Call me maybe? It's not crazy! Data collection offices are a good partner in data management

Andrew Sell & Alicia Hofelich Mohr

IASSIST 2015 Pecha Kucha Presentation

Slide 1

This is Andrew Sell, and I am Alicia Hofelich Mohr, and we both work in the College of Liberal Arts Research Support Services here at the University of Minnesota. We are going to talk about how University offices that support research data collection, as ours does, can be good partners in helping data curators understand and fix oddities they may see in the data that comes their way.

Slide 2

Data are usually things of joy, but sometimes datasets can get you down. Lack of variable labels, poor documentation, missing data, and identifiers can make a data curator or re-user just plain sad. But you're not alone. Many of these indiscretions stem from the ways in which the data were collected, and data collection offices can help identify where and why things have gone wrong.

Slide 3

One trigger of data sadness is lack of metadata. Metadata may be the way to a curator's heart, but if it's constantly missing or damaged, it's nothing but heartache. Some of these metadata issues may actually be a result of quirks of the survey tool, especially if using online software.

Slide 4

For example, it is unlikely that anyone would ask a question without value labels (that's just silly) – but sometimes that's the story told by the data. Things like the type of question, structure of the responses, and even how the response options are typed can all affect what the survey software pulls.

Slide 5

And even if value labels DO appear in the data, they may be off. Depending on the tool or how data were extracted, labels may reflect the value the software assigns OR the actual label the respondent sees. This is especially annoying to researchers or re-users who try to average the data or compare to other studies.

Slide 6

Sometimes, the actual variable names cause distress. If you transfer the data to a several formats, Excel, SPSS, and R, may not agree on what the variables are actually called. This is because some survey tools don't validate question numbers, which can lead to repeats, or even ostensibly blank names.

Slide 7

Sometimes it's not the metadata, but the paradata that's lacking. Details might be missing about the procedure - how data were collected, how they were processed. One of the things we think about is how to best capture these internal processes in ways that are transparent to the researcher – for example, R scripts for merging or reshaping, procedural details on how data were downloaded from the collection tools, etc.

Slide 8

Have you ever looked at a dataset and wondered - what's up with all the missing values? There could be many reasons why so much data are missing. Perhaps respondents were feeling lazy but this is often not the case. Here some reasons why we have seen lots of missing data which could explain what happened.

Slide 9

Perhaps it is the question itself or the question format or both. Who doesn't truly not dislike kittens? And a respondent doesn't even have to click on the slider to answer. She may think she answered the question but a data collection tool does not. Upfront validation, clarifying instructions, or reframing the question can reduce this kind of missing data.

Slide 10

Now it may be that not all of the viable answer options to a question were included. Fortunately, ballot developers appreciate this problem and give us an "Other, please specify" option and Mickey Mouse has reliably received votes ever since. If a question is missing answer options, people may choose not to give an answer or perhaps worse, fabricate an answer. A lot of missing values could indicate this type of question design problem.

Slide 11

Perhaps data are missing because they SHOULD be missing, but how do we know for sure? There could have been skip logic in the survey so that respondents didn't need to answer unnecessary questions. Unfortunately, many tools don't do a good job of keeping the skip logic with the data. Sometimes it can seem as if they are trying to hide it! However, it is worth the effort to see if it exists.

Slide 12

Another trigger of data sadness may be opening a dataset, only to see identifiers or private data in the file! Many institutional repositories like ours at Minnesota only take public use data, so any kind of identifying information is bad news. But why are people collecting them? Where do all these identifiers come from?

Slide 13

One reason is that it's hard to pay or credit online survey participants if we don't know who they are. Many studies we run use student subject pools, and it is sometimes a manual process to give credit. Survey tools don't always have good ways to keep these identifiers separate from the data without a lot of effort. People may also give partial credit or performance-based bonuses, which requires identifiers.

Slide 14
Survey tools may also collect extra information about the participant without their knowledge. This information can include IP address, browser/computer information, or even unique user ids/emails. Because this information is often collected unintentionally, the researcher may not even realize it is in their data until they download it.

Slide 15
Finally, if a dataset is part of a longer, multi-part study, identifiers may be needed to merge the data later on. While there are ways to authenticate based on random identifiers, these are only reliable if a computer, rather than the participant, is entering the number.

Slide 16
Now a few more adventures in reuse. To paraphrase an infamous quote by Donald Rumsfeld, “There are the known knowns, the known unknowns and the unknown unknowns.” There are things that aren’t often captured in metadata that may be important to the next researcher who tries to reproduce your study or understand your data. Let’s see if we can convert a few “unknown unknowns” to “known unknowns”.

Slide 17
The world wide web can feel like the Wild West – mobile vs. desktop, Mac vs. Windows. We simply aren’t able to control everything. Here a program attempts to help respondents if they run into trouble. IE must run into a lot of problems, it has two troubleshooting options. And is Chrome infallible or are you just SOL? Could these browser differences systematically impact your data?

Slide 18
Software providers often push changes whether you like it or not. They mean well but at times I think Urkel’s mantra sums it up best. There can be ways to mitigate potential problems due to these changes. However, it’s a good idea to include screenshots of what participants actually saw. We see so many updates with everything we use. However, how many of us stop to read about these “improvements” and “bug fixes”?

Slide 19
And if you try to learn more, what does “Fixes an accelerometer calibration issue” even mean? When in doubt document the version of basically everything. We once found a software bug that invalidated a researcher’s experiment. And be professionally paranoid before clicking the update button! Once a "minor” version update of a tool caused a program we were using to crash and we had to revert to the previous version.

Slide 20
These are just some of the issues we are aware of and have attempted to address as a data collection service. If you are experiencing symptoms of data sadness or know someone who is,

there is hope. Your local data collection service may be able to help. Call us maybe, you’re not crazy!