gesis
Leibniz-Institut für Sozialwissenschaften

Social Media as Research Data

Dr. Katrin Weller

GESIS – Leibniz-Institute for the Social Sciences
Dept. of Computational Social Science
Cologne, Germany
●
Digital Studies Fellow at John W. Kluge Center
Library of Congress
Washington D.C.

E-Mail: katrin.weller@gesis.org ●Twitter: @kwelle ● Web: www.katrinweller.net

Slides are available at: http://de.slideshare.net/katrinweller

SERIOUSLY? DO THEY NOT REALIZE THAT 99% OF TWEETS ARE WORTHLESS BABBLE THAT READ SOMETHING LIKE ‘JUST WOKE UP. GOING TO STARBUCKS NOW. GETTING LATTE.’

READER’S COMMENT FOUND IN THE COMMENT SECTION FOR GROSS, D. (2010, APRIL 14). LIBRARY OF CONGRESS TO ARCHIVE YOUR TWEETS. CNN. RETRIEVED FROM HTTP://EDITION.CNN.COM/2010/TECH/04/14/LIBRARY.CONGRESS.TWITTER/, RETRIEVED NOVEMBER 19.
PHOTOS: HTTPS://WWW.FLICKR.COM/SEARCH/?TEXT=COFFEE&LICENSE=4%2C5%2C6%2C9%2C10

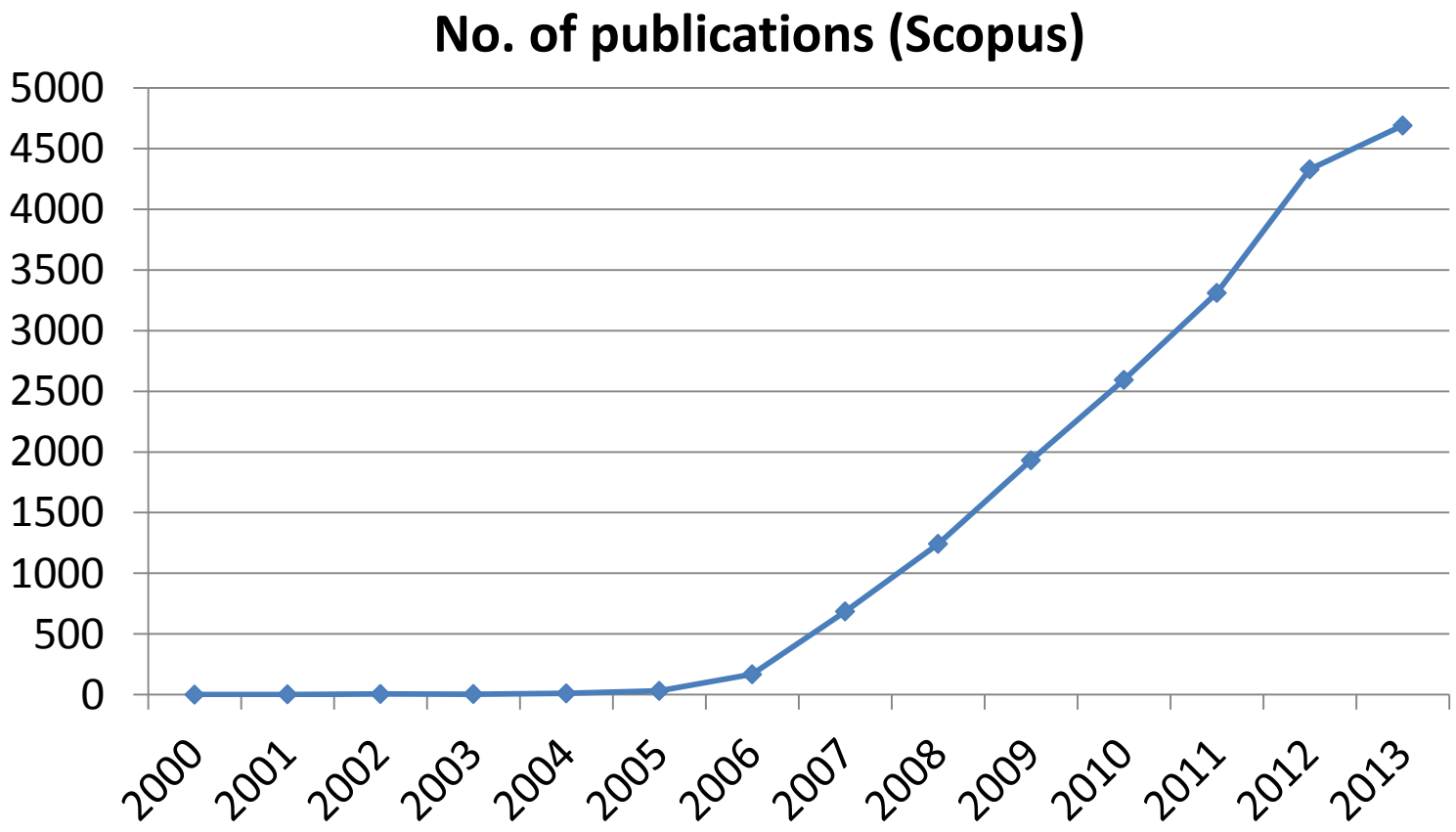
Chances in Social Media Research

- Researchers value social media as a new type of data
- Previously „ephemeral data“ become visible
- Immediate – quick reaction to events
- Structured
- „natural“ data

“What I find really interesting is that structure becomes manifest in internet communication. So it’s the first time in history actually that we can, that social structures between people become manifest within a technology. (...) They become visible, they become crawlable, they become analyzable.”

Kinder-Kurlanda, Katharina E., and Katrin Weller. 2014. "'I always feel it must be great to be a hacker!': The role of interdisciplinary work in social media research." In *Proceedings of the 2014 ACM conference on Web Science*, 91-98. New York: ACM.

Social media research 2000-today



(TITLE-ABS-KEY("social media") OR TITLE-ABS-KEY("social web") OR TITLE-ABS-KEY("social software") OR TITLE-ABS-KEY("web 2.0")) AND PUBYEAR > 1999

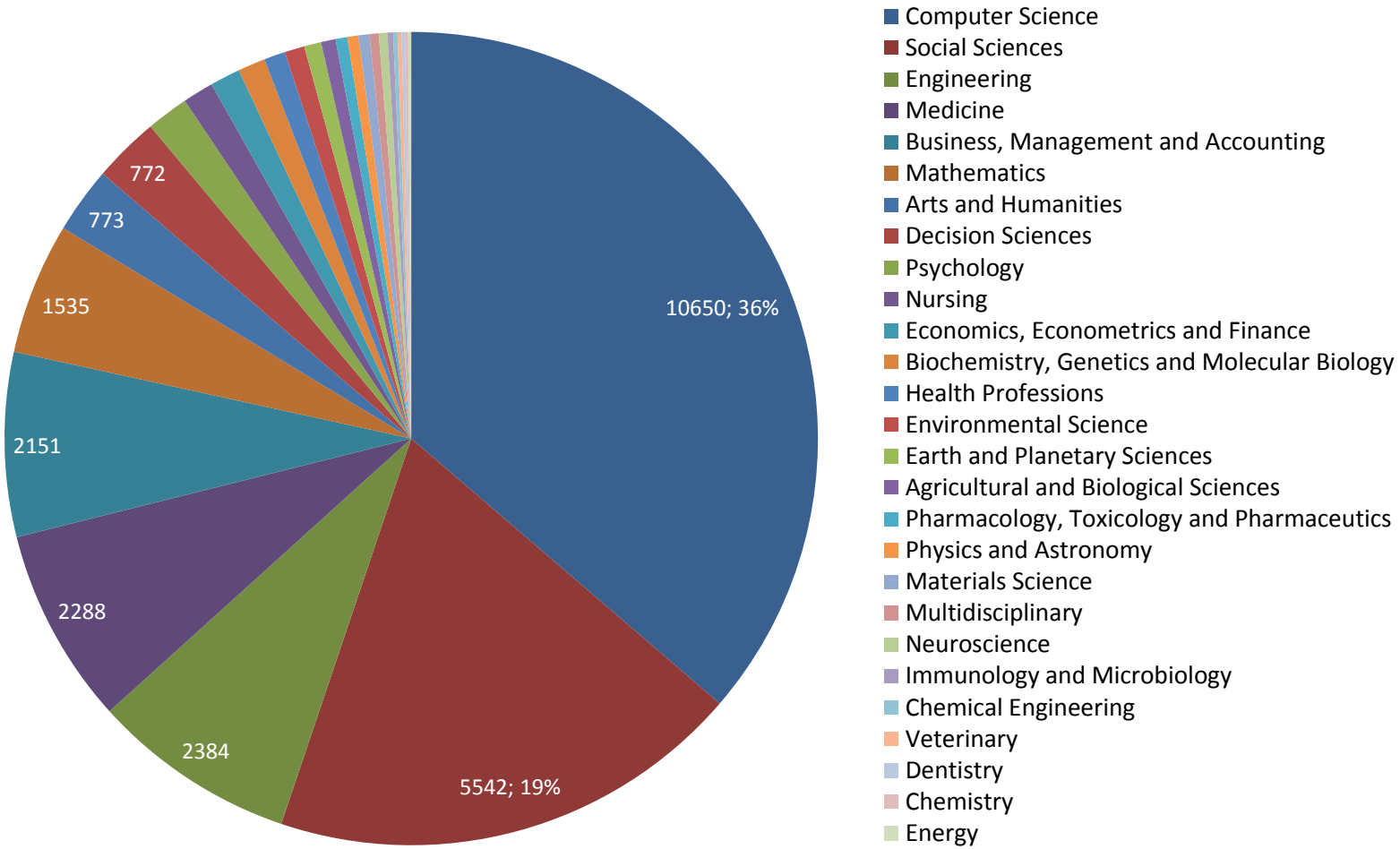
What is being studied?

- User groups
- Events
- Audiences
- Practices
- Information flow
- Influence
- Opinions and sentiments
- Networks
- Interactions
- Predictions
- Language

- Political communication
- Activism
- Crisis communication/disaster response
- E-learning
- Health
- Brand communication
- …

A new discipline?

Scopus: 2000-today by subject area



Social Media Data

- Texts
- Images
- Videos
- Mixed formats
- Connections I (friends, followers)
- Connections II (links/URLs)
- Connections/Actions (likes, favs, comments, downloads)

→ Different methods!

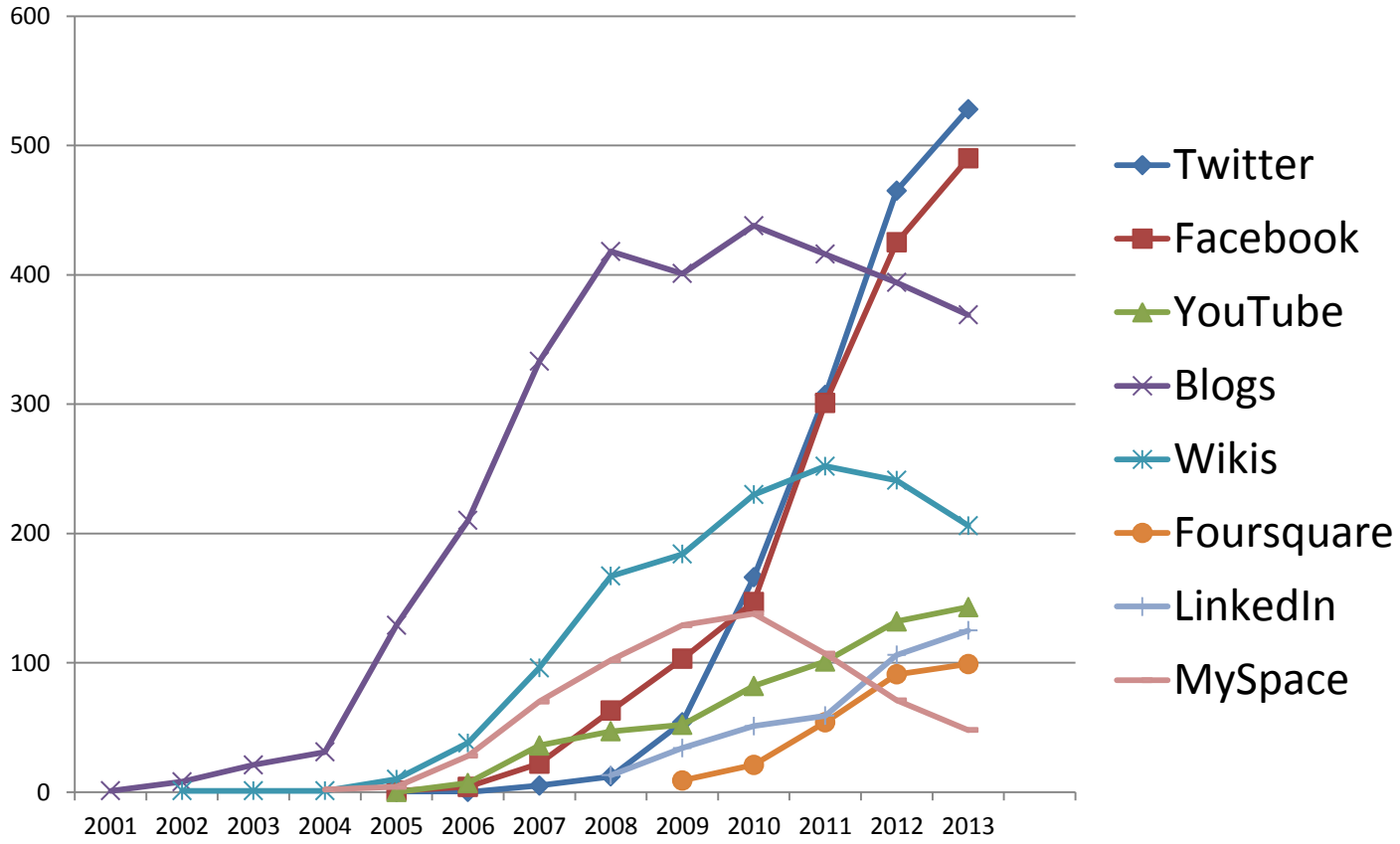
Different methods and types of datasets, examples from popular social science papers

No.	Method	Domain	Dataset
[1]	Analytic: Twitter metrics	Technical	309,740 Twitter users (with followers and tweets)
[2]	Examination: interviews	Communication	Interviews with 181 Twitter users
[3]	Examination: experiment	Education	Experiment with 125 students.
[4]	Analytic: linguistic (sentiment analysis)	Linguistics	20,000 tweets
[5]	Analytic: linguistic (event detection)	Linguistics	163,500,000 tweets
[6]	Analytic: linguistic (part of speech)	Linguistics	1,827 annotated tweets
[7]	Analytic: linguistic (sentiment analysis)	Linguistics	475,000,000 tweets
[8]	Analytic: quantitative (network analysis)	Security	17,803 tweets from 8,616 users + 1st degree network (3,048,360 directed edges, 631,416 unique followers, and 715,198 unique friends)
[9]	Analytic: linguistic (sentiment analysis)	Linguistics	200,000 annotated tweets
[10]	Analytic: linguistic (conversation structures)	Linguistics	1.3 million Twitter conversations, with each conversation containing between 2 and 243 posts
[11]	Analytic: network analysis, Twitter metrics, clustering, content analysis	Classification	One person's Twitter network (652 followers, 114 followings). 3,112 tweets.
[12]	Analytic: network analysis	Geography	481,248 tweets, 1,953 user pairs
[13]	Analytic: content analysis, Twitter metrics	Communication	102,500 tweets
[14]	Examination: experiment	Business	Experiment with 1,677 participants
[15]	Design and Development: linguistic (method development)	Linguistics	449 tweets sampled from 1.5 GB of Twitter data
[16]	Examination: survey	Classification	Survey with 505 young American adults
[17]	Design and Development: event detection (method development)	Geography	21,623,947 geo-tagged tweets
[18]	Analytic: Twitter metrics, linguistic (sentiment analysis)	Politics	104,003 tweets
[19]	Analytic: content analysis	Business	93 user profiles and 930 tweets
[20]	Analytic: content analysis, Twitter metrics Examination: survey	Education	4,574 tweets Qualitative survey with 11 participants
[21]	Analytic: content analysis	Communication	22,248 tweets
[22]	Analytic: network analysis, Twitter metrics	Geography	99,832 tweets
[23]	Analytic: Twitter metrics, linguistic	Geography	1,535,929,521 tweets from 71,273,997 users
[24]	Analytic: content analysis	Politics	4,869,264 tweets (and 43,378 YouTube URLs)
[25]	Examination: experiment	Education	Two experiments with 125 and 135 students.

Table 2. Analysis of methods, domains and datasets in the selected publications.

Weller, K. (2014). What do we get from Twitter – and what not? A close look at Twitter research in the social sciences. Knowledge Organization. 41(3), 238-248

Social Media Research



Number of publications per year, which mention the respective social media platform's name in their **title**. Scopus Title Search. For details: http://kwelle.wordpress.com/2014/04/07/bibliometric-analysis-of-social-media-research/

One of the Challenges: Data Collection and 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.

Example 2008-2013 papers on Twitter and elections: data sources

Data source	number
No information	11
Collected manually from Twitter website (Copy-Paste / Screenshot)	6
Twitter API (no further information)	8
Twitter Search API	3
Twitter Streaming API	1
Twitter Rest API	1
Twitter API user timeline	1
Own program for accessing Twitter APIs	4
Twitter Gardenhose	1
Official Reseller (Gnip, DataSift)	3
YourTwapperKeeper	3
Other tools (e.g. Topsy)	6
Received from colleagues	1

Weller, K. (2014). Twitter und Wahlen: Zwischen 140 Zeichen und Milliarden von Tweets. In: R. Reichert (Ed.), Big Data: Analysen zum digitalen Wandel von Wissen, Macht und Ökonomie (pp. 239-257). Bielefeld: transcript.

Archiving Twitter Datasets? Current approaches

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	User Name	Universal Time S	Local Time Stamp	Text	Langua	Profile Image	Source	Location	Time Zone	Geo	Hashtags	Urls	User Mention
2	333967655046373376	dimazest	5/13/2013 3:31:	5/13/2013 4:31:08 PM	Finally I've arrived to #www2013 and queuing for the	en	http://a0.twir	<a href="http	Groningen, Th	Rome		www2013		
3	333967116329971713	hinoko0529	5/13/2013 3:29:	5/14/2013 12:29:00 AM	WWW2013の昼の部は行けないくさいな	ja	http://a0.twir	<a href="http	東京	Tokyo				
4	333965108336267264	www2013rio	5/13/2013 3:21:	5/13/2013 12:21:01 PM	#www2013 Pictures on Flickr: http://t.co/TDbPwBQ3	en	http://a0.twir	web	Rio de Janeirc	Brasilia		www2013	http://www.f	
5	333962718245711872	soyuzsistemas	5/13/2013 3:11:	5/13/2013 12:11:31 PM	A Soyuz foi responsável pela codificação do site do n	pt	http://a0.twir	<a href="http	São Paulo - Br	Brasilia			http://fb.me/	
6	333962442667347968	www2013rio	5/13/2013 3:10:	5/13/2013 12:10:25 PM	#www2013 Participants, if you have any Internet pro	en	http://a0.twir	web	Rio de Janeirc	Brasilia		www2013		
7	333961564828876801	wisdelft	5/13/2013 3:06:	5/13/2013 4:06:56 PM	RT @aleboz: #WWW2013 when doing experiments w	en	http://a0.twir	<a href="http	Delft, the Net	Amsterdam		WWW2013	http://www2.	aleboz
8	333961480972152834	mdaquin	5/13/2013 3:06:	5/13/2013 3:06:36 PM	RT @stefandietze: #linkeddata and Online Learning tı	en	http://a0.twir	<a href="http	Milton Keynes	London		linkeddata linl		stefandietze
9	333961423434702848	mdaquin	5/13/2013 3:06:	5/13/2013 3:06:22 PM	MT @alexmikro: Just started: Online Learning and Lir	en	http://a0.twir	<a href="http	Milton Keynes	London		www2013		alexmikro euc
10	333961221181157378	aleboz	5/13/2013 3:05:	5/13/2013 4:05:34 PM	#WWW2013 when doing experiments with users, car	en	http://a0.twir	<a href="http	Delft, The Net	Rome		WWW2013	http://www2.	
11	333959973027594242	rtroncy	5/13/2013 3:00:	5/13/2013 4:00:36 PM	RT @mzurko: ooops! @micbuffa discovers that the p	en	http://a0.twir	<a href="http	Sophia Antipo	Paris		www2013		mzurko micbu
12	333959407513792512	mzurko	5/13/2013 2:58:	5/13/2013 9:58:22 AM	Final section of HTML5 talk slides http://t.co/I3PzIW	en	http://a0.twir	web		Quito		www2013	http://mainlin	
13	333957923514482690	jetztmoginimmer	5/13/2013 2:52:	5/13/2013 3:52:28 PM	RT @mzurko: ooops! Michel discovers that the passv	en	http://a0.twir	web		Berlin		www2013		mzurko
14	333957911611052032	thiagodini	5/13/2013 2:52:	5/13/2013 11:52:25 AM	Os amigos que já chegaram ao Rio para a #www2013	pt	http://a0.twir	web	Brasil_Rio de .	Brasilia		www2013		
15	333957828970680320	nyankomoti_64	5/13/2013 2:52:	5/13/2013 2:52:05 PM	ホープ君成長したらスノウ呼び捨てなんwww13	ja	http://a0.twir	<a href="http						
16	333957109010010112	mzurko	5/13/2013 2:49:	5/13/2013 9:49:14 AM	ooops! Michel discovers that the password is in fact	en	http://a0.twir	web		Quito		www2013		
17	333956676187205632	Debora_DG	5/13/2013 2:47:	5/13/2013 3:47:30 PM	Very interesting tutorial on how to measure User Eng	en	http://a0.twir	web	Brussels - Belg	Brussels		www2013	http://labtom	
18	333956463057833984	dirkahlers	5/13/2013 2:46:	5/13/2013 3:46:40 PM	The work-in-progress poster area at #www2013 look	en	http://a0.twir	<a href="http	Trondheim an	Berlin		www2013		
19	333956173978013697	fukutax	5/13/2013 2:45:	5/13/2013 11:45:31 PM	RT @takechan2000: WWW2013カンファレンスキ	ja	http://a0.twir	<a href="http	Hamamatsu, J	Tokyo				takechan2000
20	333956106797871104	markuslanthaler	5/13/2013 2:45:	5/13/2013 3:45:15 PM	Great #HTML5 tutorial by @micbuffa with lots of der	en	http://a0.twir	<a href="http		Rome		HTML5 www2	http://bit.ly/1	micbuffa
21	333955660553256961	mzurko	5/13/2013 2:43:	5/13/2013 9:43:28 AM	RT @www2013rio: Keynote Luis von Ahn to advise er	en	http://a0.twir	web		Quito		www2013		www2013rio
22	333955400959422464	www2013rio	5/13/2013 2:42:	5/13/2013 11:42:26 AM	Keynote Luis von Ahn to advise entrepreneurs how to	en	http://a0.twir	web	Rio de Janeirc	Brasilia		www2013	http://bit.ly/1	
23	333955131978690560	Debora_DG	5/13/2013 2:41:	5/13/2013 3:41:22 PM	Is the material of the Intr Semantic Web & Linke	en	http://a0.twir	web	Brussels - Belg	Brussels		www2013		
24	333954996469108736	tomayac	5/13/2013 2:40:	5/13/2013 3:40:50 PM	Some impressions from #WWW2013. Common them	en	http://a0.twir	<a href="http	Hamburg, Ger	Berlin		WWW2013 T	http://twitpic	
25	333954893117276160	chrisDircom	5/13/2013 2:40:	5/13/2013 3:40:25 PM	RT @ocsigen: Members of the Ocsigen team are in Ri	en	http://a0.twir	web	PMA	Paris		www2013	http://fb.me/	ocsigen
26	333954704755265536	moru1203	5/13/2013 2:39:	5/13/2013 2:39:40 PM	@rikiozu まじかwww13がラスト？	ja	http://a0.twir	<a href="http						rikiozu
27	333954630012780544	ocsigen	5/13/2013 2:39:	5/13/2013 3:39:23 PM	Excellent HTML5 tutorial by Michel Buffa #www2013	en	http://a0.twir	<a href="http		Amsterdam		www2013	http://fb.me/	
28	333954511599181827	TrinkerMedia	5/13/2013 2:38:	5/13/2013 2:38:54 PM	A good challenge: Measuring user engagement... by c	en	http://a0.twir	<a href="http	UK	London		methodology	http://www2(	
29	333954343525044224	dirkahlers	5/13/2013 2:38:	5/13/2013 3:38:15 PM	This week's topic #www2013 http://t.co/7spFCGCW	en	http://a0.twir	<a href="http	Trondheim an	Berlin		www2013		
30	333953031949398017	dirkahlers	5/13/2013 2:33:	5/13/2013 3:33:02 PM	RT @edgarmeij: Slides, code and biblio for our #WW	en	http://a0.twir	<a href="http	Trondheim an	Berlin		WWW2013	http://ejmeij.p	edgarmeij
31	333952834963927040	ocsigen	5/13/2013 2:32:	5/13/2013 3:32:15 PM	Members of the Ocsigen team are in Rio de Janeiro f	en	http://a0.twir	<a href="http		Amsterdam		www2013	http://fb.me/	

gesis
Leibniz-Institut für Sozialwissenschaften

Format supported by
Twitter Terms of services

8	339614809721528 34
9	339614234347028 48
10	339612211811573 78
11	339599730275942 42
12	339594075137925 12
13	339579235144826 90
14	339579116110520 32
15	339578289706803 20
16	339571090100101 12
17	339566761872056 32
18	339564630578339 84
19	339561739780136 97
20	339561067978711 04
21	339556605532569 61
22	339554009594224 64
23	339551319786905 60
24	339549964691087 36
25	339548931172761 60
26	339547047552655 36
27	339546300127805 44
28	339545115991818 27
29	339543435250442 24
30	339530319493980 17
31	339528349639270 40

Available datasets

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

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

Challenges in Archiving Twitter Data

Sources for Challenges

(1) the Twitter Terms of Services

(2) ethical challenges

(3) lack of standard metadata and collection methods

(4) the ever changing nature of Twitter – and Twitter users

Sources for Challenges

(1) the Twitter Terms of Services

(2) ethical challenges

(3) lack of standard metadata and collection methods

(4) the ever changing nature of Twitter – and Twitter users

The changing nature of Twitter in 5 examples

#1

Deleted content

#2

Lost context: interfaces, look and feel

#3

Lost context: stories, meanings

#4

Lost context: user names

#5

URLs and images

gesis
Leibniz-Institut für Sozialwissenschaften

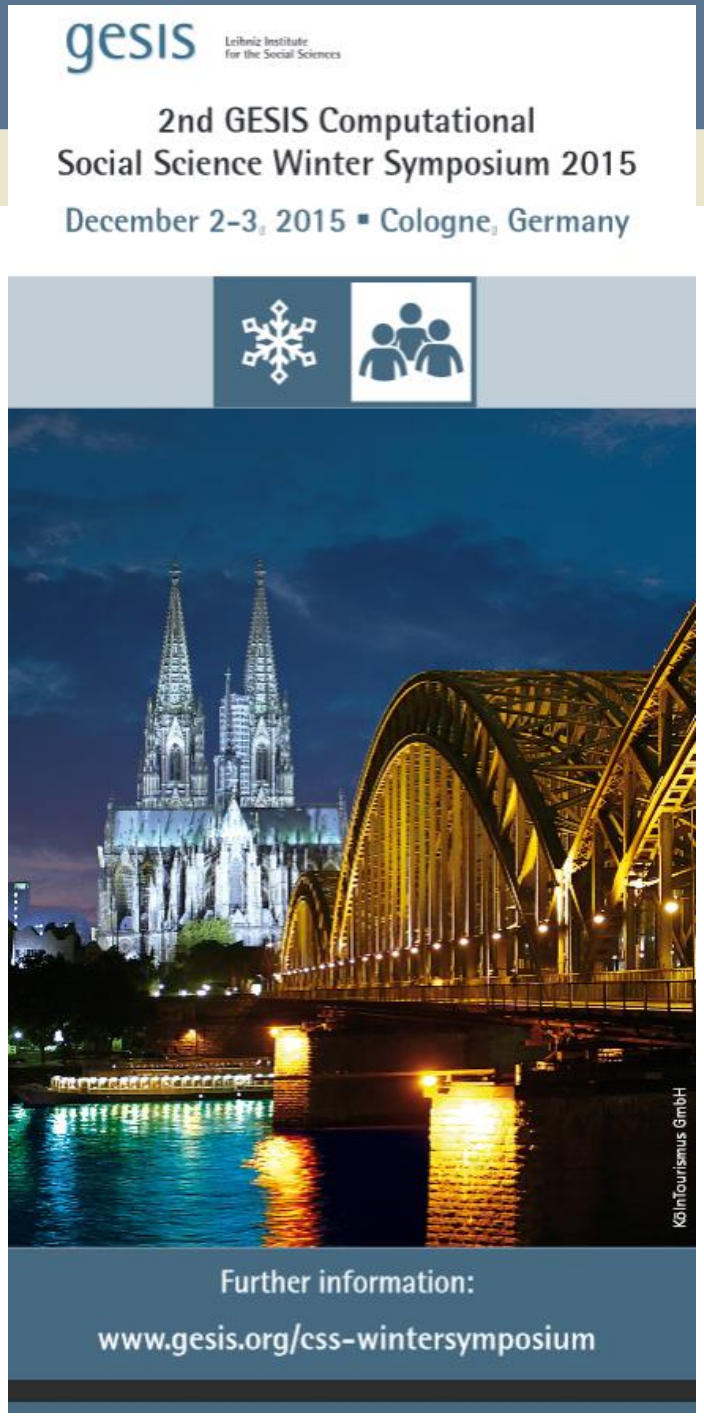
Questions and Feedback

katrin.weller@gesis.org

@kwelle

http://katrinweller.net

www.gesis.org/css-wintersymposium



Supplement: some useful references

Tools / Methods for collecting tweets:

- 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.

[There are much more tools, though. See, e.g. collection at: https://docs.google.com/document/d/1UaERzROI986HqcwrBDLaqGG8X_lYwctj6ek6ryqDOiQ/edit (curated by D. Freelon).

Supplement: some useful references

Challenges in collecting tweets / data quality:

- Bruns, A. (2011, June 21). Switching from Twapperkeeper to yourTwapperkeeper. Retrieved January 31, 2015 from http://www.mappingonlinepublics.net/2011/06/21/switching-from-twapperkeeper-to-yourtwapperkeeper/.
- Bruns, A. and Stieglitz, S. (2014), "Twitter data: what do they represent?" *IT Information Technology*, Vol. 59 No. 5, pp. 240-5, [online], available at: http://www.degruyter.com/view/j/itit.2014.56.issue-5/itit-2014-1049/itit-2014-1049.xml (accessed 28 February 2015), DOI: 10.1515/itit-2014-1049.
- Jungherr, A., Jurgens, P. and Schoen, H. (2012), "Why the Pirate Party won the German Election of 2009 or The trouble with predictions: a response to Tumasjan, A., Sprenger, T. O., Sander, P. G. and Welpe, I. M. Predicting elections with Twitter: what 140 characters reveal about political sentiment", *Social Science Computer Review*, Vol. 30 No. 2, pp. 229-34, [online], available at: http://ssc.sagepub.com/content/30/2/229 (accessed 28 February 2015), DOI: 10.1177/0894439311404119.
- Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." http://arxiv.org/abs/1306.5204.
- Sumers, E. (2015). Tweets and Deletes. Retrieved, June 9, 2015 from: https://medium.com/on-archivy/tweets-and-deletes-727ed74f84ed (see also: https://github.com/edsu/twarc)

Supplement: some useful references

Bibliometric studies of Twitter researchers:

- Williams, S. A., Terras, M. M., & Warwick, C. (2013a). What do people study when they study Twitter? Classifying Twitter related academic papers. Journal of Documentation, 69(3): 384-410.
- Williams, S. A., Terras, M. M., & Warwick, C. (2013b). How Twitter Is Studied in the Medical Professions: A Classification of Twitter Papers Indexed in PubMed. In Med 2.0 2013. doi: 10.2196/med20.2269.
- Weller, K. (2014b). What do we get from Twitter – and what not? A close look at Twitter research in the social sciences. Knowledge Organization 41(3), 238-248.
- Zimmer, M., & Proferes, J.N. (2014). A topology of Twitter research: disciplines, methods, and ethics. Aslib Journal of Information Management, 66(3), 250–261. doi:10.1108/AJIM-09-2013-0083

Supplement: some useful references

Critical perspectives on data access and inequalities:

- Boyd, D. and Crawford, K. (2012), "Critical questions for Big Data: provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Society*, Vol. 15 No. 5, pp. 662-79, [online], available at: http://www.tandfonline.com/doi/full/10.1080/1369118X.2012.678878#abstract (accessed 28 February 2015), DOI: 10.1080/1369118X.2012.678878.

Supplement: some useful references

Legal and ethical challenges:

- Beurskens, M. (2014). Legal Questions of Twitter Research. In K. Weller, A. Bruns, J. Burgess., M. Mahrt and C. Puschmann (Eds.), Twitter and Society (pp. 123-133). New York: Peter Lang.
- Markham, A. and Buchanan, E. (2012), "Ethical decision-making and internet research 2.0: recommendations from the AoIR Ethics Working Committee", available at: http://www.aoir.org/reports/ethics2.pdf (accessed 28 February 2015).
- Weller, Katrin, and Katharina E. Kinder-Kurlanda. 2014. "I love thinking about ethics: Perspectives on ethics in social media research." In *Selected Papers of Internet Research (SPIR). Proceedings of ir15 - Boundaries and Intersections*, http://spir.aoir.org/index.php/spir/article/view/997.
- Zimmer, M. & Proferes, J.N. (2014). Privacy on Twitter, Twitter on privacy. In Weller, A. Bruns, J. Burgess., M. Mahrt and C. Puschmann (Eds.), Twitter & Society (pp. 169-182), New York: Peter Lang.
- Zimmer, M. (2010), "But the data is already public: on the ethics of research in Facebook", *Ethics and Information Technology*, Vol. 12 No. 4, pp. 313-25, DOI: 10.1007/s10676-010-9227-5.

Supplement: some useful references

Twitter's activities:

- Krikorian, R. (2014a), "Introducing Twitter Data Grants", [online], available at: https://blog.twitter.com/2014/introducing-twitter-data-grants (accessed 28 February 2015).
- Krikorian, R. (2014b), "Twitter #DataGrants selections", available at: https://blog.twitter.com/2014/twitter-datagrants-selections (accessed 28 February 2015).
- Stone, B. (2010). Tweet Preservation. Blog post, April 14, 2010. Retrieved from https://blog.twitter.com/2010/tweet-preservation
- Twitter (2014). Developer Agreement & Policy. Twitter Developer Agreement, retrieved January 31, 2015 from https://dev.twitter.com/overview/terms/agreement-and-policy.
- Twitter (no date). Guidelines for using Tweets in broadcast, retrieved January 31, 2015, from https://support.twitter.com/articles/114233.

Supplement: some useful references

Library of Congress' activities:

- Allen, E. (2013, January 4). Update on the Twitter Archive at the Library of Congress. Retrieved January 31, 2015 from http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/
- McLemmee, S. (2015). The Archive is closed. Inside Higher Education. Retrieved June 9, 2015 from: https://www.insidehighered.com/views/2015/06/03/article-difficulties-social-media-research
- Raymond, M. (2010). How Tweet It Is! Library Acquires Entire Twitter Archive. Retrieved January 31, from http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/

Supplement: some useful references

Examples of Twitterdatasets shared publicly:

- CrisisLex on Github: https://github.com/sajao/CrisisLex/tree/master/data/CrisisLexT26/
- Hadgu & Jäschke 2014 dataset on Github: https://github.com/L3S/twitter-researcher
- ICWSM 2012 datasets: http://www.icwsm.org/2012/submitting/datasets/ ICWSM 2014 datasets: http://www.icwsm.org/2014/datasets/datasets/
- MPI-SWS (no date). The Twitter Project Page at MPI-SWS. Retrieved January 26, 2015 from http://twitter.mpi-sws.org/ (Archived by WebCite® at http://www.webcitation.org/6VsuuxQlU)
- TREC 2011: http://trec.nist.gov/data/tweets/
- sananalytics (2011). Public domain twitter sentiment corpus. Post in Twitter Developers Forums. Retrieved Jan 31, 2015 from https://twittercommunity.com/t/public-domain-twitter-sentiment-corpus/13290
- Kaczmirek, Lars; Mayr, Philipp (2015): German Bundestag Elections 2013: Twitter usage by electoral candidates. GESIS Data Archive, Cologne. ZA5973 Data file Version 1.0.0,doi:10.4232/1.12319