Data handling

ARZU ÇÖLTEKİN
ARZU@GEO.UZH.CH
What were the best and worst examples of data handling in your research career?

What are some challenges you face relating to (private) data storage (e.g. large datasets, degradation of samples...) which could impede the reproducibility of your research?

Have you ever shared datasets publicly, for instance alongside a publication? Why or why not (e.g. too large, privacy restrictions...)?

What kind of collaborative tools have you used and what is your experience with them?

What would be needed from your point of view to use such tools more efficiently (or at all)?

Do you think your research would become better by using such tools?

Does the research system encourage and enable (i.e. provide time) the use of such tools?
Best: Astronomers at Harvard

- [http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003542](http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003542)
- [http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0104798](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0104798)

**How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers**

Alberto Pepe, Alyssa Goodman, August Muench, Merce Crosas, Christopher Erdmann

Published: August 28, 2014  •  [http://dx.doi.org/10.1371/journal.pone.0104798](http://dx.doi.org/10.1371/journal.pone.0104798)
What were the best and worst examples of data handling in your research career?

What are some challenges you face relating to (private) data storage (e.g. large datasets, degradation of samples...) which could impede the reproducibility of your research?

Have you ever shared datasets publicly, for instance alongside a publication? Why or why not (e.g. too large, privacy restrictions...)?

What kind of collaborative tools have you used and what is your experience with them?

What would be needed from your point of view to use such tools more efficiently (or at all)?

Do you think your research would become better by using such tools?

Does the research system encourage and enable (i.e. provide time) the use of such tools?
- **Large data sets**
  - 120 *data points* per second with eye tracking data, i.e. 7 200 per minute, 432 000 per hour
  - + Lots of video

- The necessary effort / lack of time to properly annotate the collected data in relation to the experiment design

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timestamp</td>
<td>DateTimeStamp, DateTimestamp, StampStartOffset</td>
</tr>
<tr>
<td>GazePointYLeft</td>
<td>CamXLeft, PupilLeft, CamYLeft</td>
</tr>
<tr>
<td>DistanceLeft</td>
<td>ValidityLeft, CamYRight</td>
</tr>
<tr>
<td>GazePointYRight</td>
<td>GazePointXRight, DistanceRight</td>
</tr>
<tr>
<td>PupilRight</td>
<td>FixationIndex, GazePointX</td>
</tr>
<tr>
<td>Event</td>
<td>EventKey, Data1</td>
</tr>
<tr>
<td>Data2</td>
<td>Descriptor, StimuliName</td>
</tr>
<tr>
<td>MediaWidth</td>
<td>MediaHeight, MediaPosX</td>
</tr>
<tr>
<td>MappedFixationPointX</td>
<td>MediaPosY, AoiIds</td>
</tr>
<tr>
<td>AoiNames</td>
<td>MappedGazeDataPointX</td>
</tr>
<tr>
<td>WebGroupImage</td>
<td>MappedGazeDataPointY</td>
</tr>
</tbody>
</table>
What were the best and worst examples of data handling in your research career?

What are some challenges you face relating to (private) data storage (e.g. large datasets, degradation of samples...) which could impede the reproducibility of your research?

Have you ever shared datasets publicly, for instance alongside a publication? Why or why not (e.g. too large, privacy restrictions...)?

What kind of collaborative tools have you used and what is your experience with them?

What would be needed from your point of view to use such tools more efficiently (or at all)?

Do you think your research would become better by using such tools?

Does the research system encourage and enable (i.e. provide time) the use of such tools?
- Shared twice: opened data for a data challenge, journal required them

- Complications: the effort (time) involved. Without a proper explanation of the experimental design and what each column in the data means, it is not useful. Then people still ask questions.
What were the best and worst examples of data handling in your research career?

What are some challenges you face relating to (private) data storage (e.g. large datasets, degradation of samples...) which could impede the reproducibility of your research?

Have you ever shared datasets publicly, for instance alongside a publication? Why or why not (e.g. too large, privacy restrictions...)?

What kind of collaborative tools have you used and what is your experience with them?

What would be needed from your point of view to use such tools more efficiently (or at all)?

Do you think your research would become better by using such tools?

Does the research system encourage and enable (i.e. provide time) the use of such tools?
- Standard stuff: MS word track changes, Google docs, Authorea (https://www.authorea.com/), dropbox (has a commenting feature), switch drive, wordpress, ...
What were the best and worst examples of data handling in your research career?

What are some challenges you face relating to (private) data storage (e.g. large datasets, degradation of samples...) which could impede the reproducibility of your research?

Have you ever shared datasets publicly, for instance alongside a publication? Why or why not (e.g. too large, privacy restrictions...)?

What kind of collaborative tools have you used and what is your experience with them?

What would be needed from your point of view to use such tools more efficiently (or at all)?

Do you think your research would become better by using such tools?

Does the research system encourage and enable (i.e. provide time) the use of such tools?
- time

- user testing of the said tools
What were the best and worst examples of data handling in your research career?

What are some challenges you face relating to (private) data storage (e.g. large datasets, degradation of samples...) which could impede the reproducibility of your research?

Have you ever shared datasets publicly, for instance alongside a publication? Why or why not (e.g. too large, privacy restrictions...)?

What kind of collaborative tools have you used and what is your experience with them?

What would be needed from your point of view to use such tools more efficiently (or at all)?

Do you think your research would become better by using such tools?

Does the research system encourage and enable (i.e. provide time) the use of such tools?
A reproducibility crisis?

The headlines were hard to miss: Psychology, they proclaimed.

In August, the journal *Science* published the results of an am coordinated by the nonprofit Center for Open Science. Particip psychology studies that had been published in three prominent journals in 2008, 2009, and 2010 were reanalyzed by new sets of researchers. The results — widely reported in the media — were sobering. Psychology, it seemed, had a credibility problem.

But in many ways, the media missed the point. Brian Nosek, of the Center for Open Science, and his colleagues chose to study something fishy going on in the field. Reproducibility is a concern for many reasons.

- Going back to data that was collected years ago (even months)
- Having to work with other people’s data
- Having to explain your data to another person
- Enabling others to replicate or build on your results

In the case of human subject experiments, tight documentation of all conditions (including subtle ones) — characteristics of participants, time of day, temperature, noise — and many many factors can confound the results.
What were the best and worst examples of data handling in your research career?

What are some challenges you face relating to (private) data storage (e.g. large datasets, degradation of samples...) which could impede the reproducibility of your research?

Have you ever shared datasets publicly, for instance alongside a publication? Why or why not (e.g. too large, privacy restrictions...)?

What kind of collaborative tools have you used and what is your experience with them?

What would be needed from your point of view to use such tools more efficiently (or at all)?

Do you think your research would become better by using such tools?

Does the research system encourage and enable (i.e. provide time) the use of such tools?
- Kind of.

We do not account for such time in grant applications (though ‘annotated literature’ is not unheard of, ‘annotating data’ is hardly there). Some institutions do encourage the publishing of data and/or open source and open content outputs though (e.g., SNF, ZORA).