DMDS: Social Media Research Data Ethics and Management

Andrea Zeffiro: zeffiroa@mcmaster.ca
Jay Brodeur: brodeujj@mcmaster.ca
Sherman Centre for Digital Scholarship
05-March, 2020

Outline

- Ethical considerations [20 mins]
- Case studies & discussion [30 mins]

**********Break********** [10 mins]

- Managing & sharing SM materials [20 mins]
- Evaluating frameworks & wrap-up [25 mins]

Ethical & methodological considerations

Social Media

Websites and applications that enable users to create and share content or to participate in social networking.

Sharing information, ideas, personal messages and other content such as images and videos.

Types of Platforms

- Networking, information sharing, content curation
  - (i.e. Facebook, Twitter, Instagram, Reddit)
- Online forums for specific communities
  - (i.e. PatientsLikeMe, Mumsnet, BaristaExchange)
- Private collaborative tools
  - (i.e. Trello, Slack, Teams)
- Crowdsourcing platforms
  - (i.e. GoFundMe, Kickstarter, etc.)

(Taylor and Pagliari 2017)

Social Media
… as an enabler of research

Informal and formal modes of scholarly exploration.

- Gathering opinions
- Recruiting participants
- Fostering stakeholder involvement

(Taylor and Pagliari 2017)

Social Media
… as a source of data for research

‘Secondary uses’ include studies seeking to profile or understand users’ behaviours, demographics, interactions and networks, or to assess their responses or sentiments towards particular topics, products or policies.

(Taylor and Pagliari 2017)

Benefits of social media research

- Reach larger numbers of participants
- Reduce cost
- Analyse trends and associations within large corpuses of data
- Interaction across extended time periods
- Less prone to bias than approaches involving direct contact between researchers and participants
- Involvement of citizens in research process
- Create new channels for research dissemination

(Taylor and Pagliari 2017)

Methodological Considerations

- Representativeness
- Inequalities in access
- Heterogeneous data
- Non-traditional sampling approaches
- Social media service provider

Ethical Considerations

The complexity of interactions between individuals, groups, and technical systems present a number of challenges for scholars seeking to use social media data in research.

Recommendation: Ethical considerations guide the research design and methodological considerations.

Contextual

It's impossible to adopt a 'one size fits model':

- Every social media context is unique
- Ethical considerations are grounded in the specifics of the social media community, the methodology and research questions
- Ethical decision making is a deliberative and iterative process

Possible Approaches

… Computational

… Quantitative

… Qualitative

Research Design

Research Questions: What are questions driving the research? What conceptual and/or theoretical frameworks are shaping these questions? How have other disciplines explored similar questions and to what end?

Research Data: What are my data sources? How will I acquire them? Is REB approval required? If not, will I seek out a consultation? How will data be managed and by whom? How and where will data be stored? Who will be responsible for handling sensitive data? Who will have access to the data and in what form?

(Adapted from Zeffiro 2019)

Research Design

Research Tools: What, if any, computational tools and techniques will be used for research? Why these in particular? What skills and expertise are required? Who will conduct this portion of the research and how will they be acknowledged? What are other possible approaches to doing the research?

Research Relations: What are some of (negotiated) relationships forged through research? To whom do I feel accountable towards? With whom do I share this accountability? Where am I in the research and what is my situated perspective?

(Adapted from Zeffiro 2019)

Research Design

Research Participants: Who and/or what constitute my research participants? Is REB approval required? If not, will I seek out a consultation? How will participants be made aware of their involvement in the research? If this is not practical or feasible, then how will participation be made transparent? What is my responsibility or duty to these participations (and to their data)? How will I safeguard contextual integrity? How will I uphold participant autonomy? What are possible ways in which research participants may feel let down? Are there ways to mitigate disappointment?

(Adapted from Zeffiro 2019)

Research Design

Research Beneficiaries: For whom is this research for? Who and/or what is the driving force? Why do I care about it? How will it benefit me? How will it be of benefit to others? Who will derive benefits from it?

Research Dissemination: How do I intend to share research results? In what forms and with whom? How will I uphold contextual integrity when sharing results? Will a 'no guarantee' clause accompany research (including dataset) dissemination?

(Adapted from Zeffiro 2019)

Case studies

Some considerations

Is the data private?

Can the subject matter be considered sensitive?

Are any of the subjects vulnerable? What is the risk of harm?

Is consent necessary? Is it given? How to obtain it?

How (if at all) should source information be presented in publications?

How (if at all) should the data be shared?

Should the researcher identify themselves?

Is the research exploitative?

Is the data representative?

Is there a need to account for bots, trolls, and spam?

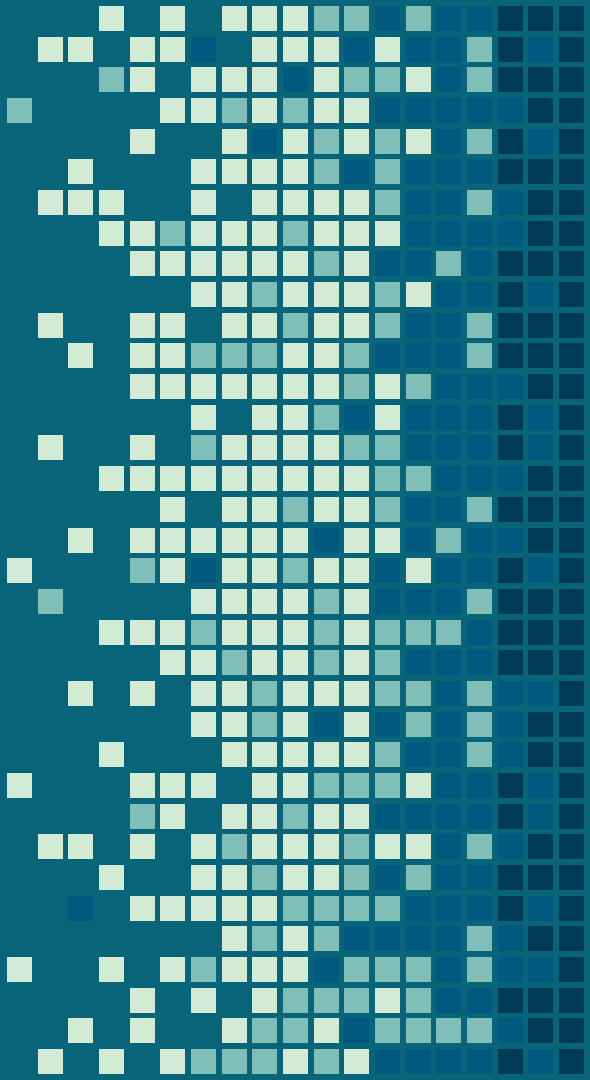
Is ethics approval necessary?

Are there other ethical and methodological considerations?

“

A researcher wishes to conduct a content analysis of tweets related to the 2020 US Presidential Election, to explore how Trump supporters argue for their candidate on Twitter.

They have paid a third-party service to provide data related to tweets using the hashtags #DonaldTrump, #TrumpTrain, #VoteTrump2020, #AlwaysTrump, #KeepAmericaGreat, and #Trump2020 that span the period leading up to and shortly after the election.

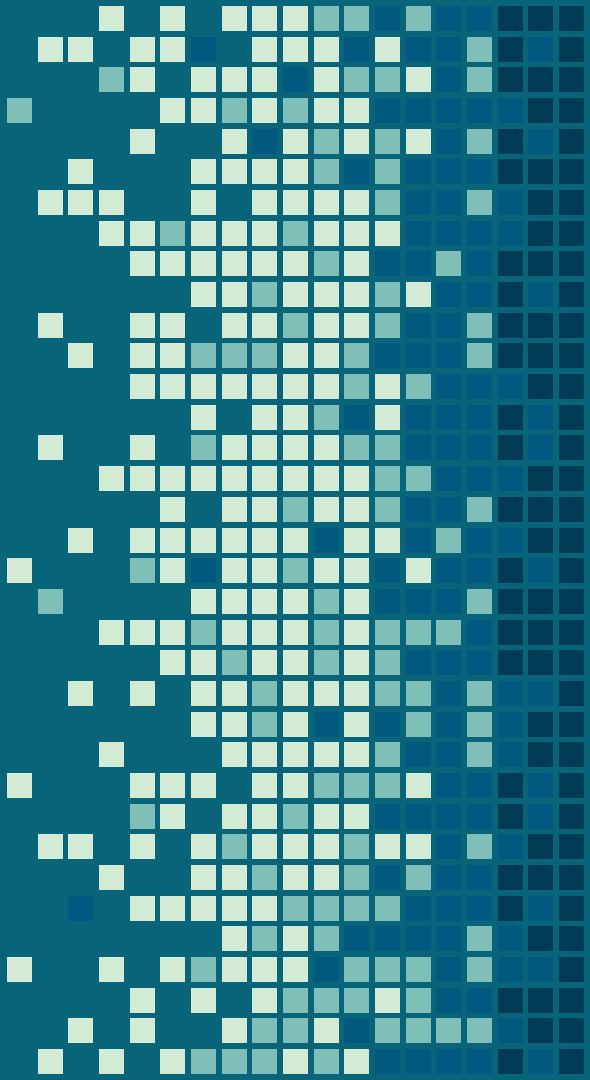


Scenario 1

“

A researcher wishes to study support mechanisms and discourse amongst members of a discussion forum which deals with mental health issues such as depression and feelings of suicide.

The forum is closed and password-protected, and registration must be approved by a gatekeeper (a site admin).



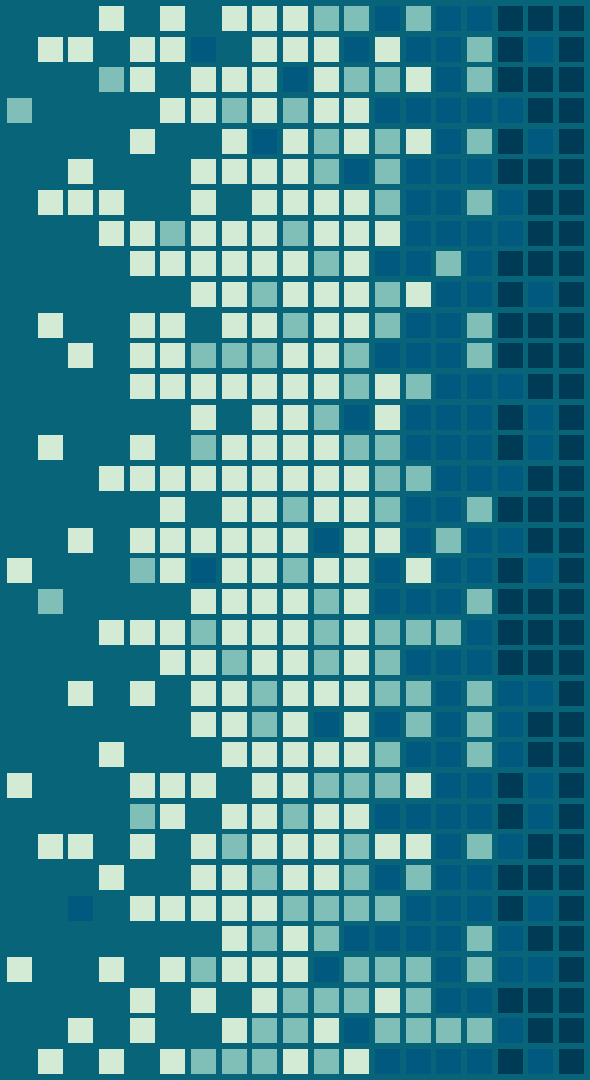
Scenario 2

“

Researchers wish to better understand how Facebook is being used by people in Puerto Rico in the aftermath of Hurricane Maria. As part of a wider study, they are planning to conduct an in-depth qualitative analysis of a number of Facebook pages that were used by local people to communicate and organize in the aftermath of the disaster. Some of the Facebook pages are private (though anyone can request to join them), and some pages are public.

There is a wide range of topics being discussed on the boards including people searching for lost family and friends…

The researchers want to join the private groups, and then observe how different types of public and private Facebook pages are being used by people as they respond to the disaster.

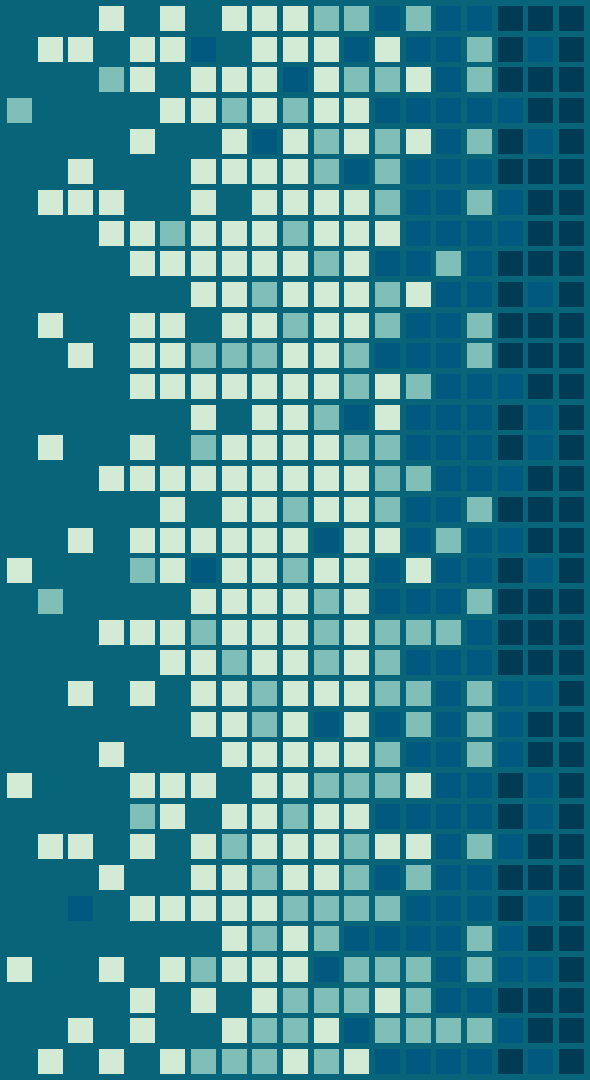


Scenario 3

“

A researcher wishes to use Tinder to study public interactions on social dating platforms. Although the posts being studied are public (rather than through private messaging), she needs to sign up to Tinder to view them.

By signing up, she has to fill in a registration form including questions such as “I am a woman looking for a man/woman” etc. It is, therefore, reasonable to think that users of the platform expect that other people viewing their profile might be doing so for similar (dating) reasons. The users of the platform are aware that there is a very large number of people using the platform and potentially able to access their profile.



Scenario 4

“

A researcher wishes to perform a discourse analysis of pipeline protests by examining interactions between environmental activists, government and state agencies, members of Indigenous populations, and corporations on Twitter through close readings of tweets for selected (less than 10) individuals and a number of prominent public groups.

They wish to share excerpts of the interactions in an upcoming publication.

Scenario 5

“

A group of researchers aim to conduct network and sentiment analysis of tweets related to the COVID-19 outbreak (#Wuhan, #COVID-19, #Coronavirus, etc.). They plan to use an online commercial tool to collect Tweets. The data provider operates its service legally and in line with the terms and conditions of Twitter. The Twitter data they gather will be fully identifiable.

They plan to create network visualizations that show how particular Tweets and hashtags became popular through retweeting practices. They also want to visualize how sentiment about the events emerged over time amongst different networks of Twitter users. They want to make an interactive online visualization in which users will be able to zoom in on particular areas of the network to view specific tweets, hashtags used, and their submitting users.

Scenario 6

“

A research group comprised of engineering and computer science researchers has used the WIDER FACE[1] dataset--a face detection benchmark consisting of approximately 400,000 images of faces--to develop a machine learning algorithm for detecting sentiment in humans.

The research group plans to publish their algorithm and results in a journal requiring supporting data to be deposited in a trusted data repository. The researchers have contacted an open data repository to inquire about depositing the WIDER FACE dataset. .

[1]http://bit.ly/WIDER-FACE

Scenario 7

Some considerations

Is the data private?

Can the subject matter be considered sensitive?

Are any of the subjects vulnerable? What is the risk of harm?

Is consent necessary? Is it given? How to obtain it?

How (if at all) should source information be presented in publications?

How (if at all) should the data be shared?

Should the researcher identify themselves?

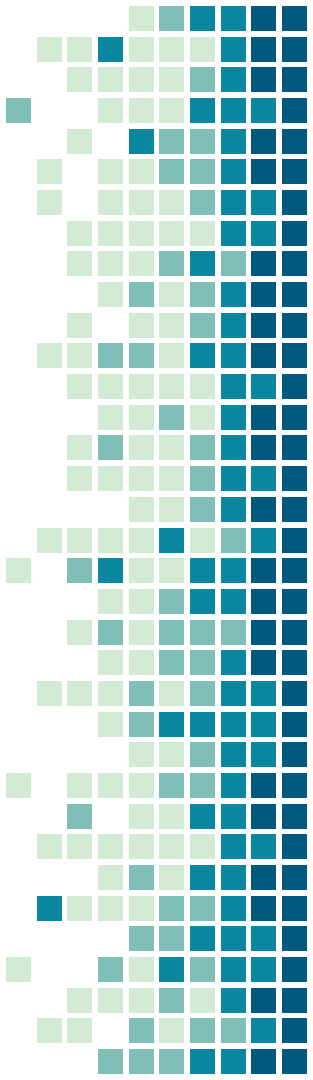
Is the research exploitative?

Is the data representative?

Is there a need to account for bots, trolls, and spam?

Is ethics approval necessary?

Are there other ethical and methodological considerations?



Common (Ethical) Challenges

I. Public vs Private

II. Informed Consent

III. Risk of Harm

IV. Anonymity

i. Public vs Private

Terms and Conditions are written in legal discourse and contain clauses on how one's data is managed and used by a platform and accessed by third parties, including researchers.

ii. Informed Consent

Terms of Service vs Informed Consent

Clicking “I agree” as informed consent?

When might access to data be mitigated by other concerns?

iii. Risk of Harm

Republishing quotes verbatim and/or using screen grabs can expose the identity and profile of the social media participant.

- Paraphrase
- Seek informed consent for research output
- Consider more traditional approaches

iv. Anonymity

- Anonymising social media data is still a complex process
- Researchers need to consider the data, metadata, and related data and contexts
- Different issues arise for different data
  - Text-based units of data
  - Interoperability of datasets

Managing & sharing social media research materials

Forms of dissemination & sharing

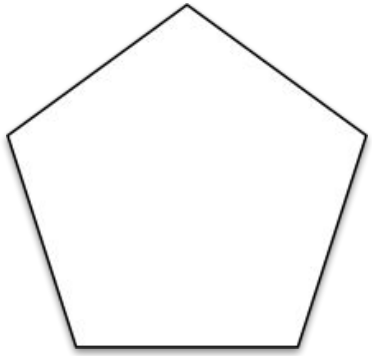
- Disseminating materials through publications, presentations, blog posts, visualizations
  - Text, images, video, audio, etc.
  - Aggregated results
  - Small units (excerpts)

- Sharing research datasets with collaborators / reviewers / research community / public

If, how, and where to share depend on:

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)

Considerations for dissemination

- Read thoroughly (and revisit!) the terms and conditions for both ***users*** and ***data users***
  - Who maintains (copy)right to the information?
  - Can direct excerpts be published?
- Seek consent where required, appropriate, and possible
- Protect participants' identity
  - Anonymize by removing/treating direct (handles, usernames, emails) & indirect (gender, location) identifiers
  - Fictionalize aspects of the research
  - Paraphrase materials

Considerations for sharing datasets

Why to (and also not to) share social media research datasets:

1. To support research transparency; i.e. reproducibility and verification
   - (Risk of harm may outweigh value of transparency)
2. To enable broad access to data
   - (Enabling broad access may violate terms of use)
3. To benefit research efficiency through reuse
   - (Reuse may not be appropriate or permitted without platform and REB clearance)
4. To satisfy publisher / funder requirements
   - (Publishers & funders prioritize privacy over sharing)

Weller and Kinder-Kurlanda p. 166

Considerations for sharing datasets

- Read the terms and conditions!
- Anonymize datasets by removing handles, usernames, and other direct identifiers
- Consider (and minimize) potential for re-identification through indirect identifiers
- Control access to datasets
  - Restrict access by accounts, groups, domain
  - Require potential reusers to request data or notify author

TCPS 2

"The easiest way to protect participants is through the collection and use of anonymous or anonymized data, although this is not always possible or desirable. For example, after information is anonymized, it is not possible to link new information to individuals within a data set, or to return results to participants.

A "next best" alternative is to use de-identified data: the data are provided to the researcher in deidentified form and the existing key code is accessible only to a custodian or trusted third party who is independent of the researcher. The last alternative is for researchers to collect data in identifiable form and take measures to de-identify the data as soon as possible"

TCPS2 (2018). p67

Considerations for sharing datasets

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.

How to Cite Twitter (MLA)

@Username. "Full text of tweet." *Twitter*, Day month year posted, time posted, URL.

@joshshepperd. “Flying under the radar today: Trump’s people just bought Twitter.” *Twitter*, 29 February 2020, 5:04 pm, twitter.com/joshshepperd/status/1233875601074925568.

(@joshshepperd)

How to Cite Facebook (MLA)

Lastname, Firstname [or username or page name]. "first several words of a facebook post..." *Facebook*, Day month year posted, time posted [if available], URL.

Penguin, Oscar. "Root beer floats are in honor of National Library Week..." *Facebook*, 18 Apr 2016, facebook.com/openguin/posts/10154065808067067.

How to Cite Instagram (MLA)

Lastname, Firstname [or single username]. (handle). "First several words of Instagram post (if any)..." *Instagram*, Day month year posted, URL.

Indigenousrising. “Much love to our @ItTakesRoots alliance members in NYC…” Instagram, 19 Feb 2020. instagram.com/p/B8wueVPjVfS/?hl=en

Resources for ethical management & sharing of social media data

- Planning
- Storing
- Sharing

Draft Tri-Agency Research Data Management Policy[2] (2018)

Government of Canada | Gouvernement du Canada

Home ➔ Collaboration between Federal Research Funding Organizations ➔ Policies and Guidelines ➔ Research Data Management

DRAFT
Tri-Agency Research Data Management Policy
For Consultation

1. Preamble

The Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC) (the agencies) are federal granting agencies that promote and support research, research training, knowledge transfer and innovation within Canada.

The agencies expect the research they fund to be conducted to the highest professional and domain standards, domestically and internationally. These standards support research excellence by ensuring that research is performed ethically and makes good use of public funds, experiments and studies are replicable, and research results are as accessible as possible.

Research data are data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content have the potential of becoming research data. Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data or repurposed data. [1] Research data enable researchers to ask new questions, pursue novel research programs, test alternative hypotheses,

[2] http://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html [shortened: bit.ly/TA-RDM-Policy]

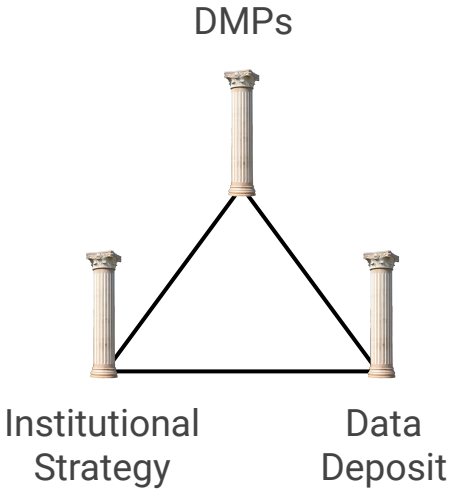
Draft Tri-Agency RDM Policy[2]

- Released in June 2018; consultation period from June-Sep, 2018
- Will apply to Tri-Agency grant recipients and institutions administering tri-agency funds.

Three Pillars:

1. Institutional Strategy
2. Data Management Plans
3. Data Deposit

Planned launch ~~in winter 2019~~ eventually.
Phased and incremental implementation



[2] http://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html [shortened: bit.ly/TA-RDM-Policy]

Data Deposit

*"Grant recipients are required to deposit into a **recognized digital repository** all digital research **data, metadata and code** that **directly support the research conclusions** in journal publications, pre-prints, and other research outputs that arise from agency-supported research…"*

- The repository will ensure safe data storage, preservation, and curation
- The agencies encourage researchers to provide access to the data where ethical, legal, and commercial requirements [e.g. TCPS 2] allow, and in accordance with the standards of their disciplines.
- Whenever possible, these data, metadata and code should be linked to the publication with a persistent digital identifier.

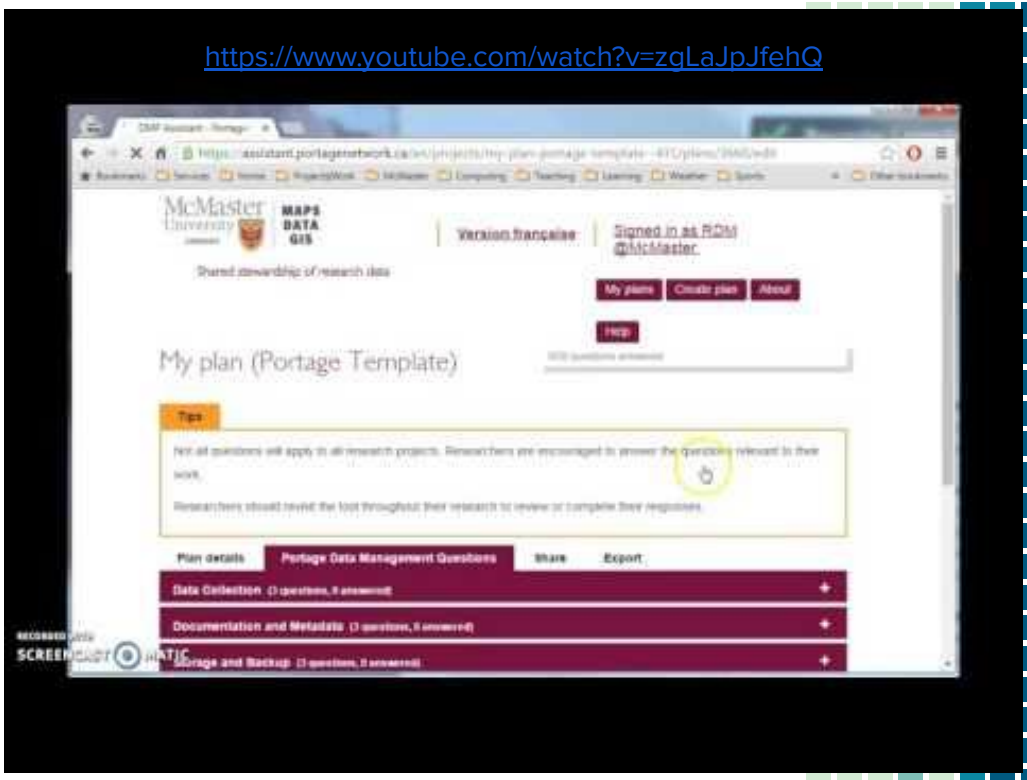
[2] http://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html [shortened: bit.ly/TA-RDM-Policy]

Portage DMP Assistant

DMP ASSISTANT
DATA MANAGEMENT PLANS
PLANS DE GESTION DES DONNÉES
PGD

- 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/

Considerations for managing data

What types of data (and how much) will you collect?

How will you organize, secure, and backup your data?

Are there ethical or commercial conditions?

- Should your data be encrypted?

How will you describe your data so that others understand it?

How will you control access to your data?

How will you manage data versions?

MREB Data Storage and Security Tools:

Research Data Management Matrix
Last Revised: 2018-06-19
Version: 0.1

The following Research Data Management Matrix is a guideline from the McMaster Research Ethics Board for collecting and storing data for research involving human participants.

	LOW RISK	MEDIUM RISK	HIGH RISK
TYPES OF DATA	Research data that does not contain any sensitive or identifiable information about individuals, organizations or communities (e.g. data which have been de-identified). NOTE: If in doubt, assume that data are sensitive. Non-sensitive research documentation (e.g. non-confidential protocols and information sheets) Publicly facing information. While public facing information is often considered low-risk, there are cases where informed consent/risk of harm should be closely considered. For example, information regarding racial or ethnic origin could be found on	Research data that may or does contain confidential, sensitive or identifiable information about individuals, organizations, or communities Some sensitive research-related documentation Personally identifiable information De-identified records of compensation Data and research protocols related to private or sensitive intellectual property	Research data that contains highly sensitive information about individuals, organizations, or communities (e.g. information about criminal activity) Personal health information Personal financial information such as banking information, income tax returns Data and research protocols related to highly sensitive intellectual property Identifiable data where disclosure, loss, or unauthorized modification of information may result in significant risk for the research participant including reputational damage, significant professional or

	RHPCS - Backup Services	RHPCS - Hosted Server Packages	MacDrive	Microsoft OneDrive / Teams
Storage Quota	1 TB; more available for fee	1 TB; more available for fee	300 GB per account	1 TB per account; up to 5 TB by request
Rates / cost	$500 / yr + one time set up fee ($125 / machine) Additional space: $300 / TB Restore services: $125 / hour	$500 - $4000 / yr Setup fee: $500 - $1000 Additional space: $450 / TB	No cost to users	No cost to users
Backups / versioning	Nightly, 14-day rotating cycle; Restore services through RHPCS	Nightly, 14-day rotating cycle; Restore services through RHPCS Nextcloud sync service available.	Ongoing real-time sync 4-month version history Full Library restore through UTS	Ongoing real-time sync Unlimited version history (?)
Who can use this service?	Any subscribing users or research group	Any subscribing users or research group	McMaster Faculty and Staff Graduate students can obtain zero-quota accounts	All McMaster faculty, staff and students
Server location	A.B. Bourns building	A.B. Bourns building	Replicated clusters in Gilmour Hall and JHE	**OneDrive**: Canadian servers **Teams**: Soon in Canadian servers only
Other notes			Supports encrypted libraries, file and directory sharing, Desktop client, web interface	Supports file and directory sharing, Desktop client, web interface
More info	rhpcs.mcmaster.ca/current-rates	rhpcs.mcmaster.ca/current-rates	macdrive.mcmaster.ca/ Documentation: https://goo.gl/AvRGwx	portal.office.com/ Documentation: mcmaster.ca/uts/licensing

https://research.mcmaster.ca/app/uploads/2018/08/MREB-Data-Storage-and-Security-Tools-1.pdf

Considerations for sharing datasets

How will your data products be stored in the long-term?

- How to ensure that it remains *integral* and *secure*?
- Who will assume long-term *responsibility* for your data?

How will others access your data products?

- What data (if any) can/should be shared? Who should have access?
- How will you manage legal, commercial & ethical constraints?

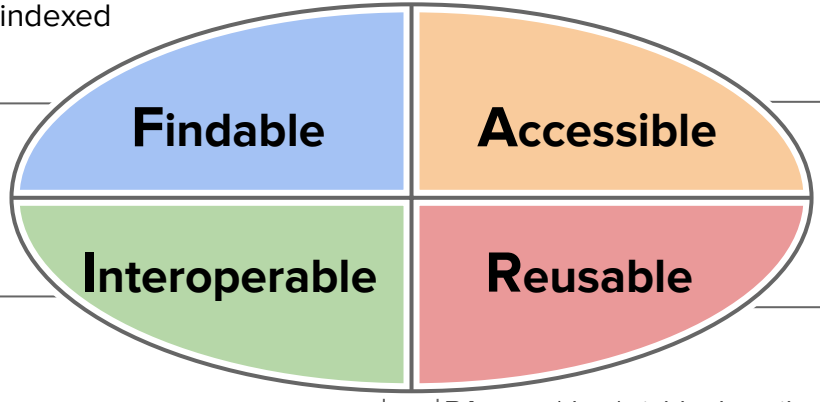
How to maximize credit for sharing your data?

- In which repository should you deposit your data?
- How to ensure that your data is FAIR (*findable*, *accessible*, *interoperable* and *reusable*)?

The FAIR Guiding Principles

F1: (meta)data have a globally unique and eternally persistent identifier
F2: data are described with rich metadata
F3: metadata clearly and explicitly includes the ID of the data it defines
F4: (meta)data are registered and indexed in a searchable resource

A1: (meta)data retrievable by their ID using a standardized protocol
A1.1: protocol is open, free and universally implementable
A1.2: protocol allows for AuthT/ AuthZ where needed
A2: metadata is always accessible



I1: (meta)data use a formal,accessible, shared, broadly applicable language for knowledge rep.
I2: (meta)data use vocabularies that follow FAIR principles
I3: (meta)data include qualified references to other (meta)data

R1: meta(data) richly described with accurate and relevant attributes
R2: (meta)data released with a clear and accessible data usage license
R3: (meta)data associated with detailed provenance
R4: (meta)data meet domain-relevant community standards

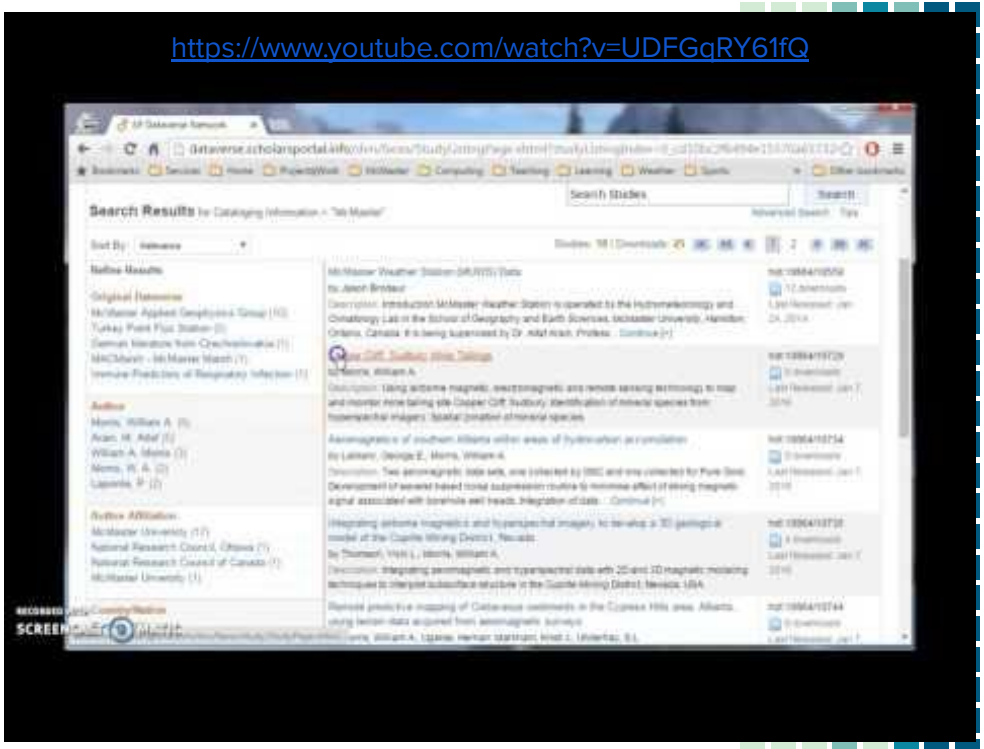
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3. *doi:10.1038/sdata.2016.18*

Scholars Portal Dataverse

Scholars Portal
A Service of the Ontario Council of University Libraries

- A data repository for researchers at Ontario's universities -- **free** and **open** for all researchers in Canada
- An online platform to share, preserve, cite, explore and analyze research data.
- Allows researchers to control how they share their data.
- Supports data DOI registration through Datacite Canada.

https://www.youtube.com/watch?v=UDFGqRY61fQ



http://dataverse.scholarsportal.info

Evaluating frameworks for ethical use of social media data

Frameworks:

Townsend & Wallace (2016): bit.ly/Townsend-2016

Williams et al. (2017): bit.ly/Williams-2017

AOIR (2016): bit.ly/AOIR-2016

Case studies: Revisited

- Revisit your case studies & re-evaluate
- Use the provided frameworks, where helpful

Follow-up discussion

- What has become clearer? What has not?
- Are the frameworks helpful?
  - Where are they lacking?
- Lingering questions?

Sources cited

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, & Social Sciences and Humanities Research Council of Canada. (2014). Tri-council policy statement: Ethical conduct for research involving humans (2018). Online: https://ethics.gc.ca/eng/documents/tcps2-2018-en-interactive-final.pdf

Conway, Mike. "Ethical issues in using Twitter for public health surveillance and research: developing a taxonomy of ethical concepts from the research literature." Journal of medical Internet research 16.12 (2014).

Moreno, Megan A., et al. "Ethics of social media research: common concerns and practical considerations." Cyberpsychology, Behavior, and Social Networking 16.9 (2013): 708-713.

Nissenbaum, Helen. "A contextual approach to privacy online." Daedalus 140.4 (2011): 32-48.

Office of the Information and Privacy Commissioner of Ontario. "Big Data Guidelines". Online: https://www.ipc.on.ca/wp-content/uploads/2017/05/bigdata-guidelines.pdf

Shilton, Katie. "Emerging Ethics Norms in Social Media Research." Online: https://bigdata.fpf.org/wp-content/uploads/2015/12/Shilton-Emerging-Ethics-Norms-in-Social-Media-Research1.pdf

Taylor, Joanna, and Claudia Pagliari. "Mining social media data: How are research sponsors and researchers addressing the ethical challenges?." Research Ethics (2017): 1747016117738559.

Townsend, Leanne, and Claire Wallace. "Social media research: A guide to ethics." University of Aberdeen (2016). Online: https://www.gla.ac.uk/media/media_487729_en.pdf

Unwin, Lindsay, and Kenny, Anita. "The Ethics of Internet-based and Social Media Research: Report of a Research Ethics Workshop held on Thursday 14 July 2016". Online: https://www.sheffield.ac.uk/polopoly_fs/1.644904!/file/Report_Ethics_of_Social_Media_Research_Jul16.pdf

Williams, Matthew L., Pete Burnap, and Luke Sloan. "Towards an ethical framework for publishing Twitter data in social research: taking into account users' views, online context and algorithmic estimation." Sociology 51.6 (2017): 1149-1168.

McMaster University | Library

Lewis & Ruth Sherman Centre for Digital Scholarship

Thank you

Andrea Zeffiro: zeffiroa@mcmaster.ca

Jay Brodeur: brodeujj@mcmaster.ca