

Tri-Agency
Data Management Policy
Initiative



Matthew Lucas, PhD
Social Sciences and Humanities
Research Council

SSHRC  CRSH

April 19 2018

PRESENTATION OUTLINE

- Rationale for Research Data Management
- Tri-Agency Data Management Policy Development
- Recent Engagement

WHAT ARE RESEARCH DATA?

Research data are contents that are used as primary sources to support research, scholarship, artistic activity or research-creation, and that are used as evidence in the research process and commonly accepted in the research community as necessary to validate research findings and results.

WHY DATA MANAGEMENT?

Research Excellence

Research Dissemination

Research Impact

Research Best Practice

RESEARCH EXCELLENCE

Project Management

Support Reproducibility

Avoid Duplication

Journal Guidelines



RESEARCH DISSEMINATION

Data Sharing

Citation

Interdisciplinarity

OPEN ACCESS Freely available online



Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma

Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

Background. Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available. **Principal Findings.** We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly ($p=0.006$) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression. **Significance.** This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

Citation: Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

INTRODUCTION

Sharing information facilitates science. Publicly sharing detailed research data—sample attributes, clinical factors, patient outcomes, DNA sequences, raw mRNA microarray measurements—with other researchers allows these valuable resources to contribute far beyond their original analysis[1]. In addition to being used to confirm original results, raw data can be used to explore related or new hypotheses, particularly when combined with other publicly available data sets. Real data is indispensable when investigating and developing study methods, analysis techniques, and software implementations. The larger scientific community also benefits: sharing data encourages multiple perspectives, helps to identify errors, discourages fraud, is useful for training new researchers, and increases efficient use of funding and patient population resources by avoiding duplicate data collection.

Believing that that these benefits outweigh the costs of sharing research data, many initiatives actively encourage investigators to make their data available. Some journals, including the *PLoS* family, require the submission of detailed biomedical data to publicly available databases as a condition of publication[2–4]. Since 2003, the NIH has required a data sharing plan for all large funding grants. The growing open-access publishing movement will perhaps increase peer pressure to share data.

However, while the general research community benefits from shared data, much of the burden for sharing the data falls to the study investigator. Are there benefits for the investigators themselves?

A currency of value to many investigators is the number of times their publications are cited. Although limited as a proxy for the

RESULTS

We studied the citations of 85 cancer microarray clinical trials published between January 1999 and April 2003, as identified in a systematic review by Ntzani and Ioannidis[7] and listed in Supplementary Text S1. We found 41 of the 85 clinical trials (48%) made their microarray data publicly available on the internet. Most data sets were located on lab websites (28), with a few found on publisher websites (4), or within public databases (6 in the Stanford Microarray Database (SMD)[8], 6 in Gene Expression Omnibus (GEO)[9], 2 in ArrayExpress[10], 2 in the NCI GeneExpression Data Portal (GEDP)[gedp.nci.nih.gov]; some datasets in more than one location). The internet locations of the datasets are listed in Supplementary Text S2. The majority of datasets were made available concurrently with the trial publication, as illustrated within the WayBackMachine internet archives (www.archive.org/web/web.php) for 25 of the datasets and mention of supplementary data within the trial publication itself for 10 of the remaining 16 datasets. As seen in Table 1, trials published in high impact journals, prior to 2001, or with US authors were more likely to share their data.

The cohort of 83 trials was cited an aggregate of 6239 times in 2004–2005 by 3133 distinct articles (median of 1.0 cohort citation per article, range 1–23). The 48% of trials which shared their data received a total of 5334 citations (85% of aggregate), distributed as shown in Figure 1.

Academic Editor: John Ioannidis, University of Ioannina School of Medicine, Greece



RESEARCH IMPACT

Within Science

Social Impact

Policy Impact

Data on bushfire fatalities helping to prevent more deaths

Who did the research?

Risk Frontiers (based at Macquarie University), RMIT University and Bushfire Cooperative Research Centre.

Melanoma study leads to ban on commercial sunbeds

What is the research about?

The **Melbourne Institute** at the University of Melbourne, funded by the Australian Government.

What is the project about?

The **Household, Income and Labour Dynamics in Australia (HILDA) survey** is an annual study of around 8000 households or around 20,000 individuals. It collects information about economic wellbeing, employment, health and family.

HILDA is a longitudinal study, meaning respondents are followed year-on-year to observe changes over time. The data is confidentialised to remove identifying

What is the real-life data impact?

The HILDA Survey is **known** as one of the best panel surveys in the world, collecting a broad range of high quality economic and social research data about Australian lives.

The data collected and analysed through HILDA have had a substantial impact on Australian policy development. Its **findings** have been used to help shape many policies, including:

- proposals in **Australia's Future Tax System Review**
- the Government's 2011 Paid Parental leave policy
- decisions about minimum wage rates



EverGraze project boosts Australian farming

Who did the research?

Charles Sturt University in partnership with **Future Farm Industries Cooperative Research Centre (FFI CRC)** and state departments in Victoria, New South Wales and Western Australia.

What is the project about?

EverGraze is a collaborative research project testing the innovative use of perennial flora

on farmland in high rainfall zones of southern Australia. It has helped farmers improve soil health (e.g. acidity and salinity), ultimately increasing livestock profitability.

More than 250 scientists, advisers and farmers have contributed to EverGraze since its inception in 2002, including experts in soil science, agronomy, environmental science, hydrology and economics.

With complex modelling across multiple

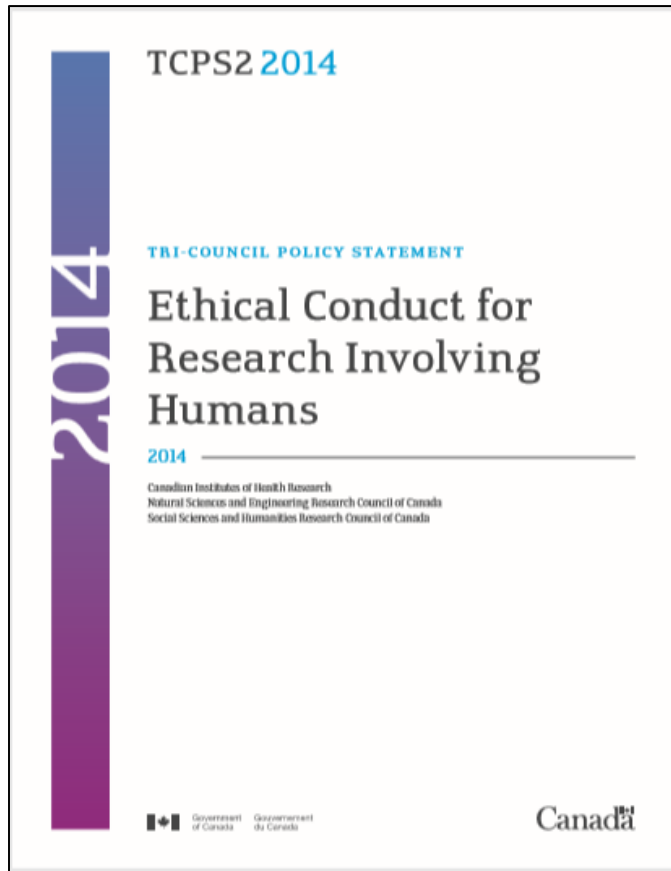
RESEARCH BEST PRACTICE

Transparency

Trust

Ethics

Responsible Use of
Public Funds



INTERNATIONAL DEVELOPMENTS

Research Funders

- National Science Foundation
- National Endowment for the Humanities - ODH
- National Institutes of Health
- UK Research Councils
- European Commission
- National Natural Science Foundation of China

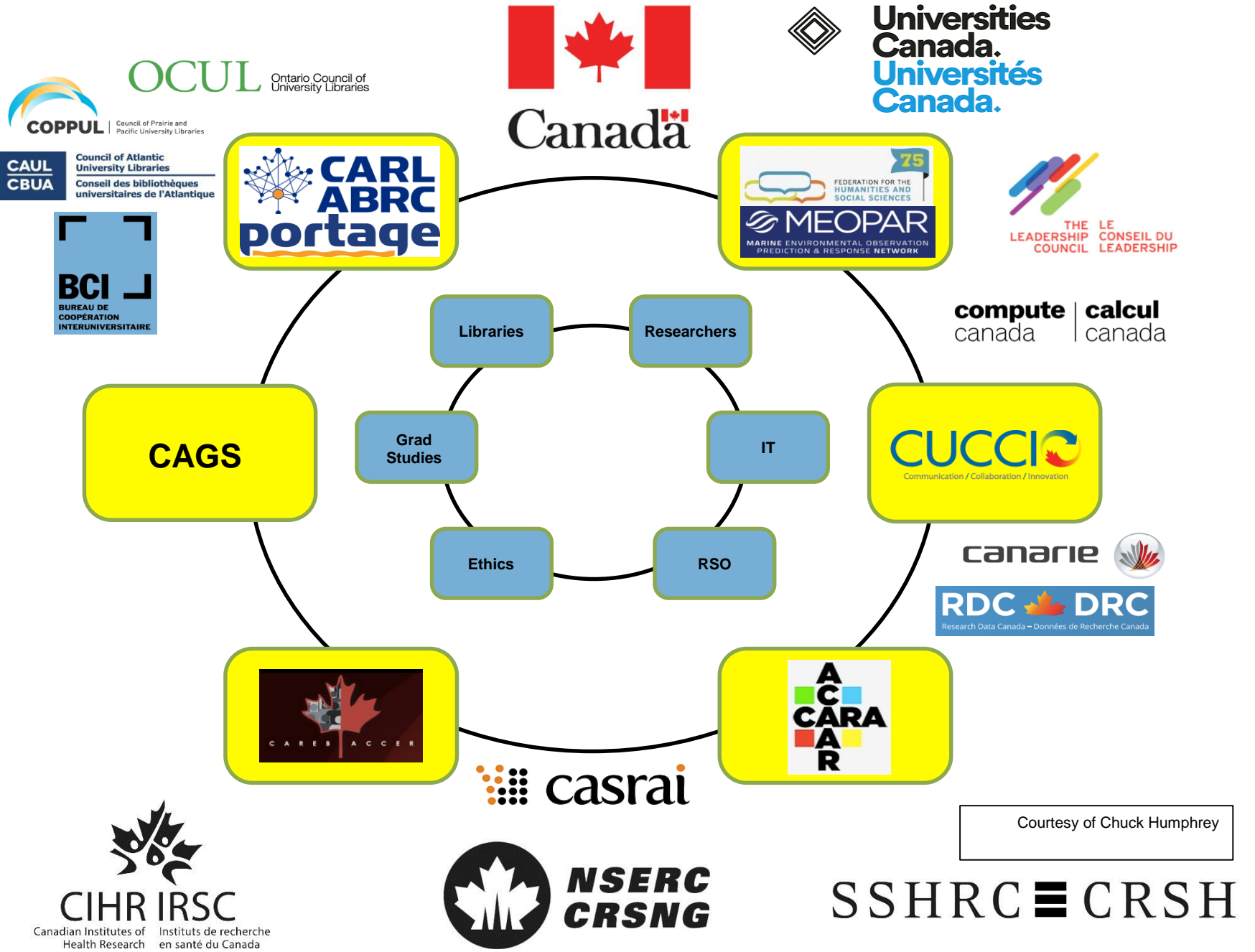
Foundations and Charities

- American Heart Association
- Bill and Melinda Gates Foundation
- Alfred P Sloan Foundation
- Gordon and Betty Moore Foundation
- Open Society Foundations
- Royal Society
- Wellcome Trust
- Cancer Research UK

Research Institutions



SSHRC  CRSH



Courtesy of Chuck Humphrey

GOVERNMENT OF CANADA DIRECTIVE ON OPEN GOVERNMENT

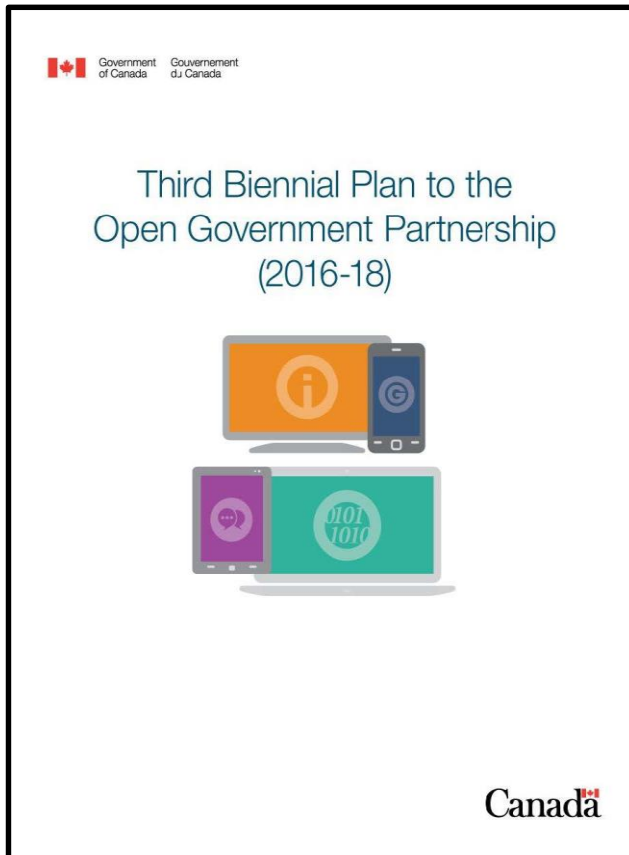


Table of Contents

- I. INTRODUCTION3
- Towards an Open and Transparent Government3**
- A Joint Effort4**
- The Open Government Partnership4**
- II. PROGRESS TO DATE5
- III. DEVELOPING CANADA'S THIRD BIENNIAL PLAN7
- IV. CANADA'S OPEN GOVERNMENT COMMITMENTS (2016-18)8
- A. Open by Default9**
- Commitment 1: Enhance Access to Information10*
- Commitment 2: Streamline Requests for Personal Information12*
- Commitment 3: Expand and Improve Open Data12*
- Commitment 4: Provide and Preserve Open Information14*
- Commitment 5: Define an Approach for Measuring Open Government Performance15*
- Commitment 6: Develop Open Government Skills across the Federal Public Service16*
- Commitment 7: Embed Transparency Requirements in the Federal Service Strategy17*
- Commitment 8: Enhance Access to Culture & Heritage Collections18*
- B. Fiscal Transparency19**
- Commitment 9: Enhance Openness of Information on Government Spending and Procurement19*
- Commitment 10: Increase Transparency of Budget and Other Department of Finance Information21*
- Commitment 11: Increase Transparency of Grants and Contributions Funding22*
- Commitment 12: Improve Public Information on Canadian Corporations23*
- C. Innovation, Prosperity, and Sustainable Development24**
- Commitment 13: Increase the Availability and Usability of Geospatial Data24*
- Commitment 14: Increase Openness of Federal Science Activities (Open Science)26*
- Commitment 15: Stimulate Innovation through Canada's Open Data Exchange (ODX)27*
- Commitment 16: Align Open Data across Canada (Open Data Canada)28*
- Commitment 17: Implement the Private Sector Transparency Measures Act29*
- Commitment 18: Support Openness and Transparency Initiatives around the World30*
- D. Engaging Canadians and the World31**
- Commitment 19: Engage Civil Society on Open Government32*
- Commitment 20: Enable Open Dialogue and Open Policy Making33*
- Commitment 21: Promote Open Government Globally34*
- Commitment 22: Engage Canadians to Improve Key Canada Revenue Agency Services35*
- V. CONCLUSION37
- APPENDIX: Summary of Canada's Third Biennial Plan to the Open Government Partnership (2016-18)38

TRI-AGENCY POLICY DEVELOPMENT BACKGROUND

2013

Capitalizing on Big Data: Toward a Policy Framework for
Advancing Digital Scholarship in Canada

2016

Tri-Agency Statement of Principles on Digital Data Management

2017-2018

Draft Tri-Agency Research Data Management Policy

INTENDED IMPACT OF A TRI-AGENCY DATA MANAGEMENT POLICY

The Tri-Agencies aim to contribute to a future research culture that sees:

- Strong data management as an accepted signifier of research excellence across disciplines, and a regular feature in the conduct of research;
- More Canadian datasets cited, and valued as a product of research in tenure, promotion and peer review processes;
- Canadian researchers equipped and ready to engage in international research collaboration where data management requirements are becoming the norm;
- Canadian research institutions ready to support the management of the data their researchers produce;
- Increased ability for research data to be archived, found and responsibly reused, to fuel new discovery and innovation.

DRAFT

TRI-AGENCY DATA MANAGEMENT POLICY

- For consultation
- Feedback will inform final policy
- Proposed policy includes 3 possible requirements:
 1. Institutions: Institutional Strategy
 2. Researchers: Data Management Plans
 3. Researchers: Data Deposit
- Implementation: Phased, incremental

DRAFT

TRI-AGENCY DATA MANAGEMENT POLICY

1. Institutions: Institutional Strategy

- Each institution administering Tri-Agency funds could be required to create an institutional research data management strategy. The strategy could outline how the institution will provide its researchers with an environment that enables and supports world class research data management practices.
- The strategy could be posted and made publicly available on the institution's website, with contact information to direct inquiries about the strategy.

DRAFT

TRI-AGENCY DATA MANAGEMENT POLICY

Why Require Institutional Strategies?

- Recognizes the role of institutions in providing supports for data management
- Provides an opportunity for institutions to think through where gaps exist, and how to address them from a campus-wide perspective
- Could aid institutions in developing an approach that works for them, while encouraging alignment and collaboration with other institutions
- Could provide information to agencies about data management capacity
- Serves as foundation for the potential requirements that follow

Support

Portage Institutional Strategy Template

DRAFT

TRI-AGENCY DATA MANAGEMENT POLICY

2. Researchers: Data Management Plans

- Grant recipients could be required to create data management plans (DMPs) for research projects supported wholly or in part by Tri-Agency funds. Grant recipients could submit these plans to their institution's research office as a condition of the release of grant funds.
- For specific funding opportunities, the agencies could require DMPs to be submitted to the appropriate agency at time of application; in these cases, they may be considered in the adjudication process.

DRAFT

TRI-AGENCY DATA MANAGEMENT POLICY

Why Require Data Management Plans?

- DMPs are an emerging international best practice
- DMPs are an excellent way for researchers to identify opportunities and challenges in managing their data, well before those opportunities and challenges emerge
- Researchers claim that the process of developing a DMP helps them to improve their research plans and methodologies
- DMPs could serve the responsible conduct of research and the research ethics approval process
- DMPs help identify and mitigate issues related to ownership of data, potential for data sharing, etc.

Support

Portage DMP Assistant

DRAFT

TRI-AGENCY DATA MANAGEMENT POLICY

3. Researchers: Data Deposit

- For all research data and code that support journal publications, pre-prints and other research outputs that arise from agency-supported research, grant recipients could be required to deposit these data and code in an appropriate public repository or other platform that will ensure safe storage, preservation, curation, and (if applicable) access to the data.

DRAFT

TRI-AGENCY DATA MANAGEMENT POLICY

Why Require Data Deposit?

- Methods, expectations and online security will change - storing in a secure location provides better chance for data to be safe and of use to the creator in the future
- Data deposit helps ensure proper use of public funds
- Facilitates reproducibility of results
- Facilitates data sharing

Support

Federated Research Data Repository

RECENT ENGAGEMENT

- Discussions with a broad array of stakeholders
 - Researchers, scholarly and scientific associations, data management advocacy and support organizations, funding agency colleagues around the globe.
- Regional stakeholder meetings
 - Meetings in Vancouver, Calgary, Toronto, Montreal and Halifax.
 - Meetings have revealed excitement and optimism about the potential for data management to contribute to research excellence, but have also demonstrated concern over challenges, such as researcher awareness, capacity and funding.

COMMUNITY FEEDBACK IS KEY

Community feedback is essential to inform the final design of the policy and the mode of its implementation.

The agencies consider the draft policy as a proposal through which to advance discussion with the many stakeholders in the research community, with a tri-agency policy as the desired end product.

The agencies plan to launch an online consultation to advance discussion with the community.

THANK YOU!

If you have any additional questions or feedback, please contact us at:

ResearchData-Donneesderecherche@sshrc-crsh.gc.ca