Advanced Data Management (CSCI 490/680)

Data Citation

Dr. David Koop
What is Data?

Marie Curie’s notebook

http://www.census.gov/population/cen2000/map02.gif

What are data?

ncl.ucar.edu

http://onlineqda.hud.ac.uk/

Intro_QDA/

Examples_of_Qualitative_Data.php

Marie Curie’s notebook

What is Data?

Zafar

He will grow old in his present house; new house is for sons + 3 sons. Not sure they want to live in village. He will only build another if they want him to. He came from Germany and did the plastering. He arranged the carpentry in Kavseri. Çok para gitti (much money went). Has a tractor.

Date: July 1980 Place: Sakaltutan

Zafar: Household now Zafar and wife; Nazlı Unal and wife and youngest son, still a boy. They run two dolmus; one with a driver from Süleymanı. Goes in and out once a day. He gets 8,000 a month. Zafar then said, keşke de 20 TL (not sharp - 20 TL not profitable) I said he did very well on 8,000 TL with only two journeys a day. Nazlı Unal has "bought" a Durak (dolmus stop) from Beledive and works all day in Kavseri.

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php

Pisa Griffin

C. Borgman
What is data?

• "Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship."
  
  [C. L. Borgman]

• Data can be digital but can also be physical (e.g. sculptures)
• Semantics are important (e.g. temperature to engineer and biologist)
• Grey Data: surveys, student records—think about privacy
Sharing Data

• Required/encouraged by universities, funding agencies, publishers
• "Publications are arguments made by authors, and data are the evidence used to support the arguments." [C. L. Borgman]

• Questions:
  - How is data maintained? Who is responsible?
  - What is the process for curating data?
  - How long should data be kept?
  - How should data collection and curation be acknowledged?
The DCC Curation Lifecycle Model

Full Lifecycle Actions
- Conceptualise
- Create or receive
- Appraise and select
- Ingest
- Preservation action
- Store
- Access, use & reuse
- Transform
- Curate
- Preservation planning
- Description
- Representation information
- Community watch & participation
- Preserve
- Reappraise & select
- Migrate
- Dispose

Occasional Actions
- Dispose
- Reappraise
- Migrate

Data, any information in binary digital form, is at the centre of the Curation Lifecycle. This includes:
- Simple digital objects are discrete digital items; such as textual files, images or sound files, along with their related identifiers and metadata.
- Complex digital objects are discrete digital objects, made by combining a number of other digital objects, such as websites, structured collections of records or data stored in a computer system.

Data Curation Lifecycle
Sequential Actions in Data Curation

- **Conceptualize**: Plan creation of data—capture method and storage options.
- **Create or Receive**: Create/receive data and make sure metadata exists.
- **Appraise and Select**: Evaluate data and select for long-term curation and preservation.
- **Ingest**: Transfer data to an archive, repository, data centre or other custodian.
- **Preservation Action**: Data cleaning, validation (ensure that data remains authentic, reliable and usable).
- **Store**: Store the data in a secure manner adhering to relevant standards.
- **Access, Use and Reuse**: Make sure is accessible to users and reusers.
- **Transform**: Create new data from the original (migrate formats, subsets, etc.).
FAIR Principles

• **Findable**: Metadata and data should be easy to find for both humans and computers

• **Accessible**: Users need to know how data can be accessed, possibly including authentication and authorization

• **Interoperable**: Can be integrated with other data, and can interoperate with applications or workflows for analysis, storage, and processing

• **Reusable**: Optimize the reuse of data. Metadata and data should be well-described so they can be replicated and/or combined in different settings
Findable: DataCite Workflow

1. Take a dataset

2. Describe it

   - Title
   - Authors
   - Year
   - Description
   - And others…

3. Assign a DOI

   - 10.1234/exampledata

   - Proxy
   - Prefix
   - Suffix

   https://doi.org/10.5438/n138-z3mk
Accessible: DOI to Landing Page with Metadata

Citation → PID resolution → Landing Page → web service → Data

- Document citing the data
- Repository housing the data
- Data store

[M. Fenner et al., 2019]
Interoperable: Standard vocabularies

[Image of a webpage showing a table with columns for Registry, Name, Abbreviation, Type, Subject, Domain, Taxonomy, Related Database, Related Standard, Related Policy, In Collection/Recommendation, and Status. The table includes entries such as "Access to Biological Collection Data" and "ACE Format".]
Reusable: Licensing

- Citation of a dataset is expected as a scholarly norm, not by law
- CC0:
  - "I hereby waive all copyright and related or neighboring rights together with all associated claims and causes of action with respect to this work to the extent possible under the law"
- CC BY: license, not a waiver as CC0
  - "You must give appropriate credit, provide a link to the license, and indicate if changes were made."
- Data Use Agreements (DUA): Used when data are restricted due to proprietary or privacy concerns.
Reusable: Data Citation & Metrics

[H. Cousijn et al., 2019]
Assignment 4

• World Education Data
• Collected/collated by UNESCO, World Bank, and OECD
• Transform World Bank Data
• Impute missing year data
• Integrate teacher and student numbers
• Fuse three datasets
Studying Data Availability

• Who mandates data sharing, and what is the impact?
  - Government
  - Funding agencies
  - Institutions
  - Journals

• How does the age of a publication/data item affect availability?
  - If not curated, how to locate?
  - What factors influence this?
Since this is a logistic model, we can readily calculate the effect that the different policy types have on the likelihood that the data will be available. We explore these odds for each type of policy below, using “no policy” as the baseline.

Having a “recommend archiving” policy made it 3.6 times more likely that the data were online compared to having no policy. However, the 95% CI overlapped with 1 (0.96–13.6); hence, this increase in the odds is not significant. Overall, recommending data archiving is only marginally more effective than having no policy at all.

The data were 17 times more likely to be available online for journals that had adopted a mandatory data archiving policy but did not require a data accessibility statement in the manuscript. This odds ratio was significantly greater (95% CI: 3.7–79.6).

For “mandate archiving” journals where a data accessibility statement is required in the manuscript, the odds of finding the data online were 974 times higher compared to having no policy. The 95% CI on these odds is very wide (97.9–9698.8), but nonetheless shows that the combination of a mandatory policy and an accessibility statement is much more effective than any other policy type.

REQUESTING DATA DIRECTLY FROM AUTHORS

A number of the “recommend archiving” policies state that the data should also be freely available from the authors by request (see the Journal Policies file at doi: 10.5061/dryad.6bs31); hence, we wanted to evaluate whether obtaining data directly from authors is an effective approach.

Part of the dataset collection for our reproducibility study (5) involved e-mailing authors of papers from two of the “recommend archiving” journals (BMC Evolutionary Biology and PLoS One) and requesting their structure input files. Here, we examine how often these requests led to us obtaining the data. We did not e-mail the authors of articles where the data were already available online.

A detailed description of our data request process appears on Dryad (doi: 10.5061/dryad.6bs31), but we essentially contacted corresponding and senior authors of each article up to 3 times over a 3-wk period, and recorded if and when the data were received.

We obtained data directly from the authors for 7 of the 12 eligible articles in BMC Evolutionary Biology, and 27 datasets from 45 articles from PLoS One (Table 1). All seven of the BMC Evolutionary Biology datasets arrived between 8 and 14 days after our initial request. Ten of the PLoS One datasets came within 1 week, 13 came between 8 and 14 days, and 4 arrived between 15 and 21 days. Unlike the online data, which could generally be obtained within a few minutes, the requested datasets took a mean of 7.7 days to arrive, with one author responding that the dataset had been lost in the year since publication. More than one e-mail had to be sent to the corresponding and/or senior author for 53% of papers, and the authors of 29% of the papers did not respond to any of our requests.

No data were received 21 days after our initial request. We also note that requesting data via e-mail did upset some authors, particularly when they were reminded of the journal’s data archiving policy or when multiple e-mails were sent.

Our average return of 59% in an average of 7.7 days is markedly better than has been reported in similar studies: Wicherts et al. (8) received only 26% of requested datasets after 6 months of effort with authors of 141 psychology articles, and Savage and Vickers (9) received only 1 of 10 papers with data available online.
We found a strong effect of article age on the availability of data from these 516 studies. The decline in data availability could arise because the authors of older papers were less likely to respond, but this was not supported by the data. Instead, researchers were equally likely to respond (Figure 1B) and to indicate the status of their data (Figure 1C) across the entire range of article ages.

The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year (Figure 1D). There was a continuum of author responses between the data being reported lost and being stored on inaccessible media, and they seemed to vary with the amount of time and effort involved in retrieving the data. Responses included authors being sure that the data were lost (e.g., on a stolen computer) or thinking that they might be stored in some distant location (e.g., their parent’s attic) to authors having some degree of certainty that the data are on a Zip or floppy disk in their possession but no longer having the appropriate hardware to access it. In the latter two cases, the authors would have to devote hours or days to retrieving the data. Our reason for needing the data (a reproducibility study) was not especially compelling for authors, and we may have received more of these inaccessible data sets if we had offered authorship on the subsequent paper or said that the data were needed for an important medical or conservation project.

The odds that we were able to find an apparently working e-mail address (either in the paper or by searching online) for any of the contacted authors did decrease by about 7% per year. This decrease was partly driven by a dearth of e-mail addresses in articles published before 2000 (0.38 per paper on average for 1991–1999) compared with those published after 2000.

### Table 1. Breakdown of Data Availability by Year of Publication

<table>
<thead>
<tr>
<th>Year</th>
<th>No Working E-Mail</th>
<th>No Response to E-Mail</th>
<th>Response Did Not Give Status of Data</th>
<th>Data Lost</th>
<th>Data Exist, Unwilling to Share</th>
<th>Data Received</th>
<th>Data Extant (Unwilling to Share + Received)</th>
<th>Number of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>9 (35%)</td>
<td>9 (35%)</td>
<td>2 (8%)</td>
<td>4 (15%)</td>
<td>1 (4%)</td>
<td>1 (4%)</td>
<td>2 (8%)</td>
<td>26</td>
</tr>
<tr>
<td>1993</td>
<td>14 (39%)</td>
<td>11 (31%)</td>
<td>3 (8%)</td>
<td>7 (19%)</td>
<td>0 (0%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
<td>36</td>
</tr>
<tr>
<td>1995</td>
<td>11 (31%)</td>
<td>9 (26%)</td>
<td>0 (0%)</td>
<td>7 (20%)</td>
<td>2 (6%)</td>
<td>6 (17%)</td>
<td>8 (23%)</td>
<td>35</td>
</tr>
<tr>
<td>1997</td>
<td>11 (37%)</td>
<td>9 (30%)</td>
<td>1 (3%)</td>
<td>2 (7%)</td>
<td>3 (10%)</td>
<td>4 (13%)</td>
<td>7 (23%)</td>
<td>30</td>
</tr>
<tr>
<td>1999</td>
<td>19 (48%)</td>
<td>13 (32%)</td>
<td>1 (2%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>6 (15%)</td>
<td>6 (15%)</td>
<td>40</td>
</tr>
<tr>
<td>2001</td>
<td>13 (30%)</td>
<td>15 (35%)</td>
<td>3 (7%)</td>
<td>4 (9%)</td>
<td>0 (0%)</td>
<td>8 (19%)</td>
<td>8 (19%)</td>
<td>43</td>
</tr>
<tr>
<td>2003</td>
<td>9 (20%)</td>
<td>20 (43%)</td>
<td>4 (9%)</td>
<td>2 (4%)</td>
<td>0 (0%)</td>
<td>11 (24%)</td>
<td>11 (24%)</td>
<td>46</td>
</tr>
<tr>
<td>2005</td>
<td>11 (24%)</td>
<td>14 (31%)</td>
<td>6 (13%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>13 (29%)</td>
<td>13 (29%)</td>
<td>45</td>
</tr>
<tr>
<td>2007</td>
<td>12 (18%)</td>
<td>31 (47%)</td>
<td>2 (3%)</td>
<td>4 (6%)</td>
<td>1 (2%)</td>
<td>16 (24%)</td>
<td>17 (26%)</td>
<td>66</td>
</tr>
<tr>
<td>2009</td>
<td>9 (13%)</td>
<td>34 (49%)</td>
<td>3 (4%)</td>
<td>5 (7%)</td>
<td>6 (9%)</td>
<td>12 (17%)</td>
<td>18 (26%)</td>
<td>69</td>
</tr>
<tr>
<td>2011</td>
<td>13 (16%)</td>
<td>29 (36%)</td>
<td>8 (10%)</td>
<td>0 (0%)</td>
<td>7 (9%)</td>
<td>23 (29%)</td>
<td>30 (38%)</td>
<td>80</td>
</tr>
<tr>
<td>Totals</td>
<td>131 (25%)</td>
<td>194 (38%)</td>
<td>33 (6%)</td>
<td>37 (7%)</td>
<td>20 (4%)</td>
<td>101 (19%)</td>
<td>121 (23%)</td>
<td>516</td>
</tr>
</tbody>
</table>

Data are displayed as n (%); the percentages are calculated by rows.
Discussion

We found a strong effect of article age on the availability of data from these 516 studies. The decline in data availability could arise because the authors of older papers were less likely to respond, but this was not supported by the data. Instead, researchers were equally likely to respond (Figure 1B) and to indicate the status of their data (Figure 1C) across the entire range of article ages.

The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year (Figure 1D). There was a continuum of author responses between the data being reported lost and being stored on inaccessible media, and they seemed to vary with the amount of time and effort involved in retrieving the data. Responses included authors being sure that the data were lost (e.g., on a stolen computer) or thinking that they might be stored in some distant location (e.g., their parent's attic) to authors having some degree of certainty that the data are on a Zip or floppy disk in their possession but no longer having the appropriate hardware to access it. In the latter two cases, the authors would have to devote hours or days to retrieving the data. Our reason for needing the data (a reproducibility study) was not especially compelling for authors, and we may have received more of these inaccessible data sets if we had offered authorship on the subsequent paper or said that the data were needed for an important medical or conservation project.

The odds that we were able to find an apparently working e-mail address (either in the paper or by searching online) for any of the contacted authors did decrease by about 7% per year. This decrease was partly driven by a dearth of e-mail addresses in articles published before 2000 (0.38 per paper on average for 1991–1999) compared with those later papers.

Table 1. Breakdown of Data Availability by Year of Publication

<table>
<thead>
<tr>
<th>Year</th>
<th>No Working E-Mail</th>
<th>No Response to E-Mail</th>
<th>Response Did Not Give Status of Data</th>
<th>Data Lost</th>
<th>Data Exist, Unwilling to Share</th>
<th>Data Received</th>
<th>Data Extant (Unwilling to Share + Received)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>9 (35%)</td>
<td>9 (35%)</td>
<td>2 (8%)</td>
<td>4 (15%)</td>
<td>1 (4%)</td>
<td>2 (8%)</td>
<td>26</td>
</tr>
<tr>
<td>1993</td>
<td>14 (39%)</td>
<td>11 (31%)</td>
<td>3 (8%)</td>
<td>7 (19%)</td>
<td>0 (0%)</td>
<td>1 (3%)</td>
<td>36</td>
</tr>
<tr>
<td>1995</td>
<td>11 (31%)</td>
<td>9 (26%)</td>
<td>0 (0%)</td>
<td>7 (20%)</td>
<td>2 (6%)</td>
<td>6 (17%)</td>
<td>35</td>
</tr>
<tr>
<td>1997</td>
<td>11 (37%)</td>
<td>9 (30%)</td>
<td>1 (3%)</td>
<td>2 (7%)</td>
<td>3 (10%)</td>
<td>4 (13%)</td>
<td>30</td>
</tr>
<tr>
<td>1999</td>
<td>19 (48%)</td>
<td>13 (32%)</td>
<td>1 (2%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>6 (15%)</td>
<td>40</td>
</tr>
<tr>
<td>2001</td>
<td>13 (30%)</td>
<td>15 (35%)</td>
<td>3 (7%)</td>
<td>4 (9%)</td>
<td>0 (0%)</td>
<td>8 (19%)</td>
<td>43</td>
</tr>
<tr>
<td>2003</td>
<td>9 (20%)</td>
<td>20 (43%)</td>
<td>4 (9%)</td>
<td>2 (4%)</td>
<td>0 (0%)</td>
<td>11 (24%)</td>
<td>46</td>
</tr>
<tr>
<td>2005</td>
<td>11 (24%)</td>
<td>14 (31%)</td>
<td>6 (13%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>13 (29%)</td>
<td>45</td>
</tr>
<tr>
<td>2007</td>
<td>12 (18%)</td>
<td>31 (47%)</td>
<td>2 (3%)</td>
<td>4 (6%)</td>
<td>1 (2%)</td>
<td>16 (24%)</td>
<td>66</td>
</tr>
<tr>
<td>2009</td>
<td>9 (13%)</td>
<td>34 (49%)</td>
<td>3 (4%)</td>
<td>5 (7%)</td>
<td>6 (9%)</td>
<td>12 (17%)</td>
<td>69</td>
</tr>
<tr>
<td>2011</td>
<td>13 (16%)</td>
<td>29 (36%)</td>
<td>8 (10%)</td>
<td>0 (0%)</td>
<td>7 (9%)</td>
<td>23 (29%)</td>
<td>80</td>
</tr>
<tr>
<td>Totals</td>
<td>131 (25%)</td>
<td>194 (38%)</td>
<td>33 (6%)</td>
<td>37 (7%)</td>
<td>20 (4%)</td>
<td>101 (19%)</td>
<td>121 (23%)</td>
</tr>
</tbody>
</table>

Data are displayed as n (%); the percentages are calculated by rows.

Figure 1. The Effect of Article Age on Four Obstacles to Receiving Data from the Authors (A) Predicted probability that the paper had at least one apparently working e-mail. (B) Predicted probability of receiving a response, given that at least one e-mail was apparently working. (C) Predicted probability of receiving a response giving the status of the data, given that we received a response. (D) Predicted probability that the data were extant (either 'shared' or 'exist but unwilling to share') given that we received a useful response. In all panels, the line indicates the predicted probability from the logistic regression, the gray area shows the 95% CI of this estimate, and the red dots indicate the actual proportions from the data.
Discussion

We found a strong effect of article age on the availability of data from these 516 studies. The decline in data availability could arise because the authors of older papers were less likely to respond, but this was not supported by the data. Instead, researchers were equally likely to respond (Figure 1B) and to indicate the status of their data (Figure 1C) across the entire range of article ages. The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year (Figure 1D). There was a continuum of author responses between the data being reported lost and being stored on inaccessible media, and they seemed to vary with the amount of time and effort involved in retrieving the data. Responses included authors being sure that the data were lost (e.g., on a stolen computer) or thinking that they might be stored in some distant location (e.g., their parent's attic) to authors having some degree of certainty that the data are on a Zip or floppy disk in their possession but no longer having the appropriate hardware to access it. Our reason for needing the data (a reproducibility study) was not especially compelling for authors, and we may have received more of these inaccessible data sets if we had offered authorship on the subsequent paper or said that the data were needed for an important medical or conservation project.

The odds that we were able to find an apparently working e-mail address (either in the paper or by searching online) for any of the contacted authors did decrease by about 7% per year. This decrease was partly driven by a dearth of e-mail addresses in articles published before 2000 (0.38 per paper on average for 1991–1999) compared with those published thereafter. Table 1. Breakdown of Data Availability by Year of Publication

<table>
<thead>
<tr>
<th>Year</th>
<th>No Working E-Mail</th>
<th>No Response to E-Mail</th>
<th>Response Did Not Give Status of Data</th>
<th>Data Lost</th>
<th>Data Exist, Unwilling to Share</th>
<th>Data Extant (Unwilling to Share + Received)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>9 (35%)</td>
<td>9 (35%)</td>
<td>2 (8%)</td>
<td>4 (15%)</td>
<td>1 (4%)</td>
<td>2 (8%)</td>
</tr>
<tr>
<td>1993</td>
<td>14 (39%)</td>
<td>11 (31%)</td>
<td>3 (8%)</td>
<td>7 (19%)</td>
<td>0 (0%)</td>
<td>1 (3%)</td>
</tr>
<tr>
<td>1995</td>
<td>11 (31%)</td>
<td>9 (26%)</td>
<td>0 (0%)</td>
<td>7 (20%)</td>
<td>2 (6%)</td>
<td>6 (17%)</td>
</tr>
<tr>
<td>1997</td>
<td>11 (37%)</td>
<td>9 (30%)</td>
<td>1 (3%)</td>
<td>2 (7%)</td>
<td>3 (10%)</td>
<td>7 (23%)</td>
</tr>
<tr>
<td>1999</td>
<td>19 (48%)</td>
<td>13 (32%)</td>
<td>1 (2%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>6 (15%)</td>
</tr>
<tr>
<td>2001</td>
<td>13 (30%)</td>
<td>15 (35%)</td>
<td>3 (7%)</td>
<td>4 (9%)</td>
<td>0 (0%)</td>
<td>8 (19%)</td>
</tr>
<tr>
<td>2003</td>
<td>9 (20%)</td>
<td>20 (43%)</td>
<td>4 (9%)</td>
<td>2 (4%)</td>
<td>0 (0%)</td>
<td>11 (24%)</td>
</tr>
<tr>
<td>2005</td>
<td>11 (24%)</td>
<td>14 (31%)</td>
<td>6 (13%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>13 (29%)</td>
</tr>
<tr>
<td>2007</td>
<td>12 (18%)</td>
<td>31 (47%)</td>
<td>2 (3%)</td>
<td>4 (6%)</td>
<td>1 (2%)</td>
<td>16 (24%)</td>
</tr>
<tr>
<td>2009</td>
<td>9 (13%)</td>
<td>34 (49%)</td>
<td>3 (4%)</td>
<td>5 (7%)</td>
<td>6 (9%)</td>
<td>12 (17%)</td>
</tr>
<tr>
<td>2011</td>
<td>13 (16%)</td>
<td>29 (36%)</td>
<td>8 (10%)</td>
<td>0 (0%)</td>
<td>7 (9%)</td>
<td>23 (29%)</td>
</tr>
<tr>
<td>Totals</td>
<td>131 (25%)</td>
<td>194 (38%)</td>
<td>33 (6%)</td>
<td>37 (7%)</td>
<td>20 (4%)</td>
<td>101 (19%)</td>
</tr>
</tbody>
</table>

Data are displayed as n (%); the percentages are calculated by rows.

Figure 1. The Effect of Article Age on Four Obstacles to Receiving Data from the Authors (A) Predicted probability that the paper had at least one apparently working e-mail. (B) Predicted probability of receiving a response, given that at least one e-mail was apparently working. (C) Predicted probability of receiving a response giving the status of the data, given that we received a response. (D) Predicted probability that the data were extant (either “shared” or “exist but unwilling to share”) given that we received a useful response. In all panels, the line indicates the predicted probability from the logistic regression, the gray area shows the 95% CI of this estimate, and the red dots indicate the actual proportions from the data.
We found a strong effect of article age on the availability of data from these 516 studies. The decline in data availability could arise because the authors of older papers were less likely to respond, but this was not supported by the data. Instead, researchers were equally likely to respond (Figure 1B) and to indicate the status of their data (Figure 1C) across the entire range of article ages.

The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year (Figure 1D). There was a continuum of author responses between the data being reported lost and being stored on inaccessible media, and they seemed to vary with the amount of time and effort involved in retrieving the data. Responses included authors being sure that the data were lost (e.g., on a stolen computer) or thinking that they might be stored in some distant location (e.g., their parent's attic) to authors having some degree of certainty that the data are on a Zip or floppy disk in their possession but no longer having the appropriate hardware to access it. In the latter two cases, the authors would have to devote hours or days to retrieving the data. Our reason for needing the data (a reproducibility study) was not especially compelling for authors, and we may have received more of these inaccessible data sets if we had offered authorship on the subsequent paper or said that the data were needed for an important medical or conservation project.

The odds that we were able to find an apparently working e-mail address (either in the paper or by searching online) for any of the contacted authors did decrease by about 7% per year. This decrease was partly driven by a dearth of e-mail addresses in articles published before 2000 (0.38 per paper on average for 1991–1999) compared with those published after 2000.

### Table 1: Breakdown of Data Availability by Year of Publication

<table>
<thead>
<tr>
<th>Year</th>
<th>No Working E-Mail</th>
<th>No Response to E-Mail</th>
<th>Response Did Not Give Status of Data</th>
<th>Data Lost</th>
<th>Data Exist, Unwilling to Share</th>
<th>Data Received</th>
<th>Data Extant (Unwilling to Share + Received)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>9 (35%)</td>
<td>9 (35%)</td>
<td>2 (8%)</td>
<td>4 (15%)</td>
<td>1 (4%)</td>
<td>1 (4%)</td>
<td>2 (8%)</td>
</tr>
<tr>
<td>1993</td>
<td>14 (39%)</td>
<td>11 (31%)</td>
<td>3 (8%)</td>
<td>7 (19%)</td>
<td>0 (0%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
</tr>
<tr>
<td>1995</td>
<td>11 (31%)</td>
<td>9 (26%)</td>
<td>0 (0%)</td>
<td>7 (20%)</td>
<td>2 (6%)</td>
<td>6 (17%)</td>
<td>8 (23%)</td>
</tr>
<tr>
<td>1997</td>
<td>11 (37%)</td>
<td>9 (30%)</td>
<td>1 (3%)</td>
<td>2 (7%)</td>
<td>3 (10%)</td>
<td>4 (13%)</td>
<td>7 (23%)</td>
</tr>
<tr>
<td>1999</td>
<td>19 (48%)</td>
<td>13 (32%)</td>
<td>1 (2%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>6 (15%)</td>
<td>6 (15%)</td>
</tr>
<tr>
<td>2001</td>
<td>13 (30%)</td>
<td>15 (35%)</td>
<td>3 (7%)</td>
<td>4 (9%)</td>
<td>0 (0%)</td>
<td>8 (19%)</td>
<td>8 (19%)</td>
</tr>
<tr>
<td>2003</td>
<td>9 (20%)</td>
<td>20 (43%)</td>
<td>4 (9%)</td>
<td>2 (4%)</td>
<td>0 (0%)</td>
<td>11 (24%)</td>
<td>11 (24%)</td>
</tr>
<tr>
<td>2005</td>
<td>11 (24%)</td>
<td>14 (31%)</td>
<td>6 (13%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>13 (29%)</td>
<td>13 (29%)</td>
</tr>
<tr>
<td>2007</td>
<td>12 (18%)</td>
<td>31 (47%)</td>
<td>2 (3%)</td>
<td>4 (6%)</td>
<td>1 (2%)</td>
<td>16 (24%)</td>
<td>17 (26%)</td>
</tr>
<tr>
<td>2009</td>
<td>9 (13%)</td>
<td>34 (49%)</td>
<td>3 (4%)</td>
<td>5 (7%)</td>
<td>6 (9%)</td>
<td>12 (17%)</td>
<td>18 (26%)</td>
</tr>
<tr>
<td>2011</td>
<td>13 (16%)</td>
<td>29 (36%)</td>
<td>8 (10%)</td>
<td>0 (0%)</td>
<td>7 (9%)</td>
<td>23 (29%)</td>
<td>30 (38%)</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>131 (25%)</strong></td>
<td><strong>194 (38%)</strong></td>
<td><strong>33 (6%)</strong></td>
<td><strong>37 (7%)</strong></td>
<td><strong>20 (4%)</strong></td>
<td><strong>101 (19%)</strong></td>
<td><strong>121 (23%)</strong></td>
</tr>
</tbody>
</table>

Data are displayed as n (%); the percentages are calculated by rows.

In all panels, the line indicates the predicted probability from the logistic regression, the gray area shows the 95% CI of this estimate, and the red dots indicate the actual proportions from the data.

**Figure 1.** The Effect of Article Age on Four Obstacles to Receiving Data from the Authors (A) Predicted probability that the paper had at least one apparently working e-mail. (B) Predicted probability of receiving a response, given that at least one e-mail was apparently working. (C) Predicted probability of receiving a response giving the status of the data, given that we received a response. (D) Predicted probability that the data were extant (either ''shared'' or ''exist but unwilling to share'') given that we received a useful response.
We found a strong effect of article age on the availability of data from these 516 studies. The decline in data availability could arise because the authors of older papers were less likely to respond, but this was not supported by the data. Instead, researchers were equally likely to respond (Figure 1B) and to indicate the status of their data (Figure 1C) across the entire range of article ages.

The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year (Figure 1D). There was a continuum of author responses between the data being reported lost and being stored on inaccessible media, and they seemed to vary with the amount of time and effort involved in retrieving the data. Responses included authors being sure that the data were lost (e.g., on a stolen computer) or thinking that they might be stored in some distant location (e.g., their parent's attic) to authors having some degree of certainty that the data are on a Zip or floppy disk in their possession but no longer having the appropriate hardware to access it. In the latter two cases, the authors would have to devote hours or days to retrieving the data. Our reason for needing the data (a reproducibility study) was not especially compelling for authors, and we may have received more of these inaccessible data sets if we had offered authorship on the subsequent paper or said that the data were needed for an important medical or conservation project.

The odds that we were able to find an apparently working e-mail address (either in the paper or by searching online) for any of the contacted authors did decrease by about 7% per year. This decrease was partly driven by a dearth of e-mail addresses in articles published before 2000 (0.38 per paper on average for 1991–1999) compared with those

<table>
<thead>
<tr>
<th>Year</th>
<th>No Working E-Mail</th>
<th>No Response to E-Mail</th>
<th>Response Did Not Give Status of Data</th>
<th>Data Lost</th>
<th>Data Exist, Unwilling to Share</th>
<th>Data Received</th>
<th>Data Extant (Unwilling to Share + Received)</th>
<th>Number of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>9 (35%)</td>
<td>9 (35%)</td>
<td>2 (8%)</td>
<td>4 (15%)</td>
<td>1 (4%)</td>
<td>1 (4%)</td>
<td>2 (8%)</td>
<td>26</td>
</tr>
<tr>
<td>1993</td>
<td>14 (39%)</td>
<td>11 (31%)</td>
<td>3 (8%)</td>
<td>7 (19%)</td>
<td>0 (0%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
<td>36</td>
</tr>
<tr>
<td>1995</td>
<td>11 (31%)</td>
<td>9 (26%)</td>
<td>0 (0%)</td>
<td>7 (20%)</td>
<td>2 (6%)</td>
<td>6 (17%)</td>
<td>8 (23%)</td>
<td>35</td>
</tr>
<tr>
<td>1997</td>
<td>11 (37%)</td>
<td>9 (30%)</td>
<td>1 (3%)</td>
<td>2 (7%)</td>
<td>3 (10%)</td>
<td>4 (13%)</td>
<td>7 (23%)</td>
<td>30</td>
</tr>
<tr>
<td>1999</td>
<td>19 (48%)</td>
<td>13 (32%)</td>
<td>1 (2%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>6 (15%)</td>
<td>6 (15%)</td>
<td>40</td>
</tr>
<tr>
<td>2001</td>
<td>13 (30%)</td>
<td>15 (35%)</td>
<td>3 (7%)</td>
<td>4 (9%)</td>
<td>0 (0%)</td>
<td>8 (19%)</td>
<td>8 (19%)</td>
<td>43</td>
</tr>
<tr>
<td>2003</td>
<td>9 (20%)</td>
<td>20 (43%)</td>
<td>4 (9%)</td>
<td>2 (4%)</td>
<td>0 (0%)</td>
<td>11 (24%)</td>
<td>11 (24%)</td>
<td>46</td>
</tr>
<tr>
<td>2005</td>
<td>11 (24%)</td>
<td>14 (31%)</td>
<td>6 (13%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>13 (29%)</td>
<td>13 (29%)</td>
<td>45</td>
</tr>
<tr>
<td>2007</td>
<td>12 (18%)</td>
<td>31 (47%)</td>
<td>2 (3%)</td>
<td>4 (6%)</td>
<td>1 (2%)</td>
<td>16 (24%)</td>
<td>17 (26%)</td>
<td>66</td>
</tr>
<tr>
<td>2009</td>
<td>9 (13%)</td>
<td>34 (49%)</td>
<td>3 (4%)</td>
<td>5 (7%)</td>
<td>6 (9%)</td>
<td>12 (17%)</td>
<td>18 (26%)</td>
<td>69</td>
</tr>
<tr>
<td>2011</td>
<td>13 (16%)</td>
<td>29 (36%)</td>
<td>8 (10%)</td>
<td>0 (0%)</td>
<td>7 (9%)</td>
<td>23 (29%)</td>
<td>30 (38%)</td>
<td>80</td>
</tr>
<tr>
<td>Totals</td>
<td>131 (25%)</td>
<td>194 (38%)</td>
<td>33 (6%)</td>
<td>37 (7%)</td>
<td>20 (4%)</td>
<td>101 (19%)</td>
<td>121 (23%)</td>
<td>516</td>
</tr>
</tbody>
</table>

Data are displayed as n (%); the percentages are calculated by rows.
Lots of Data is Shared...
Genome Sequence and Structure Data

...but how much isn't shared?

- What isn't shared?
- Who isn't sharing?
- Why not?
- How much does it matter?
- What can be done about it?
Why Share Data? Increased Citations

![Box plot showing number of citations in 2004-2005 for articles with and without data sharing. The box plot compares articles with data not shared (n=44) and articles with data shared (n=41). The increase in citations for shared data is visually represented. Source: H. Piwowar, 2013.]
What Factors Impact Sharing?

- funded by NIH?
- size of grant
- sharing plan req’d?
- funded by non-NIH?

- impact factor
- strength of policy
- open access?
- number of microarray studies published

- years since first paper
- # pubs
- # citations
- previously shared?
- previously reused?
- gender

- sector
- size
- impact rank
- country

- humans?
- mice?
- plants?
- cancer?
- clinical trial?
- number of authors
- year

[H. Piwowar, 2013]
Factors

Multivariate nonlinear regressions with interactions

Odds Ratio

Has journal policy
Count of R01 & other NIH grants
Authors prev GEOAE sharing & OA & microarray creation
NO K funding or P funding
Journal impact
Journal policy consequences & long halflife
Institution high citations & collaboration
NOT animals or mice
Institution is government & NOT higher ed
Last author num prev pubs & first year pub
Large NIH grant
Humans & cancer
NO geo reuse + YES high institution output
First author num prev pubs & first year pub

[H. Piwowar, 2013]
Why not data sharing? (self-reported)

- sharing is too much effort
- want student or jr faculty to publish more
- they themselves want to publish more
- cost
- industrial sponsor
- confidentiality
- commercial value of results

[Campbell et al., 2002 via Piwowar, 2013]
Nature data availability and data citations

• Policy as of July 2016
• http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf
The Evolution of Data Citation: From Principles to Implementation

M. Altman and M. Crosas
Data Sharing Policies

• Science:
  - "all data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science"
  - "citations to unpublished data and personal communications cannot be used to support claims in a published paper"

• Often this is only used as reason to retract work when issues arise

• Need:
  - Recognition of data authorship
  - Robust citation practices and infrastructure
There is increasing recognition that researchers are more inclined to share their data when they get credit (Borgman, 2012, p. 1072). The research community has begun to take wider notice of this. Within traditional print publishing, scholarly citation was widely formalized over a century ago. The association for the advancement of scientific computing (AAAS) published the first edition of the manual of style (1917) to provide a consistent method for recording references to archival sources, government documents, and artworks. However, many variations were used for references to archival sources, correspondence, and even digital data archives were established.

The emergence of data citation principles

This is evident in the past two years. A number of e-policies, and of these only 9% deposited the full primary raw data corresponding to the paper online (Alsheikh-Ali, et al 2011). There is increasing recognition that researchers are more inclined to share their data when they get credit (Borgman, 2012, p. 1072). The research community has begun to take wider notice of this. Within traditional print publishing, scholarly citation was widely formalized over a century ago. The association for the advancement of scientific computing (AAAS) published the first edition of the manual of style (1917) to provide a consistent method for recording references to archival sources, government documents, and artworks. However, many variations were used for references to archival sources, correspondence, and even digital data archives were established.

Data citation, which has existed for 40 years in principle, is likely to be through better scholarly recognition of data authorship. Conversely, recent studies also suggest that researchers receive more credit when they share their data (Piwowar & Vision 2013). Projects to improve reliability, reproducibility, and data availability by publishers, funders, professional associations, and organized digital data archives were established (and helped catalyze) the emergence of data citation principles.

The Chronology of Data Citation Principles and Related Systems

<table>
<thead>
<tr>
<th>Period</th>
<th>Exemplar Systems</th>
<th>Core Principles</th>
<th>Key Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>1977-1998</td>
<td>ICPA Archive Data</td>
<td>- Facilitate description &amp; information retrieval</td>
<td>[Avram 1975]</td>
</tr>
<tr>
<td></td>
<td>MARC catalog systems</td>
<td>- Describe data in archives</td>
<td>[Dodd 1979]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Describe as works not media</td>
<td>[ISBD 1990]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Provide author, title, version.</td>
<td>[ISO 1997]</td>
</tr>
<tr>
<td>1999-2003</td>
<td>NESSTAR Virtual Data Center</td>
<td>- Facilitate access &amp; persistence</td>
<td>[Altman, et al. 2001]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Cite research data in all publications that use it</td>
<td>[Ryssev &amp; Musgrave 2001]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Provide bit- or semantic- fixity</td>
<td>[Buneman 2006]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Provide granularity</td>
<td>[Altman &amp; King 2007]</td>
</tr>
<tr>
<td>2009-</td>
<td>Dataverse Network DataCite Data Dryad</td>
<td>- Facilitate integration</td>
<td>[Uhlir (ed.) 2012]</td>
</tr>
<tr>
<td></td>
<td>FigShare Data Citation Index</td>
<td>- Include data citations in standard locations in text</td>
<td>[CODATA 2013]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Index data citations in existing catalogs</td>
<td>[Data Synthesis Group 2014]</td>
</tr>
</tbody>
</table>
Phases of Data Citation (1977-2009)

1. Support description and information retrieval: what should be included in a citation? (Libraries)

2. Support data access and persistence: if citations to data in publications, need methods to discover information about data

3. Support verification and **reproducibility**: allow verification of claims based on the data (wider integration into publishing)
Joint Declaration of Data Citation Principles

1. **Importance.** Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

2. **Credit and Attribution.** Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

3. **Evidence.** In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

4. **Unique Identification.** A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
Joint Declaration of Data Citation Principles

5. **Access.** Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

6. **Persistence.** Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.
Joint Declaration of Data Citation Principles

7. **Specificity and Verifiability.** Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

8. **Interoperability and flexibility.** Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.
Generic Data Citation

- Author(s), Year, Dataset Title, Global Persistent Identifier, Data Repository or Archive, version or subset
- Authors, repository → Principle 2
- Year and title → not related to principle but consistent with other citations
- Global Persistent Identifier: Principle 4 and 6
More Information

• Provide via the web
  - Metadata
  - Fixity and provenance information
• Community Indices:
  - CrossRef
  - DataCite
• Structured Identifiers (ORCID, ISNI) preferred over unstructured metadata
Example Repositories with Citations

- Dryad, Dataverse, Figshare
- Dataverse:
  - Draft citation automatically generated
  - Includes versioning information
Remaining Challenges

- Provenance: chain of ownership
- Identity: equivalence and derivation relationships
  - Equivalence: if not bitwise equal, can data still be interchangeable?
  - Versioning: if data is updated, how to find updated version?
  - Granularity: How to describe subsets of data (deep citation)
- Attribution: ensure that the correct people and institutions receive credit
DataCite

www.datacite.org
Why Data Citation is a Computational Problem

P. Buneman, S. Davidson, and J. Frew
Computational Data Citation

• Given a database D and a query Q, generate an appropriate citation.
• Automatic Citation requires the answers to two questions:
  - Does the citation depend on both Q and D or just on the data Q(D) extracted by Q from D?
  - If we have appropriate citations for some queries, can we use them to construct citations for other queries?
• If the data is an image or numbers, cannot expect the citation to live in that data
• If the query returns an empty dataset, we still may wish to cite that
• People know how to cite certain parts of a dataset but not all…

[Buneman et al., 2016]
Computational Data Citation (GtoPdb)

Figure 1. GtoPdb family and introductory pages with independent citations.

Figure 2. The MODIS grid, with highlighted tiles (red) of spatial extent for California (green), with citation.

Computational Data Citation (GtoPdb)

Computational Data Citation (MODIS)


[Figure 1. GtoPdb family and introductory pages with independent citations.]

[Figure 2. The MODIS grid, with highlighted tiles (red) of spatial extent for California (green), with citation.]
Views and Citable Units

• Views describe "areas of responsibility" for parts of a database
• Use views to create "citable units"
• Determine which view V answers a particular query Q and generate a citation for the view
• What happens if two different views can answer the same query?
Citable Views and Partial Citations

![Diagram of hierarchical citation structure]

- **Families**
- **Introduction**
- **Targets**

**UML Diagram**

- **Root** node
- **Families** node
- **Introduction** node
- **Targets** node

**Citations**

- URI: .../family/1234
  Collaborators: Harmar, Sharman, Miller
- URI: .../intro/987
  Contributors: Miller, Drucker
- URI: .../target/1234
  Contributors: Miller, Drucker, Salvatori

**Figure 3.** The GtoPdb hierarchy showing the citable views and some partial citations.

**Figure 4.** A simple rule for generating a citation, together with the rule that produces the required citation (the head of the rule).
Hierarchies of Views

• In GtoPdb, three classes of views
• Family view:
  - /Root/Family[FamilyName=$$f]$$f$$f
• Introduction view:
  - /Root/Family[FamilyName=$$f]/ Introduction
• Target view:
  - /Root/Family[FamilyName=$$f]/ Target[TargetName=$$t]
Citation Rule and Partial Result (GtoPdb)

• Rule:
    /Root[VersionNumber: $v]/Family[FamilyName: $$f]/Introduction[Contributor-list: $a]
• Citation:

[Buneman et al., 2016]
Citation Rule and Sample Result (MODIS)

• { author: m_auth($p,$$v), m_year:($p,$$v), title: m_title($p), version: $v, bounding-box : [$$minlong, $$minlat, $$maxlong, $$maxlat], interval: [$$mint, $$maxt], organization: m_org($p), url: m_url($p), accessed: DATE(), doi = m_doi($p,$$v) } 

←
/root/product[ProdName=$p]/version[vnum=$$v]
/file[Lat ≥ $$minlat and Lat ≤ $$maxlat and Lon ≥ $$minlon and Lon ≤ $$maxlon and Time ≥ $$mint and Time ≤ $$maxt]


[Buneman et al., 2016]