

Logistics, Challenges, and Benefits of Acquiring Research Datasets in an Academic Library

Beyond the Numbers 2023

Alice Kalinowski

alicehk@stanford.edu



Why should libraries help researchers access data?

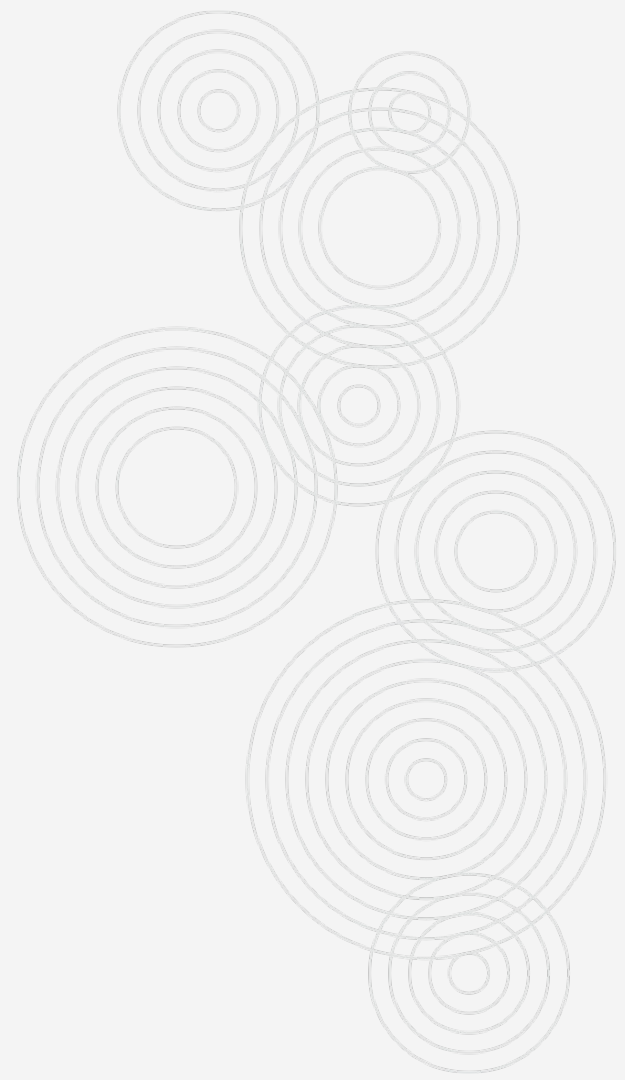
It's a core way you can support the research agenda at your institution:

- Unique and core datasets are often needed to **publish** novel and influential papers
- It is often **financially beneficial** to pool resources
- Researchers often don't know how to negotiate for necessary **license terms**

Cautionary Tale

A department licensed voter data from Catalyst, and the contract had a pre-publication clause a researcher did not follow. The clause required permission from Catalyst to publish a paper, which a researcher failed to get. When Catalyst found out, they made the researcher rescind the paper and it jeopardized the entire agreement. Stanford Libraries' stepped in to renegotiate the deal with Catalyst, removed the pre-publication clause, and mended the relationship.

Library involvement can help prevent these problems. We have the skills working with vendors to negotiate appropriate agreements that work for researchers, that general procurement/general council offices don't always have the experience doing.



Non-Academic Focused Vendors



The options for who can provide data have grown over the last ~10 years.

There are more commercial data providers who are entering the academic market, but also more researchers have been able to convince companies to share their own proprietary data that is created as part of their business operations for research purposes (e.g. a ride-sharing company giving a researcher data on user tipping behavior). For a longer explanation, see [Kalinowski & Hines \(2020\)](#).

We've noticed the following challenges arising from this, which we assume has a lot to do with the vendors making little money off of academic deals and therefore not wanting to spend many resources (aka staff time) on them.

Challenge #1

How to convince a company to give you their data

Researchers can often get high-profile papers published when using novel data sources.

But how do they get access to unique in the first place? And what should people know when trying to do this?

Challenge #1

How to convince a company to give you their data

We find that having a personal connection with the company is helpful.

One of our faculty members had a PhD student who went to work for Company A. The faculty member used that connection to open the door to start negotiating for data from Company A. Company A ended up creating an entire product line just for academics.

We've also used alumni and board member connections to help approach companies and/or help move deals along that needed some extra support.

Challenge #1

How to convince a company to give you their data

Make sure you can explain how sharing their data with you benefits the company.


The reasons may vary, but some benefits can include:

- Increased public exposure when cited in academic papers and associated outputs (e.g. news stories talking about the paper)
- The potential to also sell this data to other academics. Chances are the one researcher who wants the data is not the only one who would want to use it if given the opportunity.

A word of caution - We constantly have to educate vendors on what we can or cannot do when acquiring data to follow university rules and common practices. Big ones for us are that vendors cannot use our trademark/logo on their website, but we've also had some vendors want things like sponsored events, getting faculty to speak at their webinars, etc. that we cannot agree to.

Challenge #2

Negotiating Price

A vertical decorative bar on the right side of the slide, composed of a dark red background with several horizontal segments of varying lengths and colors, including light gray and a lighter shade of red.

Working with vendors new to the academic market can be a challenge when it comes to agreeing on a price that works for both parties.

Challenge #2

Negotiating Price

You want to negotiate the best price you can, without jeopardizing the entire agreement.

The price you agree to with a vendor first entering the academic market will often be what they expect all others to pay. Getting the best price you can also helps anyone who comes after you.

However, you don't want to push back too hard, which could jeopardize the agreement and add time, as that can diminish its benefit if your researchers are among the first to publish.

While datasets can be expensive, if they can result in papers to enhance the research reputation of your school, that may be worth it to your institution.

Negotiation Example

The Economics librarian at Stanford Libraries convinced Gallup that if they lowered the cost of their World Poll microdata, they would get more subscribers and end up with a net positive. Before this, only the World Bank subscribed at a very very high cost. Now that the cost is lower, many schools have gotten access.

The librarian and Gallup presented about this access at Charleston which helped them get more subscribers too.

Challenge #3

Contract Clauses

Each institution will likely have some unique clauses that cause issues. Here are some common things we're seeing that you'll likely want to pay attention to.

See Appendix in [Kalinowski & Hines 2020](#) for more clauses.

Adhesion Contracts

These are ‘take it or leave it’ offers from the vendor; no custom or special agreements. Because each institution has certain things they can/cannot agree to, this can cause deals to fall apart.

Example: An ESG reporting dataset now does this, and two clauses that are causing challenges for us are:

- all liability is on the subscriber (rather than it being shared)
- you have to give them a copy of every paper that used the data (normally we’d want to say we’d use “best efforts” or “reasonable efforts” to comply)

Pre-Publication Clause

This typically means you need to send the paper to the vendor for approval prior to publication . We find the rationale for this is that the vendor wants to protect their image (e.g. avoiding ‘disparaging remarks’).

Example: An unmanned vendor (due to a confidentiality clause!) which required a pre-publication review did not want a sentence in a paper that said their data was ‘less useful for the analysis in the paper than this other source’. To resolve this, the researchers changed it to ‘another distinction between the two sets of data is xyz’.

If you do sign an agreement with this clause, researchers need to be aware of this, and you may need to help them negotiate any issues that arise when the paper is reviewed.

Data Retention Clause

This determines how long (if at all) you can retain access to the data after the agreement ends and/or the researcher leaves. This is particularly important to earlier-career researchers, who need stable access to data to be able to get published.

- We try to ensure that there is a runway period where researchers can use the data for a certain amount of time to conclude their ongoing research (e.g. they should be able to continue to use the data to finish their paper or for the peer-review process).
- Some licenses will contain clauses referring to the destruction of all copies of the data after the agreement ends. From a technical perspective this can be challenging to actually do because of various backups and redundancies in many storage systems.

End-User Requirements

Many licenses have clauses like “all users must do xyz”. Given that we cannot control most of what users do, this is not something we can agree to. We try to modify this language so that instead we will make our “best (or reasonable) efforts to ensure that users do xyz”.

Data Use Agreements (DUA): For most datasets we acquire, we create Data Use Agreements that we require researchers to agree to before we provide them access to the data. The DUA will inform users of their obligations or restrictions for using the data. Depending on how you’re providing access to the data, this can be done in a semi-automatic fashion or will need some manual intervention to be sure the agreement is signed before access is given.

Challenge #4

Clarification & Misunderstandings

A few other challenges we run into when trying to acquire data.

Is it a viable option?

Basic reference fundamentals still apply.

Sometimes there can be pressures to buy data, but it is always worth asking questions as you go to make sure it makes sense. Some additional things to consider:

- Does this data actually include the information the researcher is expecting? (We once helped a researcher buy data where one variable was key, and was only filled out ~25% of the time making the dataset unusable.)
- How long will it take to get the agreement through, and does the researcher have time to wait?
- How much time/money/expertise is it going to take to get the data from the vendor and make it available? Do you have the staffing and storage to do this?
- Remember, just because a researcher asks for certain data, it doesn't necessarily mean they have already done a lot of due diligence on it. You can help with this.

New Academic Vendors

For vendors who typically sell to industry, academic needs are often new to them.

Some common things we emphasize with these vendors:

- **Datafeeds** - vendors often sell datafeeds to corporate clients who want data updated daily/weekly/monthly. We typically prefer ~annual delivery (or just an extract that isn't a feed) to make it more manageable.
- **Long backfiles** - Most corporate clients want the latest and greatest, but less so in academia. Todd Hines always jokes people want data back to the middle ages, which I thought was overly dramatic until we were actually asked for a dataset on people's tipping behavior back to the 1400s.
- **Interfaces vs. datasets** - sometimes when we are first asked about access to X data and we get in contact with a vendor, they show us their analytics platform which doesn't typically allow you to extract much data. When we've then asked for the cost of an extract it is normally \$\$\$\$\$.

Little Vendor Support

Academics often ask different questions than corporate clients, and so reps (or whoever your contact is) don't often know the answers. This, combined with little documentation, often limits the usefulness and/or increases the complexity of working with the data, given all the unknowns the researchers have to contend with.

Almost always, as soon as a researcher starts working with a new dataset that doesn't have extensive documentation, they will start to ask questions about the methodology, certain gaps, how various variables were calculated, etc., that we find challenging to get satisfactory answers to.

Questions? Want to chat?
alicehk@stanford.edu

