



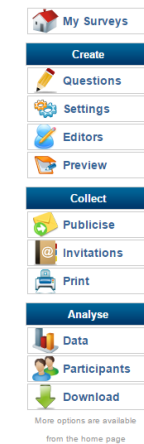
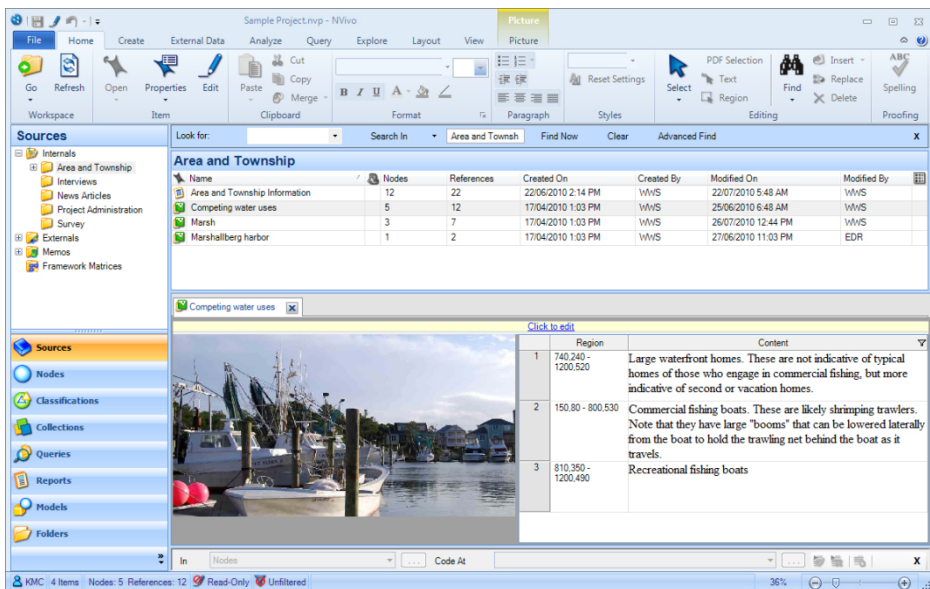
The Problem with Invisible Infrastructures: big data, social research and the challenge for curation

Susan Halford

Professor of Sociology, Director Web Science Institute

@susanjhalford

- Interviewer: working on beforehand. I suppose that there are, of course, challenges around all this, the first being cost. Especially historical datasets on Twitter or datasets across another range of other social media platforms usually have to be bought. And especially if they're – depending on the technical capacity of the research team, you often need actually commercial service to acquire and scrape and bring together all the different 20/30 social media sources of data into a single place and in a single canonical format which a research can use. So you often need subscriber access to say a platform like Datasift or Pulsar in order to get the data in ways which is reasonably convenient and consistent to be used.
- Interviewer: Sorry to interrupt. Do you have a preferred supplier then and do you have an ongoing deal?
- Respondent: No, we don't. We actually usually go to source for the datasets that we need. We also work quite a lot with the social media platforms so Facebook has funded a number of different studies that we've done and we often include our own legal architecture which allows Facebook to share data with us, and we use in ways that of course they are happy with and consistent with their privacy policies. But then we are lucky because half of our team are data scientists and software engineers, so if need to connect one of my analytical platforms up with a new API I can really pick up the phone and someone can do that for me, which I know is absolutely not necessarily the story for all researchers.
- Interviewer: Do you have a special relationship with Twitter and Facebook and other companies?
- Respondent: It depends what you mean by a special relationship. We've got an ongoing dialogue with them. I know people that work there quite well and, as I said, we have a funding relationship sometimes with Facebook. We go in and present our work fairly often and do our best to keep the platforms up to speed with what we are doing. I wouldn't say we have a special relationship though anymore than I'm sure talking to loads of researchers.
- Interviewer: Obviously there's going to be differentiation in the levels of access, different institutions and it does help to build up a relationship with them, a good working relationship. So I would be interested to compare your access to other institutions or other organisations that are trying to do similar work to you.
- Respondent: Yes, and I suppose that's the 'in principle' difficulty is that these datasets are proprietary. It's absolutely the right of a social media platform to decide to change their terms and conditions of data



Social Media Data for Social Research

1. Social Media Data for Social Research

[Return to Sections](#) [Section settings](#) [Preview this page](#)



Q1 [Edit](#) [Copy](#) [Print](#) [Delete](#)

Question 1.1 Preview

Section 1: Some Information About You [Question ID : 893831]

[Add New Question Here](#)

Q2 [Edit](#) [Copy](#) [Print](#) [Delete](#)

[Add new drop down options](#)

[Edit drop down responses](#)

Question 1.2 Preview

In which country are you mainly based? [Question ID : 894884]

Please select

[Add a question with quick links here](#)

[Add New Question Here](#)

Q3 [Edit](#) [Copy](#) [Print](#) [Delete](#)

[Add new drop down options](#)

[Edit drop down responses](#)

Question 1.3 Preview



The statistics are stunning: about 90% of all the data in the world has been generated in the past two years (a statistic holding roughly true as time passes). There are 2.7 zettabytes of data in the digital universe, where 1ZB is a billion terabytes (a typical computer hard drive these days can hold about 0.5TB...) IBM predicts that will hit 8ZB by 2015. Facebook alone stores and analyses more than 50 petabites (50 000 TB) of data

digital technologies stop being a tool for data storage and analysis and the data that they generate become the substance of research.



We are now accumulating data at a scale unprecedented in social science research, capturing social activity in real time, over time: the traces of what people do and say ‘in the wild’, rather than what they say they do in interviews and surveys, with all kinds of possibilities for high speed data mining and data linkage between diverse sources of data.

Challenges for Social Research

- The data
- Methodology
- Interdisciplinarity

How each of these challenges is addressed and resolved (or not) will shape what the data become for future generations of researchers.

Challenge 1: What are these Digital Data?

- Naturally occurring data?
- Big Data as a telescope for Social Science?



***These data are mediated by the
processes that produce them***



DATA NEVER SLEEPS 3.0

How much data is generated **every minute**?

Data is being created all the time without us even noticing it. Much of what we do every day now happens in the digital realm, leaving an ever-increasing digital trail that can be measured and analyzed. Just how much data do our tweets, likes and photo uploads really generate? For the third time, Domo has the answer—and the numbers are staggering.



THE GLOBAL INTERNET POPULATION GREW 18.5% FROM 2013-2015 AND NOW REPRESENTS

3.2 BILLION PEOPLE.

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. [Learn more at www.domo.com](http://www.domo.com).



SOURCES:

FACEBOOK, TWITTER, YOUTUBE, INSTAGRAM, PINTEREST, APPLE, NETFLIX, REDDIT, AMAZON, TINDER, BUZZFEED, STATISTA, INTERNET LIVE STATS, STATISTICBRAIN.COM

<http://thinktostart.com/how-much-big-data-is-generated-every-minute-on-top-digital-and-social-media-infographic>

Challenge 1: What are these Digital Data?

- Big data are mediated by the processes that produce them:
 - By technical infrastructures
 - By iterative social practices



... the data that we have are reflections of both the technical functionalities and the emergent social practices of the digital artefacts that produce the data ... not the unmediated reflections of a separate world that lies beyond

Challenge 1: What are these Digital Data?

- Big data are mediated by the processes that produce them:
 - By technical infrastructures & iterative data practices
 - By provenance and ownership

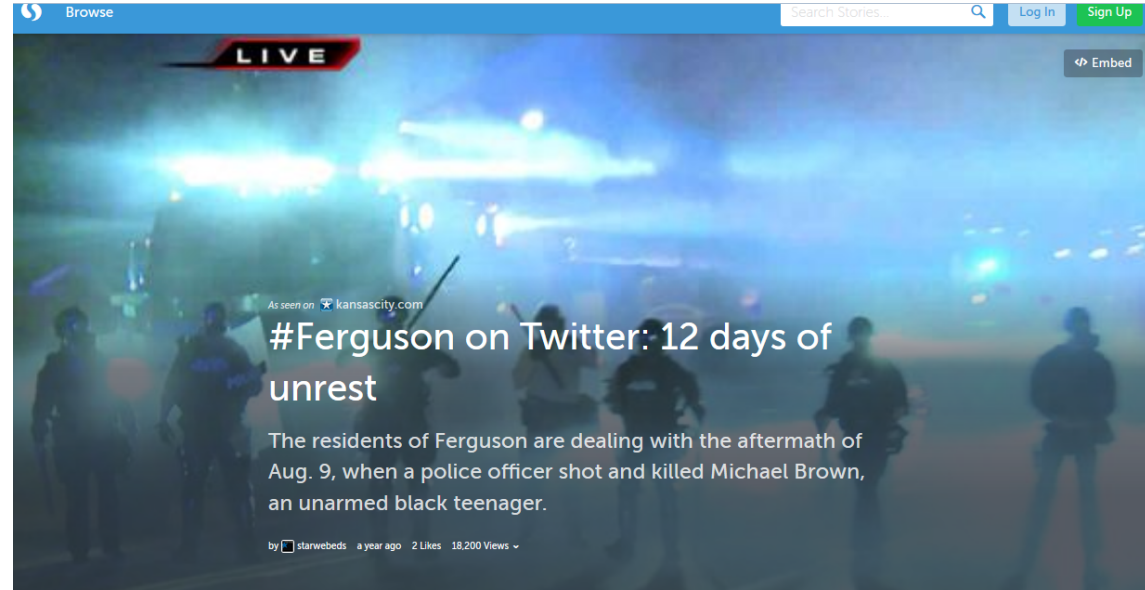
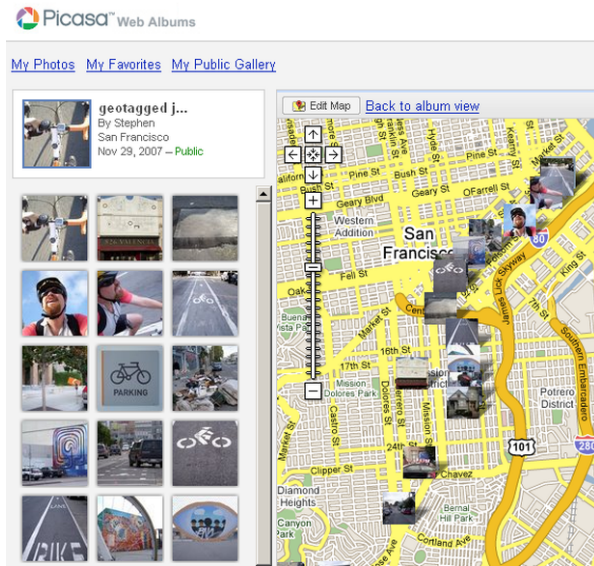


***These are thoroughly constructed data ...
as digital data become the substance of
our research so too should the
sociotechnical infrastructures that
produce them***

Challenge 2: Method



*This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear **With enough data, the numbers speak for themselves.***



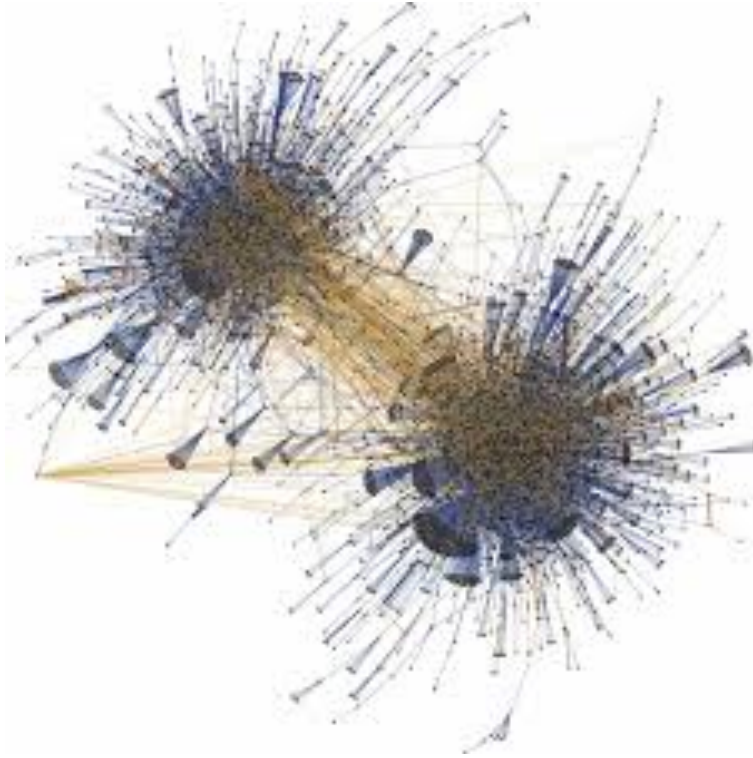
Information Age

Network Society

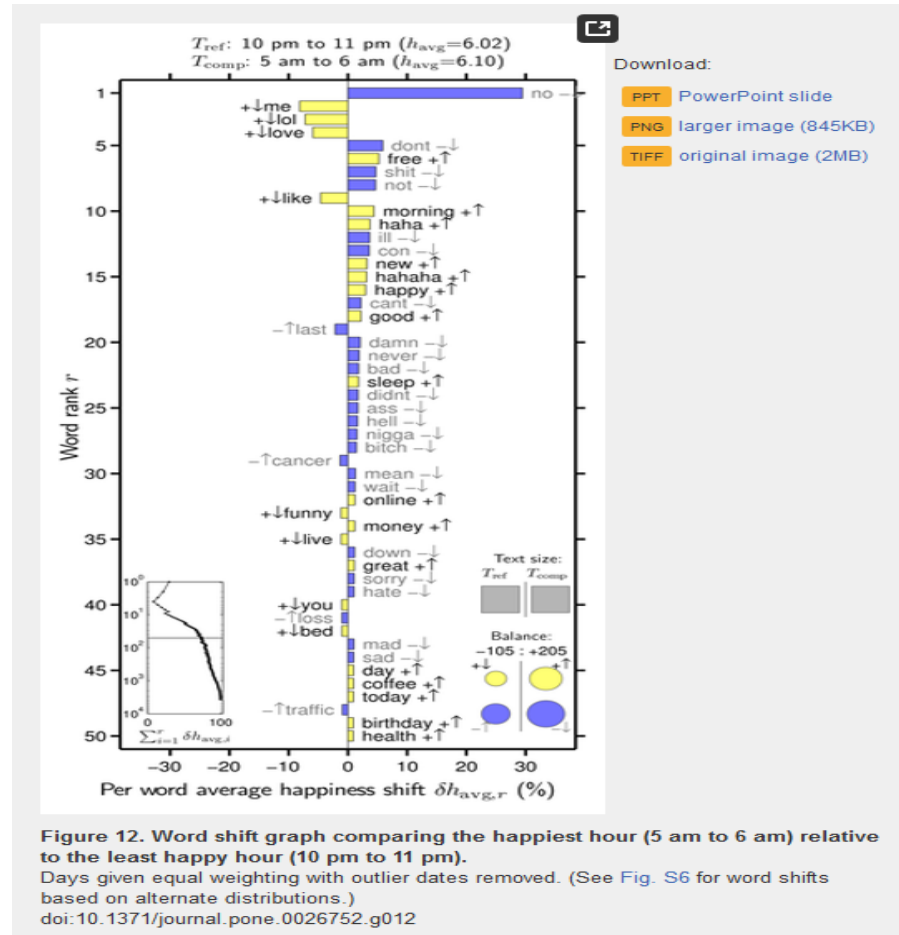
Mobilities

Flow

Methods? Small scale content &/or larger scale random/purposive sampling; static



Vespignani, A. (2012) Modelling dynamical processes in complex socio-technical systems
Nature Physics 8, 32–39



Dodds et al (2011) Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter PLoS ONE 6(12): e267

Methods – (socially) a-theoretical, largely mathematical and technical

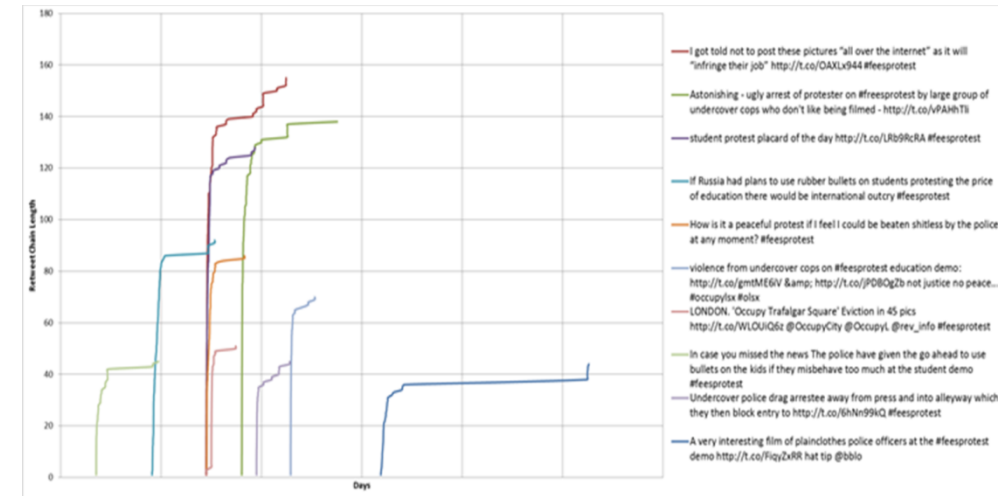



Figure 3: #feesprotest Ten Longest Retweet Chains

Challenge 3: Discipline

- Disciplines imagine their objects of study differently
- Disciplines operate with their own forms of cultural capital (... also economic and social capital!)
- Using big data for social research demands change from both sides
- We will face intransigencies
- There will be disciplinary politics

The Elephant in the Room

- ... the problem with invisible infrastructures
- Understanding sociotechnical infrastructures
- ... because good research needs good data

 **because good research needs good data**

Contact us

Search

[Home](#) [Digital curation](#) [About us](#) [News](#) [Events](#) [Resources](#) [Training](#) [Projects](#) [Community](#) [Tailored support](#)

[Home](#) > [Drupal](#) > [Events](#) > Idcc16

11th International Digital Curation Conference

"Visible data, invisible infrastructure"

22 - 25 February 2016
Mövenpick Hotel, Amsterdam City Centre, Amsterdam


Overview

Research data management, and digital curation generally, is becoming a mainstream academic activity. Universities and research institutions are ramping up support and infrastructural provision for it. Funding bodies are strengthening their requirements for it. Visionary researchers and practitioners are tackling the remaining barriers.

The success of this transition might be measured in two ways. First, is

[Register for this event](#)

Event Venue



About this event

- [Accommodation & Registration >](#)
- [Call for Papers >](#)
- [Dates >](#)
- [Day two papers >](#)
- [Demonstrations >](#)
- [Posters >](#)
- [Programme >](#)
- [Speakers >](#)
- [Submissions >](#)
- [Support IDCC16 >](#)
- [Workshop Submissions >](#)



WEB SCIENCE
Institute

UNIVERSITY OF
Southampton

The Problem with Invisible Infrastructures: big data, social research and the challenge for curation

Susan Halford

Professor of Sociology, Director Web Science Institute

@susanjhalford