Supplementary Materials for:

The Stewardship Gap: A Challenge in Long-Term Access to Data

Myron P. Gutmann[1,2,3*], Jeremy York[2], Francine Berman[4]

[1] Institute of Behavioral Science and Department of History, University of Colorado Boulder, Boulder, Colorado, United States of America

[2] School of Information, University of Michigan, Ann Arbor, Michigan, United States of America

[3] Inter-university Consortium for Political and Social Research, University of Michigan, Ann Arbor, Michigan, United States of America

[4] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York, United States of America

* Corresponding Author

E-mail: Myron.gutmann@colorado.edu

Supplementary Text. The Stewardship Gap Project Questions.

Thank you very much for participating in this interview. I have a script that I am going to go through and try to follow rather closely in order to have consistency among the interviews. There is room within the questions for flexibility, however, and follow-ups, including by you if there is something you do not understand.

This interview is part of a study led by Myron Gutmann at the University of Colorado Boulder, and Fran Berman and the Rensselaer Polytechnic Institute, that is investigating what we are calling the Stewardship Gap: the gap between the existing amount of valuable data or creative work resulting from US public- or non-profit-funded research, and the amount that is being effectively stewarded and made accessible. It is 18-month project, begun in July 2015, and it is funded by the Alfred P. Sloan Foundation.

The primary goal of the interview is to gather information in three areas we believe are critical to understanding the stewardship gap. These are the extent of the stewardship commitment that exists on research data, the extent of value that the data have, and who has responsibility for stewardship—who has the ability to act to address the stewardship gap, if it exists.

We will use the information you provide to identify patterns among researchers in a broad variety of fields that will help us begin to identify and characterize the stewardship gap. We may also make recommendations about strategies to address a gap if we find one exists. Any results we make public will be in aggregate form without identifying characteristics (for instance, we will report your discipline but not specific institution; we will report on sizes and characteristics of project data, but not identify the project or the specific data that were collected or used), or completely anonymous vignettes.

This interview is completely optional. You may end it at any time or chose not to answer specific questions without penalty. If you would like your responses to remain anonymous you can indicate this at any time and we will de-identify any information that is reported in project findings. I want to pause for a moment to see if you have any questions about how we will use information you provide, and if you would like for your responses to be anonymous.

With that, I will proceed with asking the questions. I am going to ask you a set of questions about data from a specific project of your choosing. The project needs to be one that was funded by a public- or non-profit source, one for which you were responsible for generating the digital data or creative content, and one for which you are able to speak confidently about questions of size, content characteristics, and preservation commitments related to the data.

Through the course of the interview I'm going to use variations on the words Steward and Preserve. By Steward I mean to responsibly manage data that is in your care (including a wide variety of activities that might be involved in management). By Preserve I mean to execute a set of activities with an explicit purpose to maintain the integrity of data over time.

When I use the term “data” I am referring to digital outputs resulting from your research that did not exist beforehand and have not already been deleted. Some examples are outputs from original observations, experiments, and simulations, including contextual data such as images, audio, or photographs that may not have been the primary focus of the research. We would also like to learn about contextual or administrative data or metadata that are necessary to understand the primary project data, or pre-existing datasets that may have been analyzed, cleaned, or corrected for use in the project.

1. Please tell me your full name and current institutional affiliation.

2. I would like you to think about a specific project in which you were responsible for generating digital data or creative content and for which you are able to speak confidently about questions of size, content characteristics, and preservation commitments

   Who funded the project, what was its name and purpose, what years did it run, and who was involved?
   a. What was your role in the project, and with respect to the data generated

3. Would you describe your project as one of the following:
   a. Primarily about collecting data in order to make them available to other researchers
   b. Primarily testing a hypothesis but in the process of doing so collecting data
   c. Evenly divided between the two

4. What data resulted from the project? Please give a brief description of the general content types and formats, and total size.
   a. Were any additional administrative or contextual data that are needed to understand the project data produced that you have not described?

5. Did your research involve analysis of data collected by others (secondary data)?

6. Are any of the data private, proprietary or confidential?

7. Are there any other attributes of the data that are significant?

8. What domain of content are the data in? Please indicate all that apply.

9. How much of the data are you currently responsible for stewarding?

10. Have you transferred responsibility for stewarding any of the data to someone else. If os, how much and to whom?
    a. In turning over data to [responsible entity], was your primary goal to share the data with others or preserve the data for your own future use?

11. Is anyone else stewarding the same data?

12. For the data you are currently stewarding, how much falls into each of the categories below:

    A) There is an explicit commitment to preserve the data (an agreement or MOU regarding preservation has been signed).
    B) There is an explicit intention to preserve the data (you intend to preserve the data and are or will implement preservation actions).
    C) There is no intention to preserve the data, though also no intention to delete the data.
    D) The data are temporary and will be deleted after a certain period of time.

13. For the data in each of these categories
    a. What is the remaining term of the commitment or intention?
    b. What are the reasons for the commitment or intention term (if there is no intention, why is this the case)?
    c. Where are the data located?
    d. What is the extent of the commitment or intention (what is being or will be done to take care of the data)?
    e. What is the remaining duration of the value that you believe the data have? Please indicate if there are different durations for different data, including sensitive data
    f. Why will the data have value for this duration?
    g. Do you think it would be worthwhile to reassess the value of this material in the future? At what interval?
    h. How secure do you feel in your ability or the ability of the entity stewarding the data to fulfill the commitment or intention? Would you describe yourself as Very confident, Reasonably confident, Somewhat concerned, or Very concerned?
        i. Is there a difference in your confidence in the short- and long-term?
        ii. Do you have any specific concerns (e.g., about resources, knowledge, staff)?
    i. What prospects for stewardship, if any, exist for the data when the term of the commitment or intention is over? What do you expect or assume will happen to the data? What do you think should happen to the data (who did you think should be responsible for stewarding the data)?

For each combination of commitment length and value duration (e.g., data that are valuable for 5 years and have a five-year commitment, or data that are valuable for 10 years and have a five-year commitment), please answer the following:

14. Do data in this category overlap with data in any other category? Which?

15. What are the reasons the data have value? Please indicate the degree to which you agree or disagree with the following (strongly agree, agree, neutral, disagree, strongly disagree (no value), unsure, N/A). Please also indicate if different reasons or degrees of value apply to sensitive data:
    a. The data are valuable for my own research
    b. The data are in demand in my immediate community of research (as evidenced by number of requests received, citations, or uses made)
    c. The data are in demand outside my immediate community of research (as evidenced by number of requests received, citations, or uses made)

d. The data have broadly applicable value (for instance, as cultural heritage or value for inclusion in a community knowledge base or reference collection)
e. The data have a high demonstrated or potential impact in terms of people, money, time, policy, transformative potential, or some another factor.
f. The data are valuable due to their organization, usability, discovery, accessibility, or documentation, or the time invested in these activities
g. The data are valuable because they are used in or support important services (e.g., full-text search, instrument calibrations, etc.)
h. The data have value due to their timeliness (they provide real-time data, or are particularly relevant at present)
i. The data are valuable because they would be costly or difficult to reproduce
j. The data are valuable for audit purposes or because they have been mandated to be kept
k. The data are valuable because they are needed to reproduce research results
l. The data have historic or longitudinal value
m. The data are valuable because they facilitate research, training, teaching, or outreach
n. The data are kept out of good scholarly practice
o. The data increase in value when combined with other data
p. The data only have value when combined with other data
q. The data have value for another reason (please specify)
r. The data have gained value over time
s. The data will gain value over time
t. The data have lost value over time
u. The data will lose value over time
v. The data are timeless (the data will never lose their value)

16. Which reasons for value, if any, have had the greatest impact on decisions about the likelihood of preservation, or type or duration of preservation commitment? Please rank the top three.
    a. What indication do you have that those were the reasons that had an impact?

17. What are the primary external measures or indicators of value that have or will be applied to these data?

18. Do those measures accurately represent the value that the data have?

19. Please answer these same questions for data that you transferred to someone else.

20. Did the survey allow you to describe your data in a way that is meaningful and accurate from your point of view? Is there anything else you would like to add?

21. Please describe any difficulty you had in answering the questions.

22. Would you be willing to be contacted if I have any follow-up questions about this interview?

Supplementary Table. Researcher Discipline and Research Data Domain (self-reported).

Discipline/Domain	Researcher Discipline	Research Data Area	Discipline/Domain	Researcher Discipline	Research Data Area
Aeroscience		X	Informatics and Computing	X	
Agriculture		X	Information Science	X	X
Archaeology	X	X	Management		X
Astronomy	X	X	Marine Conservation		X
Astrophysics		X	Material Science		X
Audiology		X	Medical Humanities		X
Behavioral Science		X	Medicinal Chemistry		X
Biochemistry		X	Medicine		X
Biological Oceanography		X	Meteorology		X
Biomedical engineering	X	X	Microbiology	X	X
Biomedical research		X	Migration Studies		X
Biomedical Science	X		Molecular and Integrative Physiology	X	
Business		X	Neurology		X
Cell Biology		X	Neuroscience	X	X
Chemistry		X	Nursing	X	X
Chemical Biology	X		Oncology		X
Child Development		X	Operations Research		X
Climatology		X	Parasitology		X
Computer Science	X	X	Pediatrics		X
Cybersecurity		X	Pharmacology	X	
DNA Repair		X	Physical Therapy and Human Movement Sciences	X	X
Dendrochronology		X	Physics		X
Developmental Biology		X	Plant Pathology	X	
Developmental Science		X	Political Science		X
E-Science		X	Programming Language		X
Ecology		X	Provenance		X
Economics	X	X	Psychiatry		X
Electrical Engineering and Computer Science	X		Psychology	X	X
Education		X	Public Administration		X
Environmental Science	X	X	Public Health		X
Epidemiology	X		Public Policy/Public Affairs	X	X
Fisheries		X	Scientific Computing		X
Genetics	X	X	Social Psychology		X
Geochemistry		X	Social Work		X
Geography		X	Sociology	X	X
Geology		X	Software Engineering		X
Geomorphology		X	Solar Physics		X
Geophysics	X	X	Special Education	X	
Geosciences		X	Statistical Science	X	
Health Care		X	Stem cell biology		X

Heliophysics	X	X	Structural Biology		X
Historical Geography	X	X	Urban Studies		X
History	X	X	Urban and Community Planning		X
Human Computer Interaction		X	X-Ray Astronomy	X	X
Human genetics		X			