

Social Media Research Ethics: Project Design

Emily Van Haren, Andrea Zeffiro
Isaac Pratt & Jay Brodeur

Do More with Digital Scholarship Workshop Series
February 22, 2021



McMaster University sits on the traditional Territories of the Mississauga and Haudenosaunee Nations.

McMaster University sits on the Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Social Media Research Ethics Module Series

Module 1: Preliminary Considerations

- Introduces methodological considerations related to: platforms and data types, forms of analysis, quality and availability of data, and introductory data management practices.
- Introduces ethical considerations related to: private vs. public data, informed consent, anonymity, and risk of harm.

Module 2: Project Design

- Expands on and develops methodological and ethical considerations through a project design perspective.
- Highlights a range of frameworks, tools, resources, and research terms for project framing and further self-study.

Module 3: Power and Provocations (Roundtable)

- Stay tuned!



Module Goals

By the end of this module, you will be able to:

1. Define basic terms and concepts related to social media data and research.

- Build upon introductory information in “Module 1: Preliminary Considerations”
- Highlight a range of frameworks, tools, resources, and research terms for further self-study or project framing.
- Focus on project design with guiding question: how can research data management and ethical considerations inform the design of a social media research project?



Module Goals

By the end of this module, you will be able to:

- 2. Recognize how considerations for research data management and research ethics inform a contextual, iterative, and deliberative approach to project design.**
 - Contextual → attuned to social media and research context
 - Iterative → ethical questions are returned to again and again
 - Deliberative → design and decision-making happen through consultation

Module Goals

By the end of this module, you will be able to:

- 3. Apply research data management best practices to keep your data organized, secure, and ready to reuse.**
 - Recognize information professionals (e.g. librarians, specialists) as valuable resources who can support the application of these best practices.
- 3. Use the framework presented here to identify, navigate, and integrate ethical considerations and RDM best practices into a developing social media project design.**
 - Use the “[Project Design Questionnaire](#)” to initiate this learning goal.



Designing, Assessing, Consulting, (Re)formulating

In the early formulation of a research project, ethical considerations shape the project's research questions, methods, and design.

Designing

Incorporate ethical considerations into project formation

Privacy by design (Cavoukian 2011) describes an approach in which user privacy is integrated into a project's objectives, design, and dissemination activities by default.

An example of PbD in a project that uses social media to recruit participants for a study (Bender et al. 2017):

“Seeking cancer patients for a study of nutrition and cooking”

“Does #nutrition matter to you? Tell us what you think about #cooking and #cancer”



Assessing

Determine the project's stakeholders and relationships

Early project formulation can also establish the project's **stakeholder relationships** (Suomela et al. 2017):

- **Internal:** the people actively engaged in the project (collecting, analyzing, writing, archiving)
 - e.g. primary authors, research assistants, librarians, etc.
- **Subject:** the people and relationships of study
 - e.g. participants, organizations, anonymous users, etc.
- **External:** the wider research community
 - e.g. funding bodies, publication venues, colleagues in your field(s)

Also consider
platforms as
stakeholders (e.g.
terms of service)



Assessing

Determine the project's stakeholders and relationships

Why identify the stakeholders of a project?

- to support a project design that is contextualized across a project's lifespan and can monitor evolving ethical considerations
 - e.g. mid-project research team changes; updates to platform terms and conditions

How to attend to stakeholders in project design?

- will depend on research context, but may include:
 - sharing datasets
 - citing and acknowledging stakeholders
 - linking project to pedagogy/training activities
 - outlining relationships in methodology section of a research paper



Consulting

Deliberate with stakeholders during project formulation

Collaborative construction (Bailey, citing Sample, 2015) describes a process of working with communities to determine research goals.

- This process can shift research questions so they better account for a community's identity, needs, and the possible risks of research
- Determining whether or not to consult directly with communities requires contextual assessment:
 - Is collaborative construction feasible (e.g. are you working with anonymous datasets)?
 - What are the community's previous experiences with research activities and institutions?
 - Could direct communication between researcher and individuals be harmful--for either party?

(Re)formulating

Revisit the research questions, aims, and scope

Applying a **privacy by design** approach, attending to **stakeholder relationships**, or engaging in **collaborative construction** may require adjustments to the project's scope, questions, aims, or other aspects of the design.

In this way, a **contextualized** early project design is also an **iterative** process that extends throughout the project and is supported by **deliberation** with stakeholders.

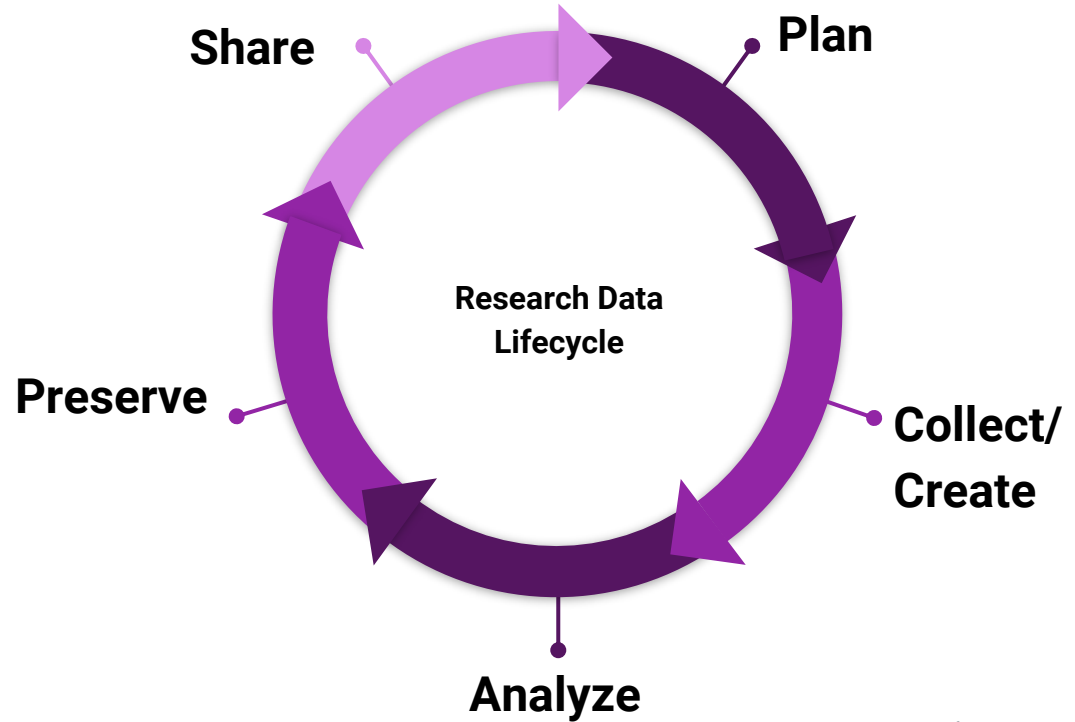
Collecting, Managing, Anonymizing, Sharing

Projects engage with the ethical tensions that arise from data collection, management, stewardship, and sharing.

Managing

What is Research Data Management anyways?

Research Data Management is the active organization & maintenance of data throughout the research data lifecycle to ensure its **security**, **accessibility**, **usability**, and **integrity**.



Planning

What is a data management plan?

A **Data Management Plan (DMP)** is a formal statement describing how research data will be managed and documented throughout a research project and the terms regarding the subsequent deposit of the data with a data repository for long-term management and preservation.

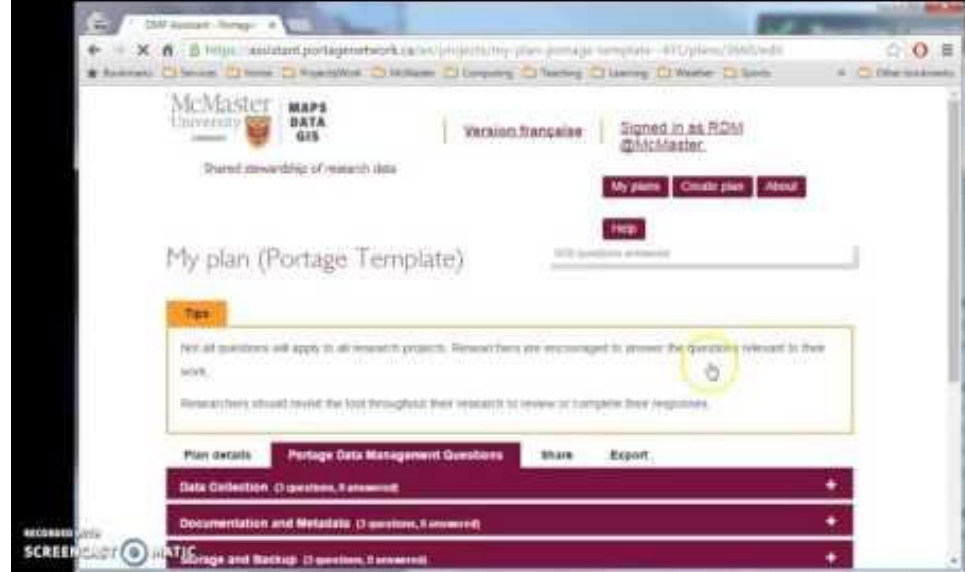
Building a DMP is a structured process that helps you plan and organize your research data.

Completing a DMP before embarking on a social media research project is a good first step towards identifying and addressing potential data management challenges.

Portage software tool for building DMPs: <http://assistant.portagenetwork.ca/>

- a web-based, bilingual data management planning tool
- available to all researchers in Canada
- a guide for best practices in data stewardship
- exportable data management plans

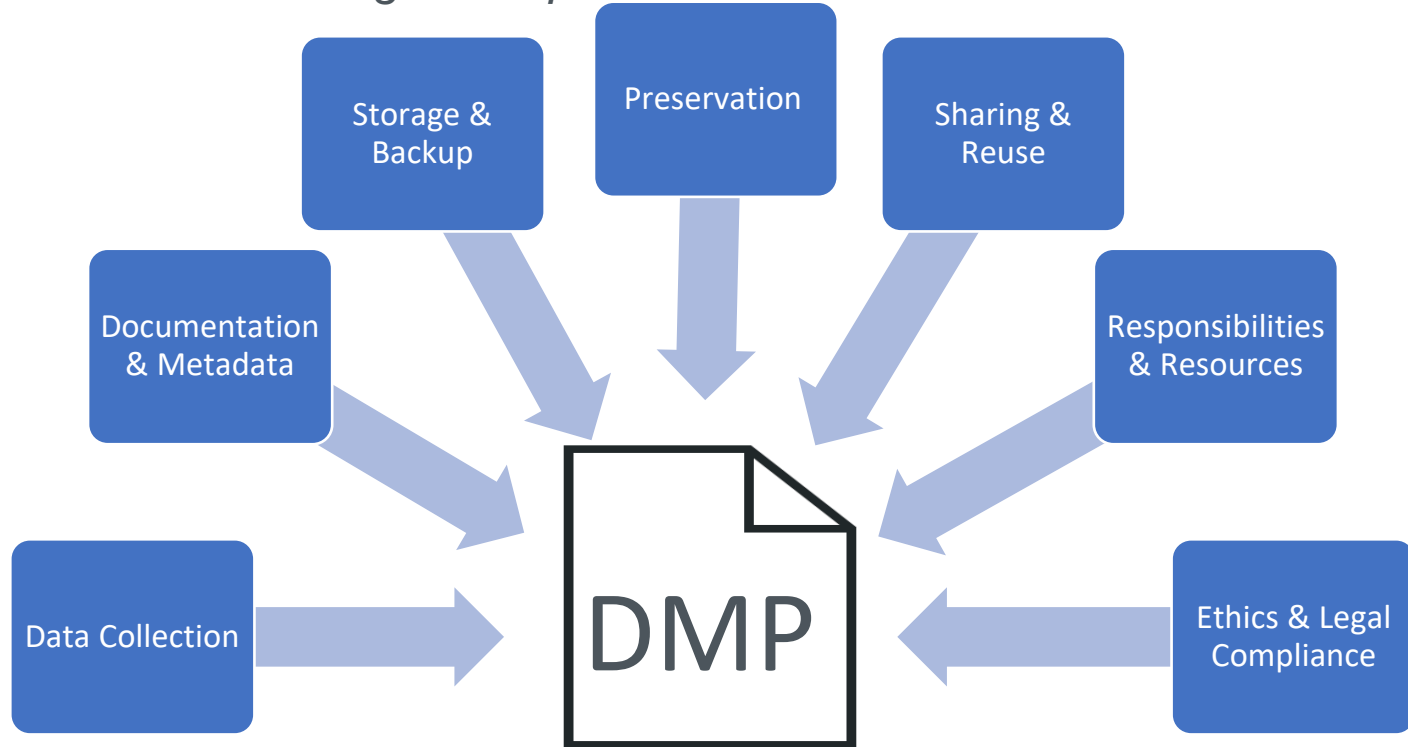
<https://www.youtube.com/watch?v=zgLaJpJfehQ>



<https://assistant.portagenetwork.ca/>

Planning

What goes in a data management plan?



Collecting

How will I collect social media data?

Methods of collecting data from social media platforms:

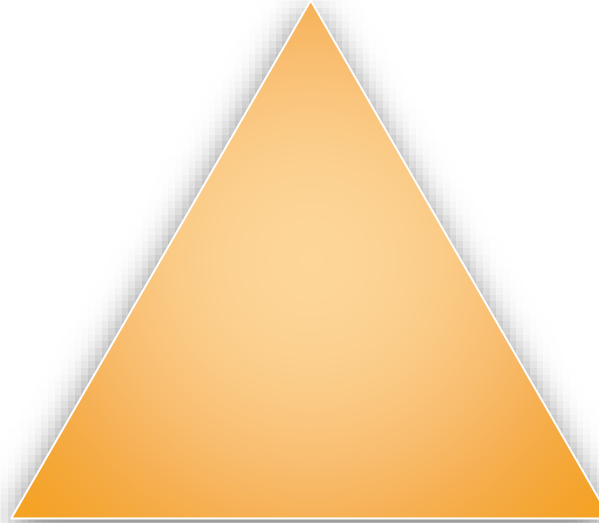
- “In-situ” manual collection and analysis of content through the platform
- Web scraping content using manual, semi-automated, or fully automated approaches
 - May or may not include third-party tools
- Direct interaction with its application programming interface (API)
 - Scripted approaches
 - Third-party apps (often with built-in analyses and visualization)



Platform Terms of Service & Access Agreements

Collecting Social
Media Data

Key considerations



Research
Design

Ethical
Implications

Storing

How should I store my data?

A good data storage plan needs to balance **accessibility** and **convenience** against **security** and **reliability**.

3-2-1 Backup Strategy:

3 copies of your data (including a backup of the raw data), each in a different location

2 copies easily accessible (production copies)

1 copy is in a trusted off site location

Storing

How should I store my data?

Features to look for when deciding on a storage platform:

- Version control
- File recovery
- Security features (2FA, encryption)
- Collaboration features
- Storage provided
- Cost
- Storage location




Storing

Where should I store my data?

Research Data Storage Finder Tool <http://u.mcmaster.ca/storagefinder>

Step 1: Answer these questions to narrow down storage provider options.

CLEAR ANSWERS

1. What risk level is your data?


☐ Low
☐ Medium
☐ High

2. What type of data storage are

Step 2: Select data storage providers you would like to compare

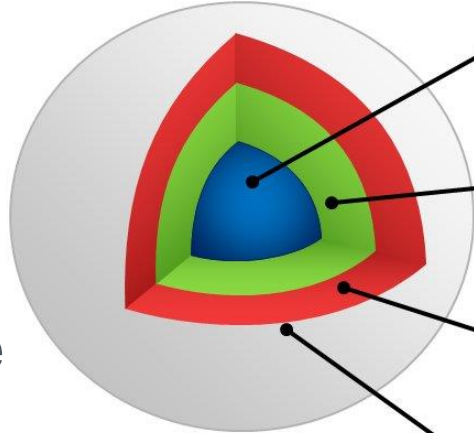
SELECT ALL **CLEAR SELECTIONS**

<p>Compute Canada</p> <p>Advanced research computing systems, storage and software</p>	<p>Compute Canada NextCloud</p> <p>Advanced research computing File hosting services</p>	<p>Dataverse</p> <p>Store, share, publish and discover research data</p>
<p>FRDR</p> <p>Find and Share Canadian Research Data</p>	<p>Github</p> <p>Distributed version control system for software code</p>	<p>MacDrive</p> <p>File Synchronization and Sharing solution</p>
<p>MacDrop</p>	<p>McMaster-based</p>	<p>QSF</p>

FAIR Data

Datasets as digital objects

Findable
Accessible
Interoperable
Reusable



Research output (data/code)

The data is surrounded by layers of information to make it FAIR

Identifiers

Persistent Unique Identifiers such as DOIs and ORCiDs help find, track, and cite data

Standards

Open standard file formats help others access and reuse data

Metadata

Rich metadata and data documentation helps others find and understand datasets

Securing

Special precautions for sensitive data

Sensitive data is any data that, if released to the public, would cause potential harm. This includes personal information, personal health information, commercial data, as well as some ecological data.

Data can be classified as:

- High Risk: contains **highly sensitive** information that would likely cause **significant harm**
- Medium Risk: contains sensitive information that would put individuals at risk of harm
- Low Risk: does not contain sensitive information

Securing

Special precautions for sensitive data

Data classified as High Risk or Medium Risk requires special precautions.

- Sensitive data must be encrypted when stored on an internet connected device.
- Sensitive data cannot be shared without being de-identified, and if it cannot be de-identified then it should not be shared openly (metadata can still be published).

The Portage Network has developed a **Sensitive Data Toolkit** for Researchers with detailed information.

<https://portagenetwork.ca/network-of-experts/sensitive-data-expert-group/>

Sharing

Movement towards openness

Sharing data openly is a critical element in pushing academia towards openness and transparency. Open and free data sharing supports research ideals of **verification**, **reproducibility**, **collaboration**, and maximizes the impact and visibility of research.



Sharing

More than just depositing data

A culture where data is open for verification is a key factor in uncovering scientific misconduct and fraud.

"When contact changes minds: An experiment on transmission of support for gay equality" was a study in 2014 by researchers at UCLA. The researchers had published their data online in OpenICPSR. A separate group of researchers (Broockman et al 2015) looking to extend the study examined the data and discovered that it had been simulated and was not real.

Sharing

More than just depositing data

Forms of sharing include traditional publications, conference presentations/seminars, media articles, blog posts, online/digital exhibits, data dashboards, social media posts.

Shared data can take the form of images, text, video, audio.

Sharing isn't necessarily just public but can be with collaborators, reviewers, and conference audiences/other researchers. Researchers need to still be mindful of preserving anonymity to these groups.

Sharing

Potential risks

When sharing social media data, researchers and data curators must measure the benefits of sharing data against the potential risks to human subjects (Mannheimer & Hull, 2017).

There is **always** a risk of re-identification which can lead to real harm to research participants.

Social media users can leave a 'trail of breadcrumbs' leading to their identity even when they appear to be anonymous behind a username or pseudonym

Sharing

Potential harms

Doxing is the public identification of an individual against their wishes, often revealing personal information about that individual such as where they live, work, and study.

This can lead to:

- Embarrassment
- Harassment (online and in-person)
- Loss of employment
- Expulsion from school
- **Swatting**, where an attacker will call the police to the individual's address on false premises, often alleging a kidnapping has taken place.

Sharing

Other considerations for sharing data

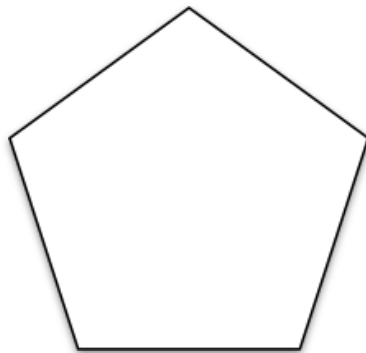
Other considerations for data sharing include:

- The terms and conditions of the social media platform,
- disciplinary norms/expectations
- requirements from funding bodies

Sharing

The subjects (vulnerability,
expectation of privacy)

The data
(privacy, sensitivity,
specificity/granularity)



Institutional, disciplinary,
funding body norms &
guidelines

The SM platform's terms of
use and conditions

The format of dissemination
(text vs. image vs, video)

Anonymizing

Identifiers

Direct Identifiers are data points that can immediately identify a person, and include things like name, addresses, phone number, username, email, detailed geographical information, IP address, etc etc.

Quasi-identifiers are pieces of data that are not themselves unique identifiers but can be combined to uniquely identify an individual. Quasi-identifiers include things like age, gender identity, income, occupation, ethnicity, geographic information, etc

Social media posts can be either direct or quasi-identifiers.

Anonymizing

Using a mathematical approach

K-anonymity is a mathematical approach to anonymizing a dataset.

A dataset has k-anonymity when a particular individual in the dataset cannot be distinguished from k other individuals in the dataset.

K is a number set by the researcher - most commonly set to 5. This means it is not possible to isolate a group of fewer than 5 identical individuals.

For a more comprehensive overview see the Portage Network's Reducing Risk Webinar:

<https://www.youtube.com/watch?v=X3MKP-FrWE>

<https://portagenetwork.ca/wp-content/uploads/2020/07/ReducingRisk-PortageWebinar.pdf>

Anonymizing

Using a mathematical approach

Two main methods for achieving k-anonymity:

- **Suppression:** where individual cases or responses are deleted.
 - For example, if there is only one individual in a particular age range of a specific ethnicity, the ethnicity response for that individual could be deleted to preserve the ethnicity category as a whole.
- **Generalization:** grouping specific values into categorized ranges.
 - For example, grouping specific ages into age ranges or merging categories into larger groups.



Anonymizing

Using a mathematical approach

Other specific approaches to reduce/minimize disclosure risk:

- **Removal** – eliminating the variable(s) from the data set
- **Bracketing** – combining the categories of a variable
- **Top-coding** – restricting the upper range of a variable
- **Collapsing** and/or combining variables – merging concepts in two or more variables into a new summary variable
- **Sampling** – releasing a random sample of sufficient size to yield reasonable inferences
- **Swapping** – matching unique cases on the indirect identifier, then exchanging the values of key variables between the cases.
- **Disturbing** – adding random variation or stochastic error to the variable.

Anonymizing

Software for anonymization

We recommend two software packages for anonymization:

Amnesia <https://amnesia.openaire.eu/>

- Amnesia is a user friendly software package that can be run online from a browser.

sdcmicro <https://cran.r-project.org/web/packages/sdcMicro/index.html>

- sdcMicro is an open source R package with a graphical interface.

Collecting, Managing, Storing, Sharing, Anonymizing

Best practices in any scenario bend towards those protecting human participants.

Federal Guidance Documents:

- Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans ([TCPS2](#))
- [Tri-Agency Statement of Principles on Digital Data Management](#)
- [Draft Tri-Agency Research Data Management Policy](#)

If questions persist, consider reaching out to the McMaster research ethics board or the research data management team at rdmgt@mcmaster.ca

Disseminating, anonymizing, visualizing, communicating

Projects consider the ethical tensions of research dissemination--its venues, forms, and consequences--throughout the project design.

Disseminating

The ways and means of sharing research

- Dissemination forms and venues are shaped by all aspects of the project design--including methods of data collection, analysis, and management.
- Disseminating research can extend or amplify ethical tensions, because user data can circulate in new and unpredictable ways.
- Communicating your research also raises additional considerations around the framing, interpretation, and presentation of findings.



Anonymizing

Designing for anonymity in qualitative research

Qualitative projects may face the tension between maintaining user anonymity and providing detail-rich examples.

Annette Markham's concept of “**fabrication as ethical practice**” (Markham 2012) shows how researchers develop strategies for ensuring that participants remain anonymous when informed consent is not possible:

- crafting composite accounts of trends in data
- restaging social media interactions so posts can be generalized, anonymized, and stripped of metadata.
- filtering, cropping, blurring, and copying images (Warfield et al. 2019)



Anonymizing

Designing for anonymity in qualitative research

Qualitative anonymization strategies can be combined with approaches that integrate privacy or anonymity into earlier project design.

Researchers can adopt an “**anonymity by design**” approach, used by the Internet of Things (IoT) industry to advocate for design practices that do not collect identifiable data in the first place.

- Identifiable data can't be shared because the researcher does not have it (Higginbotham 2020)

Combining Markham's work with new IoT design practices emphasizes an approach to project design that integrates anonymity from the start--not just at the time of dissemination.



Visualizing

The impact of dissemination design

Data visualizations (including charts, graphs, maps, etc.) are not straightforward data delivery mechanisms but rather subjective, interpretive constructions.

- “Charts are not facts. Graphs are not truths” (Mahmud et al. 2017)
- Data, and data visualizations, are tied to uneven social relations (D’Ignazio and Klein 2020)

Researchers can regard visualizations as another ‘genre’ of research dissemination- this prompts a consideration of the rhetorical impact of communication and design decisions.



Communicating

The possibilities and challenges of non-scholarly dissemination venues

Consider if, how, and when your project will share research updates or findings through **non-scholarly and/or public venues**, such as professional blogs, opinion pieces, or social media updates.

These forms of communicating results allow researchers to:

- consult with other researchers about ethical complexities
- communicate research findings to a wider audiences
- document, make visible, and reflect on the research process
- (in some cases) investigate the “usage norms” (Mannheimer and Hull 2017) and “expectations of privacy” (Townsend and Wallace 2016) of the platforms you are investigating
- engage directly with communities



Communicating

The possibilities and challenges of non-scholarly dissemination venues

Communicating research progress and findings in these ways also raises ethical complexities. For example:

- The *Tri-Council Policy Statement 2* states that “preliminary conversations” with potential research communities do not constitute research.
- However, our discussion of **collaborative construction** (Bailey 2015) noted that deliberating with communities is not necessarily ethical or unproblematic.

Being ‘in compliance’ with documents like the *TCPS 2* does not mean that researchers are absolved from ethical considerations of research, including research conducted in and shared through non-scholarly dissemination venues.

Reflecting, Questioning, Critiquing, Re-evaluating

Projects incorporate mechanisms, frameworks, and expertise for continually revisiting the ethical considerations of social media research, and for reflecting on the research(er)'s position within ethical complexities.

Reflecting, Questioning, Critiquing, Re-evaluating

An extended and ongoing ethics

The integrative approach we emphasize is:

- **Contextual:** no one-size-fits-all approach, but researchers can assess the particular social media research context in order to build ethical considerations into the project design.
- **Iterative:** no set-it-and-forget-it model, but rather ethical considerations extend throughout, and respond to changes in, the project design.
- **Deliberative:** no one-source-knows-all protocol, but consultation and collaboration with a variety of stakeholders--can support ethical project design.



Reflecting, Questioning, Critiquing, Re-evaluating

An extended and ongoing ethics

Even projects that return to ethical considerations repeatedly throughout all stages of research, or which engage with robust networks of consultation, can still exercise what Suomela et al. (2017) call “definitional power” over just which concerns and considerations get considered as ‘ethical.’

Ethics’ itself remains a fraught framework through which to conceive of our relationship with and accountabilities to the various actors (platforms, participants, researchers, etc.) within a social media research project.

Please join us to continue this conversation in

Module 3 Roundtable: Power and Provocations on March 26, 2021!



Continue the Conversation

Register for the Module 3 Roundtable: Power and Provocations:

<https://libcal.mcmaster.ca/calendar/scds/dmds-smre-power-and-provocations>

Let us know what you think about the “Project Design Questionnaire”:

<https://u.mcmaster.ca/sme-feedback>

Works Cited

- Bailey, M. (2015). #transform(ing)DH writing and research: An autoethnography of digital humanities and feminist ethics. *Digital Humanities Quarterly* 9(2). <http://www.digitalhumanities.org/dhq/vol/9/2/000209/000209.html>.
- Broockman, D, Kalla, J., & Aronow, P. M. (2015). "Irregularities in Lacour (2014)." *MetaArXiv*. <https://osf.io/preprints/metaarxiv/qy2se/>
- Cavoukian, A. (2011). *Privacy by design: The seven foundational principles*. Information and Privacy Commission of Ontario. <https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf>.
- D'Ignazio, C, and Klein, L. F. (2020). *Data feminism*. MIT Press. <https://doi.org/10.7551/mitpress/11805.001.0001>.
- Higginbotham, S. (2020, August 20). For the IoT, user anonymity shouldn't be an afterthought. It should be baked in from the start. *IEEE Spectrum*. <https://spectrum.ieee.org/telecom/security/for-the-iot-user-anonymity-shouldnt-be-an-afterthought-it-should-be-baked-in-from-the-start>.
- Mahmud, A., Hogan, M., Zeffiro, A., & Hemphill, L. (2017). Teaching students how (not) to lie, manipulate, and mislead with information visualization. In Matei, S. A., Jullien, N., & Goggins, S. P. (Eds.), *Big data factories: Collaborative approaches* (pp. 101-114). Springer. <https://doi.org/10.1007/978-3-319-59186-5>.
- Mannheimer, S. & Hull, E. A. Sharing selves: Developing an ethical framework for curating social media data. *International Journal of Digital Curation* 12(2). <https://doi-org.libaccess.lib.mcmaster.ca/10.2218/ijdc.v12i2.518>.



Works Cited

Markham, A. (2012). Fabrication as ethical practice. *Information, Communication & Society* 15(3), 334-353.
<https://doi.org/10.1080/1369118X.2011.641993>.

Townsend, L. & Wallace, C. (2016). Social media research: A guide to ethics. *The University of Aberdeen*.
https://www.gla.ac.uk/media/Media_487729_smxx.pdf.

Zeffiro, A. & Brodeur, J. (2020, March 5). *DMDs: Social Media Research Ethics and Data Management* [Workshop Powerpoint slides]. Sherman Centre for Digital Scholarship. <https://macsphere.mcmaster.ca/handle/11375/25327>.

Thank you!