- Limited number of requests per hour
- Social graph data (who is following whom)
- Trending topics
- And many more

https://dev.twitter.com/docs/using-search

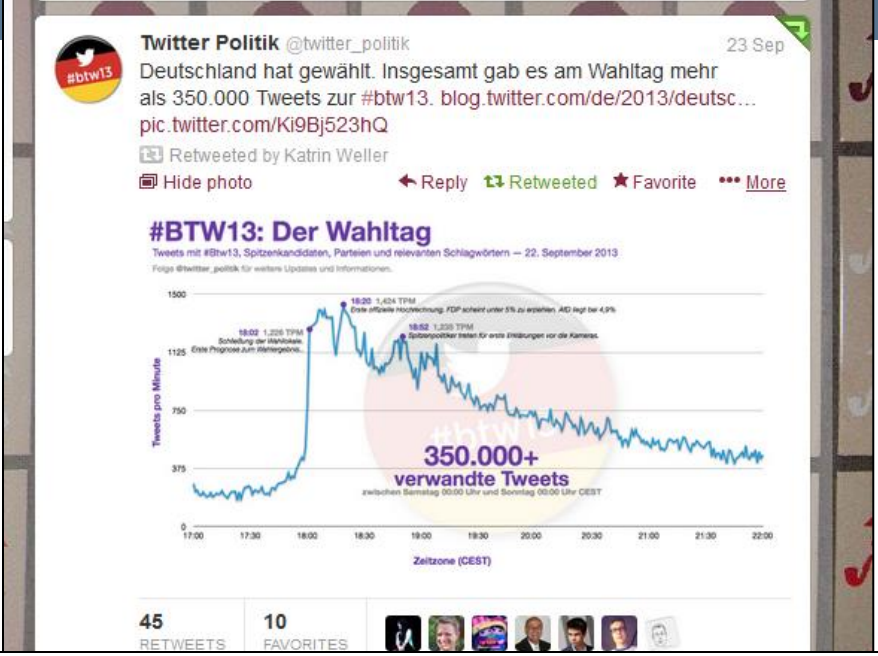


How To Search Twitter

Twitter searches have a wide range of options.

Example	How It Works
twitter search	containing both "twitter" and "search". This is the default operator.
"happy hour"	containing the exact phrase "happy hour".
love **OR** hate	containing either "love" or "hate" (or both).
beer -root	containing "beer" but not "root".
#haiku	containing the hashtag "haiku".
from:alexiskold	sent from person "alexiskold".
to:techcrunch	sent to person "techcrunch".
@mashable	referencing person "mashable".
"happy hour" **near:**"san francisco"	containing the exact phrase "happy hour" and sent near "san francisco".
superhero **since:**2010-12-27	containing "superhero" and sent since date "2010-12-27" (year-month-day).
ftw **until:**2010-12-27	containing "ftw" and sent up to date "2010-12-27".
movie -scary **:)**	containing "movie", but not "scary", and with a positive attitude.
flight **:(**	containing "flight" and with a negative attitude.
traffic **?**	containing "traffic" and asking a question.
hilarious **filter:links**	containing "hilarious" and linking to URLs.
news **source:twitterfeed**	containing "news" and entered via TwitterFeed

gesis Leibniz Institute for the Social Sciences

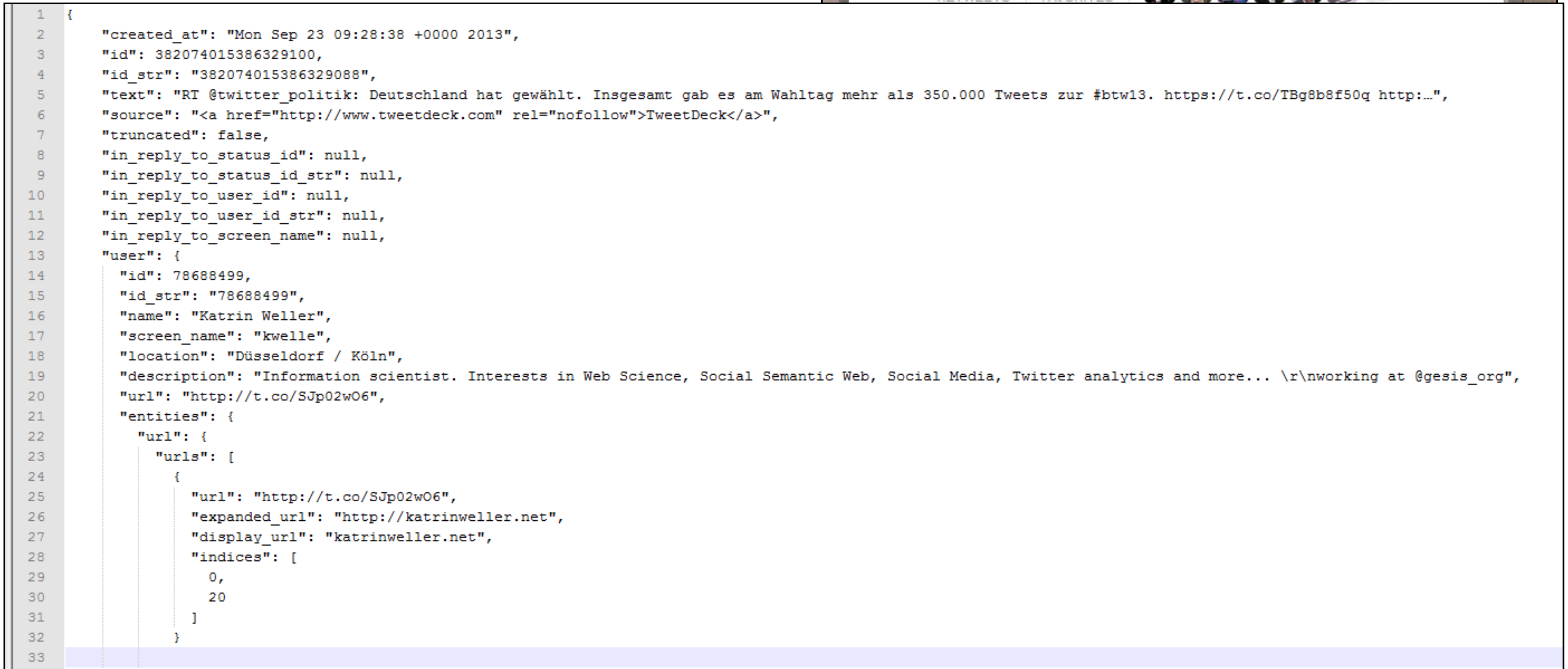


Twitter Politik @twitter_politik 23 Sep

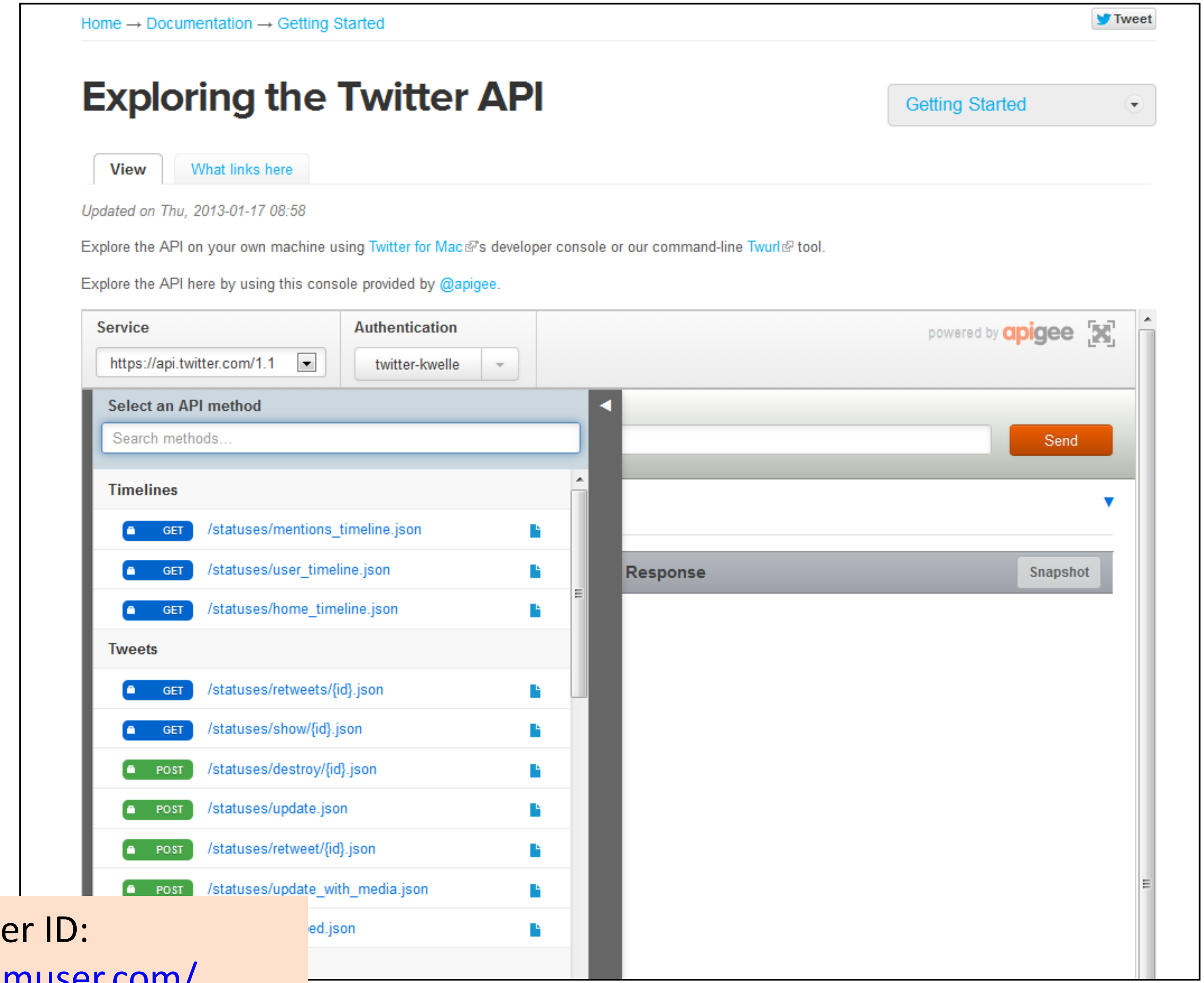
Deutschland hat gewählt. Insgesamt gab es am Wahltag mehr als 350.000 Tweets zur #btw13. blog.twitter.com/de/2013/deutsc… pic.twitter.com/Ki9Bj523hQ

Retweeted by Katrin Weller

Hide photo
Reply
Retweeted
Favorite
More
#btw13
#BTW13: Der Wahltag
Tweets mit #Btw13, Spitzenkandidaten, Parteien und relevanten Schlagwörtern — 22. September 2013
Folge @twitter_politik für weitere Updates und Informationen.
Tweets pro Minute
1500
1125
750
375
0
18:02 1,226 TPM
Schließung der Wahllokale.
Erste Prognose zum Wahlergebnis.
18:20 1,424 TPM
Erste offizielle Hochrechnung. FDP scheint unter 5% zu erzielen. AfD liegt bei 4,9%
18:52 1,235 TPM
Spitzenpolitiker treten für erste Erklärungen vor die Kameras.
350.000+
verwandte Tweets
zwischen Samstag 00:00 Uhr und Sonntag 00:00 Uhr CEST
17:00
17:30
18:00
18:30
19:00
19:30
20:00
20:30
21:00
21:30
22:00
Zeitzone (CEST)
45 RETWEETS
10 FAVORITES



```
{
    "created_at": "Mon Sep 23 09:28:38 +0000 2013",
    "id": 382074015386329100,
    "id_str": "382074015386329088",
    "text": "RT @twitter_politik: Deutschland hat gewählt. Insgesamt gab es am Wahltag mehr als 350.000 Tweets zur #btw13. https://t.co/TBg8b8f50q http:…",
    "source": "<a href="http://www.tweetdeck.com" rel="nofollow">TweetDeck</a>",
    "truncated": false,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "in_reply_to_screen_name": null,
    "user": {
      "id": 78688499,
      "id_str": "78688499",
      "name": "Katrin Weller",
      "screen_name": "kwelle",
      "location": "Düsseldorf / Köln",
      "description": "Information scientist. Interests in Web Science, Social Semantic Web, Social Media, Twitter analytics and more... \r\nworking at @gesis_org",
      "url": "http://t.co/SJp02wO6",
      "entities": {
        "url": {
          "urls": [
            {
              "url": "http://t.co/SJp02wO6",
              "expanded_url": "http://katrinweller.net",
              "display_url": "katrinweller.net",
              "indices": [
                0,
                20
              ]
            }
```



Home → Documentation → Getting Started

Exploring the Twitter API

Updated on Thu, 2013-01-17 08:58

Explore the API on your own machine using Twitter for Mac's developer console or our command-line Twurl tool.

Explore the API here by using this console provided by @apigee.

Service: https://api.twitter.com/1.1

Authentication: twitter-kwelle

Select an API method

Search methods...

Timelines

- GET /statuses/mentions_timeline.json
- GET /statuses/user_timeline.json
- GET /statuses/home_timeline.json

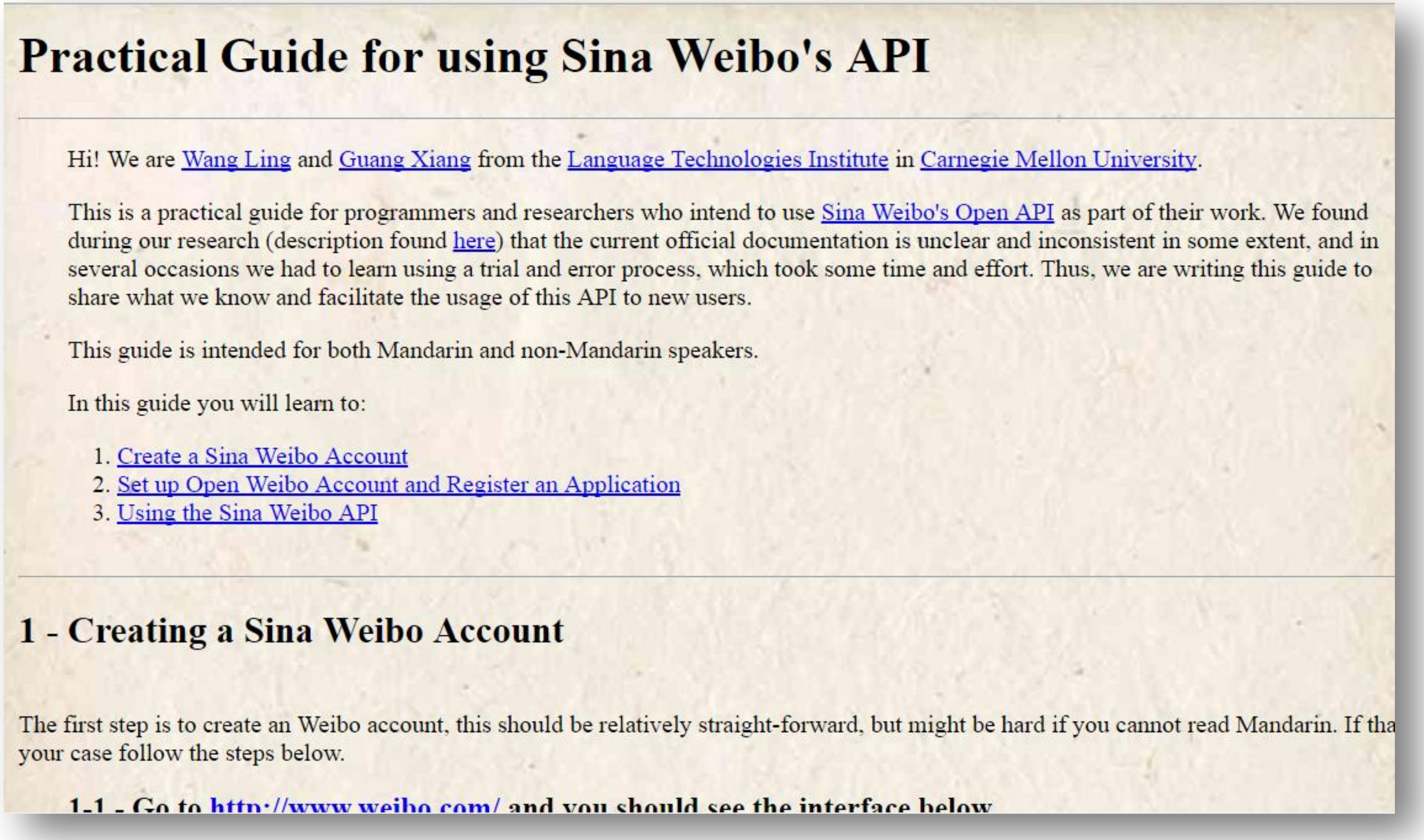
Tweets

- GET /statuses/retweets/{id}.json
- GET /statuses/show/{id}.json
- POST /statuses/destroy/{id}.json
- POST /statuses/update.json
- POST /statuses/retweet/{id}.json
- POST /statuses/update_with_media.json
- ed.json

Find your own user ID:
http://www.idfromuser.com/

https://dev.twitter.com/console

Sina Weibo



Practical Guide for using Sina Weibo's API

Hi! We are Wang Ling and Guang Xiang from the Language Technologies Institute in Carnegie Mellon University.

This is a practical guide for programmers and researchers who intend to use Sina Weibo's Open API as part of their work. We found during our research (description found here) that the current official documentation is unclear and inconsistent in some extent, and in several occasions we had to learn using a trial and error process, which took some time and effort. Thus, we are writing this guide to share what we know and facilitate the usage of this API to new users.

This guide is intended for both Mandarin and non-Mandarin speakers.

In this guide you will learn to:

1. Create a Sina Weibo Account
2. Set up Open Weibo Account and Register an Application
3. Using the Sina Weibo API

1 - Creating a Sina Weibo Account

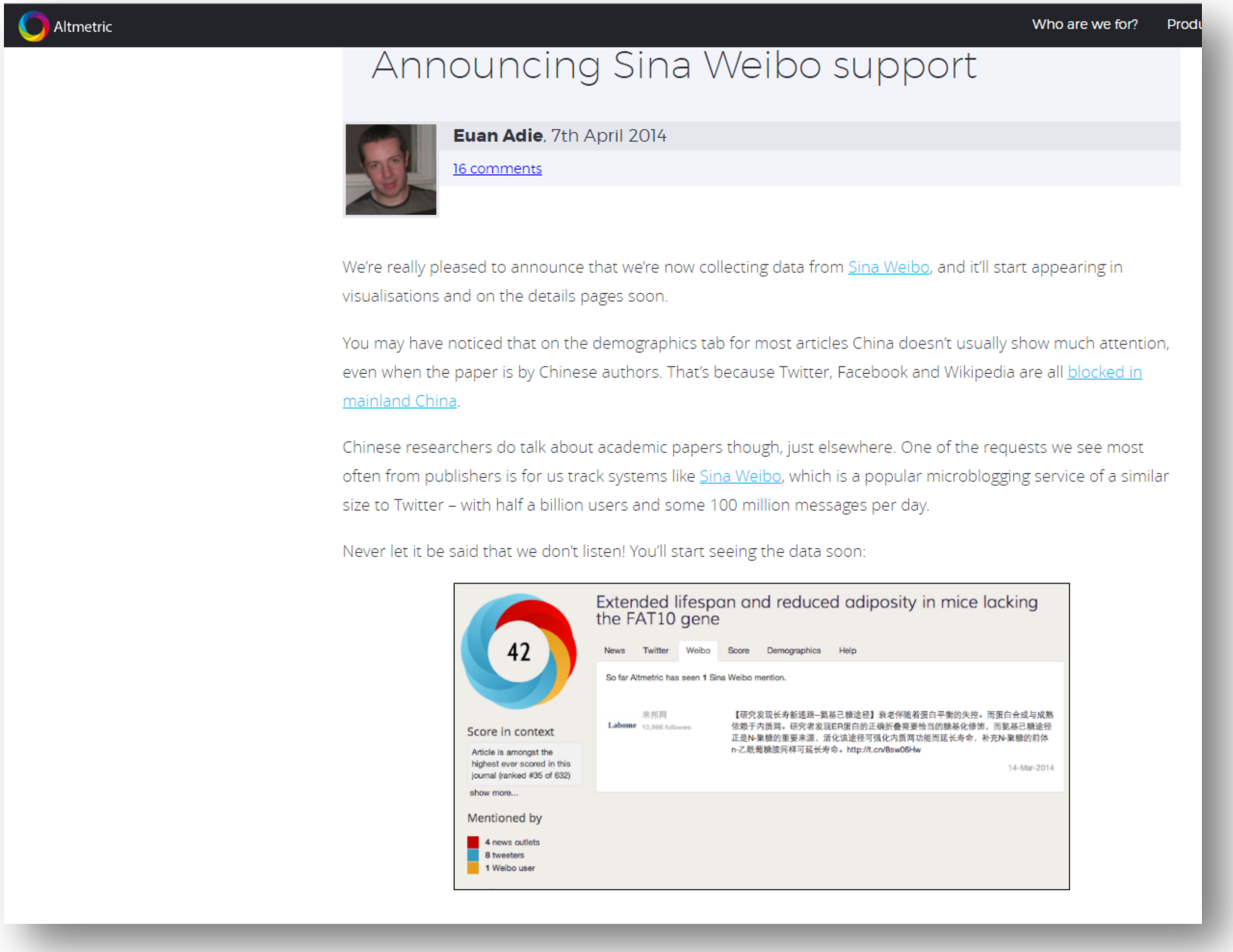
The first step is to create an Weibo account, this should be relatively straight-forward, but might be hard if you cannot read Mandarin. If tha
your case follow the steps below.

1-1 - Go to http://www.weibo.com/ and you should see the interface below

http://www.cs.cmu.edu/~lingwang/weiboguide/

https://builtvisible.com/gathering-data-sina-weibo-api/

Sina Weibo



Announcing Sina Weibo support

Euan Adie, 7th April 2014

16 comments

We're really pleased to announce that we're now collecting data from Sina Weibo, and it'll start appearing in visualisations and on the details pages soon.

You may have noticed that on the demographics tab for most articles China doesn't usually show much attention, even when the paper is by Chinese authors. That's because Twitter, Facebook and Wikipedia are all blocked in mainland China.

Chinese researchers do talk about academic papers though, just elsewhere. One of the requests we see most often from publishers is for us track systems like Sina Weibo, which is a popular microblogging service of a similar size to Twitter – with half a billion users and some 100 million messages per day.

Never let it be said that we don't listen! You'll start seeing the data soon:

Extended lifespan and reduced adiposity in mice lacking the FAT10 gene

So far Altmetric has seen **1** Sina Weibo mention.

Labome 来邦网
10,986 followers

【研究发现长寿新通路--氨基己糖途径】衰老伴随着蛋白平衡的失控。而蛋白合成与成熟依赖于内质网。研究者发现ER蛋白的正确折叠需要恰当的糖基化修饰，而氨基己糖途径正是N-聚糖的重要来源，活化该途径可强化内质网功能而延长寿命，补充N-聚糖的前体n-乙酰葡糖胺同样可延长寿命。http://t.cn/8sw06Hw

14-Mar-2014

Score in context

Article is amongst the highest ever scored in this journal (ranked #35 of 632)

show more...

Mentioned by

🟥 4 news outlets
🟦 8 tweeters
🟧 1 Weibo user

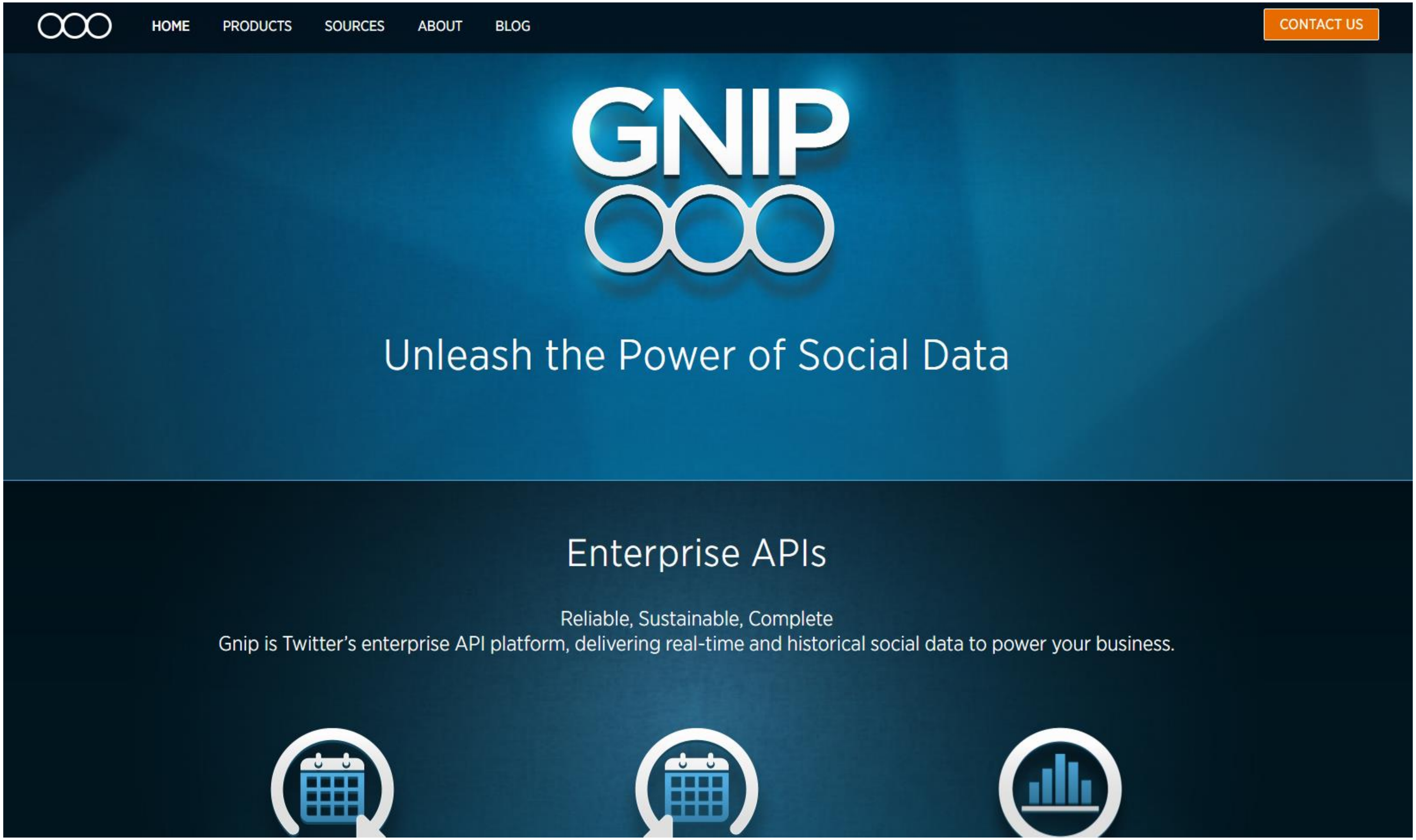
https://www.altmetric.com/blog/announcing-sina-weibo-support/

Data access options

Via offical resellers

- Rather costly fees
- Contracts may need to be made with the head of an institute rather than with individual researchers.
- Some technical understanding is required
- May prohibit sharing the data
- Main players: GNIP and Datasift

GNIP



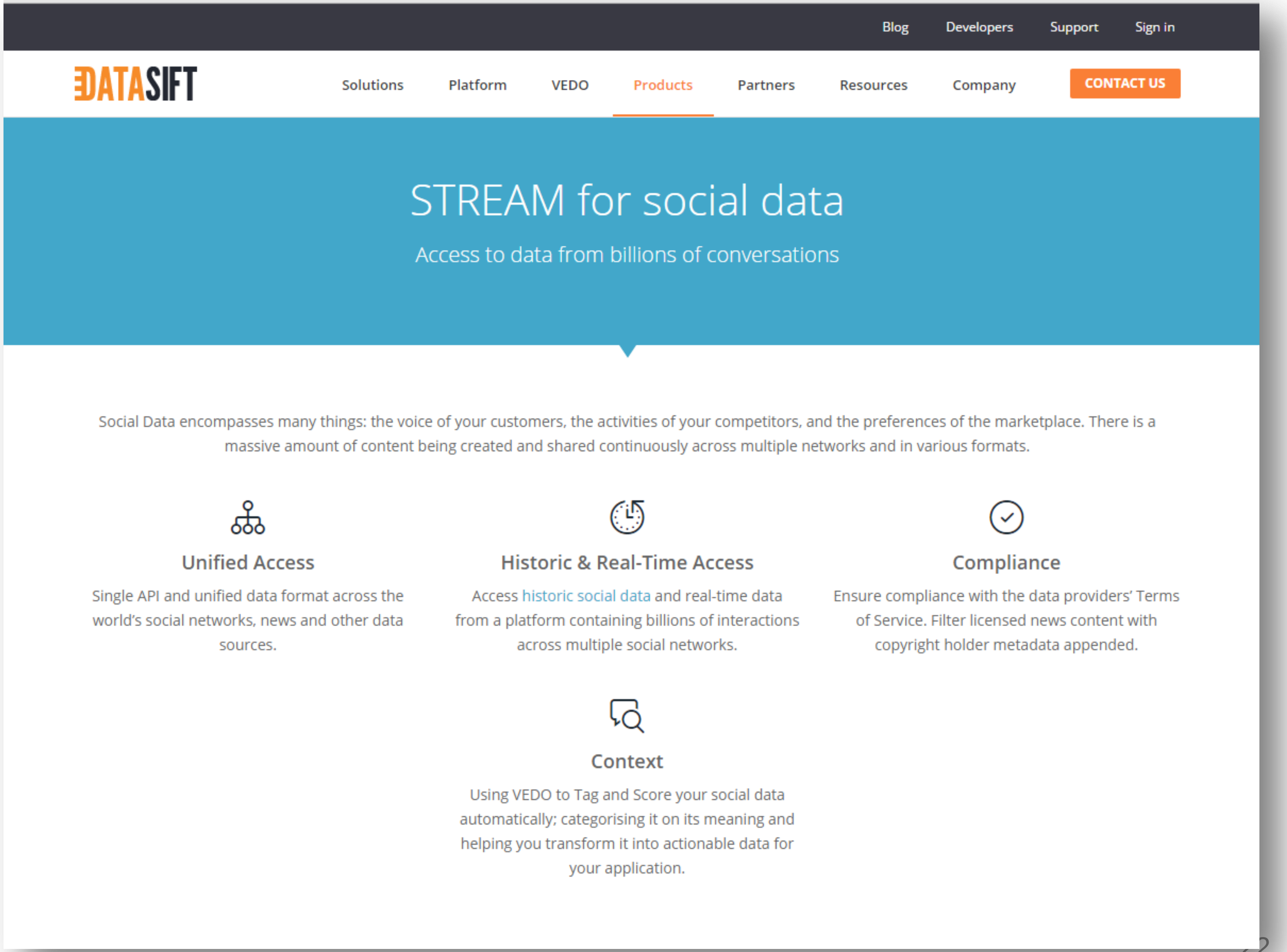
Unleash the Power of Social Data

Enterprise APIs

Reliable, Sustainable, Complete
Gnip is Twitter's enterprise API platform, delivering real-time and historical social data to power your business.

https://gnip.com/

DataSift



STREAM for social data

Access to data from billions of conversations

Social Data encompasses many things: the voice of your customers, the activities of your competitors, and the preferences of the marketplace. There is a massive amount of content being created and shared continuously across multiple networks and in various formats.

Unified Access

Single API and unified data format across the world's social networks, news and other data sources.

Historic & Real-Time Access

Access historic social data and real-time data from a platform containing billions of interactions across multiple social networks.

Compliance

Ensure compliance with the data providers' Terms of Service. Filter licensed news content with copyright holder metadata appended.

Context

Using VEDO to Tag and Score your social data automatically; categorising it on its meaning and helping you transform it into actionable data for your application.

Data access options

Via third party tools

- Black box effect – lack of transparancy
- Fees may apply
- Changes are possible
- Additional layer of legal restrictions (e.g. limiting what might be published)

Data access options

„Manual“ forms of collection

- E.g. copy-and-paste, screenshots
- Time-consuming, error-prone
- Little exchange about best practices, non standardized
- Good in addition to other methods, e.g. to capture images and as„proof“

Data access options

Reuse of existing datasets

- Lack of available datasets
- Lack of standards for documentation and formats
- Lack of persistent identifyers
- Legal and ethical restrictions

Data access options

Crowdsourcing

- Little guidance / best practices
- Via Amazon Mechanical Turk and others: ethical challenges
- Via Apps (e.g. on Facebook)

Data access options

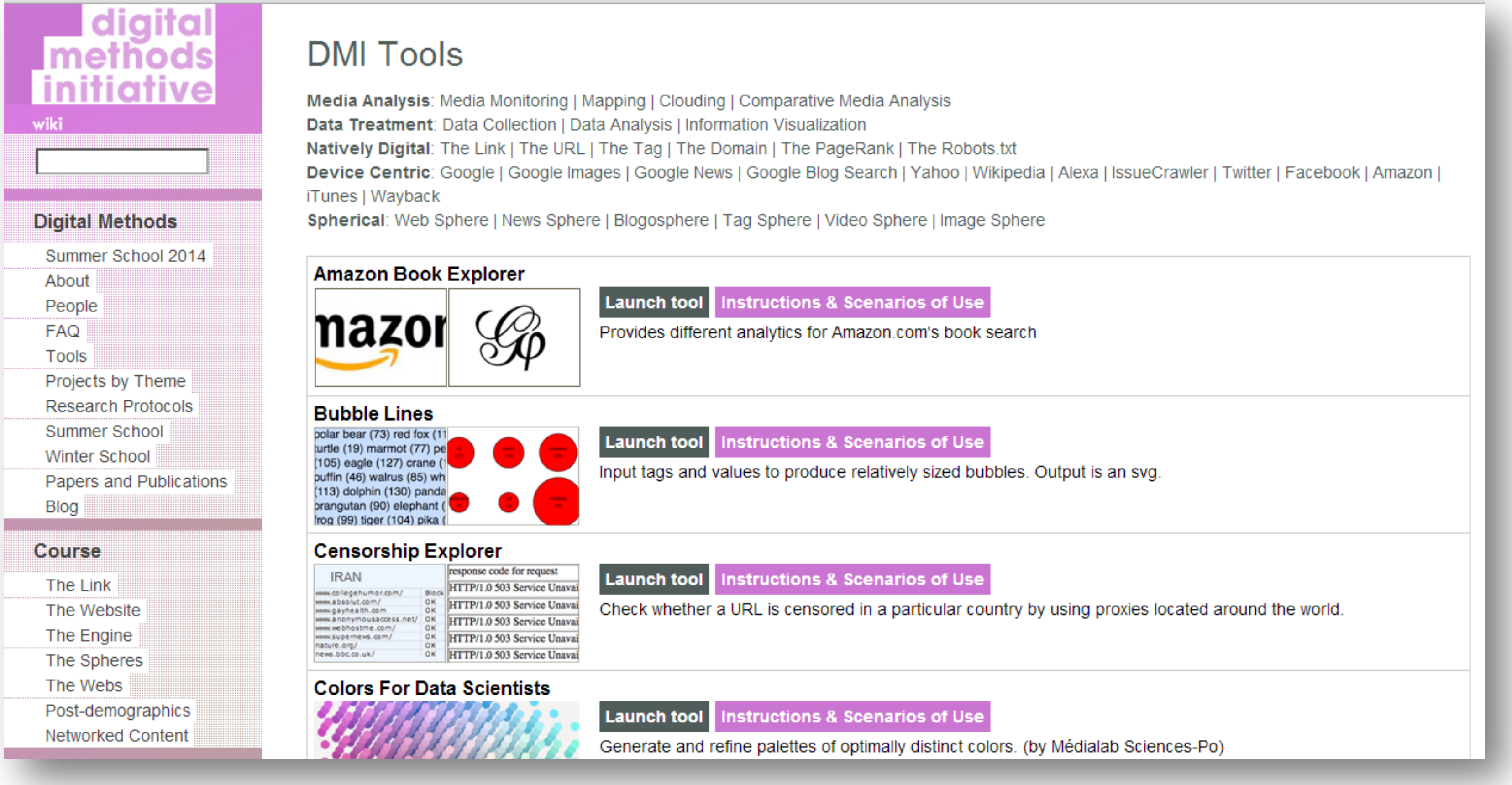
Via social media companies

- Usually in form of priviledged access (special agreements)
- High-quality data
- Limitations on usage and sharing

Research groups that offer different tools or training

Digital Methods Initiative

- https://wiki.digitalmethods.net/Dmi/ToolDatabase



digital methods initiative

wiki

Digital Methods

- Summer School 2014
- About
- People
- FAQ
- Tools
- Projects by Theme
- Research Protocols
- Summer School
- Winter School
- Papers and Publications
- Blog

Course

- The Link
- The Website
- The Engine
- The Spheres
- The Webs
- Post-demographics
- Networked Content

DMI Tools

Media Analysis: Media Monitoring | Mapping | Clouding | Comparative Media Analysis
Data Treatment: Data Collection | Data Analysis | Information Visualization
Natively Digital: The Link | The URL | The Tag | The Domain | The PageRank | The Robots.txt
Device Centric: Google | Google Images | Google News | Google Blog Search | Yahoo | Wikipedia | Alexa | IssueCrawler | Twitter | Facebook | Amazon | iTunes | Wayback
Spherical: Web Sphere | News Sphere | Blogosphere | Tag Sphere | Video Sphere | Image Sphere

Amazon Book Explorer

Launch tool | Instructions & Scenarios of Use

Provides different analytics for Amazon.com's book search

Bubble Lines

Launch tool | Instructions & Scenarios of Use

Input tags and values to produce relatively sized bubbles. Output is an svg.

Censorship Explorer

Launch tool | Instructions & Scenarios of Use

Check whether a URL is censored in a particular country by using proxies located around the world.

Colors For Data Scientists

Launch tool | Instructions & Scenarios of Use

Generate and refine palettes of optimally distinct colors. (by Médialab Sciences-Po)

Digital Methods Initiative

Includes

- Tools for extracting and resolving URLs

- Tool for scraping google images

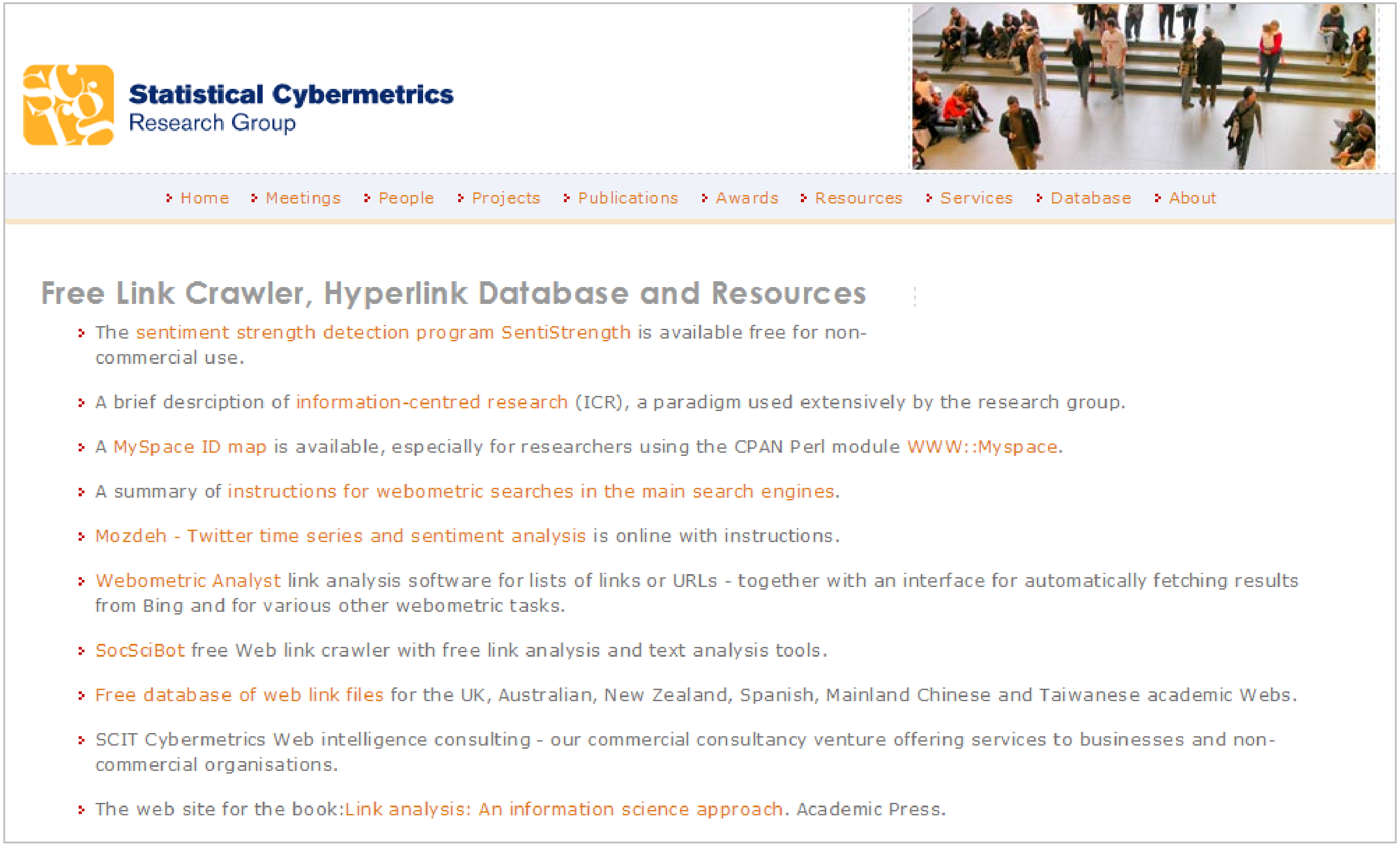
- Wikipedia Edits Scraper

- …

They also offer Summer / Winter Schools.

- https://wiki.digitalmethods.net/Dmi/ToolDatabase

Statistical Cybermetrics Research Group Wolverhampton



Free Link Crawler, Hyperlink Database and Resources

- The sentiment strength detection program SentiStrength is available free for non-commercial use.
- A brief desrciption of information-centred research (ICR), a paradigm used extensively by the research group.
- A MySpace ID map is available, especially for researchers using the CPAN Perl module WWW::Myspace.
- A summary of instructions for webometric searches in the main search engines.
- Mozdeh - Twitter time series and sentiment analysis is online with instructions.
- Webometric Analyst link analysis software for lists of links or URLs - together with an interface for automatically fetching results from Bing and for various other webometric tasks.
- SocSciBot free Web link crawler with free link analysis and text analysis tools.
- Free database of web link files for the UK, Australian, New Zealand, Spanish, Mainland Chinese and Taiwanese academic Webs.
- SCIT Cybermetrics Web intelligence consulting - our commercial consultancy venture offering services to businesses and non-commercial organisations.
- The web site for the book:Link analysis: An information science approach. Academic Press.

http://cybermetrics.wlv.ac.uk/resources.htm

Statistical Cybermetrics Research Group Wolverhampton

Includes for example:

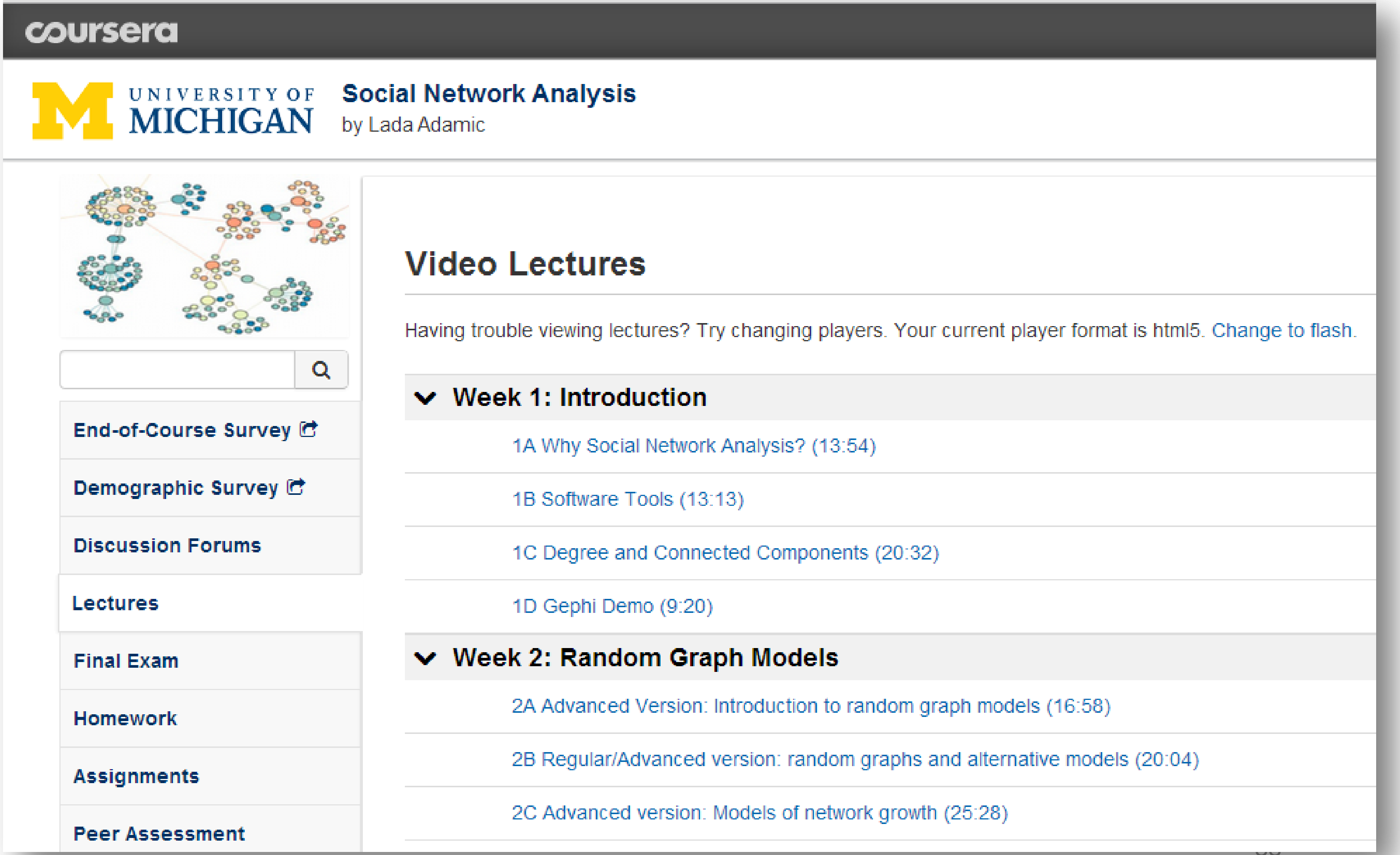
- Text analysis, Sentiment Analyse

- Webometrics: „Web citation data“

http://cybermetrics.wlv.ac.uk/resources.htm

gesis Leibniz Institute for the Social Sciences

Open Educational Ressources



coursera

UNIVERSITY OF MICHIGAN

Social Network Analysis

by Lada Adamic

- End-of-Course Survey
- Demographic Survey
- Discussion Forums
- Lectures
- Final Exam
- Homework
- Assignments
- Peer Assessment

Video Lectures

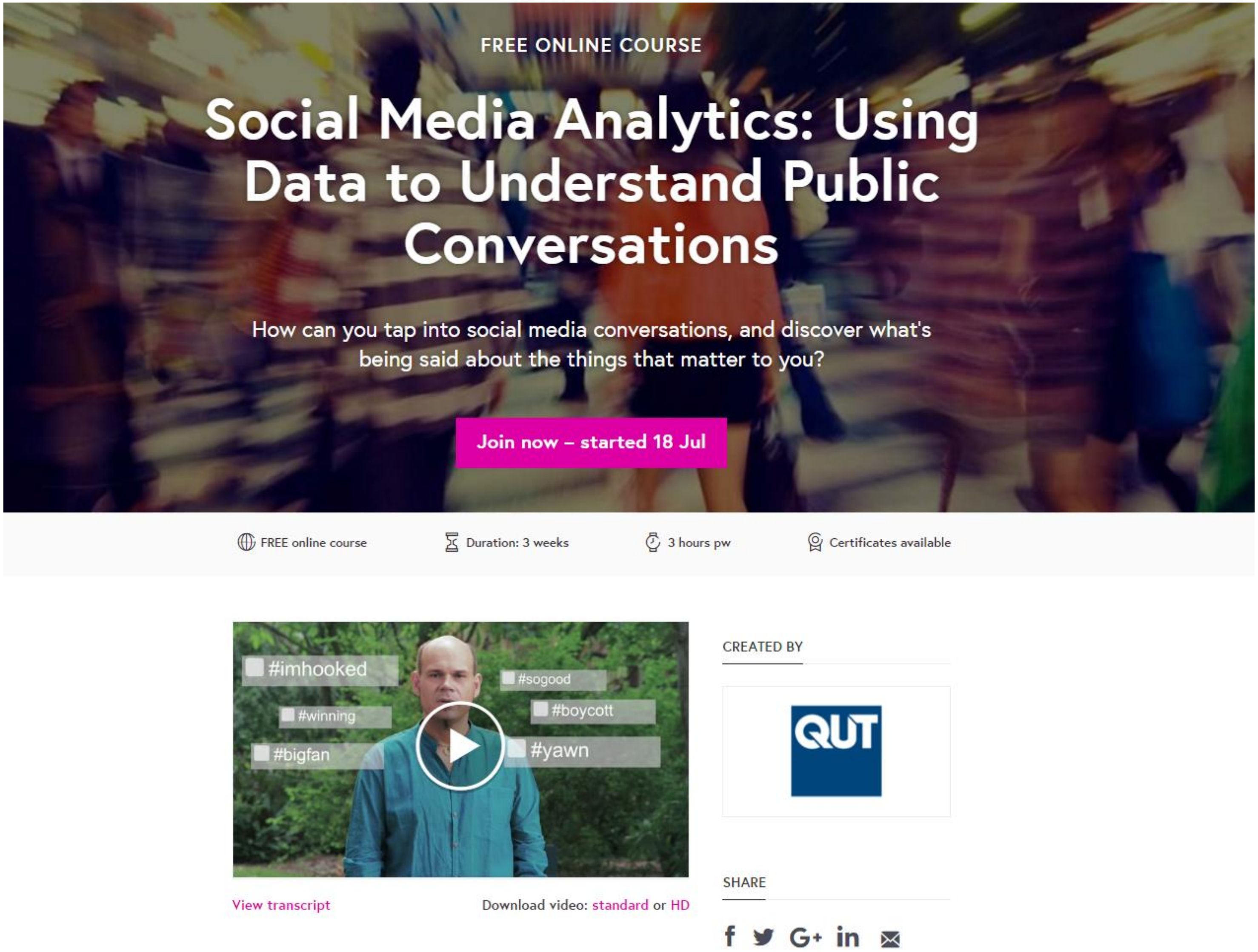
Having trouble viewing lectures? Try changing players. Your current player format is html5. Change to flash.

Week 1: Introduction

- 1A Why Social Network Analysis? (13:54)
- 1B Software Tools (13:13)
- 1C Degree and Connected Components (20:32)
- 1D Gephi Demo (9:20)

Week 2: Random Graph Models

- 2A Advanced Version: Introduction to random graph models (16:58)
- 2B Regular/Advanced version: random graphs and alternative models (20:04)
- 2C Advanced version: Models of network growth (25:28)



FREE ONLINE COURSE

Social Media Analytics: Using Data to Understand Public Conversations

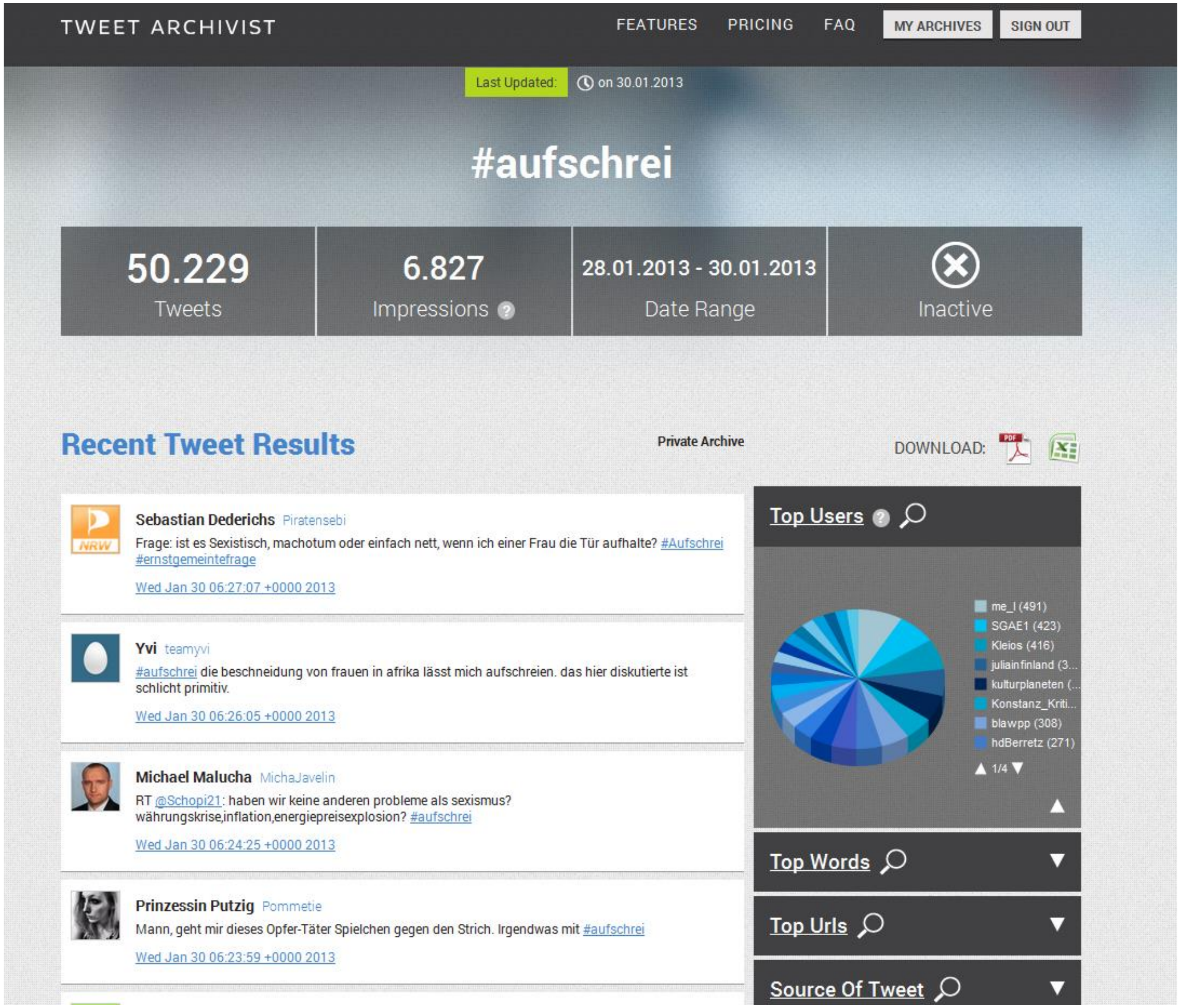
How can you tap into social media conversations, and discover what's being said about the things that matter to you?

https://www.futurelearn.com/courses/social-media-analytics

Data collection tools

Data collection tools

- Run on your own server
- Install on your own laptop
- Web-based



#aufschrei

50.229	6.827	28.01.2013 - 30.01.2013	⊗
Tweets	Impressions ?	Date Range	Inactive

Recent Tweet Results

Private Archive

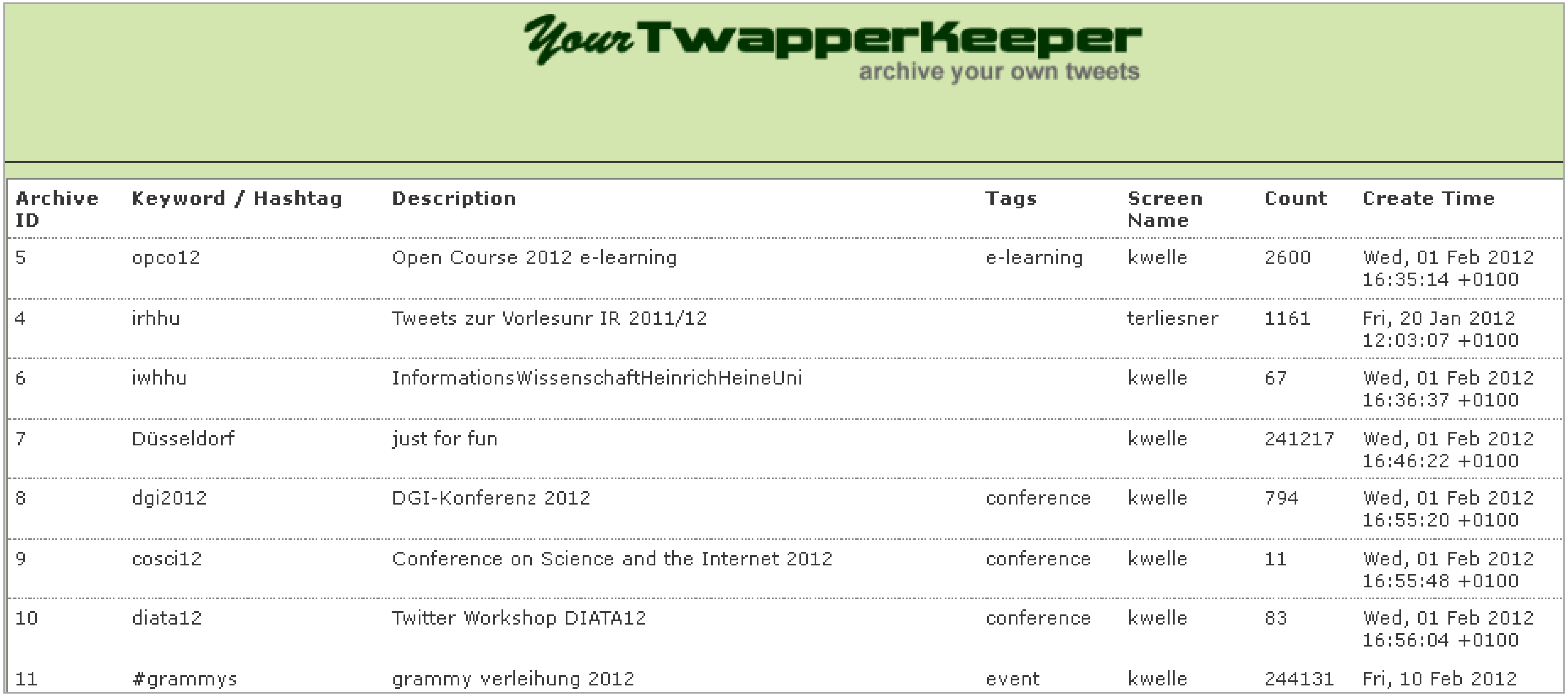
Sebastian Dederichs Piratensebi
Frage: ist es Sexistisch, machotum oder einfach nett, wenn ich einer Frau die Tür aufhalte? #Aufschrei #ernstgemeintefrage
Wed Jan 30 06:27:07 +0000 2013

Yvi teamyvi
#aufschrei die beschneidung von frauen in afrika lässt mich aufschreien. das hier diskutierte ist schlicht primitiv.
Wed Jan 30 06:26:05 +0000 2013

Michael Malucha MichaJavelin
RT @Schopi21: haben wir keine anderen probleme als sexismus? währungskrise,inflation,energiepreisexplosion? #aufschrei
Wed Jan 30 06:24:25 +0000 2013

Prinzessin Putzig Pommetie
Mann, geht mir dieses Opfer-Täter Spielchen gegen den Strich. Irgendwas mit #aufschrei
Wed Jan 30 06:23:59 +0000 2013

YourTwapperkeeper



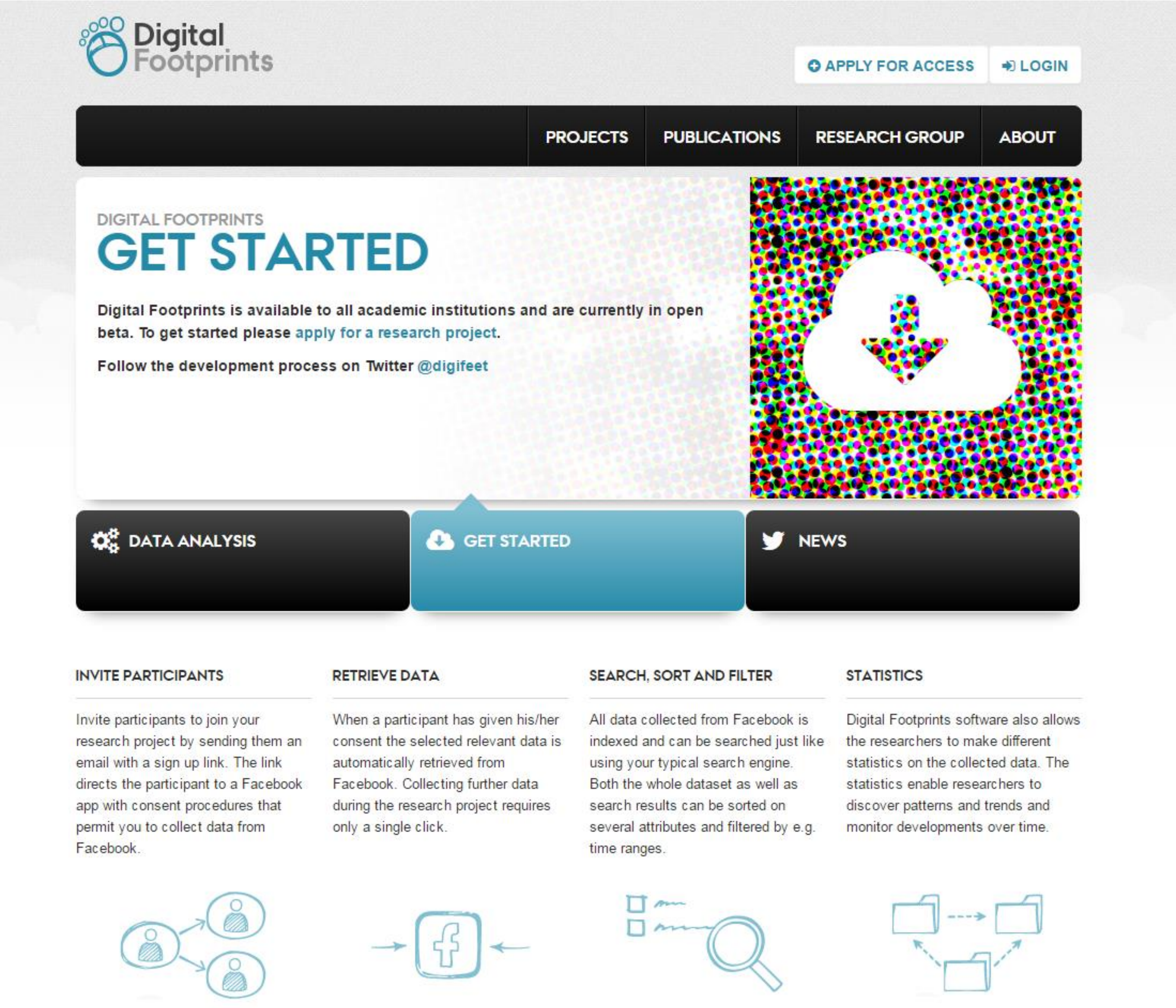
YourTwapperKeeper
archive your own tweets

Archive ID	Keyword / Hashtag	Description	Tags	Screen Name	Count	Create Time
5	opco12	Open Course 2012 e-learning	e-learning	kwelle	2600	Wed, 01 Feb 2012 16:35:14 +0100
4	irhhu	Tweets zur Vorlesunr IR 2011/12		terliesner	1161	Fri, 20 Jan 2012 12:03:07 +0100
6	iwhhu	InformationsWissenschaftHeinrichHeineUni		kwelle	67	Wed, 01 Feb 2012 16:36:37 +0100
7	Düsseldorf	just for fun		kwelle	241217	Wed, 01 Feb 2012 16:46:22 +0100
8	dgi2012	DGI-Konferenz 2012	conference	kwelle	794	Wed, 01 Feb 2012 16:55:20 +0100
9	cosci12	Conference on Science and the Internet 2012	conference	kwelle	11	Wed, 01 Feb 2012 16:55:48 +0100
10	diata12	Twitter Workshop DIATA12	conference	kwelle	83	Wed, 01 Feb 2012 16:56:04 +0100
11	#grammys	grammy verleihung 2012	event	kwelle	244131	Fri, 10 Feb 2012

Bruns, A., & Liang, Y. E. (2012). Tools and methods for capturing Twitter data during natural disasters. First Monday, 17(4). http://doi.org/10.5210/fm.v17i4.3937

http://mappingonlinepublics.net/2011/06/21/switching-from-twapperkeeper-to-yourtwapperkeeper/

Digital Footprints (for Facebook)



DIGITAL FOOTPRINTS

GET STARTED

Digital Footprints is available to all academic institutions and are currently in open beta. To get started please apply for a research project.

Follow the development process on Twitter @digifeet

INVITE PARTICIPANTS

Invite participants to join your research project by sending them an email with a sign up link. The link directs the participant to a Facebook app with consent procedures that permit you to collect data from Facebook.

RETRIEVE DATA

When a participant has given his/her consent the selected relevant data is automatically retrieved from Facebook. Collecting further data during the research project requires only a single click.

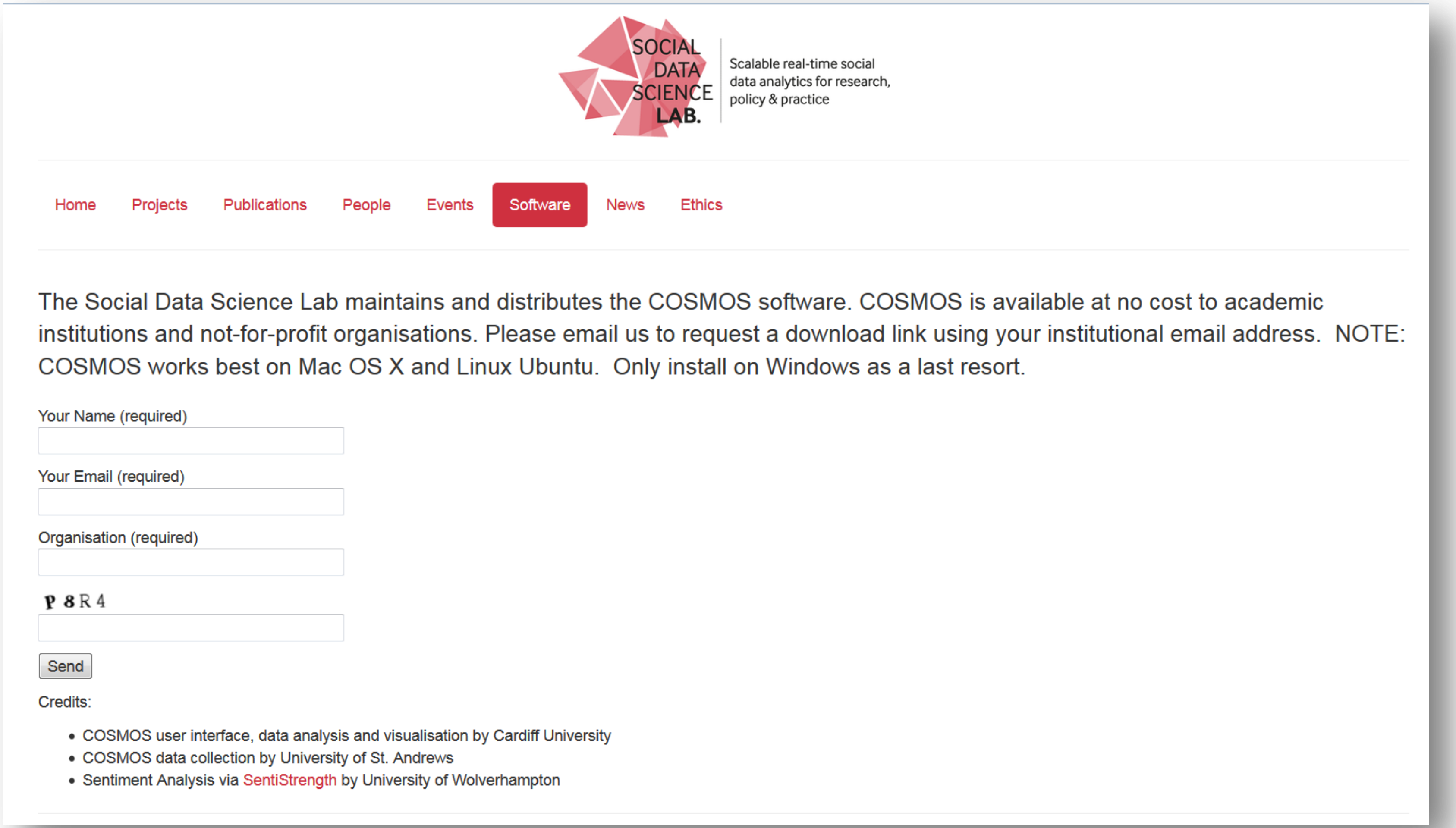
SEARCH, SORT AND FILTER

All data collected from Facebook is indexed and can be searched just like using your typical search engine. Both the whole dataset as well as search results can be sorted on several attributes and filtered by e.g. time ranges.

STATISTICS

Digital Footprints software also allows the researchers to make different statistics on the collected data. The statistics enable researchers to discover patterns and trends and monitor developments over time.

COSMOS



The Social Data Science Lab maintains and distributes the COSMOS software. COSMOS is available at no cost to academic institutions and not-for-profit organisations. Please email us to request a download link using your institutional email address. NOTE: COSMOS works best on Mac OS X and Linux Ubuntu. Only install on Windows as a last resort.

Your Name (required)

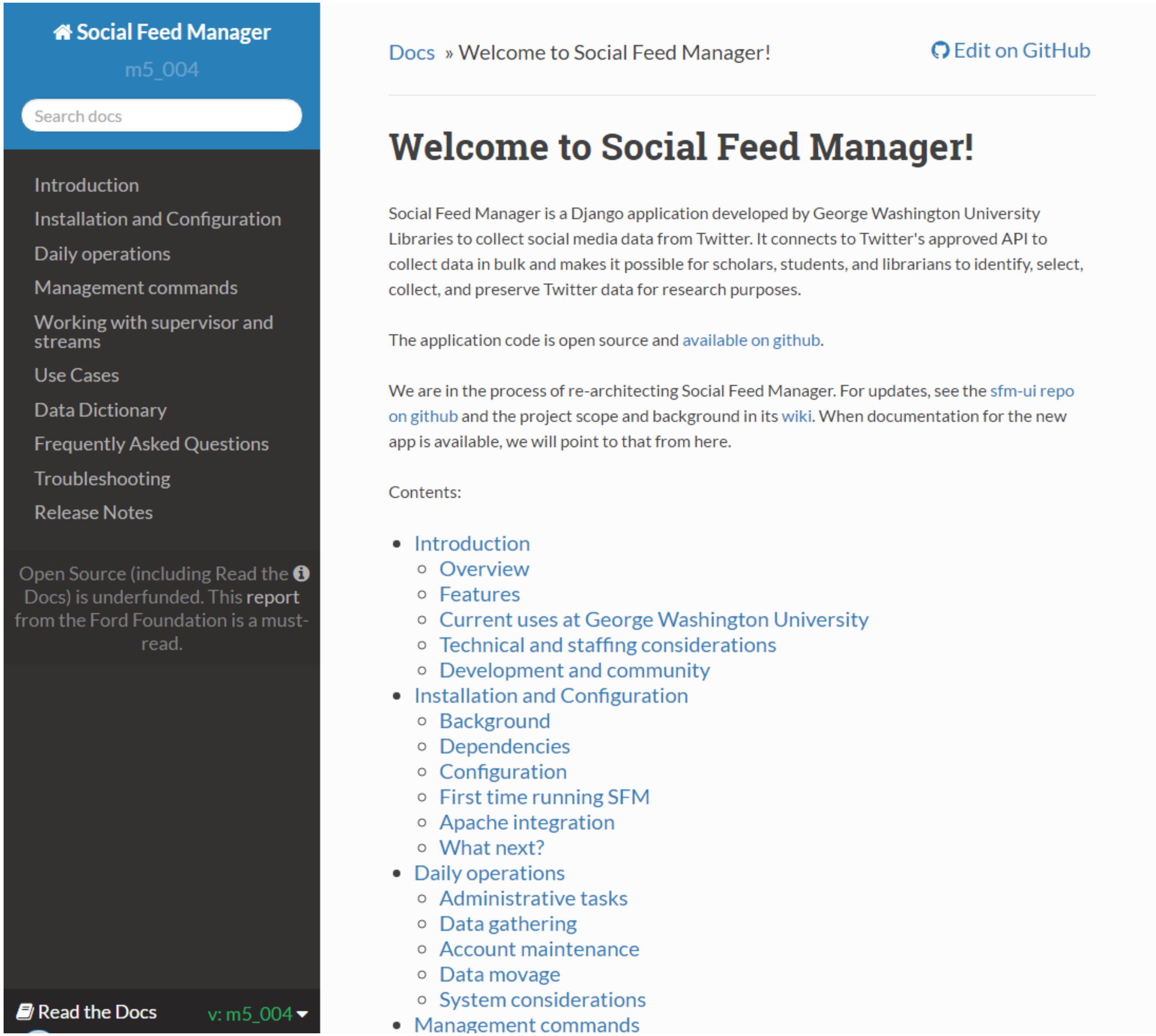
Your Email (required)

Organisation (required)

Credits:

- COSMOS user interface, data analysis and visualisation by Cardiff University
- COSMOS data collection by University of St. Andrews
- Sentiment Analysis via SentiStrength by University of Wolverhampton

Social Feed Manager



Open Source (including Read the Docs) is underfunded. This report from the Ford Foundation is a must-read.

Welcome to Social Feed Manager!

Social Feed Manager is a Django application developed by George Washington University Libraries to collect social media data from Twitter. It connects to Twitter's approved API to collect data in bulk and makes it possible for scholars, students, and librarians to identify, select, collect, and preserve Twitter data for research purposes.

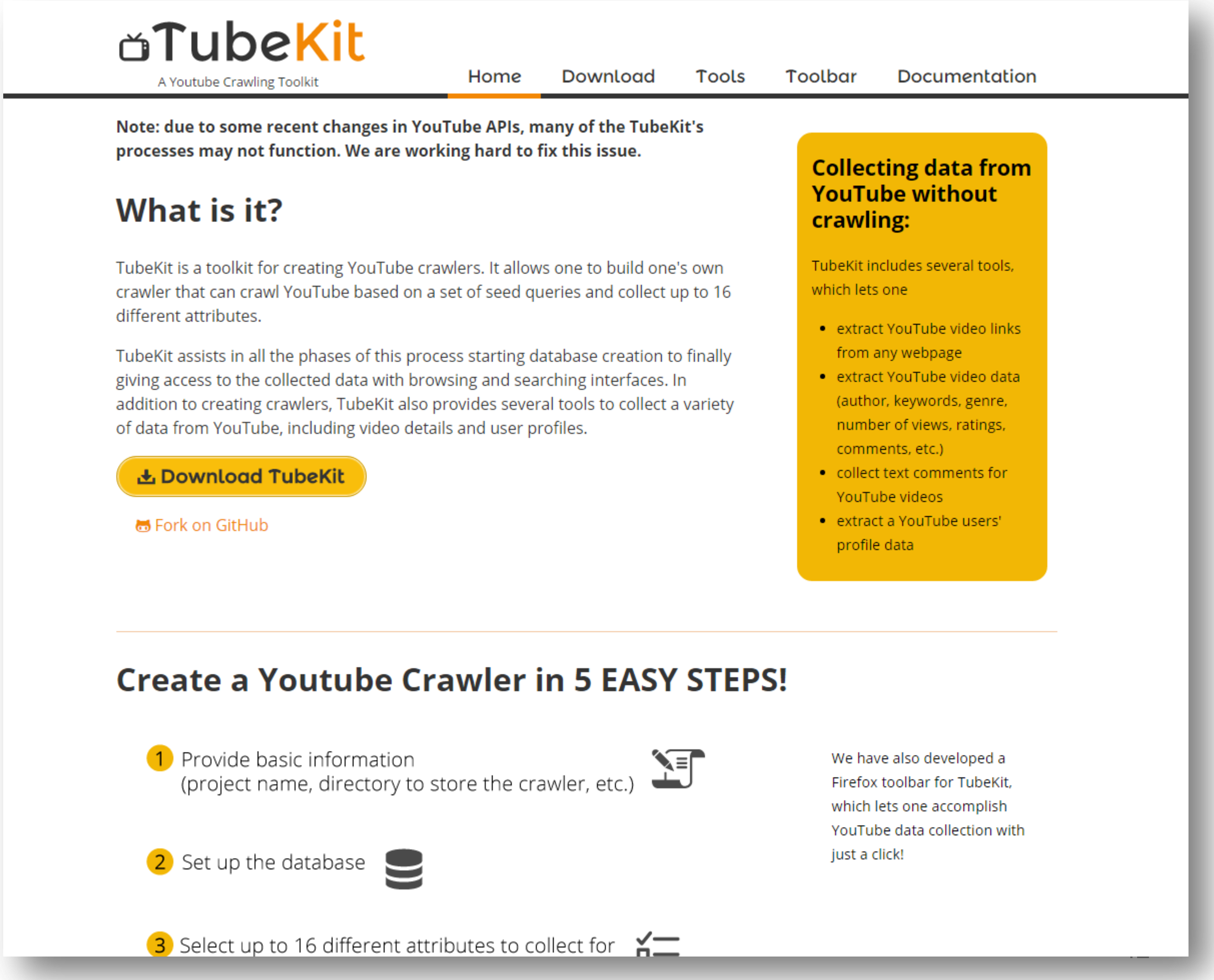
The application code is open source and available on github.

We are in the process of re-architecting Social Feed Manager. For updates, see the sfm-ui repo on github and the project scope and background in its wiki. When documentation for the new app is available, we will point to that from here.

Contents:

gesis Leibniz Institute for the Social Sciences

http://www.tubekit.org/



Note: due to some recent changes in YouTube APIs, many of the TubeKit's processes may not function. We are working hard to fix this issue.

What is it?

TubeKit is a toolkit for creating YouTube crawlers. It allows one to build one's own crawler that can crawl YouTube based on a set of seed queries and collect up to 16 different attributes.

TubeKit assists in all the phases of this process starting database creation to finally giving access to the collected data with browsing and searching interfaces. In addition to creating crawlers, TubeKit also provides several tools to collect a variety of data from YouTube, including video details and user profiles.

Collecting data from YouTube without crawling:

TubeKit includes several tools, which lets one

- extract YouTube video links from any webpage
- extract YouTube video data (author, keywords, genre, number of views, ratings, comments, etc.)
- collect text comments for YouTube videos
- extract a YouTube users' profile data

Create a Youtube Crawler in 5 EASY STEPS!

1 Provide basic information (project name, directory to store the crawler, etc.)

2 Set up the database

3 Select up to 16 different attributes to collect for

We have also developed a Firefox toolbar for TubeKit, which lets one accomplish YouTube data collection with just a click!

Manual data collection
Handling multimedia data?

Explore options for capturing and managing data

- Evernote
- Zotero
- Content Analysis Software.

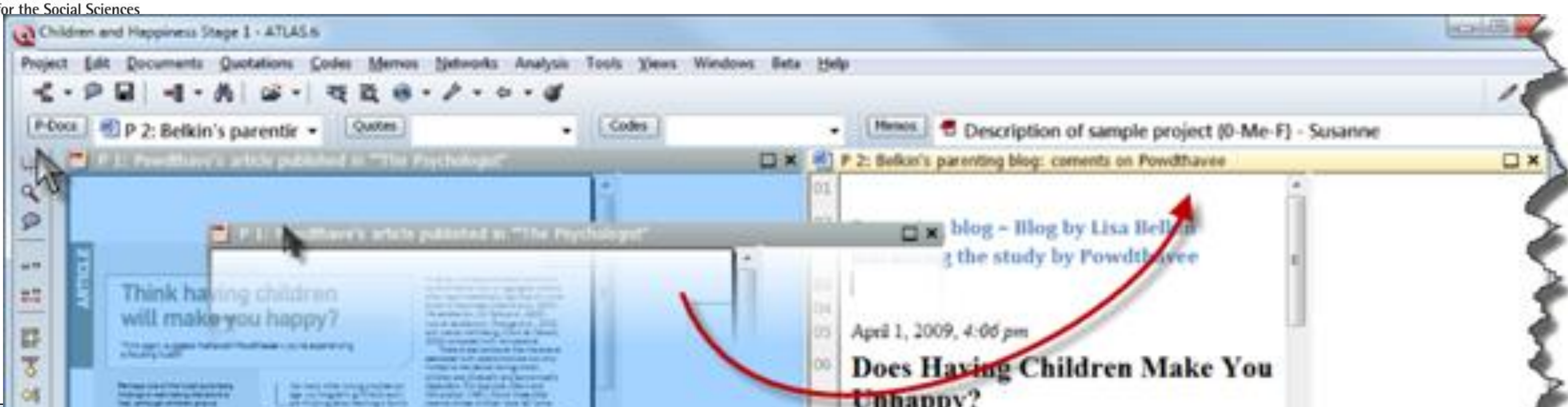
Content Analysis

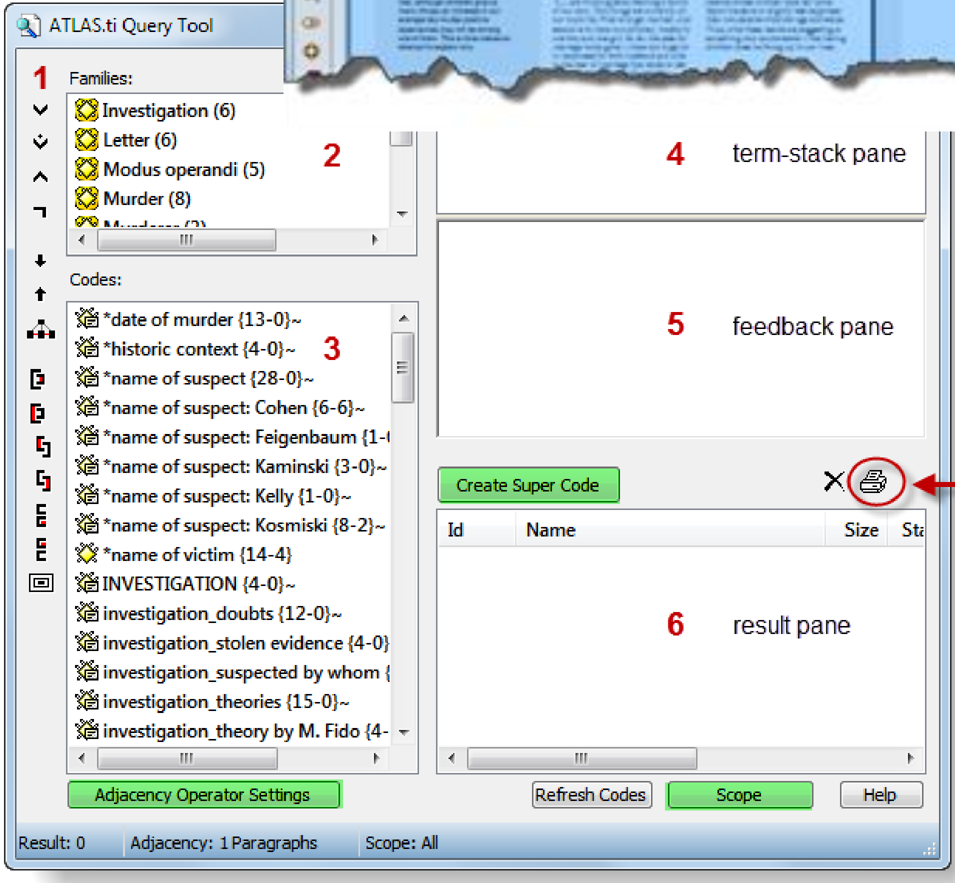
CAQDAS Computer-Assisted Qualitative Data AnalysiS

- (e.g. MAXQDA, QDAMiner, ATLAS.ti, Qualrus, Nvivo)
- Speech Act Analyse
- Statistics

- Some of them now come with import options for social media data!

ATLAS.ti


Children and Happiness Stage 1 - ATLAS.ti
Project Edit Documents Quotations Codes Memos Networks Analysis Tools Views Windows Beta Help
P-Docs
P 2: Belkin's parentir
Quotes
Codes
Memos
Description of sample project (0-Me-F) - Susanne
P 1: Powdthavee's article published in "The Psychologist"
P 2: Belkin's parenting blog: coments on Powdthavee
Think having children will make you happy?
blog – Blog by Lisa Belkin
the study by Powdthavee
April 1, 2009, 4:06 pm
Does Having Children Make You Unhappy?
By LISA BELKIN



ATLAS.ti Query Tool

1 Families:

- Investigation (6)
- Letter (6)
- Modus operandi (5)
- Murder (8)
- Murderer (2)

2

Codes:

- *date of murder {13-0}~
- *historic context {4-0}~
- *name of suspect {28-0}~
- *name of suspect: Cohen {6-6}~
- *name of suspect: Feigenbaum {1-
- *name of suspect: Kaminski {3-0}~
- *name of suspect: Kelly {1-0}~
- *name of suspect: Kosmiski {8-2}~
- *name of victim {14-4}
- INVESTIGATION {4-0}~
- investigation_doubts {12-0}~
- investigation_stolen evidence {4-0}
- investigation_suspected by whom {
- investigation_theories {15-0}~
- investigation_theory by M. Fido {4-

3

4 term-stack pane

5 feedback pane

6 result pane

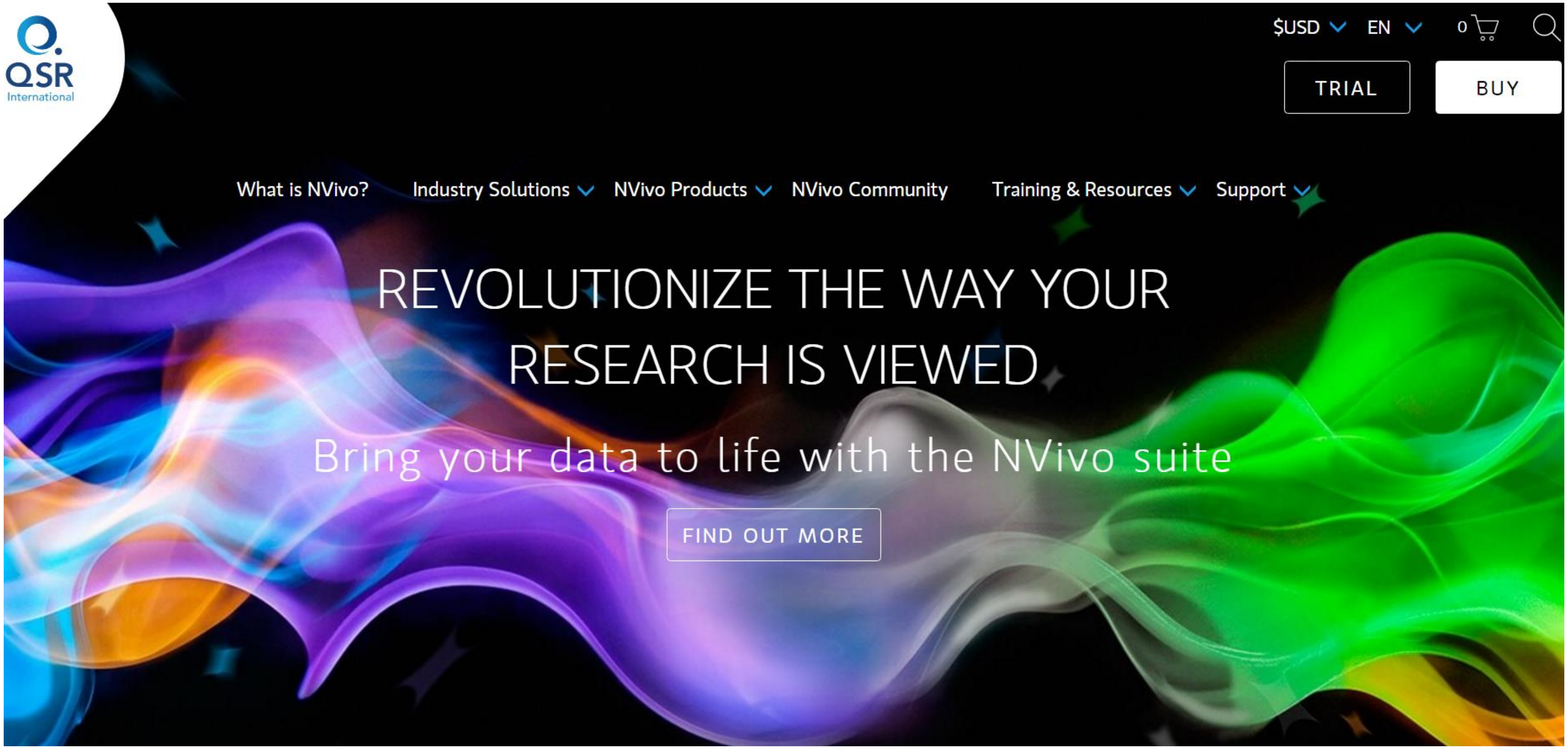
Id	Name	Size	Sta

Result: 0 | Adjacency: 1 Paragraphs | Scope: All

output (view, save or print reports)

NVIVO

- Also able to import tweets.



REVOLUTIONIZE THE WAY YOUR RESEARCH IS VIEWED

Bring your data to life with the NVivo suite

http://www.qsrinternational.com/

Data analysis

Data analysis

Network analysis

Sentiment analysis

Visualization

Content analysis

(Crowdsourcing)

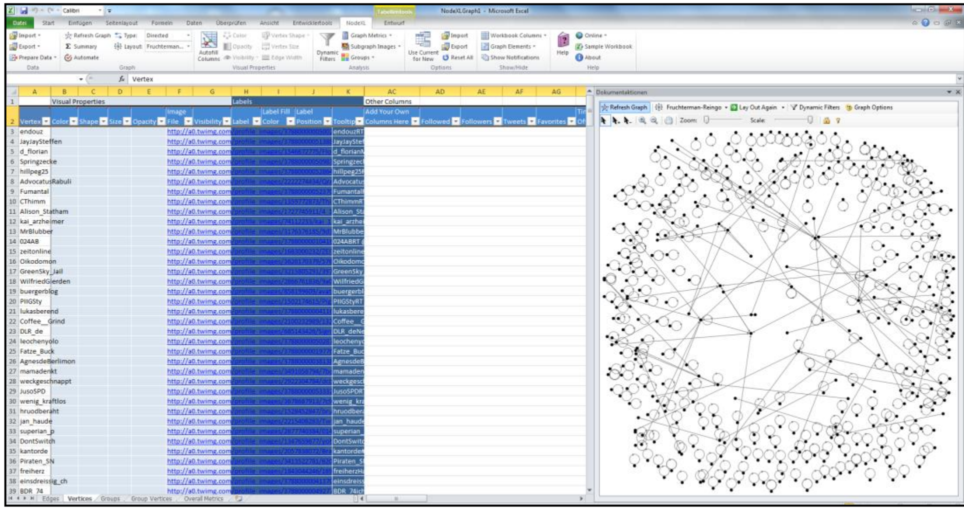
Network Analysis

gesis Leibniz Institute for the Social Sciences

NodeXL

Network Analysis

Based on Excel

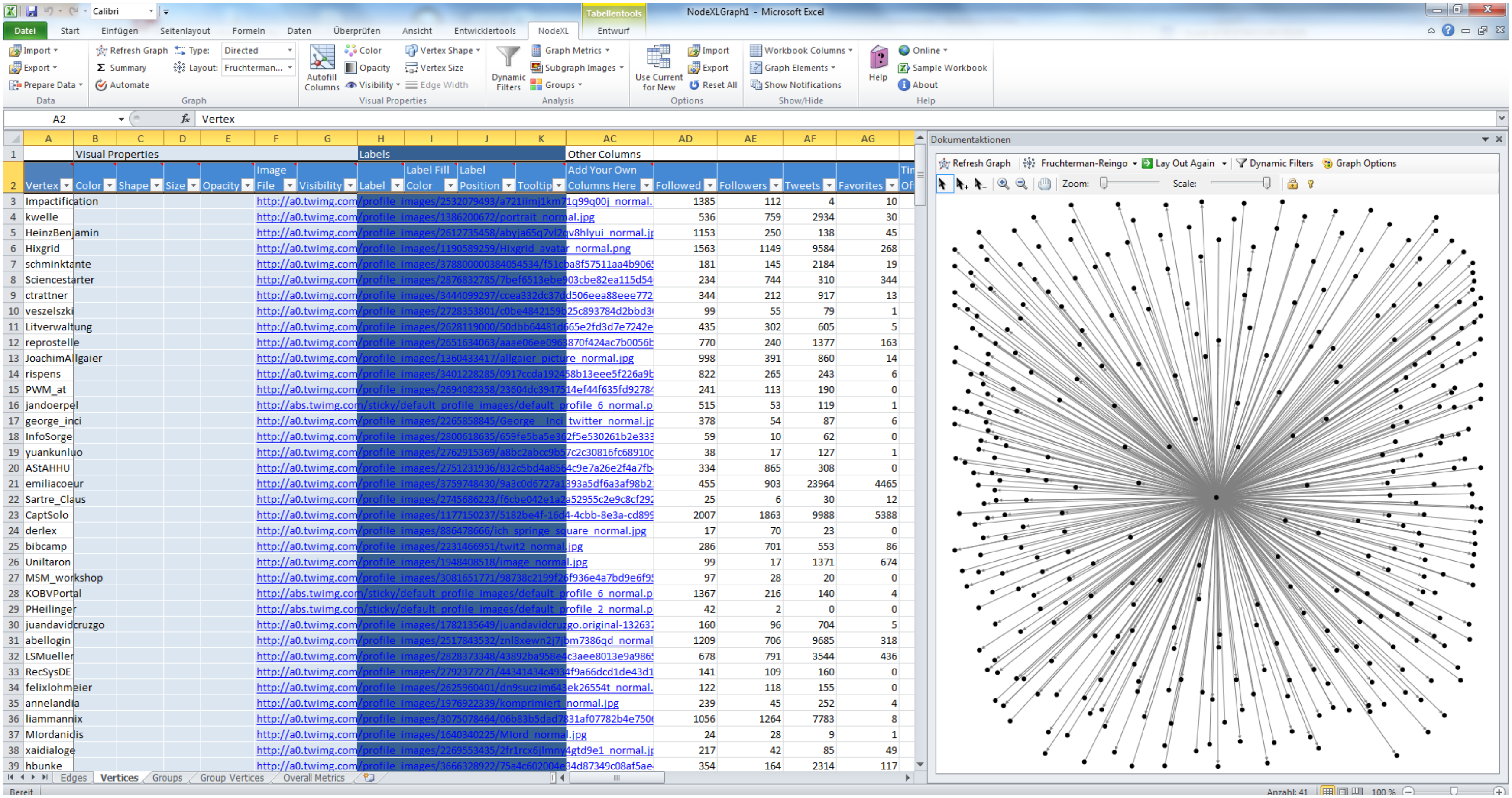


	A	B	C	D	E	F	G	H	I	J	K	AC
1		Visual Properties						Labels				Other Columns
2	Vertex	Color	Shape	Size	Opacity	Image File	Visibility	Label	Label Fill Color	Label Position	Tooltip	Add Your Own Columns Here
3	endouz					http://a0.twimg.com/profile_images/[illegible]					endouzRT	
4	JayJaySteffen					http://a0.twimg.com/profile_images/[illegible]					JayJaySte	
5	d_florian					http://a0.twimg.com/profile_images/[illegible]					d_florian	
6	Springzecke					http://a0.twimg.com/profile_images/[illegible]					Springzec	
7	hillpeg25					http://a0.twimg.com/profile_images/[illegible]					hillpeg25	
8	AdvocatusRabuli					http://a0.twimg.com/profile_images/[illegible]					Advocatus	
9	Fumantal					http://a0.twimg.com/profile_images/[illegible]					Fumantal	
10	CThimm					http://a0.twimg.com/profile_images/[illegible]					CThimmR	
11	Alison_Statham					http://a0.twimg.com/profile_images/[illegible]					Alison_St	
12	kai_arzheimer					http://a0.twimg.com/profile_images/[illegible]					kai_arzhe	
13	MrBlubber					http://a0.twimg.com/profile_images/[illegible]					MrBlubbe	
14	024AB					http://a0.twimg.com/profile_images/[illegible]					024ABRT @	
15	zeitonline					http://a0.twimg.com/profile_images/[illegible]					zeitonline	
16	Oikodomon					http://a0.twimg.com/profile_images/[illegible]					Oikodomo	
17	GreenSky_Jail					http://a0.twimg.com/profile_images/[illegible]					GreenSky	
18	WilfriedGierden					http://a0.twimg.com/profile_images/[illegible]					WilfriedG	
19	buergerblog					http://a0.twimg.com/profile_images/[illegible]					buergerbl	
20	PIIGSty					http://a0.twimg.com/profile_images/[illegible]					PIIGStyRT	
21	lukasberend					http://a0.twimg.com/profile_images/[illegible]					lukasbere	
22	Coffee__Grind					http://a0.twimg.com/profile_images/[illegible]					Coffee__G	
23	DLR_de					http://a0.twimg.com/profile_images/[illegible]					DLR_deNe	
24	leochenyolo					http://a0.twimg.com/profile_images/[illegible]					leochenyo	
25	Fatze_Buck					http://a0.twimg.com/profile_images/[illegible]					Fatze_Buc	
26	AgnesdeBerlimon					http://a0.twimg.com/profile_images/[illegible]					AgnesdeB	
27	mamadenkt					http://a0.twimg.com/profile_images/[illegible]					mamaden	
28	weckgeschnappt					http://a0.twimg.com/profile_images/[illegible]					weckgesc	
29	JusoSPD					http://a0.twimg.com/profile_images/[illegible]					JusoSPDR	
30	wenig_kraftlos					http://a0.twimg.com/profile_images/[illegible]					wenig_kra	
31	hruodberaht					http://a0.twimg.com/profile_images/[illegible]					hruodbera	
32	jan_haude					http://a0.twimg.com/profile_images/[illegible]					jan_haude	
33	superian_p					http://a0.twimg.com/profile_images/[illegible]					superian_	
34	DontSwitch					http://a0.twimg.com/profile_images/[illegible]					DontSwitc	
35	kantorde					http://a0.twimg.com/profile_images/[illegible]					kantorde	
36	Piraten_SN					http://a0.twimg.com/profile_images/[illegible]					Piraten_S	
37	freiherz					http://a0.twimg.com/profile_images/[illegible]					freiherzH	
38	einsdreissig_ch					http://a0.twimg.com/profile_images/[illegible]					einsdreis	
39	BDR_74					http://a0.twimg.com/profile_images/[illegible]					BDR_74ich	

Can collect data from (amongst others):

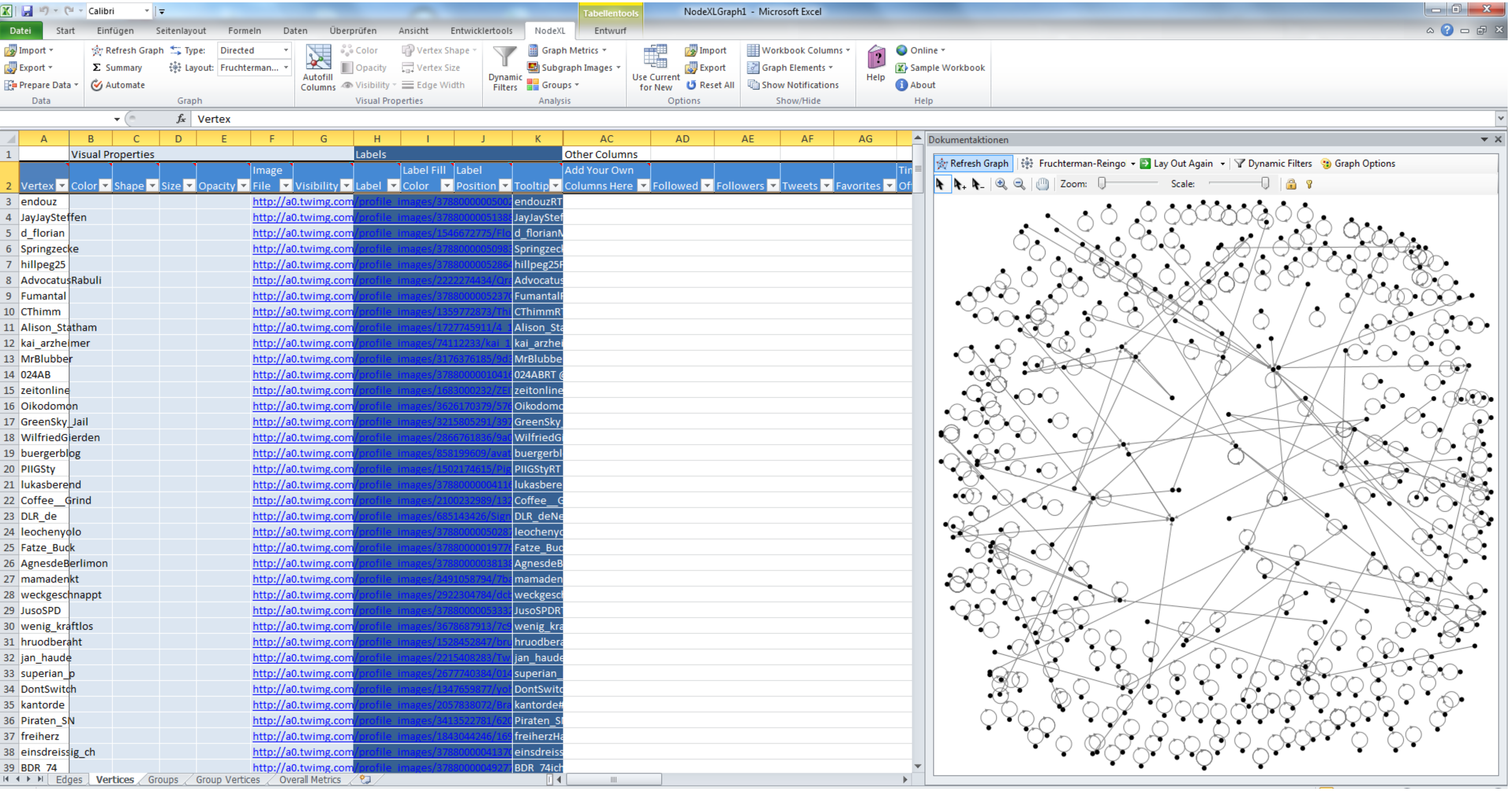
- Facebook
- Twitter
- YouTube
- Flickr
- Wikipedia
- If you can get data into excel, you can get it into NodeXL…

NodeXL



	A	B	C	D	E	F	G	H	I	J	K	AC	AD	AE	AF	AG
1		Visual Properties						Labels				Other Columns				
2	Vertex	Color	Shape	Size	Opacity	Image File	Visibility	Label	Label Fill Color	Label Position	Tooltip	Add Your Own Columns Here	Followed	Followers	Tweets	Favorites
3	Impactification					http://a0.twimg.com/profile_images/2532079493/a721iimj1km71q99q00j_normal.							1385	112	4	10
4	kwelle					http://a0.twimg.com/profile_images/1386200672/portrait_normal.jpg							536	759	2934	30
5	HeinzBenjamin					http://a0.twimg.com/profile_images/2612735458/abyja65q7vl2qv8hlyui_normal.jp							1153	250	138	45
6	Hixgrid					http://a0.twimg.com/profile_images/1190589259/Hixgrid_avatar_normal.png							1563	1149	9584	268
7	schminktante					http://a0.twimg.com/profile_images/378800000384054534/f51cba8f57511aa4b9065							181	145	2184	19
8	Sciencestarter					http://a0.twimg.com/profile_images/2876832785/7bef6513ebe903cbe82ea115d54							234	744	310	344
9	ctrattner					http://a0.twimg.com/profile_images/3444099297/ccea332dc37dd506eea88eee772							344	212	917	13
10	veszelszki					http://a0.twimg.com/profile_images/2728353801/c0be4842159b25c893784d2bbd3							99	55	79	1
11	Litverwaltung					http://a0.twimg.com/profile_images/2628119000/50dbb64481d665e2fd3d7e7242e							435	302	605	5
12	reprostelle					http://a0.twimg.com/profile_images/2651634063/aaae06ee0963870f424ac7b0056b							770	240	1377	163
13	JoachimAllgaier					http://a0.twimg.com/profile_images/1360433417/allgaier_picture_normal.jpg							998	391	860	14
14	rispens					http://a0.twimg.com/profile_images/3401228285/0917ccda192458b13eee5f226a9b							822	265	243	6
15	PWM_at					http://a0.twimg.com/profile_images/2694082358/23604dc3947514ef44f635fd92784							241	113	190	0
16	jandoerpel					http://abs.twimg.com/sticky/default_profile_images/default_profile_6_normal.p							515	53	119	1
17	george_inci					http://a0.twimg.com/profile_images/2265858845/George__Inci_twitter_normal.jp							378	54	87	6
18	InfoSorge					http://a0.twimg.com/profile_images/2800618635/659fe5ba5e362f5e530261b2e333							59	10	62	0
19	yuankunluo					http://a0.twimg.com/profile_images/2762915369/a8bc2abcc9b57c2c30816fc68910c							38	17	127	1
20	AStAHHU					http://a0.twimg.com/profile_images/2751231936/832c5bd4a8564c9e7a26e2f4a7fb							334	865	308	0
21	emiliacoeur					http://a0.twimg.com/profile_images/3759748430/9a3c0d6727a1393a5df6a3af98b2							455	903	23964	4465
22	Sartre_Claus					http://a0.twimg.com/profile_images/2745686223/f6cbe042e1a2a52955c2e9c8cf292							25	6	30	12
23	CaptSolo					http://a0.twimg.com/profile_images/1177150237/5182be4f-16d4-4cbb-8e3a-cd899							2007	1863	9988	5388
24	derlex					http://a0.twimg.com/profile_images/886478666/ich_springe_square_normal.jpg							17	70	23	0
25	bibcamp					http://a0.twimg.com/profile_images/2231466951/twit2_normal.jpg							286	701	553	86
26	Uniltaron					http://a0.twimg.com/profile_images/1948408518/image_normal.jpg							99	17	1371	674
27	MSM_workshop					http://a0.twimg.com/profile_images/3081651771/98738c2199f26f936e4a7bd9e6f9							97	28	20	0
28	KOBVPortal					http://abs.twimg.com/sticky/default_profile_images/default_profile_6_normal.p							1367	216	140	4
29	PHeilinger					http://abs.twimg.com/sticky/default_profile_images/default_profile_2_normal.p							42	2	0	0
30	juandavidcruzgo					http://a0.twimg.com/profile_images/1782135649/juandavidcruzgo.original-132637							160	96	704	5
31	abellogin					http://a0.twimg.com/profile_images/2517843532/znl8xewn2j7jbm7386qd_normal							1209	706	9685	318
32	LSMueller					http://a0.twimg.com/profile_images/2828373348/43892ba958e4c3aee8013e9a986							678	791	3544	436
33	RecSysDE					http://a0.twimg.com/profile_images/2792377271/44341434c4934f9a66dcd1de43d1							141	109	160	0
34	felixlohmeier					http://a0.twimg.com/profile_images/2625960401/dn9suczim643ek26554t_normal.							122	118	155	0
35	annelandia					http://a0.twimg.com/profile_images/1976922339/komprimiert_normal.jpg							239	45	252	4
36	liammannix					http://a0.twimg.com/profile_images/3075078464/06b83b5dad7831af07782b4e750							1056	1264	7783	8
37	MIordanidis					http://a0.twimg.com/profile_images/1640340225/MIord_normal.jpg							24	28	9	1
38	xaidialoge					http://a0.twimg.com/profile_images/2269553435/2fr1rcx6jlmny4gtd9e1_normal.jp							217	42	85	49
39	hbunke					http://a0.twimg.com/profile_images/3666328922/75a4c602004e34d87349c08af5ae							354	164	2314	117

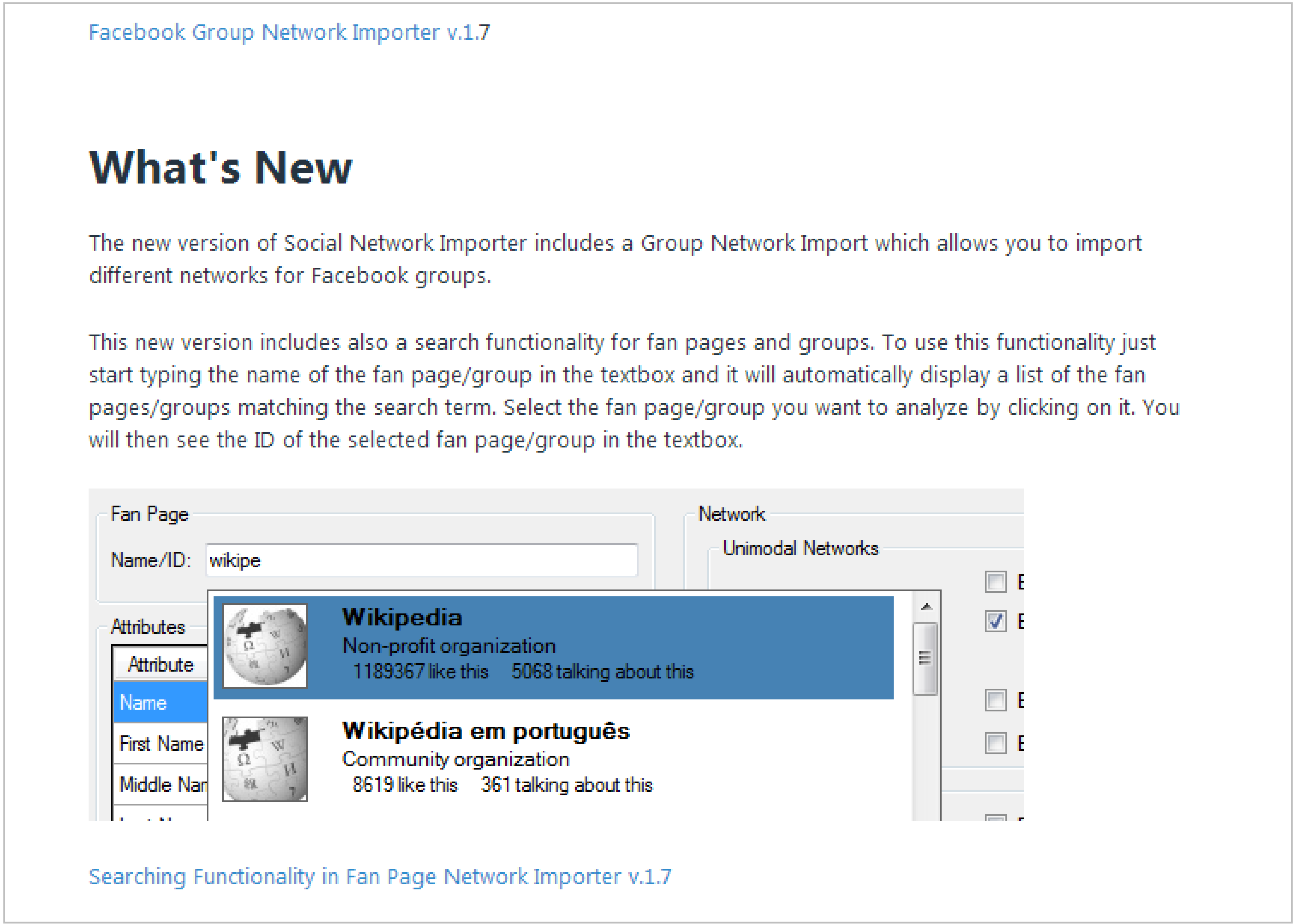
NodeXL



	A	B	C	D	E	F	G	H	I	J	K	AC	AD	AE	AF	AG	
1		Visual Properties						Labels				Other Columns					
2	Vertex	Color	Shape	Size	Opacity	Image File	Visibility	Label	Label Fill Color	Label Position	Tooltip	Add Your Own Columns Here	Followed	Followers	Tweets	Favorites	Ti Of
3	endouz					http://a0.twimg.com/profile_images/3788000005002					endouzRT						
4	JayJaySteffen					http://a0.twimg.com/profile_images/3788000051388					JayJayStef						
5	d_florian					http://a0.twimg.com/profile_images/1546672775/Flo					d_florianM						
6	Springzecke					http://a0.twimg.com/profile_images/3788000050983					Springzeck						
7	hillpeg25					http://a0.twimg.com/profile_images/3788000052864					hillpeg25R						
8	AdvocatusRabuli					http://a0.twimg.com/profile_images/2222274434/Qra					Advocatus						
9	Fumantal					http://a0.twimg.com/profile_images/3788000052370					FumantalR						
10	CThimm					http://a0.twimg.com/profile_images/1359772873/Thi					CThimmRT						
11	Alison_Statham					http://a0.twimg.com/profile_images/1727745911/4_1					Alison_Sta						
12	kai_arzheimer					http://a0.twimg.com/profile_images/74112233/kai_1					kai_arzhei						
13	MrBlubber					http://a0.twimg.com/profile_images/3176376185/9d3					MrBlubbe						
14	024AB					http://a0.twimg.com/profile_images/3788000010416					024ABRT @						
15	zeitonline					http://a0.twimg.com/profile_images/1683000232/ZEI					zeitonline						
16	Oikodomon					http://a0.twimg.com/profile_images/3626170379/576					Oikodomo						
17	GreenSky_Jail					http://a0.twimg.com/profile_images/3215805291/397					GreenSky_						
18	WilfriedGierden					http://a0.twimg.com/profile_images/2866761836/9a0					WilfriedGi						
19	buergerblog					http://a0.twimg.com/profile_images/858199609/avat					buergerbl						
20	PIIGSty					http://a0.twimg.com/profile_images/1502174615/Pig					PIIGStyRT						
21	lukasberend					http://a0.twimg.com/profile_images/3788000004116					lukasbere						
22	Coffee__Grind					http://a0.twimg.com/profile_images/2100232989/132					Coffee__G						
23	DLR_de					http://a0.twimg.com/profile_images/685143426/Sign					DLR_deNe						
24	leochenyolo					http://a0.twimg.com/profile_images/3788000050287					leochenyo						
25	Fatze_Buck					http://a0.twimg.com/profile_images/3788000019776					Fatze_Buc						
26	AgnesdeBerlimon					http://a0.twimg.com/profile_images/3788000038138					AgnesdeB						
27	mamadenkt					http://a0.twimg.com/profile_images/3491058794/7ba					mamaden						
28	weckgeschnappt					http://a0.twimg.com/profile_images/2922304784/dcb					weckgesch						
29	JusoSPD					http://a0.twimg.com/profile_images/3788000053332					JusoSPDRT						
30	wenig_kraftlos					http://a0.twimg.com/profile_images/3678687913/7c9					wenig_kra						
31	hruodberaht					http://a0.twimg.com/profile_images/1528452847/bru					hruodbera						
32	jan_haude					http://a0.twimg.com/profile_images/2215408283/Tw					jan_haude						
33	superian_p					http://a0.twimg.com/profile_images/2677740384/014					superian_						
34	DontSwitch					http://a0.twimg.com/profile_images/1347659877/yoh					DontSwitc						
35	kantorde					http://a0.twimg.com/profile_images/2057838072/Bra					kantorde#						
36	Piraten_SN					http://a0.twimg.com/profile_images/3413522781/620					Piraten_SI						
37	freiherz					http://a0.twimg.com/profile_images/1843044246/169					freiherzHa						
38	einsdreissig_ch					http://a0.twimg.com/profile_images/3788000041370					einsdreiss						
39	BDR_74					http://a0.twimg.com/profile_images/3788000049277					BDR_74ich						

NodeXL: SocialNetImporter

http://socialnetimporter.codeplex.com/



Facebook Group Network Importer v.1.7

What's New

The new version of Social Network Importer includes a Group Network Import which allows you to import different networks for Facebook groups.

This new version includes also a search functionality for fan pages and groups. To use this functionality just start typing the name of the fan page/group in the textbox and it will automatically display a list of the fan pages/groups matching the search term. Select the fan page/group you want to analyze by clicking on it. You will then see the ID of the selected fan page/group in the textbox.

Name/ID: wikipe

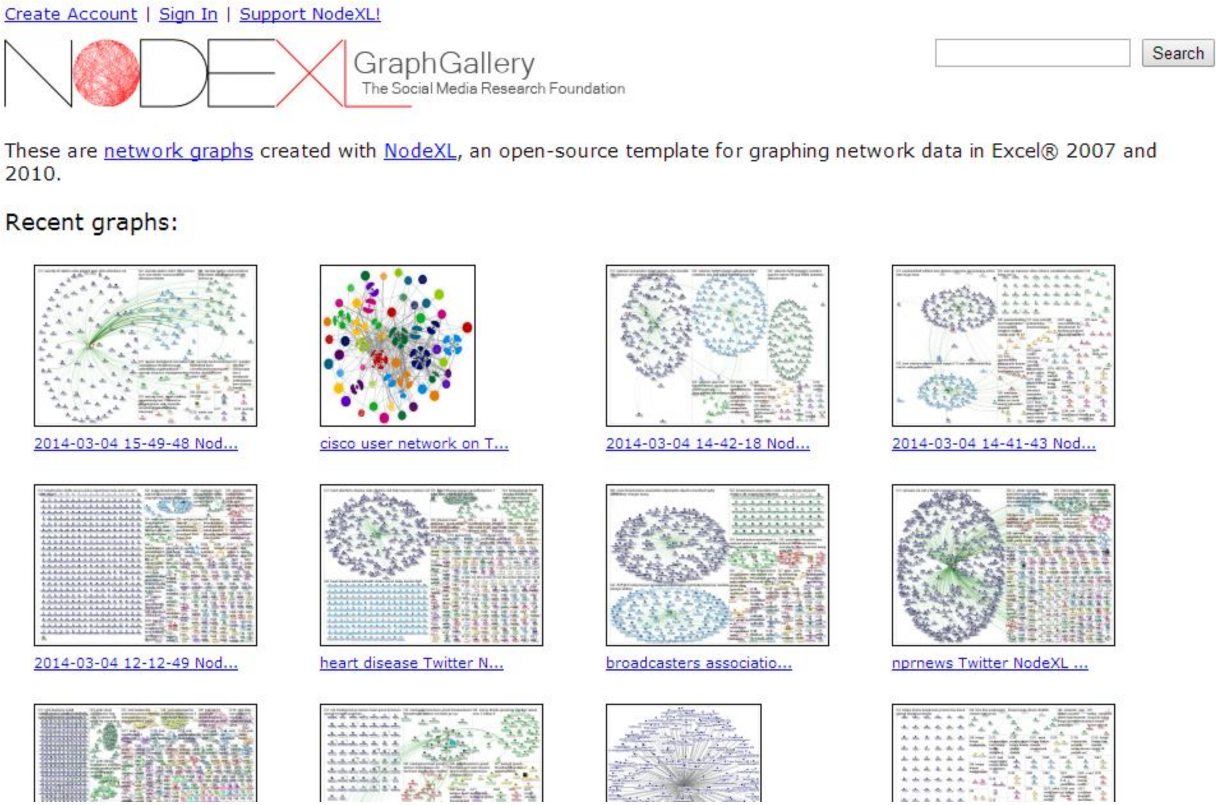
Wikipedia
Non-profit organization
1189367 like this 5068 talking about this

Wikipédia em português
Community organization
8619 like this 361 talking about this

Searching Functionality in Fan Page Network Importer v.1.7

NodeXL

- NodeXLGraphGallery.org



Create Account | Sign In | Support NodeXL!

These are network graphs created with NodeXL, an open-source template for graphing network data in Excel® 2007 and 2010.

Recent graphs:

2014-03-04 15-49-48 Nod...

cisco user network on T...

2014-03-04 14-42-18 Nod...

2014-03-04 14-41-43 Nod...

2014-03-04 12-12-49 Nod...

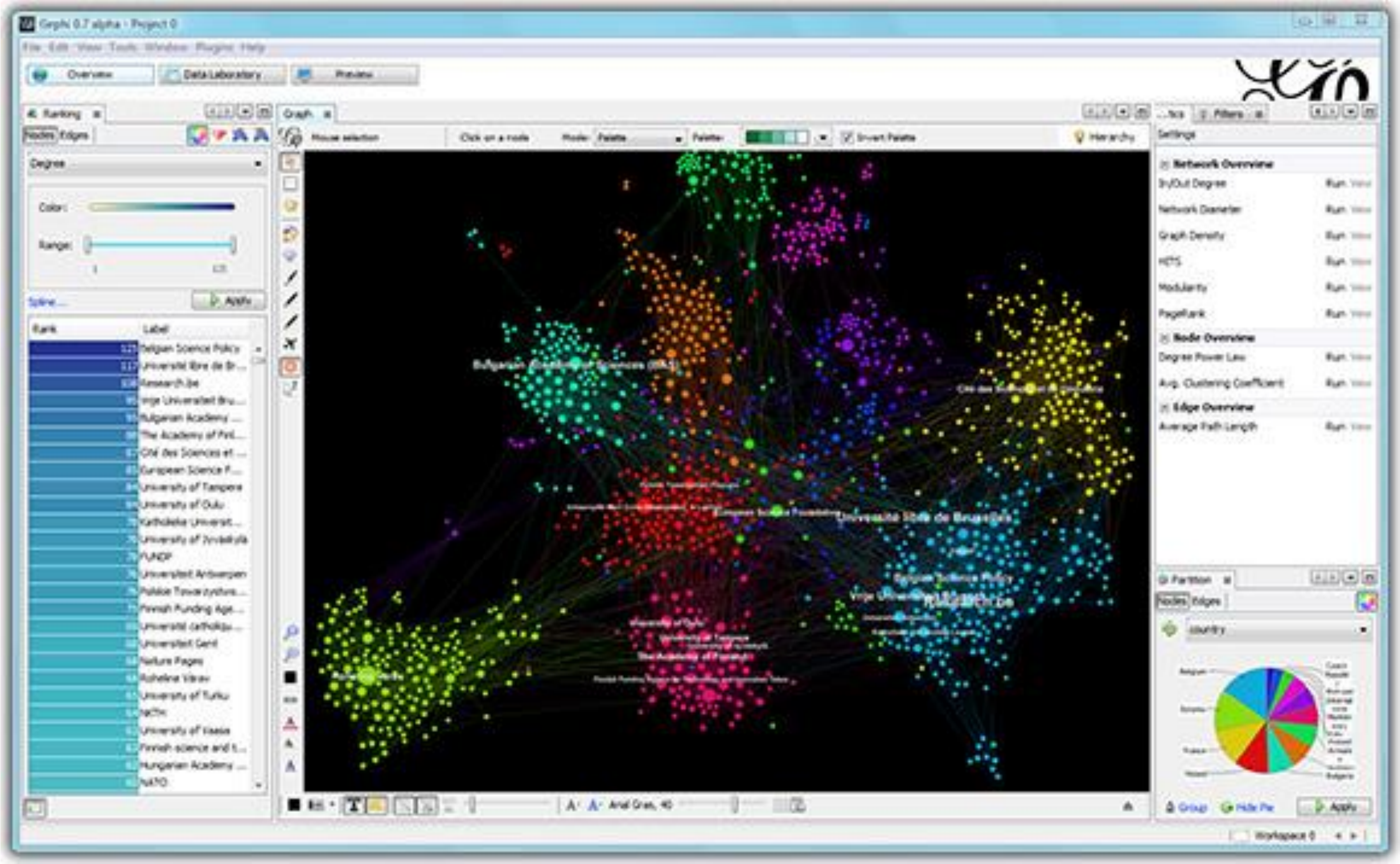
heart disease Twitter N...

broadcasters associatio...

nprnews Twitter NodeXL ...

Visualization

Gephi



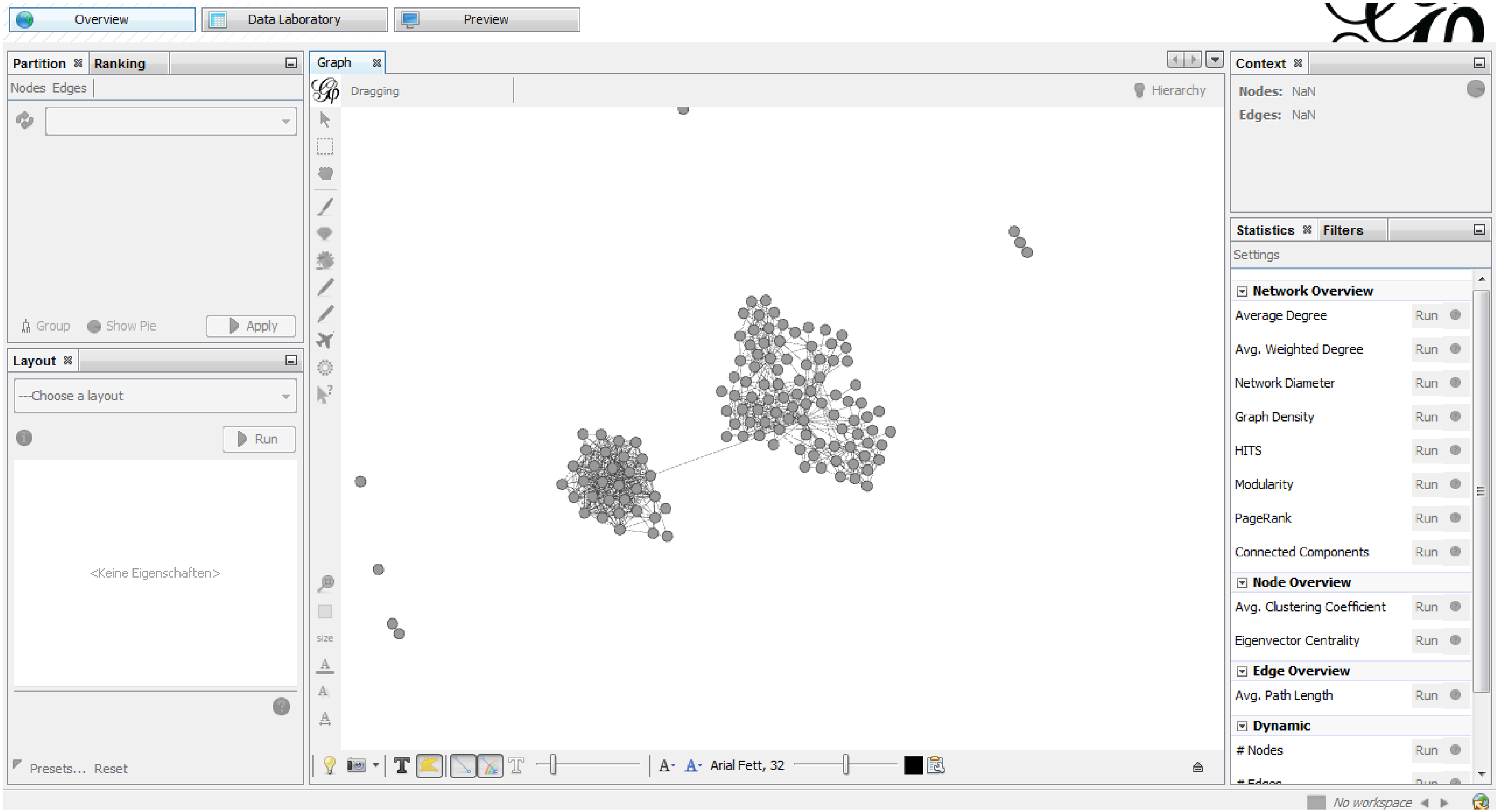
Rank	Label
125	Belgian Science Policy
117	Université libre de Br...
108	Research.be
95	Vrije Universiteit Bru...
95	Bulgarian Academy ...
88	The Academy of Finl...
87	Cité des Sciences et ...
85	European Science F...
84	University of Tampere
[illegible]	University of Oulu
79	Katholieke Universit...
79	University of Jyväskylä
79	FUNDP
78	Universiteit Antwerpen
76	Polskie Towarzystwo...
[illegible]	Finnish Funding Age...
[illegible]	Université catholiqu...
[illegible]	Universiteit Gent
[illegible]	Nature Pages
[illegible]	Rohelina Varav
[illegible]	University of Turku
[illegible]	NCTH
[illegible]	University of Vaasa
[illegible]	Finnish science and t...
[illegible]	Hungarian Academy ...
[illegible]	NATO

Network Overview
In/Out Degree Run View
Network Diameter Run View
Graph Density Run View
HITS Run View
Modularity Run View
PageRank Run View

Node Overview
Degree Power Law Run View
Avg. Clustering Coefficient Run View

Edge Overview
Average Path Length Run View

Gephi



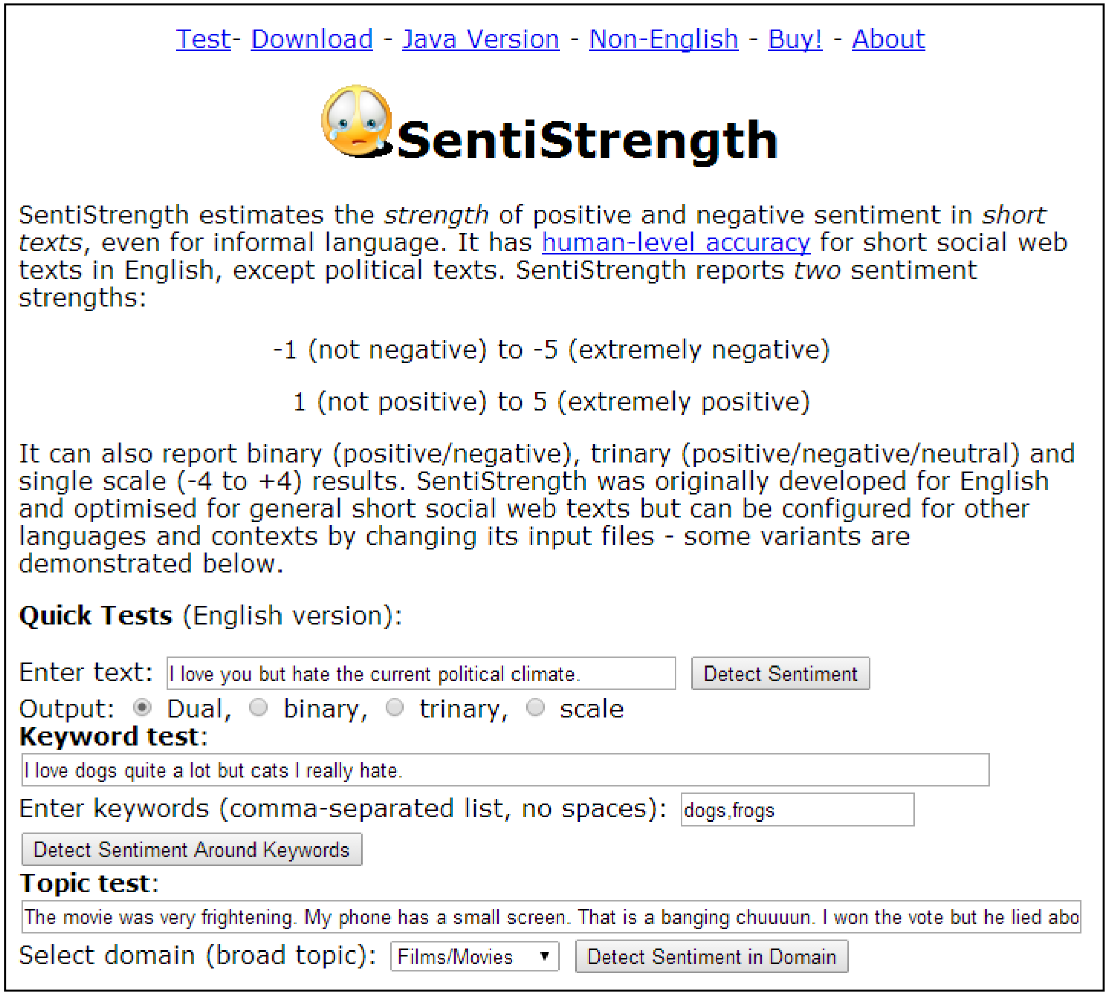
Nodes: NaN
Edges: NaN

Sentiment analysis

NLP / Sentiment Analysis

- Discover text: http://discovertext.com/
- Twitter NLP: http://www.ark.cs.cmu.edu/TweetNLP/
- Internet Community Text Analyzer (ICTA) http://textanalytics.net
- Twitter NLP and Part-of-Speech Tagging: http://www.ark.cs.cmu.edu/TweetNLP/
- SentiStrength http://sentistrength.wlv.ac.uk/

Sentiment Analysis



Test- Download - Java Version - Non-English - Buy! - About

SentiStrength

SentiStrength estimates the *strength* of positive and negative sentiment in *short texts*, even for informal language. It has human-level accuracy for short social web texts in English, except political texts. SentiStrength reports *two* sentiment strengths:

-1 (not negative) to -5 (extremely negative)

1 (not positive) to 5 (extremely positive)

It can also report binary (positive/negative), trinary (positive/negative/neutral) and single scale (-4 to +4) results. SentiStrength was originally developed for English and optimised for general short social web texts but can be configured for other languages and contexts by changing its input files - some variants are demonstrated below.

Quick Tests (English version):

Enter text: I love you but hate the current political climate. [Detect Sentiment]

Output: ● Dual, ○ binary, ○ trinary, ○ scale

Keyword test:

I love dogs quite a lot but cats I really hate.

Enter keywords (comma-separated list, no spaces): dogs,frogs

[Detect Sentiment Around Keywords]

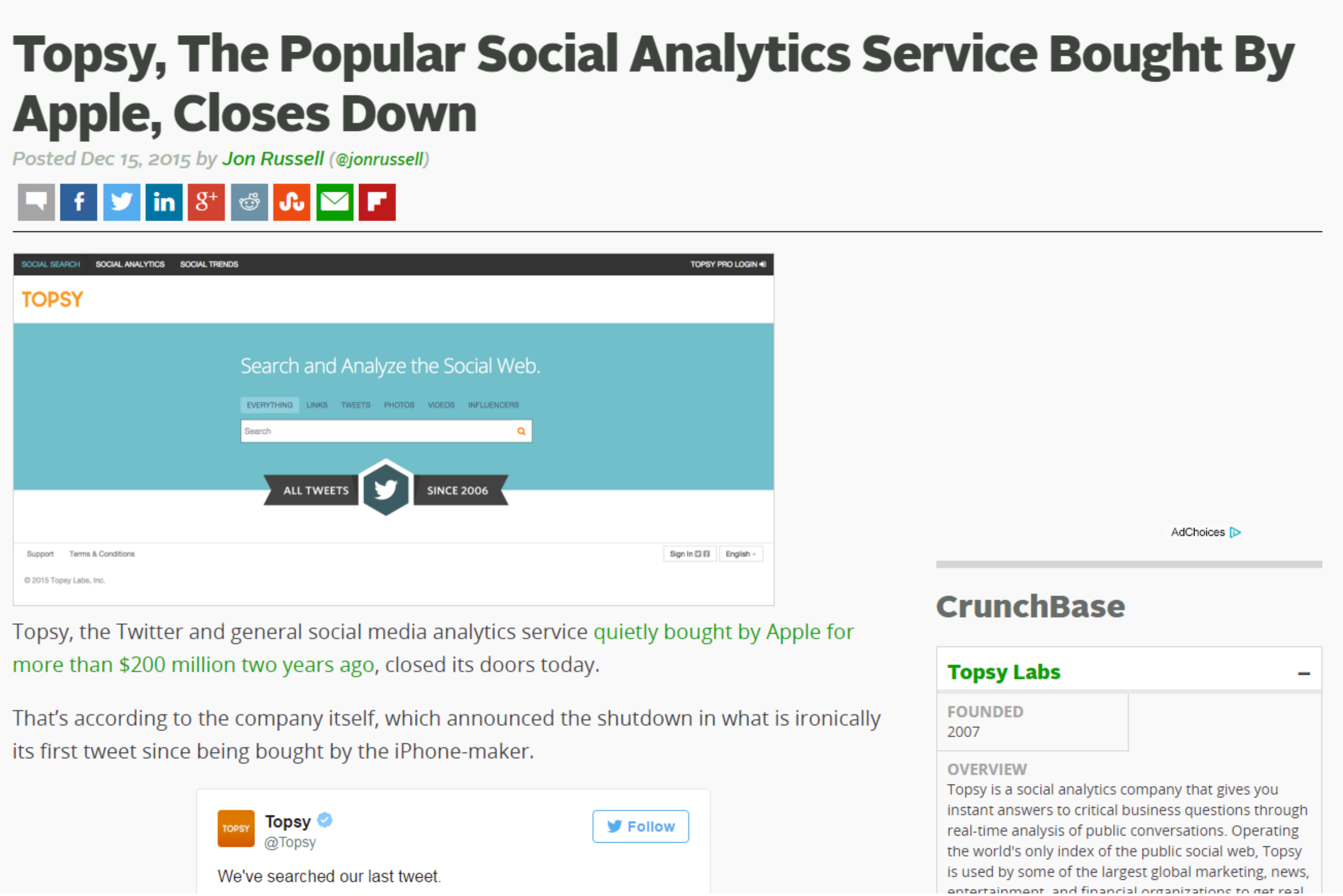
Topic test:

The movie was very frightening. My phone has a small screen. That is a banging chuuuun. I won the vote but he lied abo

Select domain (broad topic): Films/Movies ▼ [Detect Sentiment in Domain]

Others

RIP: TOPSY



Topsy, The Popular Social Analytics Service Bought By Apple, Closes Down

Posted Dec 15, 2015 by Jon Russell (@jonrussell)

Topsy, the Twitter and general social media analytics service quietly bought by Apple for more than $200 million two years ago, closed its doors today.

That's according to the company itself, which announced the shutdown in what is ironically its first tweet since being bought by the iPhone-maker.

Topsy
@Topsy
Follow

We've searched our last tweet.

CrunchBase

Topsy Labs

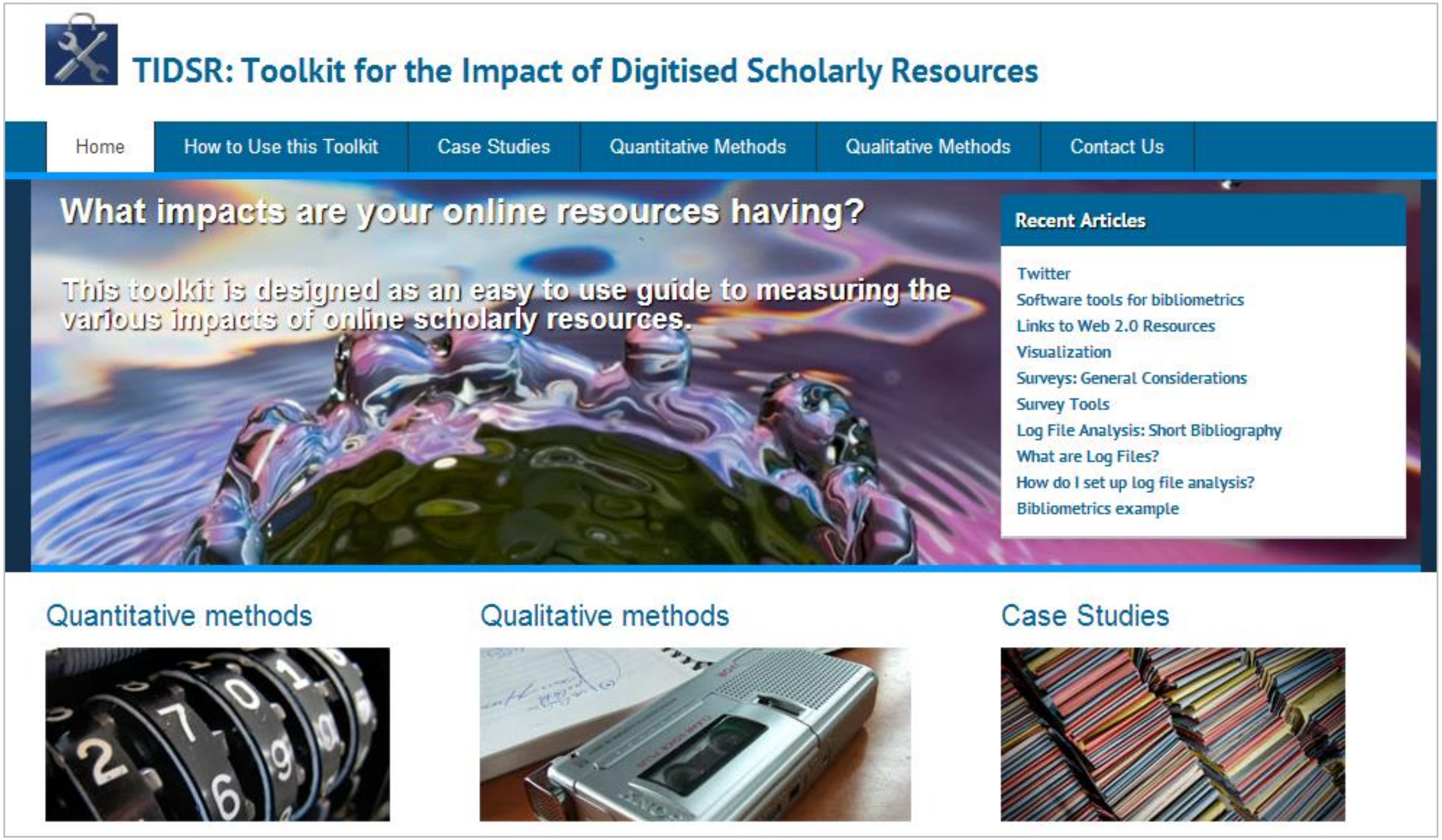
FOUNDED
2007

OVERVIEW
Topsy is a social analytics company that gives you instant answers to critical business questions through real-time analysis of public conversations. Operating the world's only index of the public social web, Topsy is used by some of the largest global marketing, news, entertainment, and financial organizations to get real

https://techcrunch.com/2015/12/15/rip-topsy/

Oxford Internet Institute

Measuring impact online:
http://microsites.oii.ox.ac.uk/tidsr/welcome



TIDSR: Toolkit for the Impact of Digitised Scholarly Resources

What impacts are your online resources having?

This toolkit is designed as an easy to use guide to measuring the various impacts of online scholarly resources.

Recent Articles

- Twitter
- Software tools for bibliometrics
- Links to Web 2.0 Resources
- Visualization
- Surveys: General Considerations
- Survey Tools
- Log File Analysis: Short Bibliography
- What are Log Files?
- How do I set up log file analysis?
- Bibliometrics example

Quantitative methods

Qualitative methods

Case Studies

More

- Mapbox (add data to maps): https://www.mapbox.com/
- classify websites into categories? http://diffbot.com/products/automatic/classifier/

More?

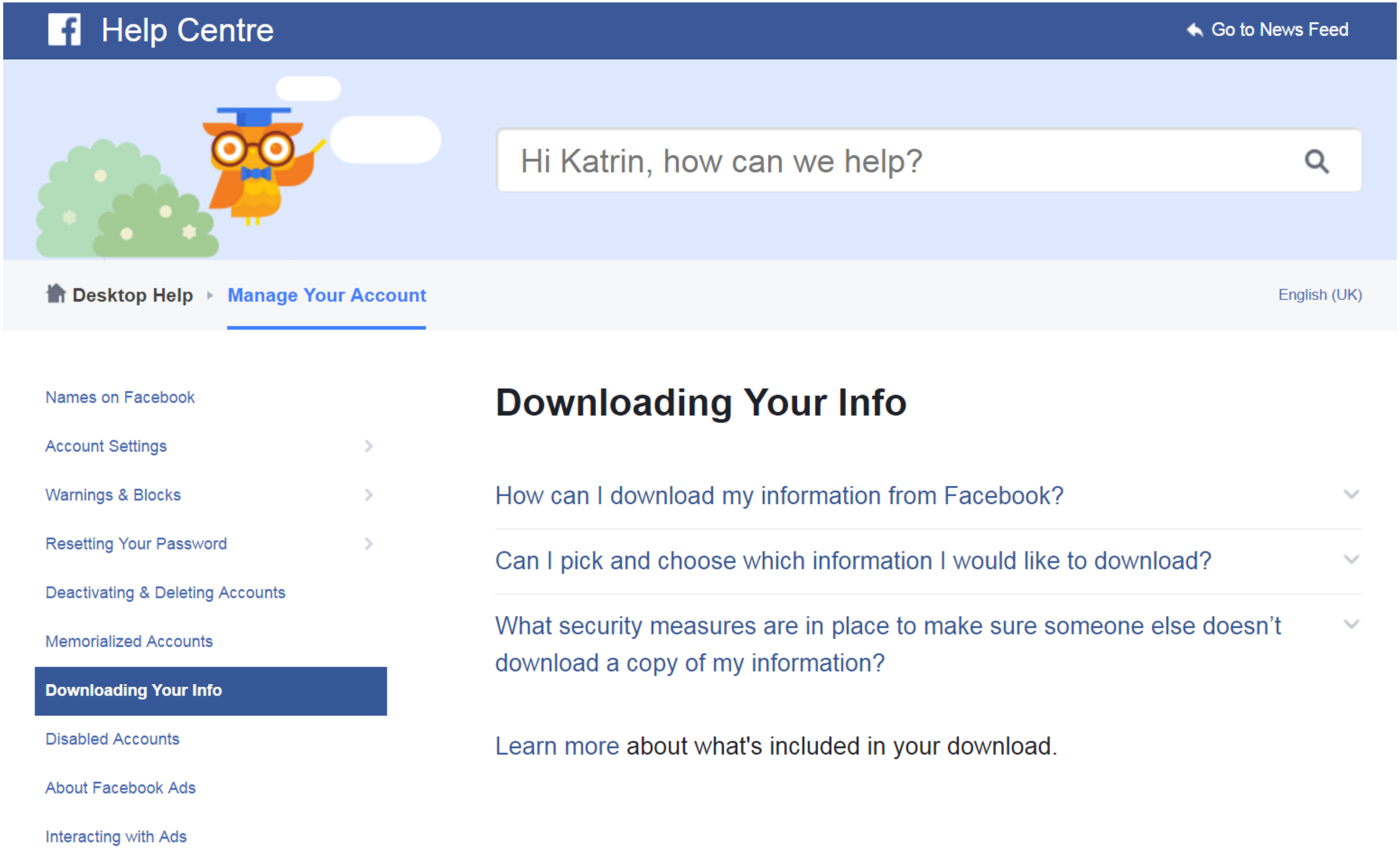
- Keep looking for the following keywords:
  - Web Science
  - Data Science
  - Computational Social Science
  - Digital Humanities
  - Digital Methods
  - Internet Research
  - Social Network Analysis

Open Educational Ressources:

- Coursera
- EdX
- …

Crowdsourcing
– for data collection and analysis

Get your own Facebook data



Help Centre

Go to News Feed

Hi Katrin, how can we help?

Desktop Help ▸ Manage Your Account

English (UK)

- Names on Facebook
- Account Settings ›
- Warnings & Blocks ›
- Resetting Your Password ›
- Deactivating & Deleting Accounts
- Memorialized Accounts
- **Downloading Your Info**
- Disabled Accounts
- About Facebook Ads
- Interacting with Ads

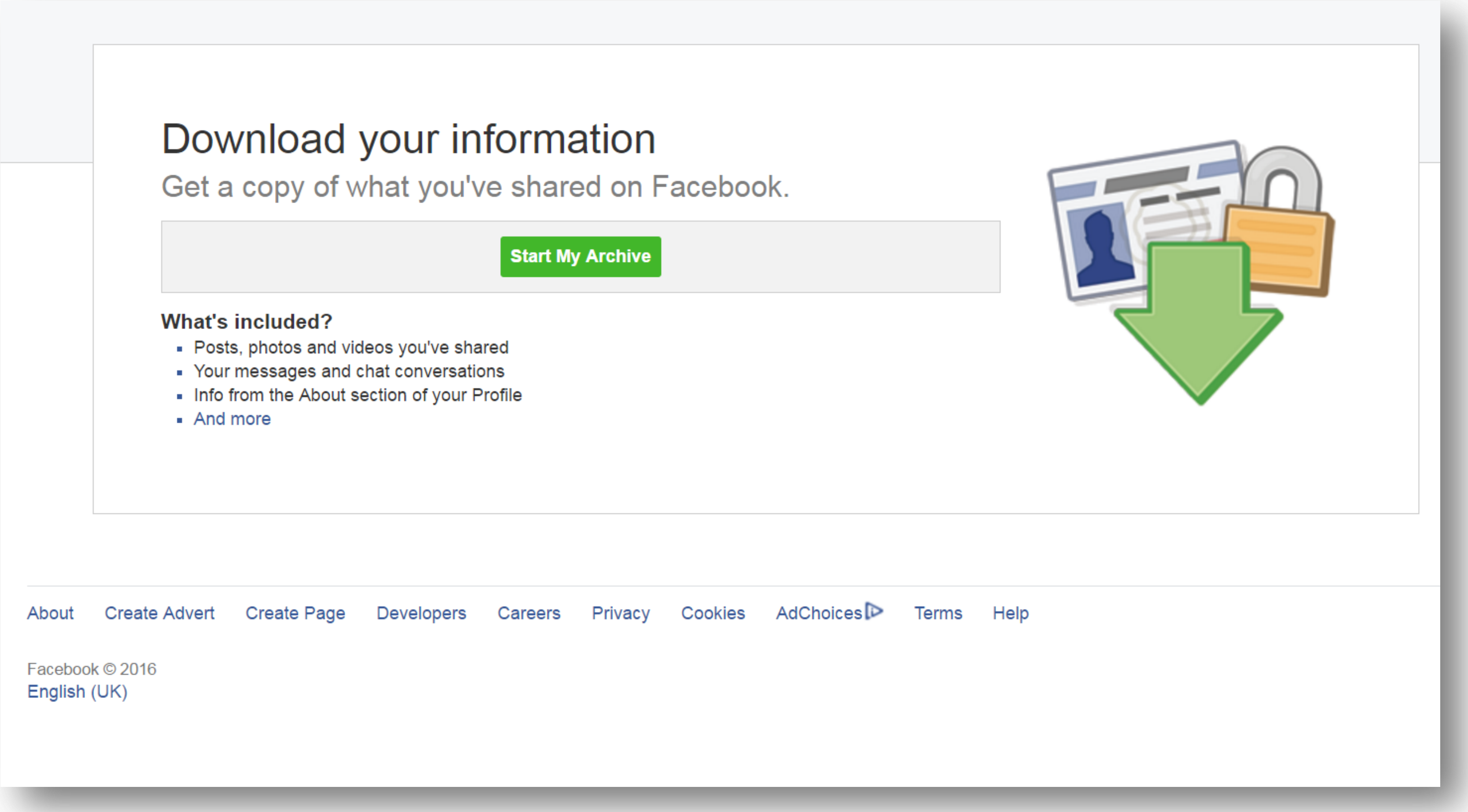
Downloading Your Info

How can I download my information from Facebook?

Can I pick and choose which information I would like to download?

What security measures are in place to make sure someone else doesn’t download a copy of my information?

Learn more about what's included in your download.



Download your information

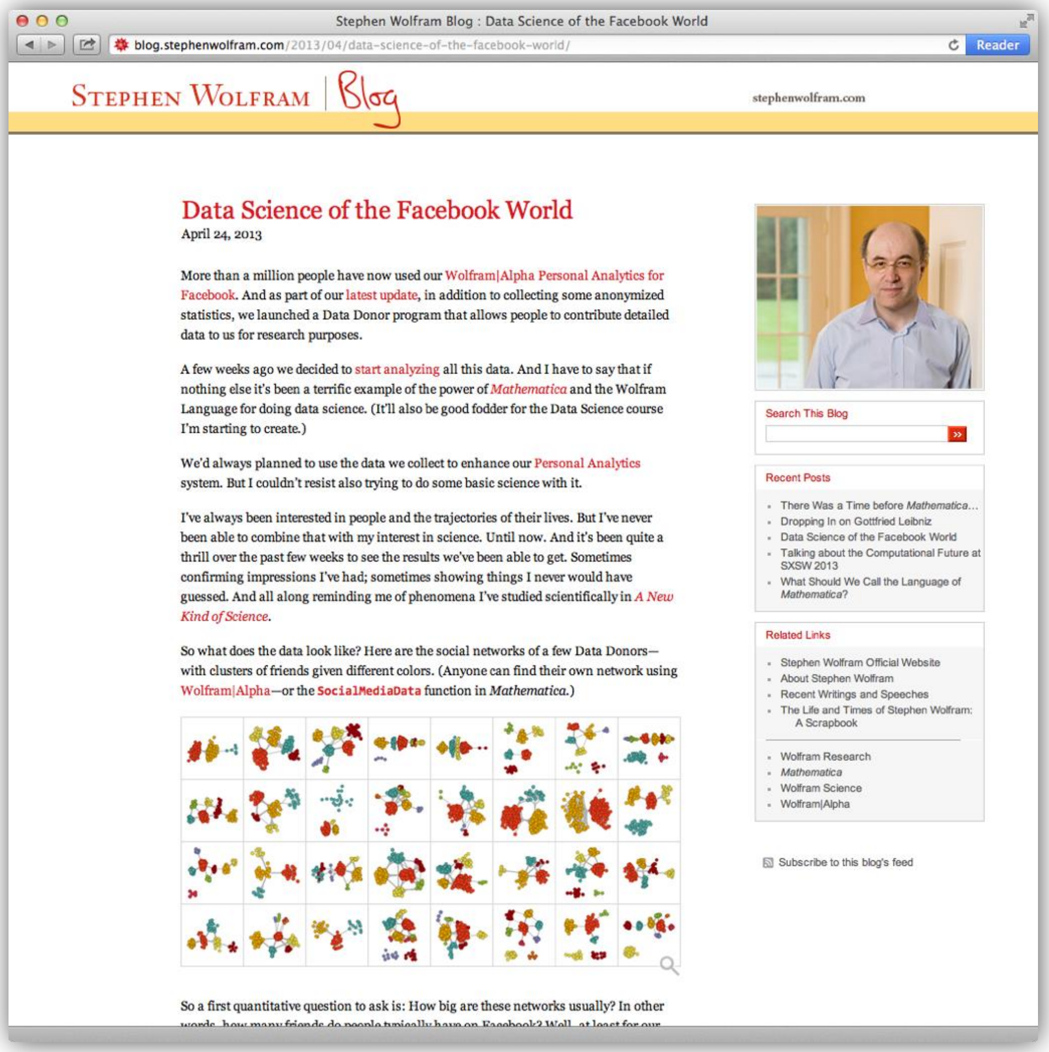
Get a copy of what you've shared on Facebook.

What's included?

- Posts, photos and videos you've shared
- Your messages and chat conversations
- Info from the About section of your Profile
- And more

Facebook © 2016
English (UK)

Crowdsourcing


Stephen Wolfram Blog : Data Science of the Facebook World
blog.stephenwolfram.com/2013/04/data-science-of-the-facebook-world/
Reader
Stephen Wolfram | Blog
stephenwolfram.com

Data Science of the Facebook World

April 24, 2013

More than a million people have now used our Wolfram|Alpha Personal Analytics for Facebook. And as part of our latest update, in addition to collecting some anonymized statistics, we launched a Data Donor program that allows people to contribute detailed data to us for research purposes.

A few weeks ago we decided to start analyzing all this data. And I have to say that if nothing else it's been a terrific example of the power of *Mathematica* and the Wolfram Language for doing data science. (It'll also be good fodder for the Data Science course I'm starting to create.)

We'd always planned to use the data we collect to enhance our Personal Analytics system. But I couldn't resist also trying to do some basic science with it.

I've always been interested in people and the trajectories of their lives. But I've never been able to combine that with my interest in science. Until now. And it's been quite a thrill over the past few weeks to see the results we've been able to get. Sometimes confirming impressions I've had; sometimes showing things I never would have guessed. And all along reminding me of phenomena I've studied scientifically in *A New Kind of Science*.

So what does the data look like? Here are the social networks of a few Data Donors—with clusters of friends given different colors. (Anyone can find their own network using Wolfram|Alpha—or the `SocialMediaData` function in *Mathematica*.)

So a first quantitative question to ask is: How big are these networks usually? In other words, how many friends do people typically have on Facebook? Well, at least for our

Search This Blog

»

Recent Posts

- There Was a Time before *Mathematica*…
- Dropping In on Gottfried Leibniz
- Data Science of the Facebook World
- Talking about the Computational Future at SXSW 2013
- What Should We Call the Language of *Mathematica*?

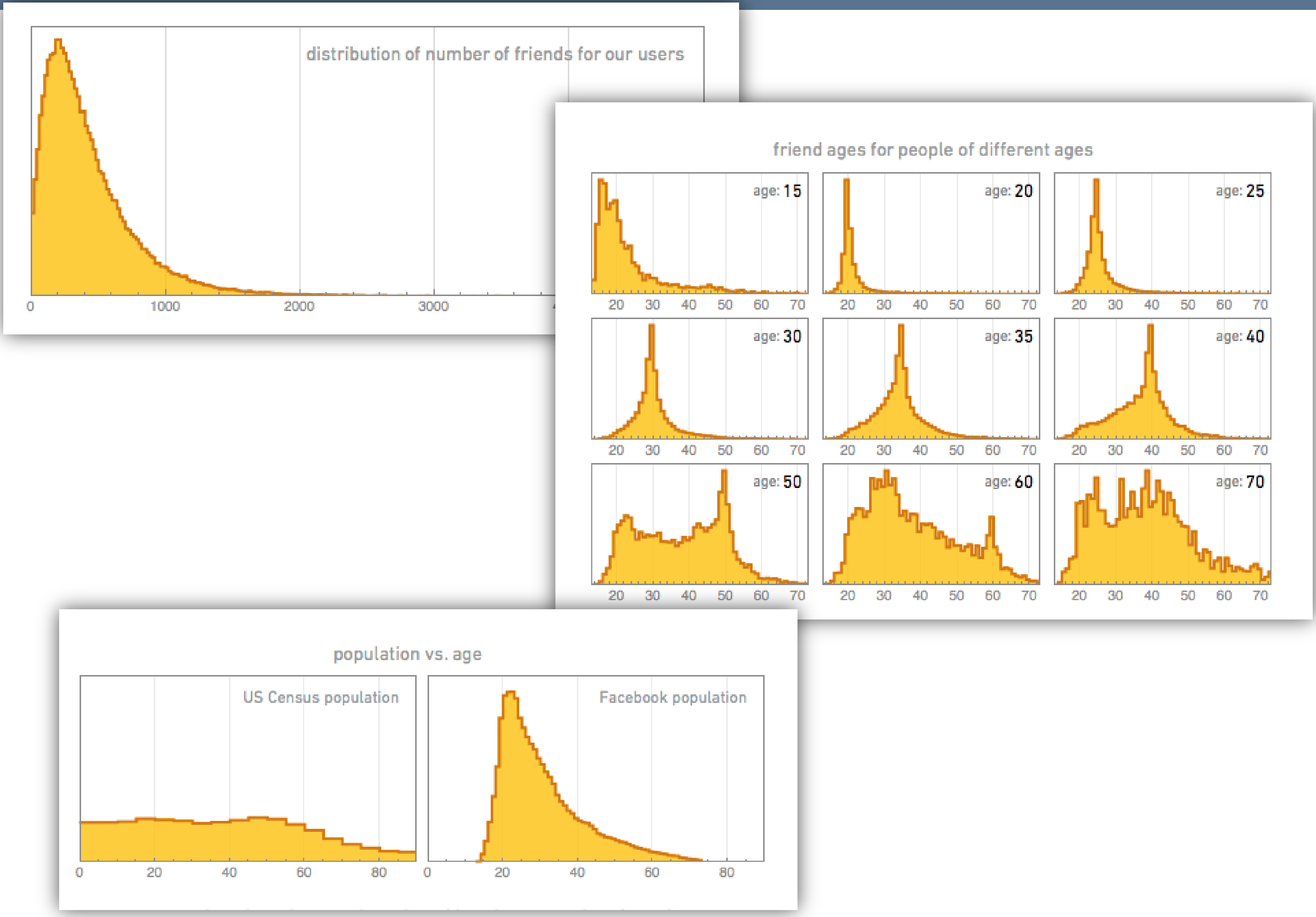
Related Links

- Stephen Wolfram Official Website
- About Stephen Wolfram
- Recent Writings and Speeches
- The Life and Times of Stephen Wolfram: A Scrapbook

- Wolfram Research
- *Mathematica*
- Wolfram Science
- Wolfram|Alpha

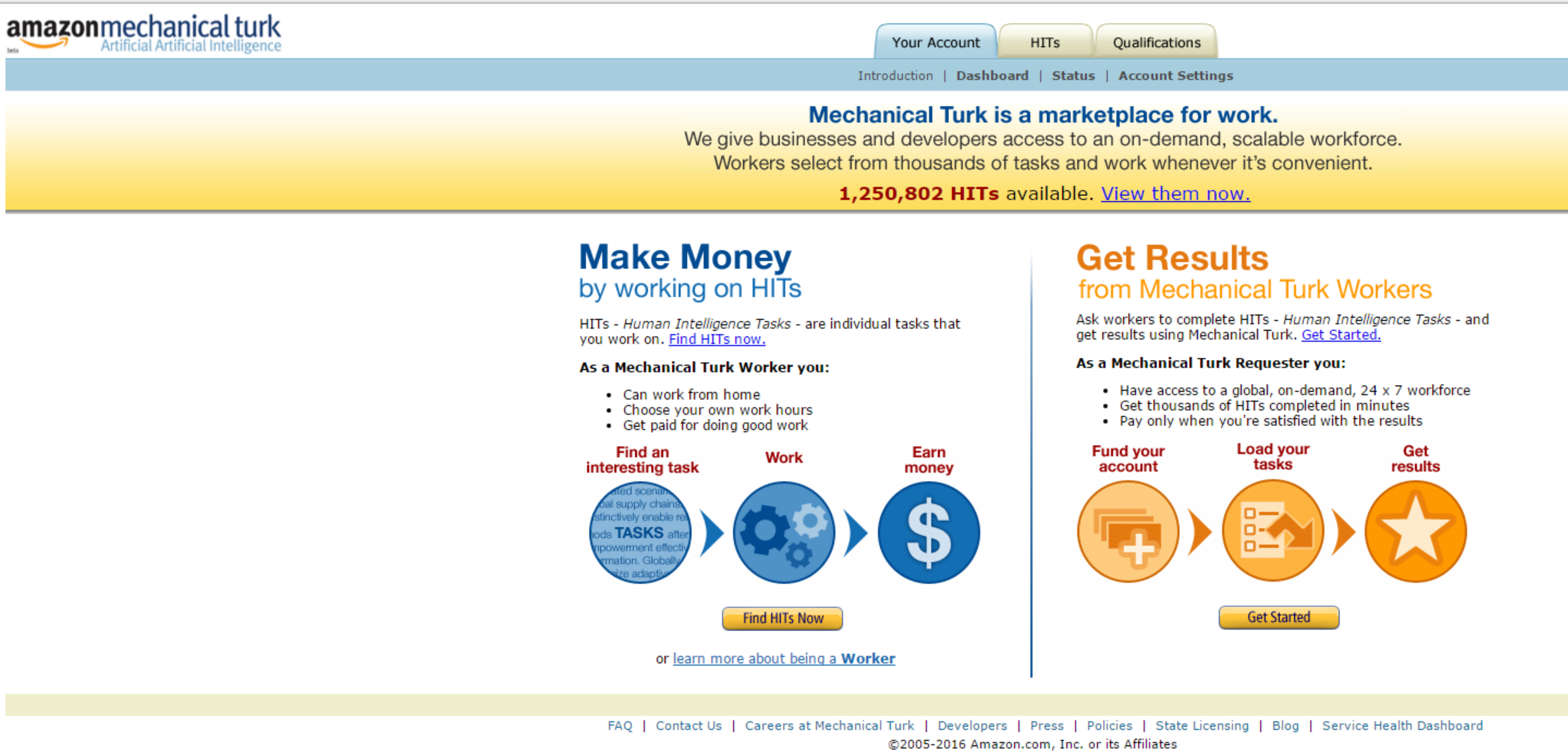
Subscribe to this blog's feed

Stephen Wolfram
Daten: Facebook
http://blog.stephenwolfram.com/2013/04/data-science-of-the-facebook-world/



Stephen Wolfram
Daten: Facebook
http://blog.stephenwolfram.com/2013/04/data-science-of-the-facebook-world/

Amazon Mechanical Turk



amazonmechanical turk
Artificial Artificial Intelligence
beta
Your Account
HITs
Qualifications
Introduction | Dashboard | Status | Account Settings

Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

1,250,802 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

or learn more about being a **Worker**

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get Started.

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

FAQ | Contact Us | Careers at Mechanical Turk | Developers | Press | Policies | State Licensing | Blog | Service Health Dashboard

https://www.mturk.com/mturk/welcome

http://www.wearedynamo.org/forum/actions/topics/call-for-comments-turker-authored-guidelines-for-research-on-amt?page=1

Research ethics

Legal Framework?

Terms of Services

Ethics

Privacy and protecting subjects

- Anonymisation?
- Informed consent?
- Guidelines for handling sensitive user data?

Data Sharing

One of the Challenges: Data Sharing

“But you can’t make your data available for others to look at, which means both your study can’t really be replicated and it can’t be tested for review. But also it just means your data can’t be made available for other people to say, Ah you have done this with it, I’ll see what I can do with it, (…) There is no open data.”

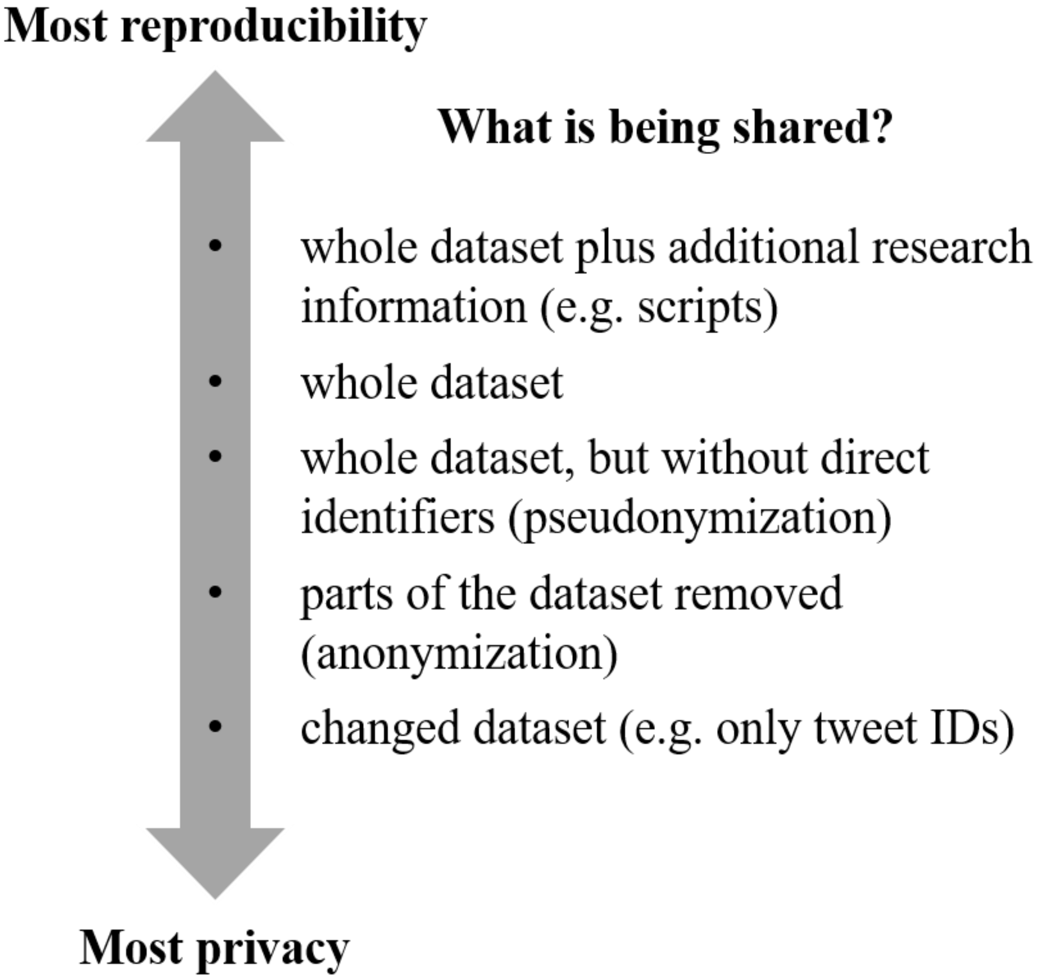
Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2015. "Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research?." In Standards and Practices in Large-Scale Social Media Research: Papers from the 2015 ICWSM Workshop. Proceedings Ninth International AAAI Conference on Web and Social Media Oxford University, May 26, 2015 – May 29, 2015, 28-37. Ann Arbor, MI: AAAI Press.

Reusing datasets

“I actually only use [other researcher’s datasets] where I’m very sure about where it comes from and how it was processed and analyzed. There is too much uncertainty in it.”

Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2015. "Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research?." In Standards and Practices in Large-Scale Social Media Research: Papers from the 2015 ICWSM Workshop. Proceedings Ninth International AAAI Conference on Web and Social Media Oxford University, May 26, 2015 – May 29, 2015, 28-37. Ann Arbor, MI: AAAI Press.

How much should I share?



Most reproducibility

What is being shared?

- whole dataset plus additional research information (e.g. scripts)
- whole dataset
- whole dataset, but without direct identifiers (pseudonymization)
- parts of the dataset removed (anonymization)
- changed dataset (e.g. only tweet IDs)

Most privacy

Examples of current practices

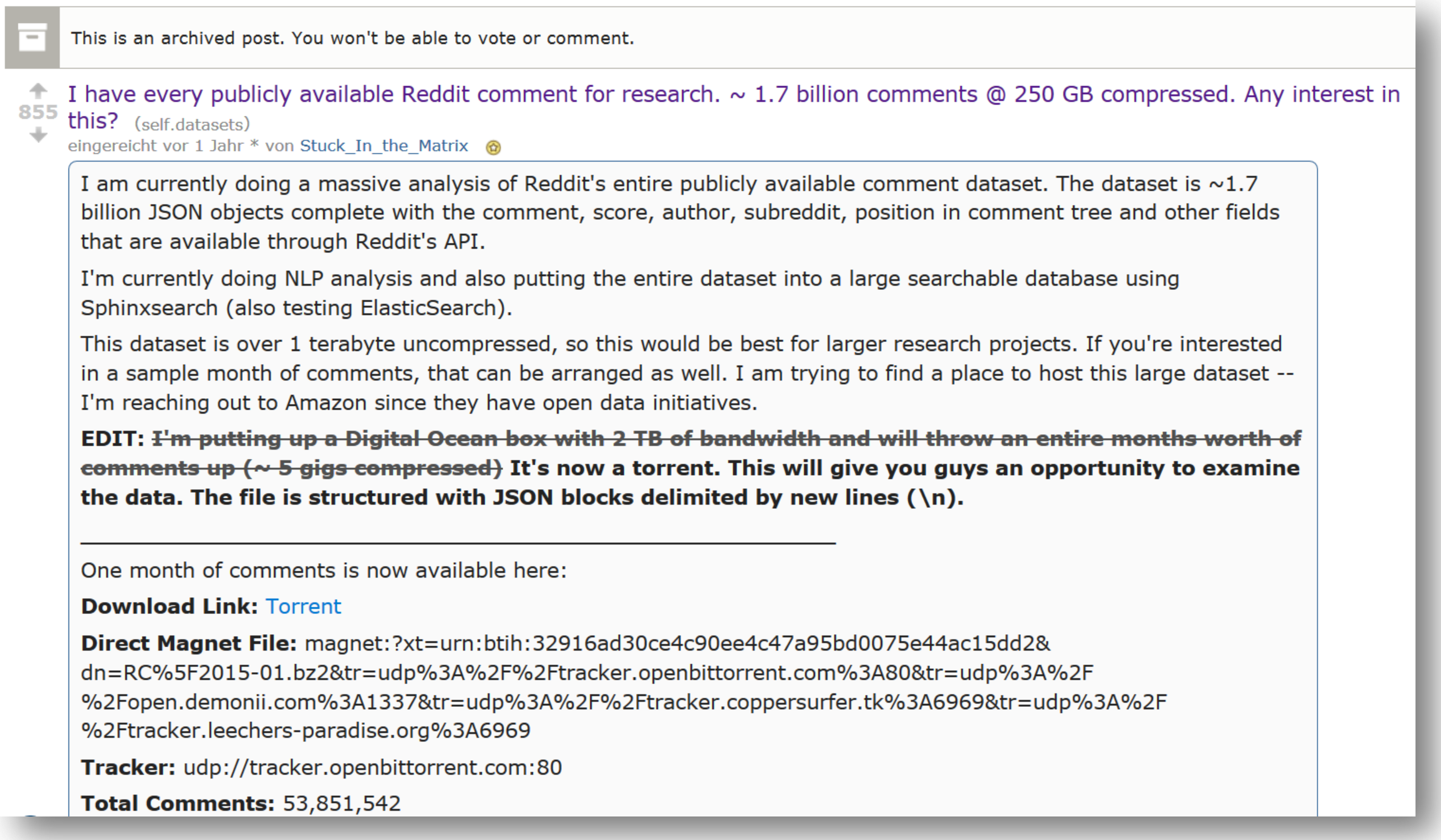
Approach	Example	Retrievability	Documentation + standards	Long-term perspective
„Grey market“	Data shared with colleagues (often upon request)	Low, via personal connections	Low, no standardized documentation or data format	Low, no guaranteed long term availability
Researchers‘ personal/ professional websites		Medium, URLs may be referenced	Low, no standardized documentation or data format	Low, no guaranteed long term availability
Social media providers	e.g. Wikipedia dumps	High, if directly provided from social media platform	Medium/high, depending on the producer	Medium/high, depending on the producer
Project-based or thematic collections	e.g. KONECT, CrisisLex	Medium, URLs may be referenced	Medium, if same principles are applied within entire collection	Low, no guaranteed long term availability
Conferences & journals	e.g. ICWSM datasets with conference papers	High, usually related to accepted publications	Medium/high, depending on the publisher	Medium/high, depending on the publisher
Professional archives	e.g. datasets at GESIS data archive	Medium/high, datasets may be referenceable with DOIs.	Medium/high, depending on the archive’s requirements	High, guaranteed availability for different time spans

- For additional examples see our paper: http://dx.doi.org/10.1145/2908131.2908172
- See also: Thomson, S.D. 2016. Preserving Social Media. DPC Technology Watch Report. Retrieved from http://dpconline.org/publications/technology-watch-reports

Available datasets

- From individual researchers/groups (sometimes „grey market").
- From conferences: e.g. ICWSM
- Archival institutions: e.g. GESIS (*doi:10.4232/1.12319*)

Reddit dataset



This is an archived post. You won't be able to vote or comment.

855

I have every publicly available Reddit comment for research. ~ 1.7 billion comments @ 250 GB compressed. Any interest in this? (self.datasets)

eingereicht vor 1 Jahr * von Stuck_In_the_Matrix

I am currently doing a massive analysis of Reddit's entire publicly available comment dataset. The dataset is ~1.7 billion JSON objects complete with the comment, score, author, subreddit, position in comment tree and other fields that are available through Reddit's API.

I'm currently doing NLP analysis and also putting the entire dataset into a large searchable database using Sphinxsearch (also testing ElasticSearch).

This dataset is over 1 terabyte uncompressed, so this would be best for larger research projects. If you're interested in a sample month of comments, that can be arranged as well. I am trying to find a place to host this large dataset -- I'm reaching out to Amazon since they have open data initiatives.

EDIT: ~~I'm putting up a Digital Ocean box with 2 TB of bandwidth and will throw an entire months worth of comments up (~ 5 gigs compressed)~~ It's now a torrent. This will give you guys an opportunity to examine the data. The file is structured with JSON blocks delimited by new lines (\n).

One month of comments is now available here:

Download Link: Torrent

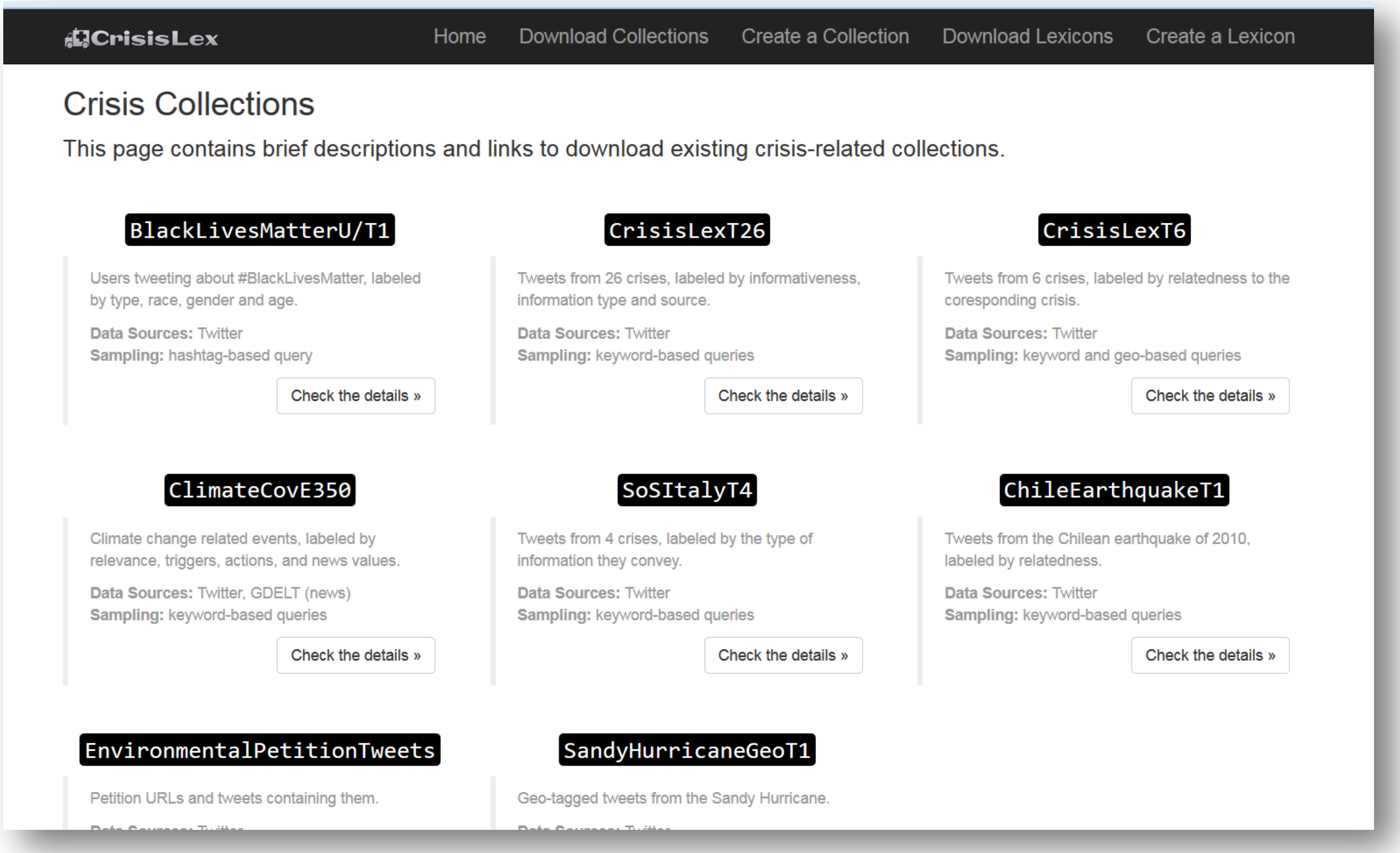
Direct Magnet File: magnet:?xt=urn:btih:32916ad30ce4c90ee4c47a95bd0075e44ac15dd2&dn=RC%5F2015-01.bz2&tr=udp%3A%2F%2Ftracker.openbittorrent.com%3A80&tr=udp%3A%2F%2Fopen.demonii.com%3A1337&tr=udp%3A%2F%2Ftracker.coppersurfer.tk%3A6969&tr=udp%3A%2F%2Ftracker.leechers-paradise.org%3A6969

Tracker: udp://tracker.openbittorrent.com:80

Total Comments: 53,851,542

https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

CrisisLex



Crisis Collections

This page contains brief descriptions and links to download existing crisis-related collections.

BlackLivesMatterU/T1

Users tweeting about #BlackLivesMatter, labeled by type, race, gender and age.

Data Sources: Twitter
Sampling: hashtag-based query

CrisisLexT26

Tweets from 26 crises, labeled by informativeness, information type and source.

Data Sources: Twitter
Sampling: keyword-based queries

CrisisLexT6

Tweets from 6 crises, labeled by relatedness to the coresponding crisis.

Data Sources: Twitter
Sampling: keyword and geo-based queries

ClimateCovE350

Climate change related events, labeled by relevance, triggers, actions, and news values.

Data Sources: Twitter, GDELT (news)
Sampling: keyword-based queries

SoSItalyT4

Tweets from 4 crises, labeled by the type of information they convey.

Data Sources: Twitter
Sampling: keyword-based queries

ChileEarthquakeT1

Tweets from the Chilean earthquake of 2010, labeled by relatedness.

Data Sources: Twitter
Sampling: keyword-based queries

EnvironmentalPetitionTweets

Petition URLs and tweets containing them.

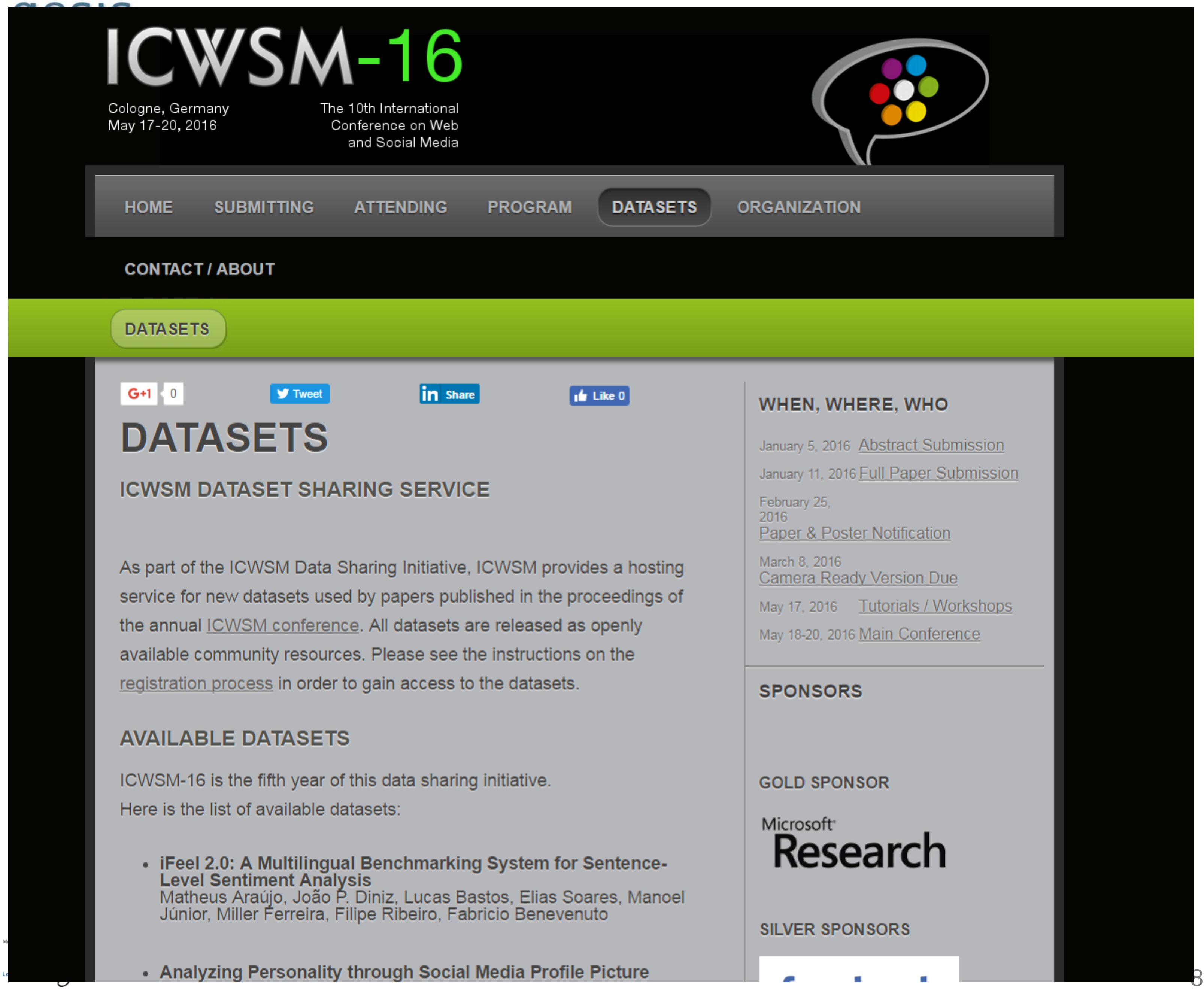
Data Sources: Twitter

SandyHurricaneGeoT1

Geo-tagged tweets from the Sandy Hurricane.

Data Sources: Twitter

http://crisislex.org/data-collections.html



Cologne, Germany
May 17-20, 2016

The 10th International
Conference on Web
and Social Media

DATASETS

ICWSM DATASET SHARING SERVICE

As part of the ICWSM Data Sharing Initiative, ICWSM provides a hosting service for new datasets used by papers published in the proceedings of the annual ICWSM conference. All datasets are released as openly available community resources. Please see the instructions on the registration process in order to gain access to the datasets.

AVAILABLE DATASETS

ICWSM-16 is the fifth year of this data sharing initiative.
Here is the list of available datasets:

- **iFeel 2.0: A Multilingual Benchmarking System for Sentence-Level Sentiment Analysis**
  Matheus Araújo, João P. Diniz, Lucas Bastos, Elias Soares, Manoel Júnior, Miller Ferreira, Filipe Ribeiro, Fabricio Benevenuto
- **Analyzing Personality through Social Media Profile Picture**

WHEN, WHERE, WHO

January 5, 2016	Abstract Submission
January 11, 2016	Full Paper Submission
February 25, 2016	Paper & Poster Notification
March 8, 2016	Camera Ready Version Due
May 17, 2016	Tutorials / Workshops
May 18-20, 2016	Main Conference

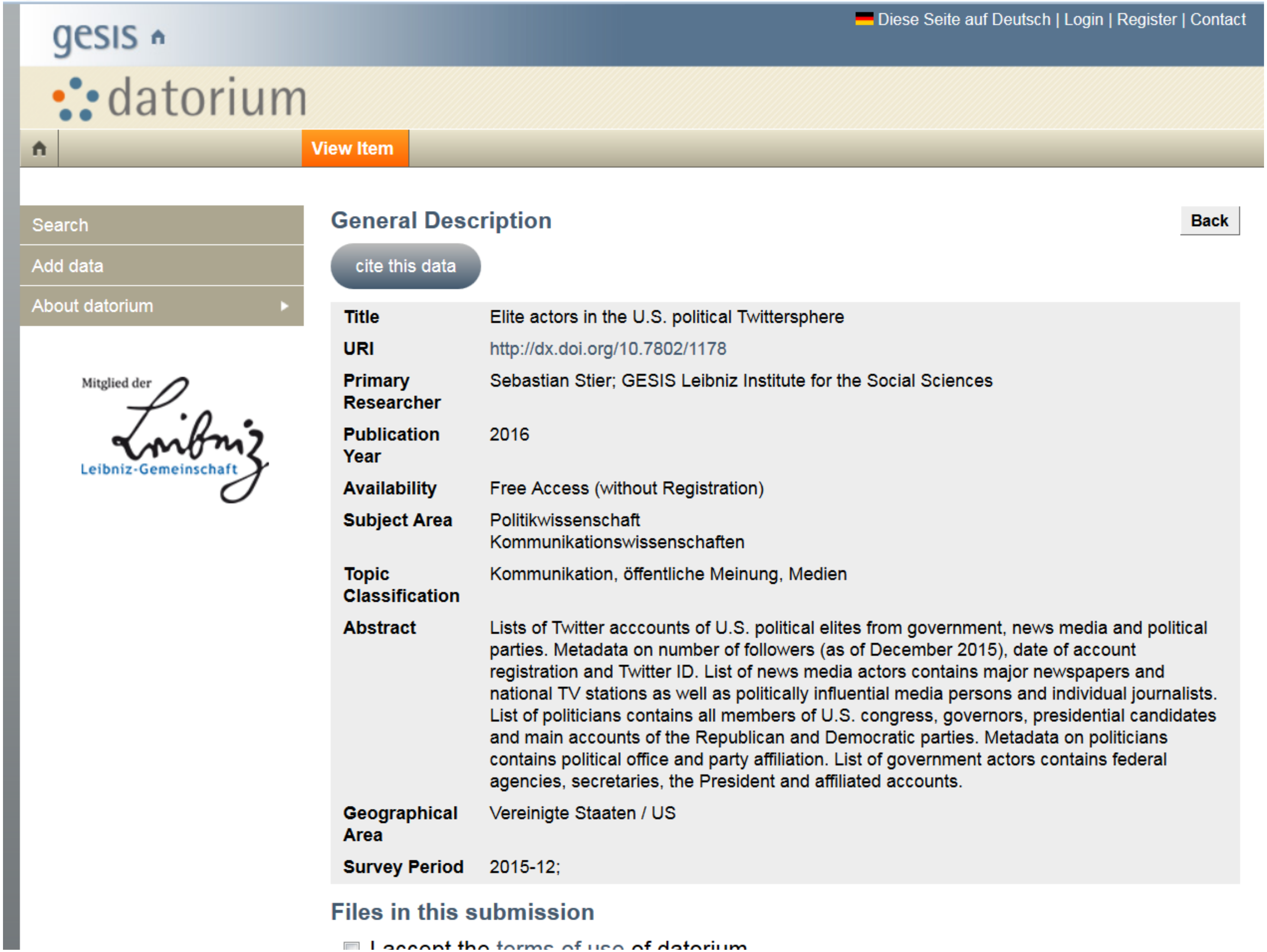
SPONSORS

GOLD SPONSOR

Microsoft Research

SILVER SPONSORS

https://datorium.gesis.org/xmlui/handle/10.7802/1178?locale-attribute=en


Diese Seite auf Deutsch | Login | Register | Contact
gesis
datorium
View Item
Search
Add data
About datorium
Mitglied der Leibniz-Gemeinschaft
Back
cite this data

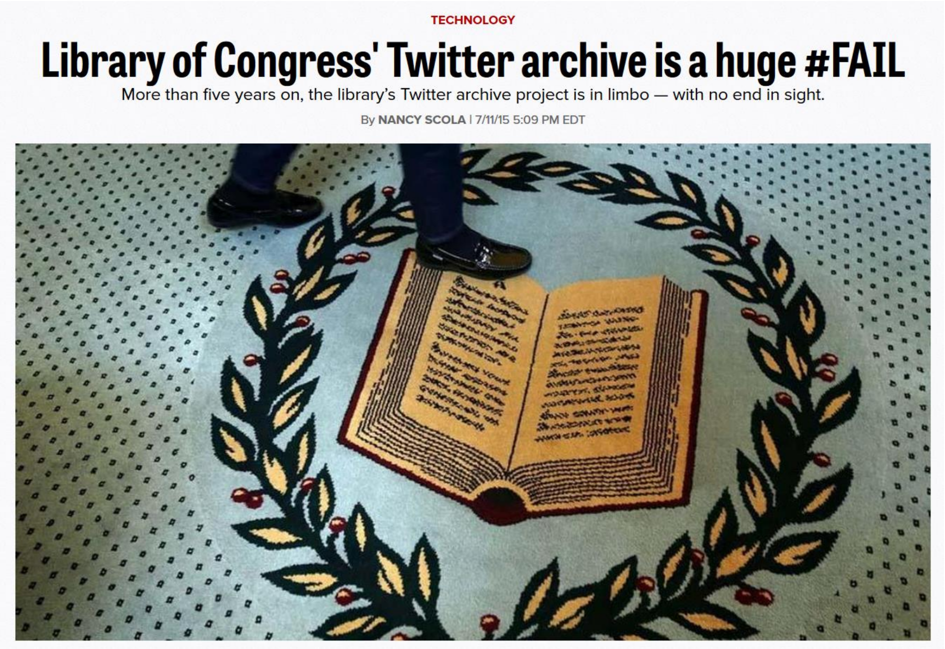
General Description

Title	Elite actors in the U.S. political Twittersphere
URI	http://dx.doi.org/10.7802/1178
Primary Researcher	Sebastian Stier; GESIS Leibniz Institute for the Social Sciences
Publication Year	2016
Availability	Free Access (without Registration)
Subject Area	Politikwissenschaft Kommunikationswissenschaften
Topic Classification	Kommunikation, öffentliche Meinung, Medien
Abstract	Lists of Twitter acccounts of U.S. political elites from government, news media and political parties. Metadata on number of followers (as of December 2015), date of account registration and Twitter ID. List of news media actors contains major newspapers and national TV stations as well as politically influential media persons and individual journalists. List of politicians contains all members of U.S. congress, governors, presidential candidates and main accounts of the Republican and Democratic parties. Metadata on politicians contains political office and party affiliation. List of government actors contains federal agencies, secretaries, the President and affiliated accounts.
Geographical Area	Vereinigte Staaten / US
Survey Period	2015-12;

Files in this submission

□ I accept the terms of use of datorium

Unvailable datasets



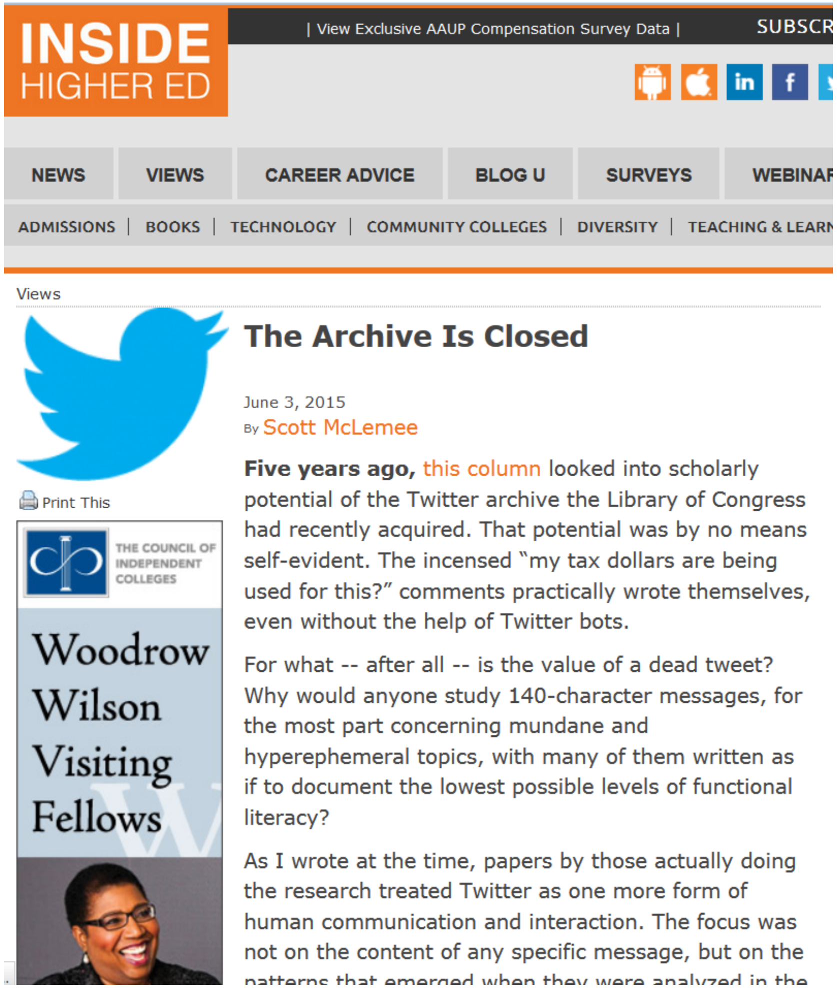
TECHNOLOGY

Library of Congress' Twitter archive is a huge #FAIL

More than five years on, the library's Twitter archive project is in limbo — with no end in sight.

By NANCY SCOLA | 7/11/15 5:09 PM EDT

http://www.politico.com/story/2015/07/library-of-congress-twitter-archive-119698.html



INSIDE HIGHER ED

| View Exclusive AAUP Compensation Survey Data |

SUBSCR

The Archive Is Closed

June 3, 2015

By Scott McLemee

Five years ago, this column looked into scholarly potential of the Twitter archive the Library of Congress had recently acquired. That potential was by no means self-evident. The incensed "my tax dollars are being used for this?" comments practically wrote themselves, even without the help of Twitter bots.

For what -- after all -- is the value of a dead tweet? Why would anyone study 140-character messages, for the most part concerning mundane and hyperephemeral topics, with many of them written as if to document the lowest possible levels of functional literacy?

As I wrote at the time, papers by those actually doing the research treated Twitter as one more form of human communication and interaction. The focus was not on the content of any specific message, but on the patterns that emerged when they were analyzed in the

https://www.insidehighered.com/views/2015/06/03/article-difficulties-social-media-research

Background: LoC Twitter archive

- 2006: Twitter was made public
- 14.04.2010: Twitter donates archive to the Library of Congress: *"Recently, the Library of Congress signaled to us that the public tweets we have all been creating over the years are important and worthy of preservation."* https://blog.twitter.com/2010/tweet-preservation
- 2010-2012: nothing new on the "Twitter archive"…
- 04.01.2013: Update from the Library of Congress: "We now have an archive of approximately 170 billion tweets and growing. The volume of tweets the Library receives each day has grown from 140 million beginning in February 2011 to nearly half a billion tweets each day as of October 2012." http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/

→ But still no access to the Twitter archive at the LoC and no information. *"Although the Library has been building and stabilizing the archive and has not yet offered researchers access, we have nevertheless received approximately 400 inquiries from researchers all over the world."*

Conclusions 5

Lessons learned

- There is no perfect tool. But you may find solutions that are sufficient for different usage scenarios.
- Tools come and go, you need to keep track of what is available.

If you have time to read 3 papers…

- Borra, E., & Rieder, D. (2014). Programmed method: developing a toolset for capturing and analyzing tweets, Aslib Journal of Information Management, 66(3), 262 – 278. DOI: http://dx.doi.org/10.1108/AJIM-09-2013-0094
- Bruns, A., & Liang, Y. E. (2012). Tools and methods for capturing Twitter data during natural disasters. First Monday, 17(4). doi:10.5210/fm.v17i4.3937
- Gaffney, D., & Puschmann, C. (2014). Data collection on Twitter. In Weller, A. Bruns, J. Burgess., M. Mahrt and C. Puschmann (Ed.), Twitter and Society (pp. 55–68). New York: Peter Lang.