Qualitative Transparency Deliberations

hosted by the Social Science Research Institute at Duke University on behalf of the APSA Section for Qualitative and Multi-Method Research
https://www.qualtd.net/

Access to datasets - when, how?

https://www.qualtd.net/viewtopic.php?f=24&t=152

Page **1** of **1**

Access to datasets - when, how?

Posted: **Mon Sep 19, 2016 9:23 am**
by **jonastall**
What are appropriate standards for other researchers' access to datasets built on hand-coded material (possibly including source materials and coding decisions)?

When: How long is it appropriate to wait to share a very intensive hand-coded research project, if one is planning multiple publications and possibly a book project? Might a scholar provide a description of the data and coding methods, while withholding the actual data and document collections for 2 years post publication? 3 years? 5 years?

How: Should datasets be freely available online or access contingent on registration and permission? What kind of material should be made available next to the dataset - the full source material, all coding decisions, or less? How should the data be maintained after its publication?

Re: Access to datasets - when, how?

Posted: **Thu Oct 20, 2016 4:24 am**
by **Guest**
[quote="jonastall"]What are appropriate standards for other researchers' access to datasets built on hand-coded material (possibly including source materials and coding decisions)?

When: How long is it appropriate to wait to share a very intensive hand-coded research project, if one is planning multiple publications and possibly a book project? Might a scholar provide a description of the data and coding methods, while withholding the actual data and document collections for 2 years post publication? 3 years? 5 years?

How: Should datasets be freely available online or access contingent on registration and permission? What kind of material should be made available next to the dataset - the full source material, all coding decisions, or less? How should the data be maintained after its publication?[/quote]

There are two purposes of giving access: (1) replication of existing studies and (2) use of data for further studies.

(1) There is a legitimate interest of the scientific community in checking/repolicating published studies as soon as they come out. Whatever is needed for this purpose, should be made available on request: the data used in the study, the methods used to produce the results (do-files or whatever the equvialent protocols are), codebooks, and the sources used. If these are publicly available (such as parliamentary debates or legal texts), they do not need to be provided. If they are not public, the sources need to be made available, too (in anonymous form if appropriate). This does not include the right to use the data beyond replication (unless granted by the authors of the publication and dataset). If journals request transfer of the replication material to a website of the journal, access must be regulated through registration and password protection.
(2) There is a legitimate interest of the researchers in first-use of their hand-collected data and a legitimate interest of the scientific community that data enters the public domain after a certain period. I don't think that researchers should sit on their data for more than 2-3 years after the first publication. From my experience, that is the usual project cycle after which researchers will have produced their most important papers or monographs and move on to new projects. It is tue that this books and papers might not actually be published during this period but other researchers will also need time to check the data, conduct their analyses, and getting their work published. So 2-3 years is still a sufficient head start.

Re: Access to datasets - when, how?

Posted: **Thu Oct 20, 2016 8:05 am**
by **Rocabert**
What are appropriate standards for other researchers' access to datasets built on hand-coded material (possibly including source materials and coding decisions)?

One should not distinguish between manually and automatically generated quantitative datasets in terms of materials needed for replication. However, all data use beyond replication should be subject to approval. In general, researchers should keep records where they justify their coding decisions, and replication materials should include a record of problematic cases so they can be challenged. No access to original material should be provided if this is public, and restricted material should only be provided upon challenging of a coding decision.

When: How long is it appropriate to wait to share a very intensive hand-coded research project, if one is planning multiple publications and possibly a book project? Might a scholar provide a description of the data and coding methods, while withholding the actual data and document collections for 2 years post publication? 3 years? 5 years?

Notwithstanding the material needed for replication, datasets which have been costly to manually code should be hold sufficiently long to ensure that researchers have an advantage in publishing. I believe 2 to 3 years should be enough.

How: Should datasets be freely available online or access contingent on registration and permission? What kind of material should be made available next to the dataset - the full source material, all coding decisions, or less? How should the data be maintained after its publication?

Maintaining data online is in the interests of both the community and the owners. In this regard, publishers should maintain replication data and institutions should encourage publication of datasets online and (ideally) provide cloud resources for its maintenance. However, access should always be contingent on registration and permission, as long as there is an institutional system to challenge a negative decision to access data.

Re: Access to datasets - when, how?

Posted: **Sun Oct 23, 2016 2:51 pm**
by **Guest**
[quote="jonastall"]When: How long is it appropriate to wait to share a very intensive hand-coded research project, if one is planning multiple publications and possibly a book project? Might a scholar provide a description of the data and coding methods, while withholding the actual data and document collections for 2 years post publication? 3 years? 5 years?[/quote]

I agree with the earlier posts in that it is important to distinguish between the release of data for replication purposes and the release of data for new studies that go beyond the original. It is extremely important that data are made available for replication purposes immediately upon publication of the first article/book using these data. However, it should be up to the scholar(s) that gathered the data to decide whether they also make these data available for new studies. In most cases, it will make sense for the scholar(s) to do so as soon as possible, to attract citations to their work. But if they decide not to permit such usage, I don't think they should be forced to do so (e.g. via requirements at the publication stage).

[quote="jonastall"]How: Should datasets be freely available online or access contingent on registration and permission? What kind of material should be made available next to the dataset - the full source material, all coding decisions, or less? How should the data be maintained after its publication?[/quote]

It does not make sense to publish data without a detailed code book: not releasing a code book would not least create problems for the scholars that produced the dataset, as they would be confronted with repeated questions by other scholars who are using their data. In my experience, even with a detailed code book, one can get many requests for clarifications.

There shouldn't be any requirements on researchers on how to "maintain" datasets after their publication: they should be published in a place where they can be accessed in the future, and whatever served to replicate the published findings at the time of submission should also be sufficient in the future. It is important that this move towards data transparency does not make the process of producing data more costly than really necessary, because otherwise we will simply see fewer novel datasets.

Andreas Dür
University of Salzburg

Re: Access to datasets - when, how?

Posted: **Tue Oct 25, 2016 7:01 am**
by **Guest**
[quote="Guest"][quote="jonastall"]When: How long is it appropriate to wait to share a very intensive hand-coded research project, if one is planning multiple publications and possibly a book project? Might a scholar provide a description of the data and coding methods, while withholding the actual data and document collections for 2 years post publication? 3 years? 5 years?[/quote]

I agree with the earlier posts in that it is important to distinguish between the release of data for replication purposes and the release of data for new studies that go beyond the original. It is extremely important that data are made available for replication purposes immediately upon publication of the first article/book using these data. However, it should be up to the scholar(s) that gathered the data to decide whether they also make these data available for new studies. In most cases, it will make sense for the scholar(s) to do so as soon as possible, to attract citations to their work. But if they decide not to permit such usage, I don't think they should be forced to do so (e.g. via requirements at the publication stage).

[quote="jonastall"]How: Should datasets be freely available online or access contingent on registration and permission? What kind of material should be made available next to the dataset - the full source material, all coding decisions, or less? How should the data be maintained after its publication?[/quote]

It does not make sense to publish data without a detailed code book: not releasing a code book would not least create problems for the scholars that produced the dataset, as they would be confronted with repeated questions by other scholars who are using their data. In my experience, even with a detailed code book, one can get many requests for clarifications.

There shouldn't be any requirements on researchers on how to "maintain" datasets after their publication: they should be published in a place where they can be accessed in the future, and whatever served to replicate the published findings at the time of submission should also be sufficient in the future. It is important that this move towards data transparency does not make the process of producing data more costly than really necessary, because otherwise we will simply see fewer novel datasets.

Andreas Dür
University of Salzburg[/quote]

I agree with Andreas Dür (Manfred Elsig, University of Bern), just to add that sometimes younger researchers (doing a PhD), involved in large scale data collection exercises, might need some more time to get their study published in a good journal and they may use the data for 2-3 papers. Important is not too put too much pressure on them to release too quickly.

Re: Access to datasets - when, how?

Posted: **Wed Oct 26, 2016 5:44 am**
by **TobiasLenz**
What are appropriate standards for other researchers' access to datasets built on hand-coded material (possibly including source materials and coding decisions)?

One very important standard is transparency, i.e. it should be clear how researchers arrived at the individual scores of a dataset, before aggregation. This includes transparency about (a) the specific sources used, (b) the questions asked and the categories provided (code book), and (c) coding decisions, i.e. procedures on how to adjudicate grey cases. The most transparent datasets will justify each individual coding decisions, but this might be an unfeasible standard to expect of every dataset. This raises a question about minimal standards vs. an ideal world. The more transparent a dataset is, the better. Ideally, datasets would be transparent in the sense outlined before. Minimally, datasets give sufficient detail on each of the three aforementioned dimensions.

I don't think that there should be a generalised expectation that scholars make accessible the actual source materials of their datasets. This would require a rather extensive online infrastructure -- a website or something similar -- that not all scholars constructing datasets are able to provide. However, it would undoubtedly enhance transparency and constitute a major resource for other scholars.

In this context, I want to caution against treating certain types of sources, e.g. the ones that are "publicly available", different from others, e.g. those that are not publicly available. Public availability is an ambiguous term. Does it mean accessible on the internet? Or does it mean accessible by the public in principle, i.e. available at a library to which any researcher has access? Moreover, we know that the information available on the internet is highly volatile. For example, you may find a specific treaty text online at one moment in

time, but it has disappeared at another. Having a standard that a source was available online "when I last checked" is not sufficient.

When: How long is it appropriate to wait to share a very intensive hand-coded research project, if one is planning multiple publications and possibly a book project? Might a scholar provide a description of the data and coding methods, while withholding the actual data and document collections for 2 years post publication? 3 years? 5 years?

I agree with others that there is a legitimate interest of the researchers in first-use of their hand-collected data and a legitimate interest of the scientific community that data enters the public domain after a certain period. At the latest, this should be when a scholar, or groups of scholars, complete a project and move on to other topics. However, defining an appropriate period in terms of a specific number of years seems difficult. Much depends on the scope of the dataset. In my experience, especially large data gathering efforts tend to evolve over time, which is reflected in the scope and nature of the dataset that results. We all know that research processes are messy, and it is seldom the case that a scholar finalises a dataset completely before starting with some initial analyses. This means that one may publish a paper on the basis of a dataset several years before the dataset is completely finished, and even more years before some of the core, or final, analyses may be published. Therefore, determining a specific time period between the first publication based on a dataset and publicising the dataset itself for general use is too rigid.

Personally, I would weigh the standard of transparency (and thoroughness) of a dataset higher than the interest in its early accessibility.

How: Should datasets be freely available online or access contingent on registration and permission? What kind of material should be made available next to the dataset - the full source material, all coding decisions, or less? How should the data be maintained after its publication?

Once a dataset is public, it is public. This means that scholars should not need to register nor to request permission to use it. This should be the case when a research project terminates (often associated with the publication of a book), and it does not apply to datasets used for (exploratory) analyses published in journals beforehand. Here I would maintain the distinction, as others in this forum do, between dataset access for purposes of replication (immediate) vs. for purposes of own further analyses.

No specific requirements on the maintenance of datasets are needed.

Re: Access to datasets - when, how?

Posted: **Fri Oct 28, 2016 5:40 am**
by **dianapanke**
Q What are appropriate standards for other researchers' access to datasets built on hand-coded material (possibly including source materials and coding decisions)?

In my opinion, a database is valuable if it can be used for replication purposes and/or if some of the variables it entails can be used in other research project. This requires that databases are accompanied by good codebooks. Good codebooks not only list the variables, categories/unit, but also provide information on the coding decisions as well as a list of data sources for publicly accessible materials (but not the materials of non-public sources (e.g. interviews etc. – see blog debate below) etc.).

When: How long is it appropriate to wait to share a very intensive hand-coded research project, if one is planning multiple publications and possibly a book project? Might a scholar provide a description of the data and coding methods, while withholding the actual data and document collections for 2 years post publication? 3 years? 5 years?

I agree with the other posts that one needs to distinguish between publishing databases for replication purposes and publishing the entire database. The former needs to be made publicly accessible directly after publication of the corresponding article, book, chapter etc. and needs to be accompanied by a codebook (that also entails a list of data sources and code decisions).
Whether and if so when to make the entire database publicly available should be up to the researcher (who as others in this blog have already pointed out should have self-interests in doing so).

How: Should datasets be freely available online or access contingent on registration and permission?

In this respect, I would also distinguish between databases for replication purposes and entire database.
For the former I do not think that registration/permission should be necessary, as replicability is a core feature of science and should be made as easy as possible.
For the latter, registration would be acceptable (although I personally don't see much added value in knowing who is working with which data). However, requesting permission (which implies that this can be denied) is in my opinion not acceptable. Once a dataset is published it should not be up to the researcher to decide who has access and who hasn't.

Re: Access to datasets - when, how?

Posted: **Tue Nov 01, 2016 6:14 am**
by **ChristophK**

> *jonastall wrote:*What are appropriate standards for other researchers' access to datasets built on hand-coded material (possibly including source materials and coding decisions)?
>
> Datasets that build the cornerstone of a publication should be made accessible on request in order to ensure replicability. This should also include necessary material (detailed codebook). On request means that the data need not yet be subject to unrestricted public access. Rather data - for replication purposes - are given to a researcher requesting it. This also means that the researcher asking for the data should only use the data for the purpose of replication.
>
> At a certain point (see below) data sets should be made publicly available for broader research purpose. Access should be given to the full data set and the codebook. Access should be given without any particular request or other restriction.
>
> When: How long is it appropriate to wait to share a very intensive hand-coded research project, if one is planning multiple publications and possibly a book project? Might a scholar provide a description of the data and coding methods, while withholding the actual data and document collections for 2 years post publication? 3 years? 5 years?
>
> I find it difficult to define a concrete time period; this very much depends on the specific work plan of the project team. As a general rule, data should be made publicly available as soon as the central research questions guiding a certain project have been analyzed on the basis of the

collected data and these analyses have been published in international outlets (journals / books).

There is a trade-off between data exploitation and data publication - where to define the tipping point should be up to the researchers who have collected the data.

How: Should datasets be freely available online or access contingent on registration and permission? What kind of material should be made available next to the dataset - the full source material, all coding decisions, or less? How should the data be maintained after its publication?

see above: availability for wider research purpose with out restriction, publication of full source material and codebook.
Data maintenance: Ideally, data sets should be published and accessible in a such a way that allows the continuous work (update of data in terms of longer time periods, more countries) etc. by the scientific community. Anyone who builds on a published dataset along these line should do so in a publicly accessible way.

Re: Access to datasets - when, how?

Posted: **Thu Nov 03, 2016 9:41 am**
by **rmitchel**
Prompt #1: "What documentation in terms of source materials and coding decisions is it reasonable for journals to demand? What is it unreasonable to demand?"
I think the core requirement is that the materials provided should provide “relatively clear” breadcrumbs between the source materials and the coding decisions. In my view, this requires having the original “warrant” (or evidence) for the coding, the codebook that identifies how “warrants” are turned into codes, and the final coding. Essentially, a new person should have the evidence in front of them, the rules for converting evidence into codes … so they can do it on their own and see if it matches … the actual coding of the initial researcher.
So, for example, you might have something like this:
a) Warrant (text from treaty): “In the event of a dispute between any two or more Parties concerning the interpretation or application of the Convention, the Parties concerned shall seek a settlement of the dispute through negotiation or any other peaceful means of their own choice.”
b) Source of warrant: The original coder should also say where the original warrant comes from: Text of UNFCCC Agreement at http://unfccc.int/files/essential_backg ... onveng.pdf
c) Codebook rule on whether a treaty contains a dispute settlement clause: “A treaty is considered to contain a dispute settlement clause if it contains a clause that describes what will happen if a dispute arises between or among Parties to the treaty.”
d) Code: “Contains dispute settlement clause” (or simply “yes” or “1”)
Armed with the warrant and the codebook rule, the new researcher should be able to generate the original coder’s coding. That is what I think is reasonable to expect. That requires considerable care on the part of the original researcher but seems what would be the requirement to allow replication and verification of original results.

Prompt #2: "What are appropriate standards for other researchers' access to datasets built on hand-coded material (possibly including source materials and coding decisions)?"

--- I think we should distinguish between 2 types of data: a) replication data, and b) the full dataset from which the replication data is drawn. Think of the distinction between 1) evidence, 2) coding, and 3) analysis.

Replication data should consist of the evidence and coding for each variable in the analysis. Armed with that, another author can check that codings correspond to the evidence and coding manual, and that the analysis of the coded data corresponds to those generated by the author. That allows others to "check the authors work" and seems central to the enterprise. This replication data should be made available simultaneous with publication. The author should create a “hand-offable” dataset of all such evidence needed to evaluate if the evidence supports the claims in the published work. If the author has met that standard, then others should not be allowed to demand more than that, even if they know there is more underlying data.
But, in most cases, authors generate considerably more data while doing research that ends up not being used in any particular paper. Thus, each article may use only X% of the variables and observations (fields and records) from the overall dataset they have created. The author(s) should not need to hand off this additional data until they have published on it or determined that they do not want to do so.

Prompt #3: "How long is it appropriate to wait to share a very intensive hand-coded research project, if one is planning multiple publications and possibly a book project? Might a scholar provide a description of the data and coding methods, while withholding the actual data and document collections for 2 years post publication? 3 years? 5 years?

-- I think there is something to be said for a longer period, say 2-3 years but not 5. The reason is that collecting all the SOURCE information in one place is often the hardest part of the task. With my IEA Database, it now has over 1200 agreement texts and that took literally a decade to establish. Now that it is there, its quite easy to code and manipulate but that is because I have spent way too much time creating the database and not enough time coding and writing articles from it. But there is value to the community of urging people to not hold things too closely.
-- In my experience, my data has been available for years but the database is so complex that relatively few people have gotten the really hard nuggets of knowledge that it contains simply because it takes a while to get to know the database. So, it would be the rare case that the database developer would not have a huge advantage over others in using the database appropriately.
-- There might be value to a rule of “adding database developer as last author” --- I am not sure how this works in the natural sciences but I think that is, at least in part, how things work there. That is, I believe many database developers would be FAR more willing to hand off data if they are offered 3rd or 4th or 5th authorship on articles. And, of course, this would involve the database developer contributing to the article by gathering, manipulating, and interpreting data in ways that facilitate the publication of the article. So, that is something I haven’t seen discussed but might warrant consideration and development, perhaps in conversation with natural scientists to see how they address similar situations.

Prompt #4: "Should datasets be freely available online or access contingent on registration and permission? What kind of material should be made available next to the dataset - the full source material, all coding decisions, or less? How should the data be maintained after its publication?"
-- I think having a central repository or at least a “single meta-database” like ICPSR(?) but more agreed to as a standard expectation that ALL journals require that a meta-database tag be created at ICPSR(?) where each journal might keep its own data but anybody could identify where datasets were via a single meta=repository. If there was one stop shopping for data, that would help a lot.
-- I am not sure whether registration/permission means much or matters much. At the end of the day, I think there needs to be a disclosure / “permission rights” page that people should have to “accept” and those permission rights should be attached in a file that accompanies any database download. Then the person who downloads has the moral obligation to follow through on what is expected (Citation, terms of use, etc). My sense is that generally people want to do the right thing and will if one makes it easy enough for them to do so. Material to include would be the same answer as above: Warrant, Source, Codebook, Codes.

Re: Access to datasets - when, how?

Posted: **Fri Nov 25, 2016 6:41 pm**
by **saranied**
I agree with some of the previous posts in that data for replication should be made available immediately after publication. However, the decision about what data to share should depend on the type of information it contains. For non-sensitive data such as the coding of newspapers, public speeches, or treaties, I think full information including the original sources and coding decisions should in principle be made available. In the case of the coding of interviews conducted by the researcher, particularly when these contain sensitive information about the respondent, coding rules should be clearly explained without necessarily including the complete transcript of the interviews to avoid exposing the respondents.

Re: Access to datasets - when, how?

Posted: **Sun Nov 27, 2016 4:43 pm**
by **pstapleton**
I also agree with Andreas Dur and Sara Niedzwiecki in their earlier posts (below), and would also like to add my concerns about earlier stage researchers and the need to protect their data. Who will be responsible to ensure that shared data will only be used for replication studies until a project is complete? Will it be up to book editors, journal editors, and/or reviewers to check that publications are appropriately using data? will authors of datasets be consulted when their data are used? I could see early stage researchers also feeling pressured to provide permissions to use data for publications that would undermine their own ability to publish from it. I do think that it is important to provide data publicly to replicate studies or to check analyses. But the current tenure system in the United States places earlier stage academics in difficult positions with regard to publications and collaborations. This must be acknowledged in however we develop a system for sharing datasets.

Guest wrote:

> *Guest wrote:*
>
> > *jonastall wrote:*When: How long is it appropriate to wait to share a very intensive hand-coded research project, if one is planning multiple publications and possibly a book project? Might a scholar provide a description of the data and coding methods, while withholding the actual data and document collections for 2 years post publication? 3 years? 5 years?
>
> I agree with the earlier posts in that it is important to distinguish between the release of data for replication purposes and the release of data for new studies that go beyond the original. It is extremely important that data are made available for replication purposes immediately upon publication of the first article/book using these data. However, it should be up to the scholar(s) that gathered the data to decide whether they also make these data available for new studies. In most cases, it will make sense for the scholar(s) to do so as soon as possible, to attract citations to their work. But if they decide not to permit such usage, I don't think they should be forced to do so (e.g. via requirements at the publication stage).

> *jonastall wrote:*How: Should datasets be freely available online or access contingent on registration and permission? What kind of material should be made available next to the dataset - the full source material, all coding decisions, or less? How should the data be maintained after its publication?

It does not make sense to publish data without a detailed code book: not releasing a code book would not least create problems for the scholars that produced the dataset, as they would be confronted with repeated questions by other scholars who are using their data. In my experience, even with a detailed code book, one can get many requests for clarifications.

There shouldn't be any requirements on researchers on how to "maintain" datasets after their publication: they should be published in a place where they can be accessed in the future, and whatever served to replicate the published findings at the time of submission should also be sufficient in the future. It is important that this move towards data transparency does not make the process of producing data more costly than really necessary, because otherwise we will simply see fewer novel datasets.

Andreas Dür
University of Salzburg

> I agree with Andreas Dür (Manfred Elsig, University of Bern), just to add that sometimes younger researchers (doing a PhD), involved in large scale data collection exercises, might need some more time to get their study published in a good journal and they may use the data for 2-3 papers. Important is not too put too much pressure on them to release too quickly.

Re: Access to datasets - when, how?

Posted: **Thu Dec 01, 2016 6:01 am**
by **resodihardjo**
I'm not against the publication of datasets online (one of my blame game datasets (aggregate level) is already available online through a journal in which we published & our university is working on creating a system where, for instance, you can link your codebook or dataset to a publication), but I'm hesitant about making publication of datasets compulsory and within a set deadline as I'm having some concerns/questions.

appropiate standards
It's important (whether you do quantitative or qualitative content coding) that your study can be checked/replicated. This does not necessarily mean putting the dataset online. Instead, it means that you should have a proper codebook/coding protocol where the major decisions are clearly stated so everyone who wants to can replicate the study. Appropriate standards thus refer to good, thorough coding procedures that we all agree on and a publicly available codebook.

Access when?
Access to the codebook should immediately follow upon publication. Access to the datasets,however, is a different story and should not be mandatory for a number of reasons, including
- *time*: not everyone will be able to finish all publications they envision within the 2-3 years currently supported in this forum. I'm not just talking about PhD candidates here. I'm also talking about those working at universities where, for instance, they have to teach 5-6 courses and supervise 30 MA theses a year, making it difficult to find time to work with the datasets they created. All the work they put into creating the dataset will become moot if someone else can start using the dataset before they have completed their research (and accompanying publications). Moreover, there might be personal reasons why you do not have the time to work on a new article once you've published your first one. You might have a kid who is sick, parents who need help etc. As women are often the ones who take on the role as caretaker combined with the fact that women are often (in my country) working part-time when they have kids, they (but also male caretakers, anyone facing a tremendous teaching load and PhD candidates) will be disproportionally affected by a requirement to publish the dataset soon after the first publication.
- copyright issues: if you make it mandatory to publish the dataset, you will make it impossible for some academics to publish their research in your journal as that would infringe on copyright. One of my datasets, for instance, contains text from publications I gathered and then coded. The Dutch copyright law makes it impossible for me to put this dataset online.
- confidentiality issues: same goes for confidential information. You might have signed a confidentiality contract to get access to the data. But that also means you cannot put the dataset online.

How?
Codebooks can be easily published online and everyone has free access to these codebooks. Where is a different matter as not all universities/national government research institutes have the capacity to put this on the internet for eternity.

Additionally, I have some questions
- how long should something stay on the internet?
- will you keep control over the dataset (i.e. you can actually remove it or update it if you want to)?
- who is responsible for it staying on the internet? (people move to different jobs, they die; what happens with all their data?)
- who is responsible for ethically managing the data? (my university is discussing this at the moment and one of the options is that when you leave this job, the department's/faculty's data/reserach manager will decide what to do with the data linked to your publications. This includes data that you as a researcher might have uploaded in the data research management software but have listed as confidential. What happens if this manager hates your guts and puts it online? Or accidentally puts it online?)
- which dataset are we talking about? The raw dataset? The aggregate one?
- what if your codebook is written in a language which is not English? Having to translate it will put an additional burden on an academic interested in investigating non-English sources.

All in all, I see complications when making it mandatory to publish datasets. I therefore propose that publishing the original (i.e. untranslated) codebook does become mandatory, but it is up to the researcher to decide (1) if and when the codebook will be translated and (2) if and when the accompanying dataset will be published online (with the option, of course, of sending it confidentially to reviewers when reviewing the publication or sending it later on to scholars who are interested in the research).

Re: Access to datasets - when, how?

Posted: **Sun Dec 04, 2016 2:57 pm**
by **pstapleton**

I think Sandra makes excellent, salient points that need to be considered seriously before any decisions are made. And, I think she provides a reasonable solution - it should be required for researchers to publish their codebooks, and the decision to publish datasets should be made by the researcher.

> *resodihardjo wrote:*I'm not against the publication of datasets online (one of my blame game datasets (aggregate level) is already available online through a journal in which we published & our university is working on creating a system where, for instance, you can link your codebook or dataset to a publication), but I'm hesitant about making publication of datasets compulsory and within a set deadline as I'm having some concerns/questions.
>
> *appropiate standards*
> It's important (whether you do quantitative or qualitative content coding) that your study can be checked/replicated. This does not necessarily mean putting the dataset online. Instead, it means that you should have a proper codebook/coding protocol where the major decisions are clearly stated so everyone who wants to can replicate the study. Appropriate standards thus refer to good, thorough coding procedures that we all agree on and a publicly available codebook.
>
> *Access when?*
> Access to the codebook should immediately follow upon publication. Access to the datasets,however, is a different story and should not be mandatory for a number of reasons, including
> - *time*: not everyone will be able to finish all publications they envision within the 2-3 years currently supported in this forum. I'm not just talking about PhD candidates here. I'm also talking about those working at universities where, for instance, they have to teach 5-6 courses and supervise 30 MA theses a year, making it difficult to find time to work with the datasets they created. All the work they put into creating the dataset will become moot if someone else can start using the dataset before they have completed their research (and accompanying publications). Moreover, there might be personal reasons why you do not have the time to work on a new article once you've published your first one. You might have a kid who is sick, parents who need help etc. As women are often the ones who take on the role as caretaker combined with the fact that women are often (in my country) working part-time when they have kids, they (but also male caretakers, anyone facing a tremendous teaching load and PhD candidates) will be disproportionally affected by a requirement to publish the dataset soon after the first publication.
> - copyright issues: if you make it mandatory to publish the dataset, you will make it impossible for some academics to publish their research in your journal as that would infringe on copyright. One of my datasets, for instance, contains text from publications I gathered and then coded. The Dutch copyright law makes it impossible for me to put this dataset online.
> - confidentiality issues: same goes for confidential information. You might have signed a confidentiality contract to get access to the data. But that also means you cannot put the dataset online.
>
> *How?*
> Codebooks can be easily published online and everyone has free access to these codebooks. Where is a different matter as not all universities/national government research institutes have the capacity to put this on the internet for eternity.
>
> Additionally, I have some questions
> - how long should something stay on the internet?
> - will you keep control over the dataset (i.e. you can actually remove it or update it if you want to)?

- who is responsible for it staying on the internet? (people move to different jobs, they die; what happens with all their data?)
- who is responsible for ethically managing the data? (my university is discussing this at the moment and one of the options is that when you leave this job, the department's/faculty's data/reserach manager will decide what to do with the data linked to your publications. This includes data that you as a researcher might have uploaded in the data research management software but have listed as confidential. What happens if this manager hates your guts and puts it online? Or accidentally puts it online?)
- which dataset are we talking about? The raw dataset? The aggregate one?
- what if your codebook is written in a language which is not English? Having to translate it will put an additional burden on an academic interested in investigating non-English sources.

All in all, I see complications when making it mandatory to publish datasets. I therefore propose that publishing the original (i.e. untranslated) codebook does become mandatory, but it is up to the researcher to decide (1) if and when the codebook will be translated and (2) if and when the accompanying dataset will be published online (with the option, of course, of sending it confidentially to reviewers when reviewing the publication or sending it later on to scholars who are interested in the research).

Re: Access to datasets - when, how?

Posted: **Sat Mar 11, 2017 8:50 pm**
by **ZachElkins**

I understand the costs of releasing data, in terms of time and increased competition. I'd say that once a significant article or book is published, any embargo is difficult to justify, and may well be counterproductive (in terms of the impact of the work).

Powered by phpBB® Forum Software © phpBB Limited
https://www.phpbb.com/