BIG DATA FOR ALL

Toward ethical big data sharing

Sara Mannheimer, Montana State University
Elizabeth Hull, Dryad Digital Repository

#RDAP17 Seattle, Washington

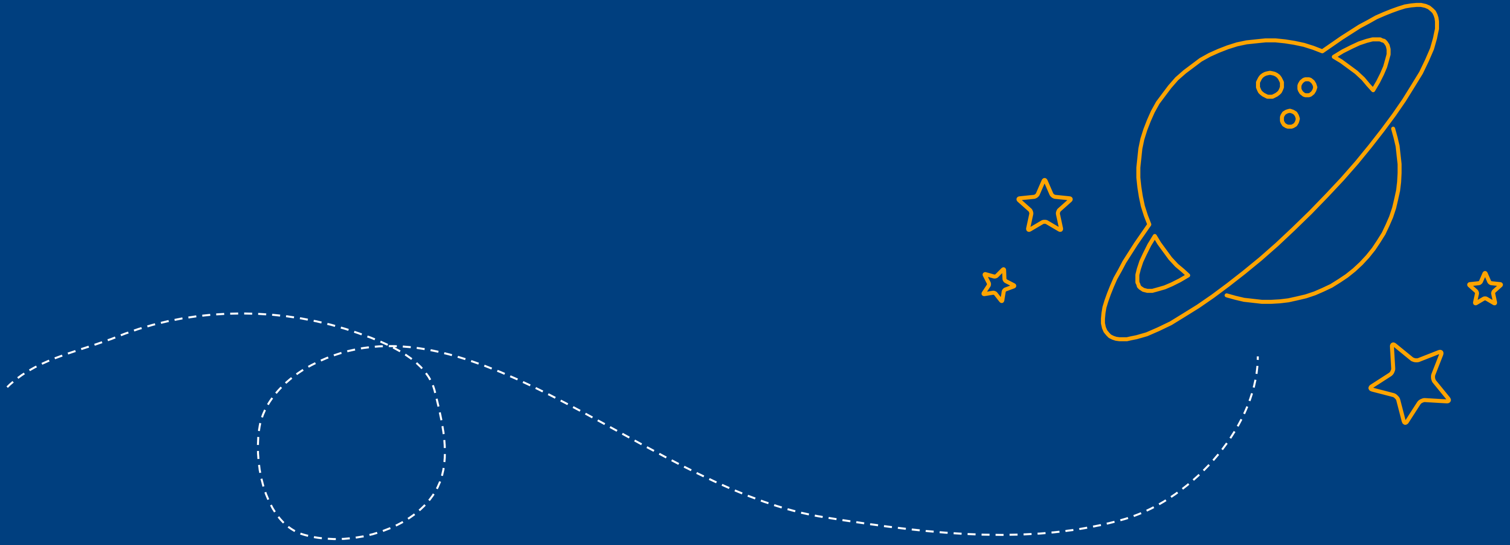
Sara Mannheimer
Data Management Librarian
Montana State University
@saramannheimer

Elizabeth Hull
Operations Manager
Dryad Digital Repository
@datadryad

GOALS FOR TODAY'S TALK

Pattern by Penelope Dullaghan

Present our framework and case studies



Present our framework and case studies

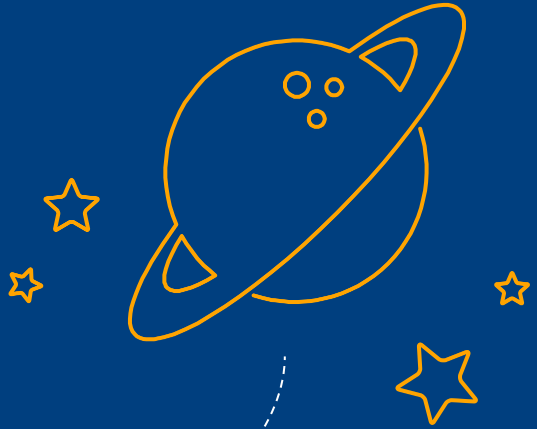
Encourage
big data sharing

Present our framework and case studies

Encourage *ethical* big data sharing

Present our framework and case studies

Encourage *ethical* big data sharing



Spark conversation

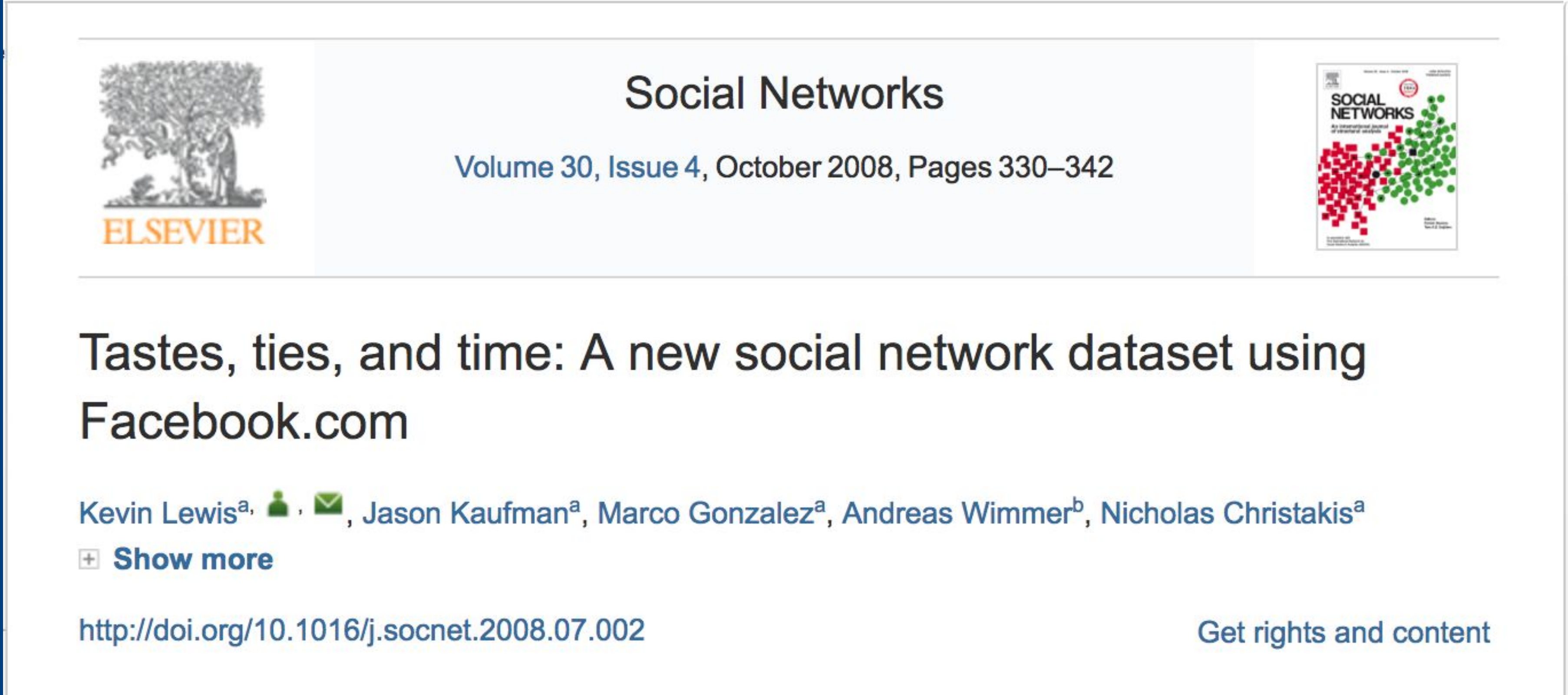
TODAY'S TALK

1. Some big data sharing missteps
2. Navigating big data research
3. Toward ethical big data sharing
4. Key takeaways

1.

SOME BIG DATA SHARING MISSTEPS

Pattern by Penelope Dullaghan



ELSEVIER

Social Networks

Volume 30, Issue 4, October 2008, Pages 330–342

Tastes, ties, and time: A new social network dataset using Facebook.com

Kevin Lewis[a], ✉, Jason Kaufman[a], Marco Gonzalez[a], Andreas Wimmer[b], Nicholas Christakis[a]

⊞ Show more

http://doi.org/10.1016/j.socnet.2008.07.002

Get rights and content

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook. com. Social networks, 30(4), 330-342. https://doi.org/10.1016/j.socnet.2008.07.002

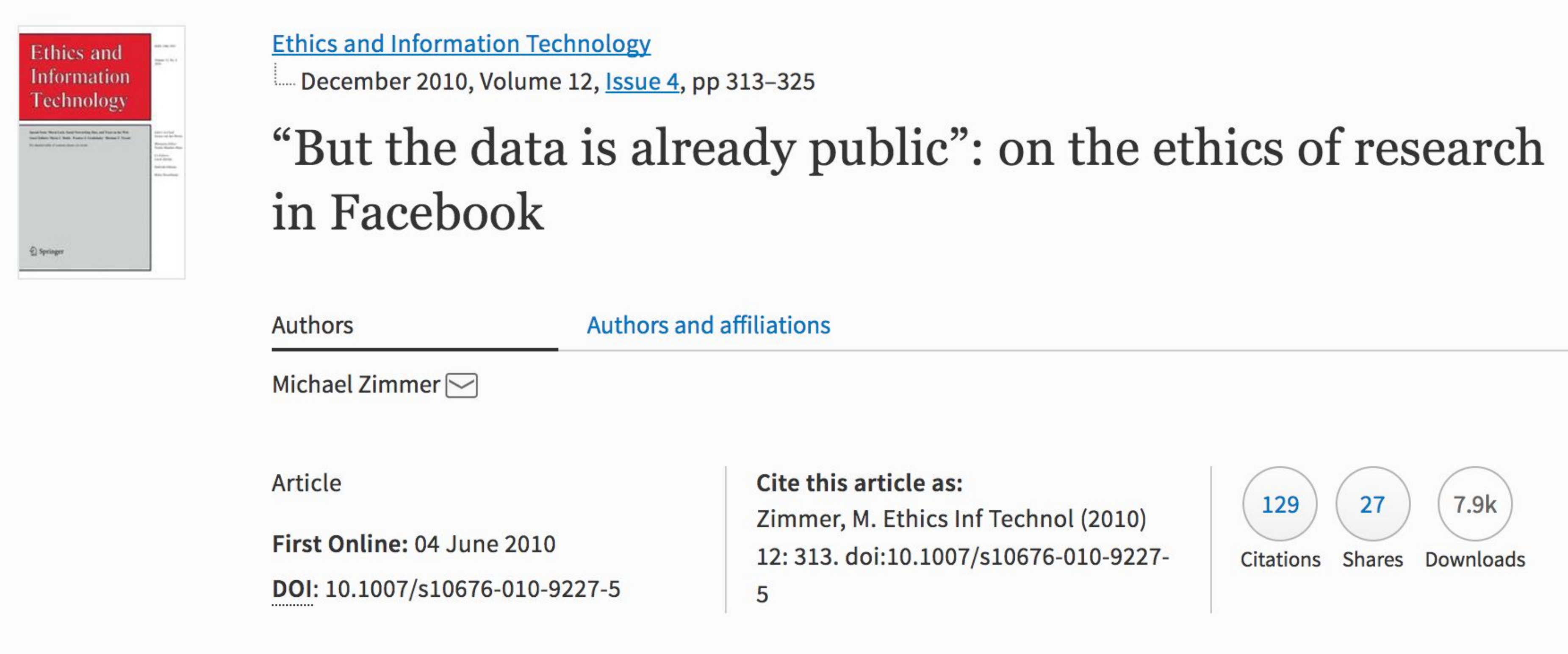
Data collection

- With permission from Facebook and the university, accessed Facebook and downloaded the profile and network data provided by one cohort of college students.
- Cohort of students agreed to a "Terms of Use" statement

“

Student privacy was assured by converting all names to numerical identifiers and promptly removing or encoding all other information that could be traced back to individual students.

—Lewis et al., 2008



Ethics and Information Technology

December 2010, Volume 12, Issue 4, pp 313–325

“But the data is already public”: on the ethics of research in Facebook

Authors | Authors and affiliations

Michael Zimmer

Article

First Online: 04 June 2010

DOI: 10.1007/s10676-010-9227-5

Cite this article as:
Zimmer, M. Ethics Inf Technol (2010) 12: 313. doi:10.1007/s10676-010-9227-5

129	27	7.9k
Citations	Shares	Downloads

Zimmer, M. (2010). “But the data is already public”: on the ethics of research in Facebook. Ethics and information technology, 12(4), 313-325. https://doi.org/10.1007/s10676-010-9227-5

Anonymization

- Description of college was too specific
- Dataset included each subject’s gender, race, ethnicity, home state, and major
- Only a single student each from Delaware, Louisiana, Mississippi, Montana, and Wyoming
- Only a single student each identified as Albanian, Hungarian, Malaysian, Nepali, Filipino, and Romanian

Data collection

- "Terms of Use" statement was not enough for informed consent

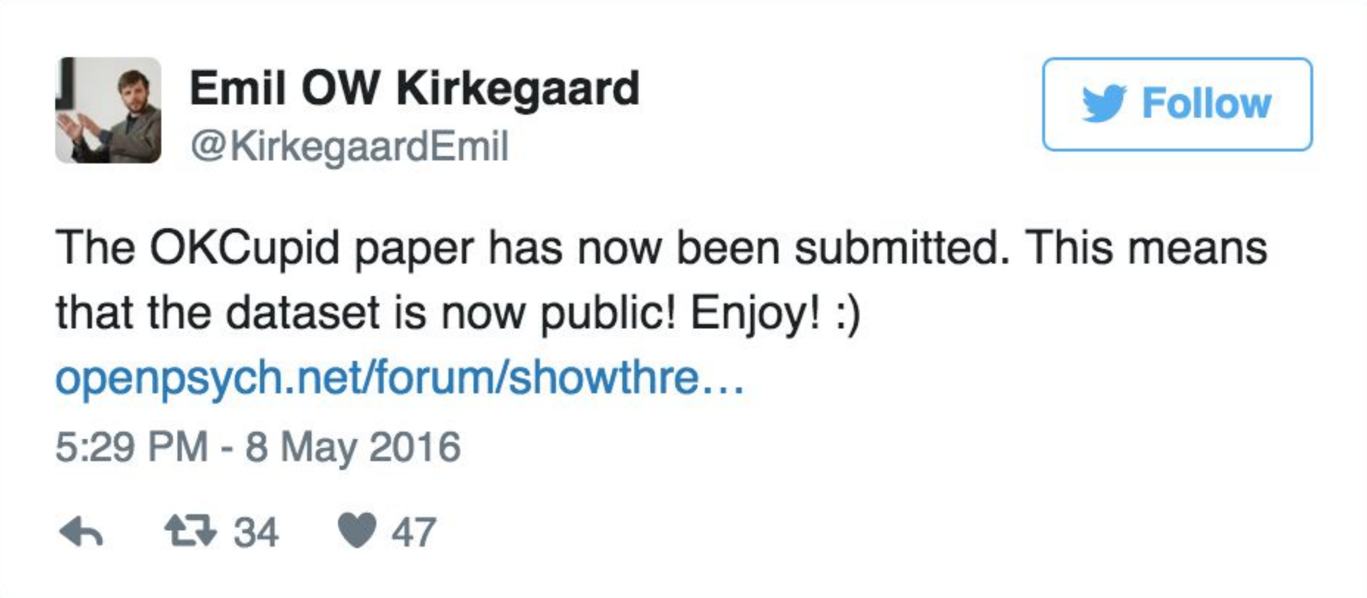
“

Our dataset contains almost no information that isn’t on Facebook. (Privacy filters obviously aren’t much of an obstacle to those who want to get around them.)

—Kaufman, 2008

“

> *Concerns over consent, privacy and anonymity do not disappear simply because subjects participate in online social networks; rather, they become even more important.*
>
> —Zimmer, 2010

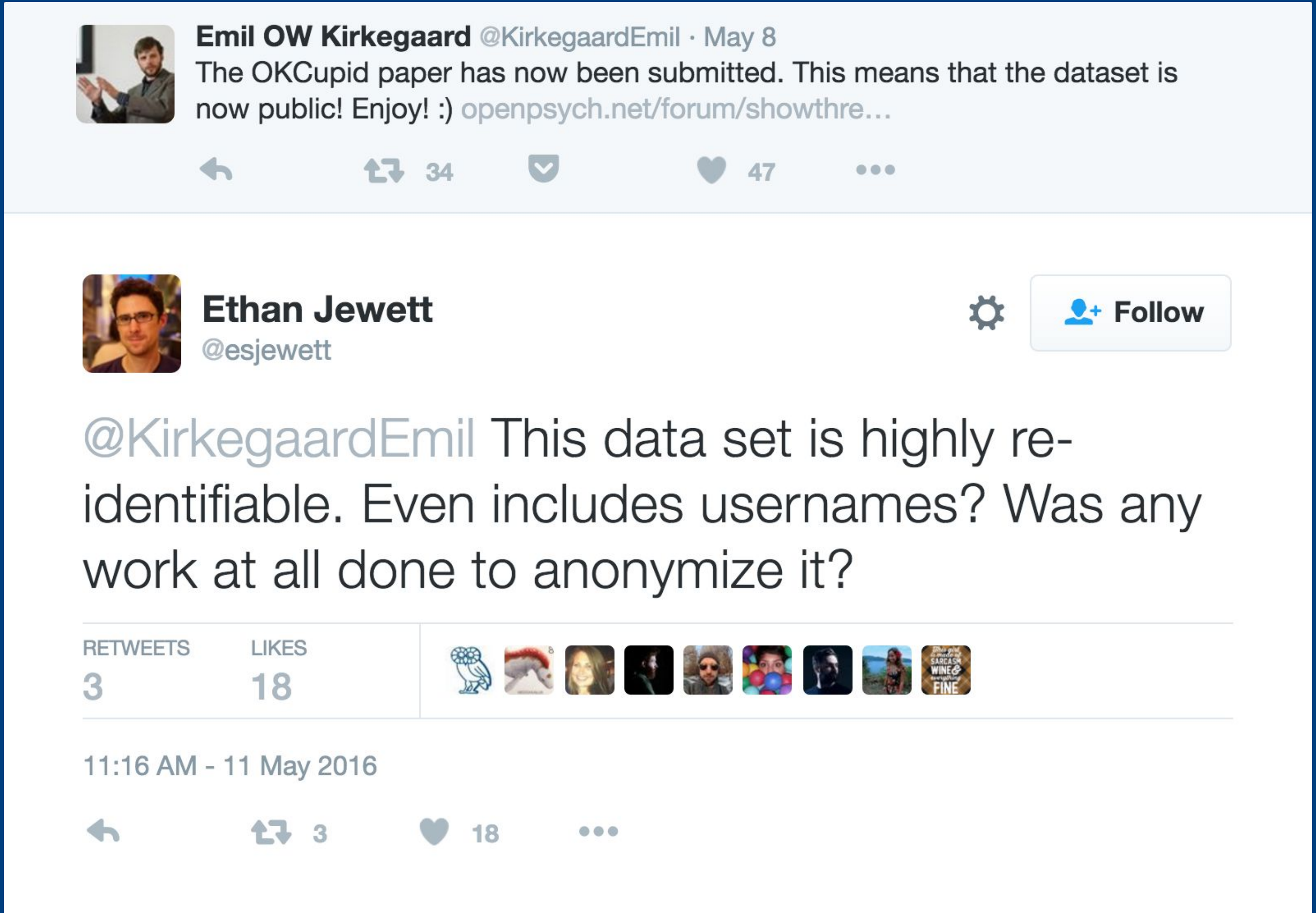


Emil OW Kirkegaard
@KirkegaardEmil

The OKCupid paper has now been submitted. This means that the dataset is now public! Enjoy! :)
openpsych.net/forum/showthre…

5:29 PM - 8 May 2016

https://twitter.com/KirkegaardEmil/status/729453122221936640



Emil OW Kirkegaard @KirkegaardEmil · May 8

The OKCupid paper has now been submitted. This means that the dataset is now public! Enjoy! :) openpsych.net/forum/showthre…

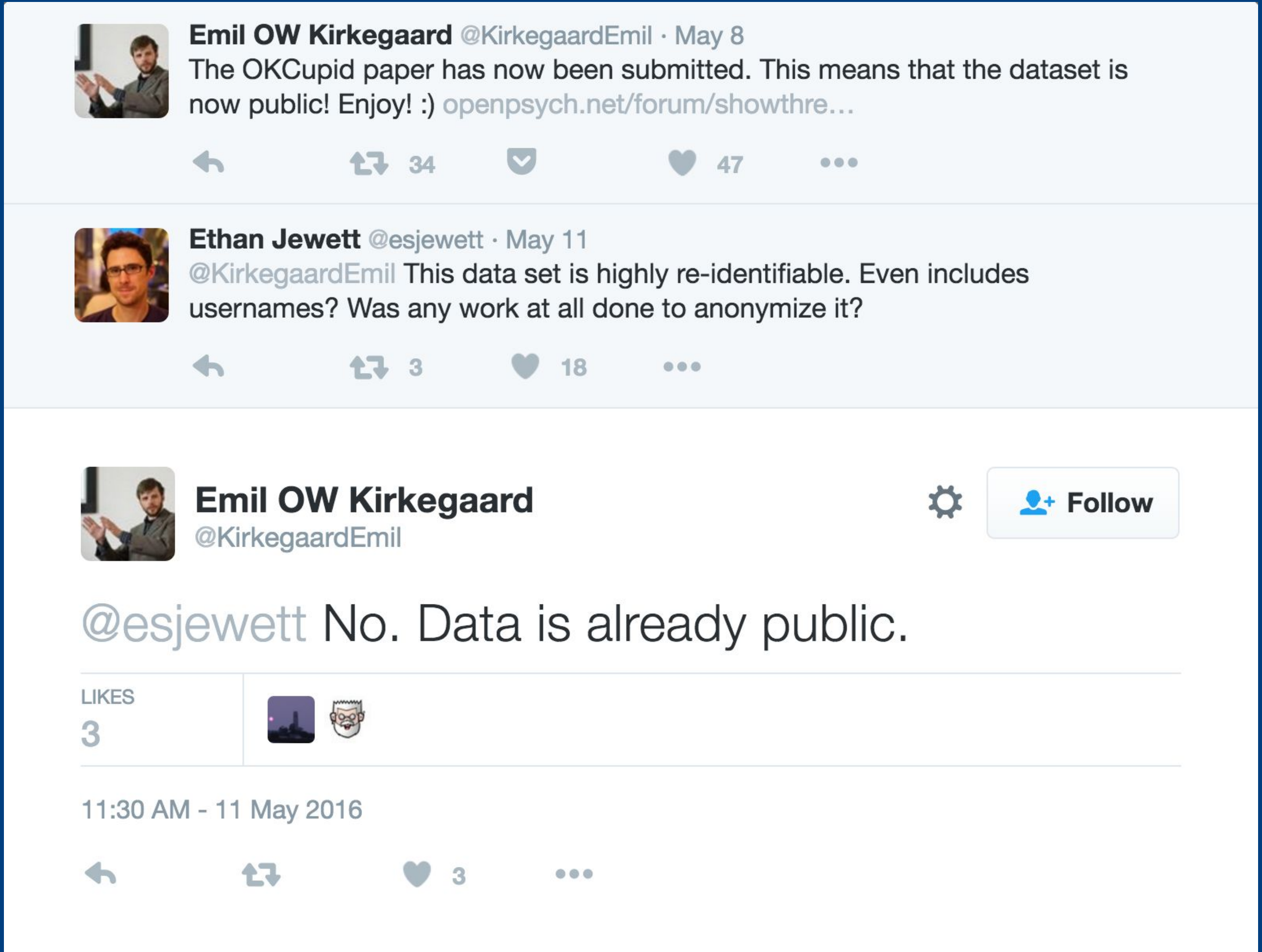
Ethan Jewett
@esjewett

@KirkegaardEmil This data set is highly re-identifiable. Even includes usernames? Was any work at all done to anonymize it?

RETWEETS 3

LIKES 18

11:16 AM - 11 May 2016



Emil OW Kirkegaard @KirkegaardEmil · May 8

The OKCupid paper has now been submitted. This means that the dataset is now public! Enjoy! :) openpsych.net/forum/showthre…

Ethan Jewett @esjewett · May 11

@KirkegaardEmil This data set is highly re-identifiable. Even includes usernames? Was any work at all done to anonymize it?

Emil OW Kirkegaard
@KirkegaardEmil

@esjewett No. Data is already public.

LIKES
3

11:30 AM - 11 May 2016

2.

NAVIGATING BIG DATA RESEARCH

Pattern by Penelope Dullaghan

PLOS | COMPUTATIONAL BIOLOGY

OPEN ACCESS

EDITORIAL

Ten simple rules for responsible big data research

Matthew Zook, Solon Barocas, danah boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A. Koenig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson, Frank Pasquale

Published: March 30, 2017 • https://doi.org/10.1371/journal.pcbi.1005399

Zook, M., Barocas, S., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., ... & Nelson, A. (2017). Ten simple rules for responsible big data research. PLOS Computational Biology, 13(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399

“

One of the most fundamental rules of responsible big data research is the steadfast recognition that ***most data represent or impact people*****.**

— Zook et al., 2017

> *We exhort researchers to … make grappling with ethical questions part of their standard workflow.*
>
> — Zook et al., 2017

Key issues

Informed consent

Privacy

Ownership

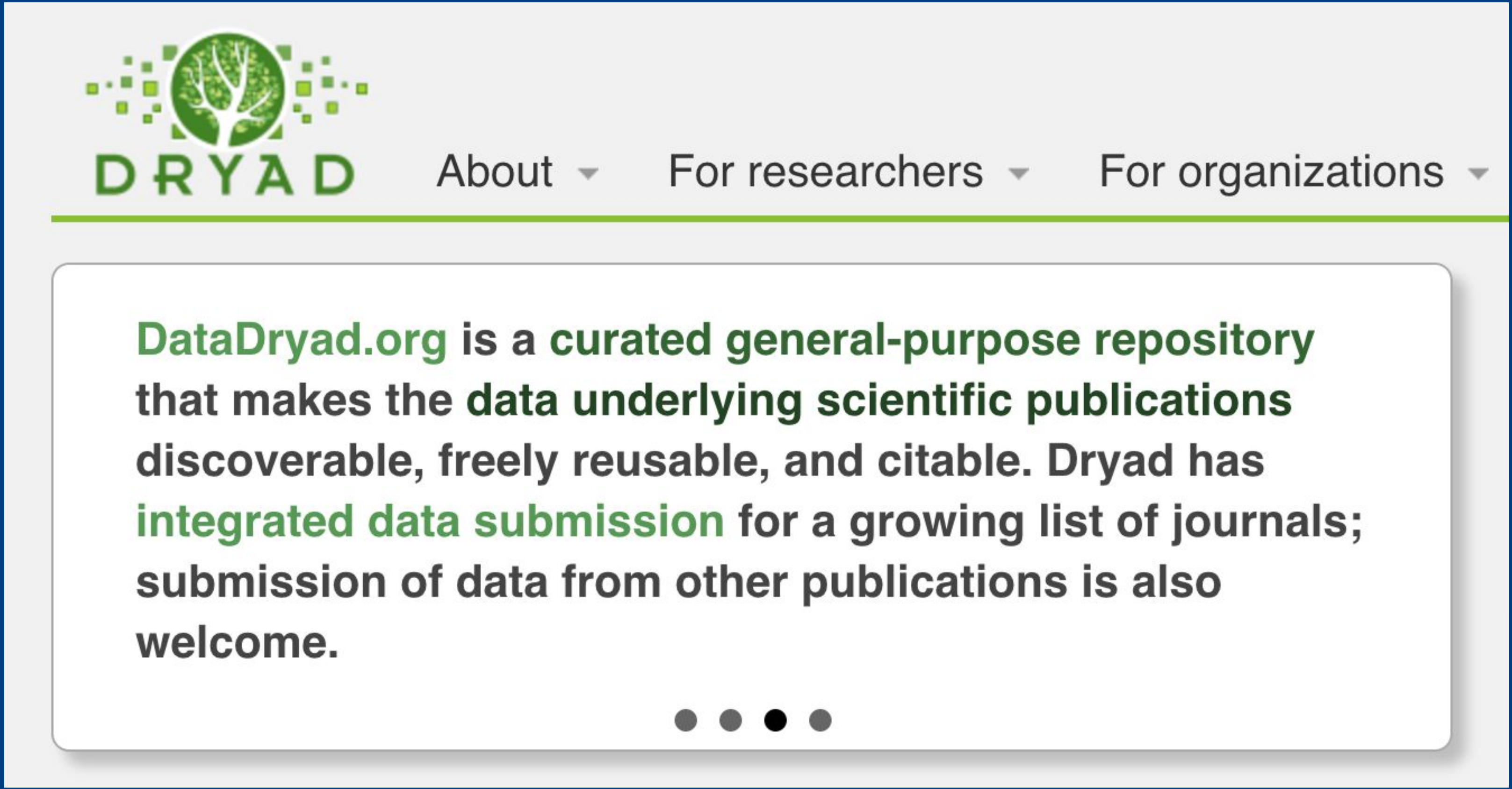
Big data disparity

Mittelstadt, B. D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. Science and Engineering Ethics, 22(2), 303-341. https://doi.org/10.1007/s11948-015-9652-2

3.

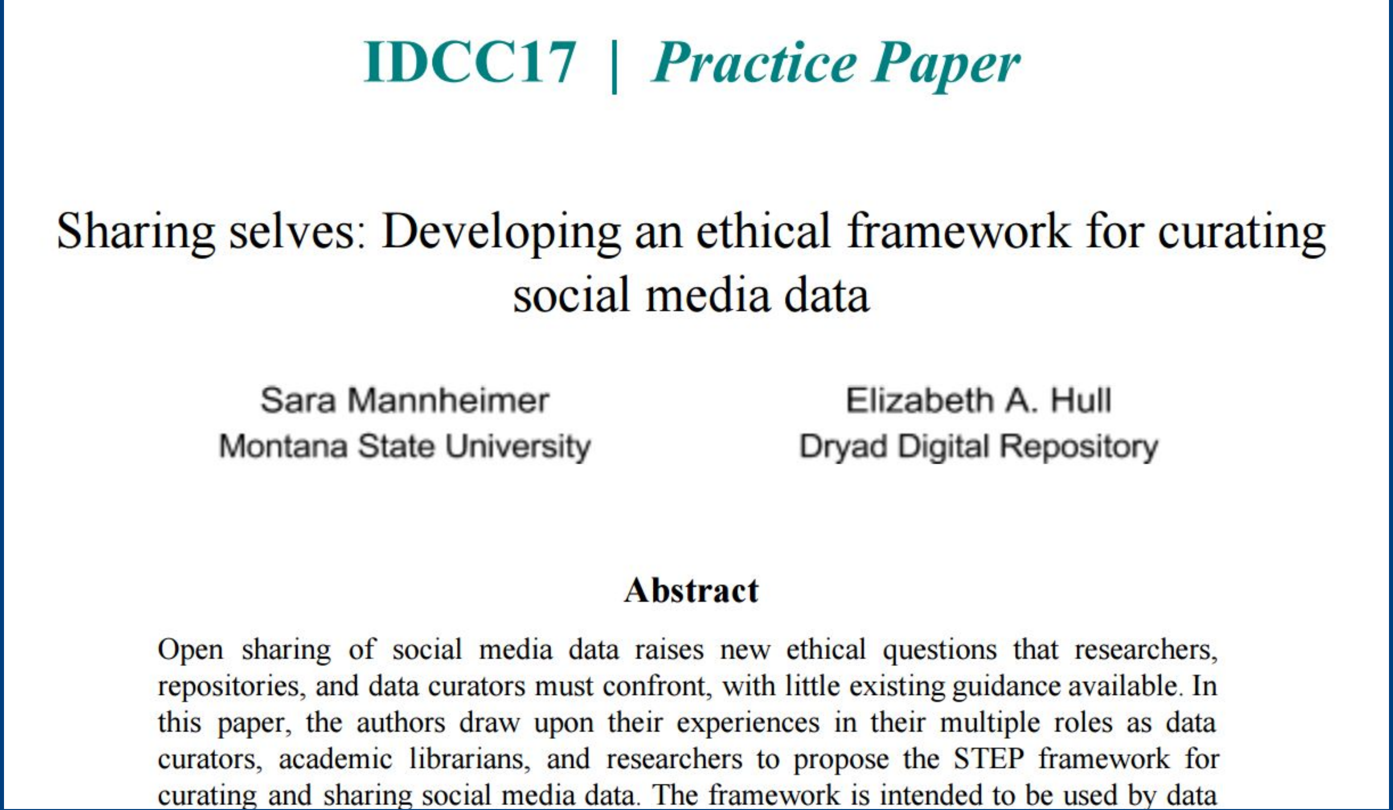
TOWARD ETHICAL BIG DATA SHARING

Pattern by Penelope Dullaghan



DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad has integrated data submission for a growing list of journals; submission of data from other publications is also welcome.

http://datadryad.org



IDCC17 | *Practice Paper*

Sharing selves: Developing an ethical framework for curating social media data

Sara Mannheimer
Montana State University

Elizabeth A. Hull
Dryad Digital Repository

Abstract

Open sharing of social media data raises new ethical questions that researchers, repositories, and data curators must confront, with little existing guidance available. In this paper, the authors draw upon their experiences in their multiple roles as data curators, academic librarians, and researchers to propose the STEP framework for curating and sharing social media data. The framework is intended to be used by data

Mannheimer S, Hull, EA (2017) Sharing selves: developing an ethical framework for curating social media data. International Digital Curation Conference, Edinburgh, February 20-23. http://scholarworks.montana.edu/xmlui/handle/1/12661

Guiding Principles

for social media data sharing

Value analysis

Measure the benefits of sharing data against the potential risks to human subjects

Guiding Principles

for social media data sharing

Value analysis

Measure the benefits of sharing data against the potential risks to human subjects

Responsibility

Data curators can help educate researchers about ethical data sharing, but researchers themselves are ultimately responsible for the data they share

Guiding Principles

for social media data sharing

Value analysis

Measure the benefits of sharing data against the potential risks to human subjects

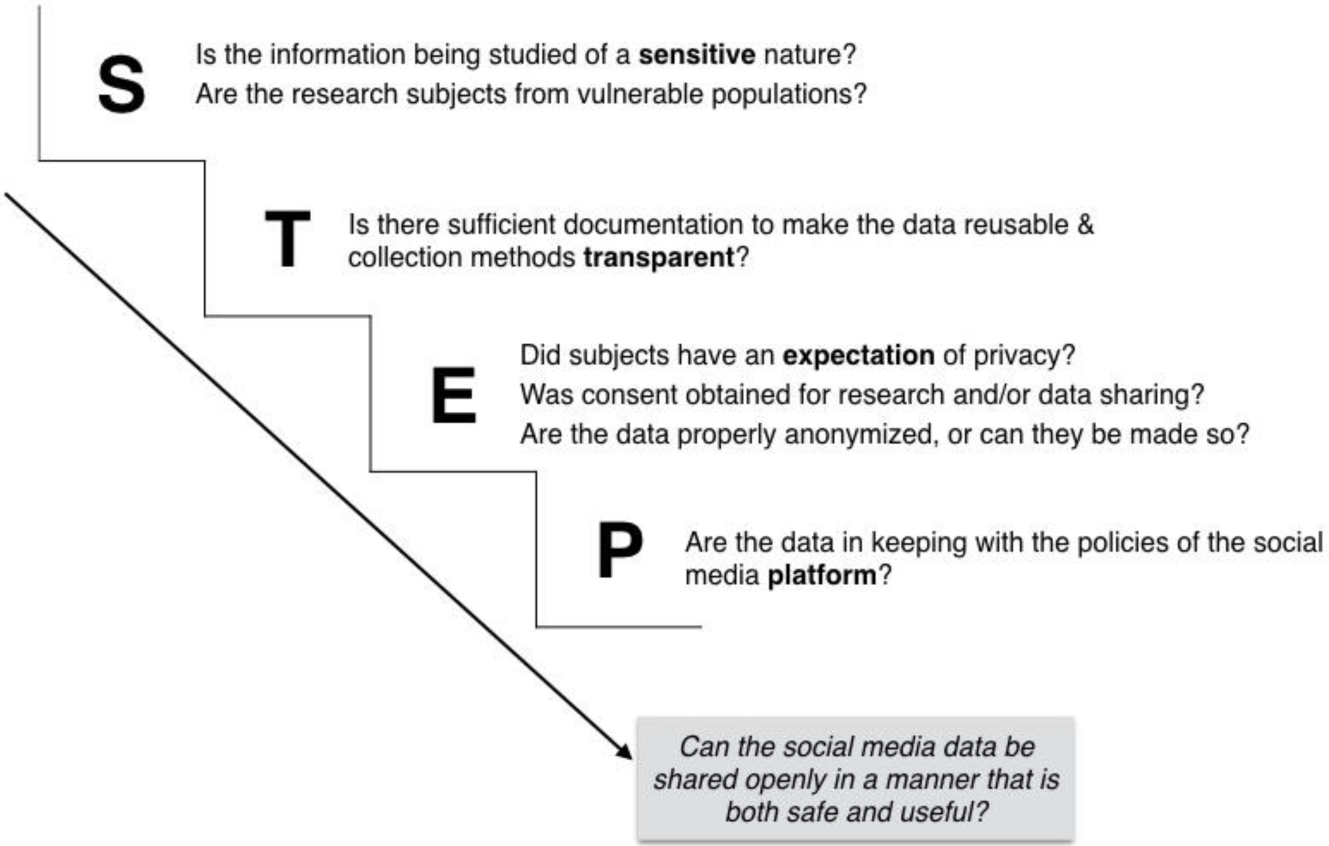
Responsibility

Data curators can help educate researchers about ethical data sharing, but researchers themselves are ultimately responsible for the data they share

Continual inquiry

Ethical practice requires ongoing dialogue and examination

STEP Framework



S Is the information being studied of a **sensitive** nature?
Are the research subjects from vulnerable populations?

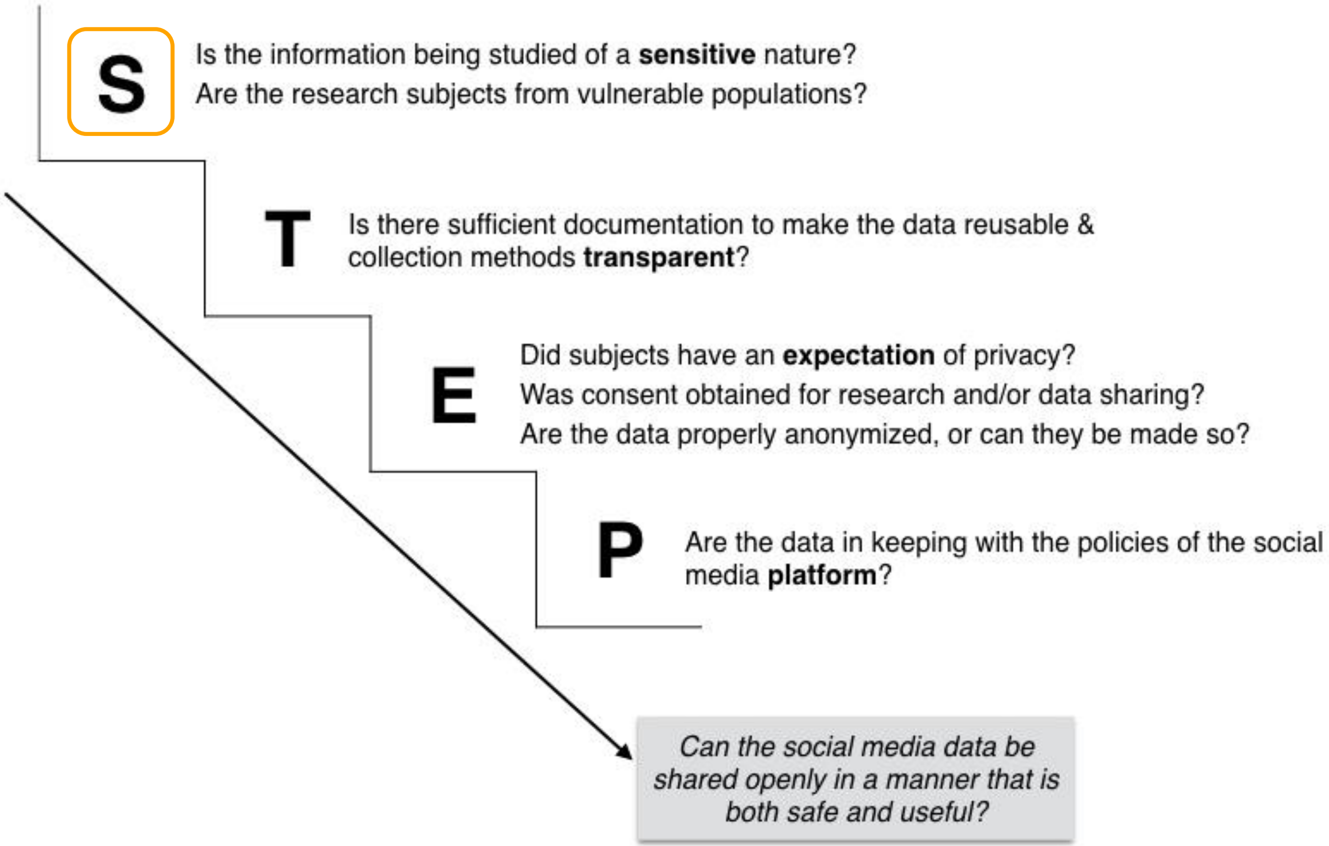
T Is there sufficient documentation to make the data reusable & collection methods **transparent**?

E Did subjects have an **expectation** of privacy?
Was consent obtained for research and/or data sharing?
Are the data properly anonymized, or can they be made so?

P Are the data in keeping with the policies of the social media **platform**?

Can the social media data be shared openly in a manner that is both safe and useful?

STEP Framework



S Is the information being studied of a **sensitive** nature?
Are the research subjects from vulnerable populations?

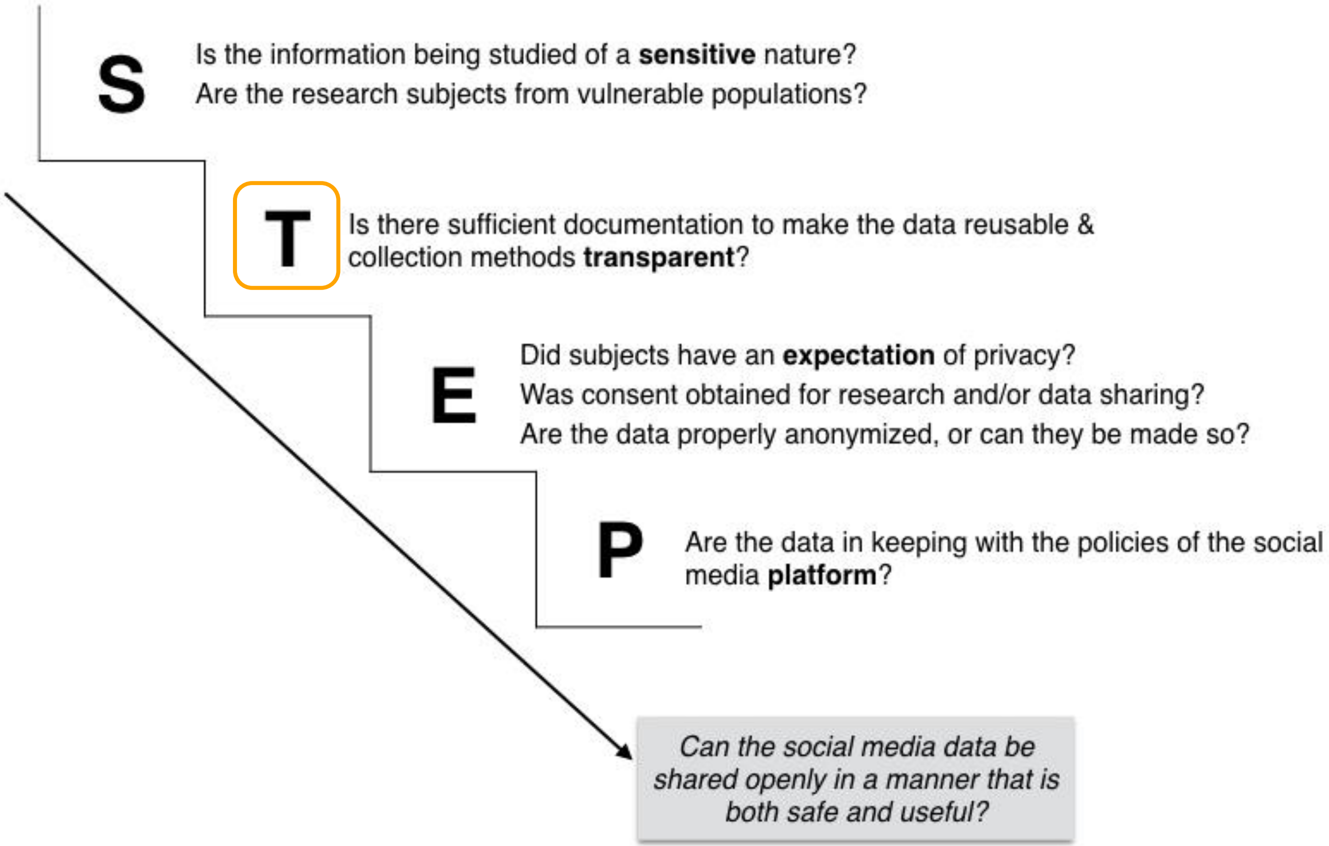
T Is there sufficient documentation to make the data reusable & collection methods **transparent**?

E Did subjects have an **expectation** of privacy?
Was consent obtained for research and/or data sharing?
Are the data properly anonymized, or can they be made so?

P Are the data in keeping with the policies of the social media **platform**?

Can the social media data be shared openly in a manner that is both safe and useful?

STEP Framework



S

Is the information being studied of a **sensitive** nature?
Are the research subjects from vulnerable populations?

T

Is there sufficient documentation to make the data reusable & collection methods **transparent**?

E

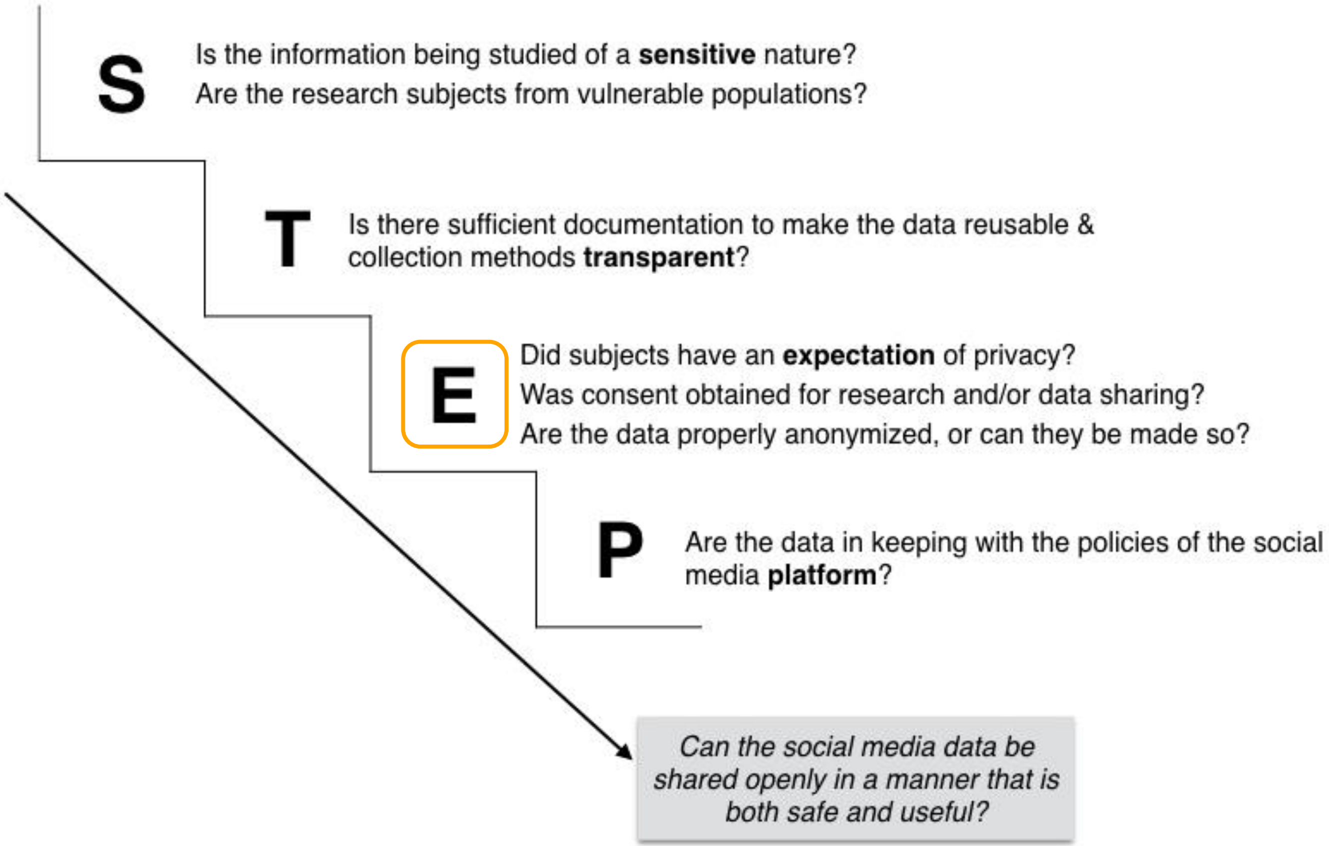
Did subjects have an **expectation** of privacy?
Was consent obtained for research and/or data sharing?
Are the data properly anonymized, or can they be made so?

P

Are the data in keeping with the policies of the social media **platform**?

Can the social media data be shared openly in a manner that is both safe and useful?

STEP Framework



S Is the information being studied of a **sensitive** nature?
Are the research subjects from vulnerable populations?

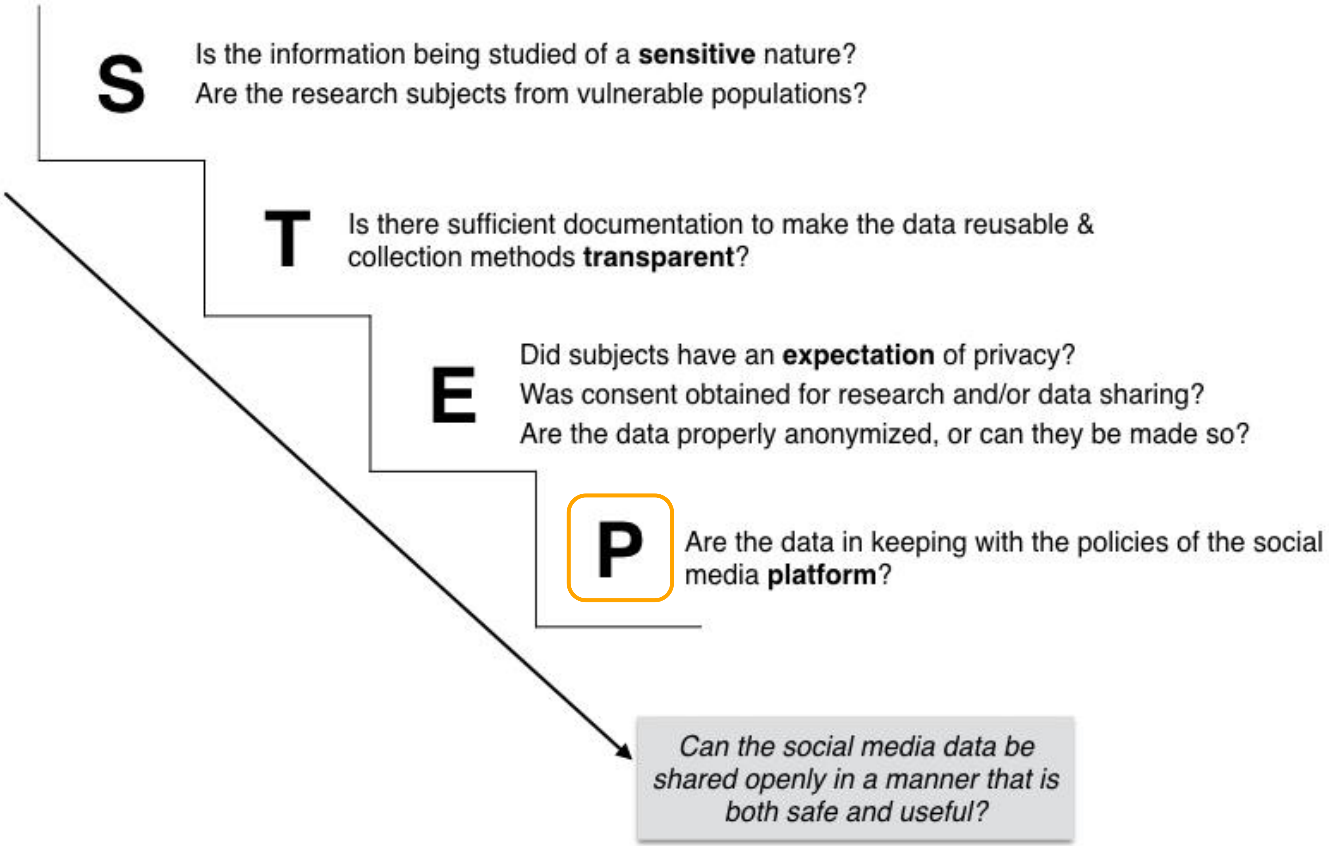
T Is there sufficient documentation to make the data reusable & collection methods **transparent**?

E Did subjects have an **expectation** of privacy?
Was consent obtained for research and/or data sharing?
Are the data properly anonymized, or can they be made so?

P Are the data in keeping with the policies of the social media **platform**?

Can the social media data be shared openly in a manner that is both safe and useful?

STEP Framework



S Is the information being studied of a **sensitive** nature?
Are the research subjects from vulnerable populations?

T Is there sufficient documentation to make the data reusable & collection methods **transparent**?

E Did subjects have an **expectation** of privacy?
Was consent obtained for research and/or data sharing?
Are the data properly anonymized, or can they be made so?

P Are the data in keeping with the policies of the social media **platform**?

Can the social media data be shared openly in a manner that is both safe and useful?

Dryad Case Study 1

the #occupy case

Network analysis of Twitter users and hashtags was used to study the evolution of political discussion during and after the Occupy Wall Street movement.

The Dryad data package includes one CSV file containing three variables: user, hashtag, and time.

Gargiulo F, Bindi J, Apolloni A (2015) The topology of a discussion: the #occupy case. PLOS ONE 10(9): e0137191. https://doi.org/10.1371/journal.pone.0137191

Gargiulo F, Bindi J, Apolloni A (2015) Data from: The topology of a discussion: the #occupy case. Dryad Digital Repository. https://doi.org/10.5061/dryad.q1h04

S

T

E

P

Sensitivity of the data

The research deals with participation in a social movement

The research does not focus on a particular population

S
T
E
P

Transparency of the data

No documentation provided

Some information about data collection in associated article.

Data analysis method detailed in the article, facilitating reproducibility

S
T
E
P

Expectations of users

Using hashtags on Twitter generally indicates desire to participate in a larger conversation and/or be identified with a concept or cause

S
T
E
P

Platform policies

CSV file contains hashtags

Dryad Case Study 1

Conclusion

Low sensitivity of research

Some concern about transparency

Some concern about platform policies

Data properly anonymized

Benefits of publication outweigh risks

Dryad Case Study 2

"in the mood"

Twitter networks were studied to determine relationship between users' sentiment and the network structure created by @-mentions.

The Dryad data package contains several networks. Variables include tweet ID, anonymised user IDs, and timestamps of tweets.

Charlton N, Singleton C, Greetham DV (2016) In the mood: the dynamics of collective sentiments on Twitter. Royal Society Open Science 3(6): 160162. https://doi.org/10.1098/rsos.160162

Charlton N, Singleton C, Greetham DV (2016) Data from: In the mood: the dynamics of collective sentiments on Twitter. Dryad Digital Repository. https://doi.org/10.5061/dryad.5302r

S

T

E

P

Sensitivity of the data

Topics discussed are wide-ranging—from “dogs” to “Islam versus atheism” to “Gamergate”

STEP

Transparency of the data

ReadMe accompanying the data package explains the content of each file

Article details how data were obtained.

S

T

E

P

Expectations of users

@-mentions indicate communications intended for specific people, imply expectation of privacy within the user's specific network

S
T
E
P

Platform policies

Tweet IDs okay

Twitter policies unclear on whether timestamps may be distributed to third parties

Dryad Case Study 2

Conclusion

Some sensitive topics

@-mentions emphasize user expectation of privacy

Research presented in a transparent and reproducible way

Benefits of publication outweigh risks

Building on the STEP framework

More case studies and testing with a variety of repositories and platforms

Framework should evolve over time

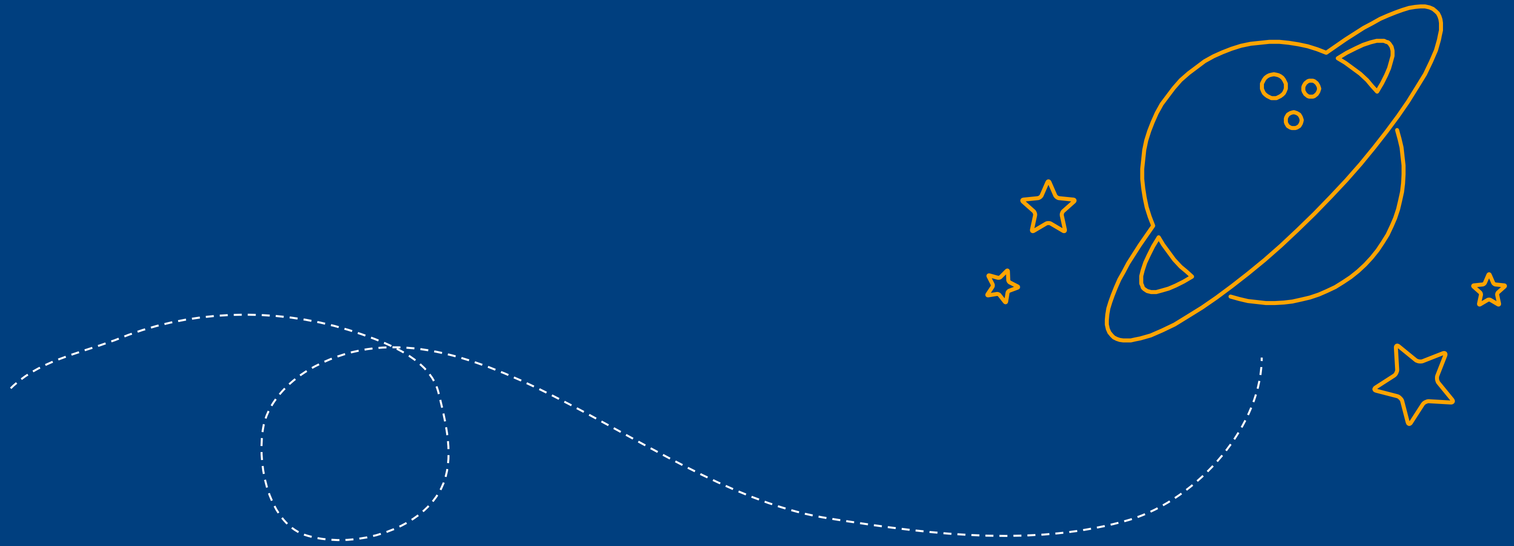
Adapt for other big data research, social science data journalism

4.

KEY TAKEAWAYS

Pattern by Penelope Dullaghan

Consider and discuss ethical gray areas



Consider and discuss ethical gray areas

Use existing frameworks

Consider and discuss ethical gray areas

Use existing frameworks

Create new frameworks

Thank you !

Sara Mannheimer
@saramannheimer
saramannheimer.com

Elizabeth Hull
@datadryad
datadryad.org

Pattern by Penelope Dullaghan