Good morning; my name is Jamene Brooks-Kieffer. I’m the Data Services Librarian at the University of Kansas. We’re here this morning to learn more about – and then to discuss – where researchers are storing their confidential data. I’m going to argue that it’s going under their mattresses, figuratively speaking.
Here is our agenda this morning. I’m going to share a little bit about how this talk came about and offer a few statements I hope you’ll find, if not alarming, then at least uncomfortable. I’ll define our terms and then dig in to what the literature says about how and where researchers store their data. Then we’ll pull a page from the SWOT analysis technique (Strengths, Weaknesses, Opportunities, Threats) and look at some imminent threats and opportunities, all the while asking, “Where does the library fit in?”
This talk came about as a result of some inquiry I was doing on our campus to locate available storage options and find out whether any were okay for confidential data. I also talked to a number of faculty in the social sciences – many of whom were working with confidential data and had their own practices for storing it. I wondered about their reluctance to use institutional storage and wondered how widespread this attitude was.

As it turns out, quite widespread. Researchers by and large are storing their data in ways that make the data extremely vulnerable to loss and theft, which puts your institution at risk when these data are confidential. I hope this makes you squirm a bit – it should. You should not be surprised to hear that researchers need help with their data.
Just one more provocation for you. I want to emphasize here that the intersection we’re looking at – of researchers, their data, and storage – is not necessarily where your security office is watching for data breaches, but it is a very big problem sitting right out there in the open.

Let’s look at these three elements more closely.
The researchers we’re talking about can come from any discipline, but the most fruitful sources of confidential data are going to be researchers who study living people – so social sciences, life sciences, and education, as examples. We’re leaving medical research aside today, since the health science disciplines don’t show up much in the literature I reviewed.

We’re also talking about employees, including full- and part-time faculty and non-teaching researchers, and graduate students. Grad students can have real problems with confidential data because they’re still subject to IRB rules and other regulations, but they seldom have access to the same level of storage services that employee researchers do. It’s important to catch grad students now because they’re learning the data management practices that they’ll carry with them throughout their careers.
The data we’re talking about is small – although that’s a judgment call, since confidential audio and video recordings can be quite large and there could be many for one project. The data can be in any format, not just numeric, tabular data. It’s not well-funded, so there’s little ability for researchers to write storage funding into a grant application. Finally, in this presentation we’re calling it confidential or sensitive by whether it contains direct or indirect identifiers – pieces of information that could be used to directly or indirectly identify a research subject – or whether the data is caught by your institution’s data classification policy. The policy will catch things that the identifiers don’t, such as data that are subject to export control regulations or are under a non-disclosure agreement.
For our purposes, we’re contrasting ideal storage situations – these conditions on the screen – with the “under the mattress” options that don’t have the same protections in place. We’re also talking about active, working storage, not archival storage.

Ideally, all researchers would use networked storage that is professionally managed and provisioned for confidential or sensitive data, but in the real world networked storage isn’t always available and people use storage devices inappropriate for the data they’re storing. This fact creeps into the other face of storage, which is...
...storage practices – how researchers interact with available storage that’s either controlled by their institutions or by the researchers themselves. This is where the “under the mattress” metaphor comes into play – it’s not a productive or safe place to store either money or data. Identifying this split between storage facilities and storage practices is really where I started asking questions about why researchers are storing data the way they are – particularly confidential data – and, if they are using sub-optimal practices, is there anything that libraries and IT professionals can do about it?
We’re looking at this three-way intersection of researchers, their data, and storage. As it turns out, the literature is pretty lean when it comes to all three of these variables. As you can see on the screen, there are a lot of recent reports about data management and data services, but very few specifically mention confidential data and only two break down researchers’ storage practices by discipline.

So what we get from these 17 reports is a picture of all kinds of researchers storing all kinds of data using all kinds of practices. We’ll look at what’s available from these reports and then refocus on what we can take away that can help the researchers at our own institutions.
We’re going to take things out of order and talk about data in the literature first, since researchers and their storage practices tend to bleed into each other. Relatively few studies discuss the types of data that researchers worked regularly with, beyond generally surveying the file types and formats that they stored. Two that specifically mention sensitive data are here on the screen – you’ll see that at the University of Iowa the percentage of researchers indicating they work with sensitive data is pretty high and that researchers at the University of Nottingham feel they need more help with sensitive data. Respondents from the University of Houston indicated they were not working with “big science,” and also indicated they planned to store their data indefinitely – which thought may give you heartburn and may indicate an educational opportunity.

<table>
<thead>
<tr>
<th>University of Iowa</th>
<th>University of Nottingham</th>
<th>University of Houston</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over half of respondents indicate they work with confidential or sensitive data. (Averkamp, Gu, &amp; Rogers, 2014, p. 18)</td>
<td>Among top 5 areas needing help: • “Greater data storage capacity” • “Support regarding sensitive data” (Parsons, Grimshaw, &amp; Williamson, 2013, p. 33)</td>
<td>No respondents working on “big science” (p. 396). Most plan on storing data indefinitely (p. 394). (Peters &amp; Dryden, 2011)</td>
</tr>
</tbody>
</table>
Let’s move on to researchers.
Only two studies, from Emory University and the University of Kansas (disclaimer: I was not involved in the research or writing of the article), attempt any in-depth study of researchers’ data practices and needs broken down by discipline and, in the case of the KU study, also by research methodology. So broadly:

- Humanities researchers were most likely to rely on local and third-party cloud storage (Emory)
- Sciences researchers were worried about backup needs, file sizes, and costs (KU), and tend to store more data than other disciplines (Emory)
- (continued next slide)
• Social scientists were most interested by space requirements and privacy concerns (KU), and were the least likely to share their data, likely also due to privacy concerns (Emory)

Overall, across both studies, networked institutional storage – our ideal storage facility – ranked no higher than third in researcher storage practices. These specific findings are supported by other studies in the review, which by and large found that institutionally-sponsored networked storage wasn’t even close to researchers’ first choice for storage and backup.

So even for our focus – social scientists working with confidential data – there seems to be a disconnect, possibly:
• Between what they worry about (security) and how they store their data and/or
• Between what they believe is best for their data and what a neutral party might say is best
So what DO researchers want in their storage?
One word: Easy.
This finding is from the KU study but it’s echoed in other articles I reviewed. Researchers need their primary storage method to be easy, so depending on what networked storage is like at their institutions, they may not always make the best choices with regard to storage security.
This is a list of the storage facilities available to and used by researchers that are mentioned in the literature I reviewed. It’s roughly sorted by the frequency with which researchers reported using the facility or media for data storage – roughly, because this is across many papers and they’re all a little different in the details and remarkably similar in the aggregate. Once again, you’ll see institution-provided networked storage ranking pretty far down on the list. We just saw that in 1 study “easy” is the most influential factor in storage choice, and that’s really coming through here in how this list is sorted.
Here is another view of that list of storage facilities, just to drive the point home that many of the methods researchers seem to prefer are also some of the least secure. The most common storage methods overall appear on your left, with less common methods appearing on your right.

Some common storage practices across the literature include:

- Researchers used a mix of storage media, not just one, but they seldom used networked storage as their primary method for either primary storage or backups. It would be nice to know whether researchers were implementing security practices on their personal storage devices, such as encryption or keeping media under lock and key, but there’s no information about this in the literature, so we just don’t know.

- Most respondents used third party cloud storage (e.g.: Dropbox, Google) for some storage – primarily to access easy sharing functions unavailable through institutional resources.

And some specific practices that tell us a bit more about how researchers interact with networked storage:

- Respondents at the University of Bath reported keeping their working copy on a local machine and using networked storage as the backup (Pink, Cope, Jordan, & Jones, 2013).
• At the University of Iowa, respondents used a variety of non-networked storage to save space on their networked storage quota and/or to get some measure of synced storage across devices (Averkamp, Gu, & Rogers, 2014).
• At Cal Poly, only 28% saw university or department servers as the best data storage space; over half were unsure how to back up to protect data from accidental overwrite or other loss (Scaramozzino, Ramirez, & McGaughey, 2012).

These practices point out some of the problems that researchers have with networked storage. Let’s take a look at some of those.
This slide could also be titled “Why the Mattress?” Researchers’ behaviors here come down to a mismatch between the way networked storage must be used at their institution and the way they need to work.

These were the most common issues reported from most of the studies:

- **Networked storage is, among other things, difficult to use.** Of course, we know that researchers want “easy” but let’s give people credit – there are many burdens on researchers’ time and expecting them to use a storage solution with a steep learning curve isn’t realistic.
- **Insufficient support for sensitive data** – a problem that really gets at what we’re talking about today.
- **Difficult to sync and share files** and collaborate. Cross-departmental and cross-institutional collaboration is increasingly necessary for getting shrinking grant funding, so being unable to share files even with someone from a different department is pretty crippling.
- **Many researchers** we’re focusing on do fieldwork where there is no internet or in countries where commonly-used VPN ports are blocked, so while the idea of a network copy of their notes, pictures, videos, etc. sounds great, it doesn’t work

<table>
<thead>
<tr>
<th>Problems with networked storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Low capacity, high cost, difficult to use</td>
</tr>
<tr>
<td>(Averkamp, Gu, &amp; Rogers, 2013)</td>
</tr>
<tr>
<td>• Insufficient support for sensitive data</td>
</tr>
<tr>
<td>(Parsons, Grimshaw, &amp; Williamson, 2013)</td>
</tr>
<tr>
<td>• Difficult to sync and share files and collaborate across institutions</td>
</tr>
<tr>
<td>(Pirk, Cope, Jordan, &amp; Jones, 2013)</td>
</tr>
<tr>
<td>(Averkamp, Gu, &amp; Rogers, 2014)</td>
</tr>
<tr>
<td>• Incompatible working conditions:</td>
</tr>
<tr>
<td>(Pirk, Cope, Jordan, &amp; Jones, 2013)</td>
</tr>
<tr>
<td>• Off campus with slow/unreliable VPN</td>
</tr>
<tr>
<td>• Using mobile devices</td>
</tr>
<tr>
<td>• In locations with no Internet access</td>
</tr>
</tbody>
</table>
in practice. I’ve heard stories from qualitative researchers who could not, for the safety of their research subjects, risk exposing their interview tapes to customs, so they loaded the recordings onto their phones, took backup pictures of their notes, and hoped for the best.

What I’m trying to get across here is that not using networked storage isn’t necessarily the researcher’s choice – each is dealing with a unique set of working conditions and doing the best they can with limited time. They are also working in the wider world’s computing environment, where services like cloud storage, syncing files across devices, and applications that are compatible with many operating systems are presented as easy and free – so it’s frustrating when the professional environment can’t match that. With that said, I hope we can take this discussion to a place where we’re aware of these challenges and can start to find ways of meeting researchers where they are.
We’ve already seen that many researchers use 3\textsuperscript{rd} party cloud services for some of their storage needs – likely to meet their work requirements for syncing across devices, sharing files, and collaborating with colleagues in other departments or institutions. So these individual accounts are distinct from any storage resources provided by their institutions and, for many, may meet their immediate needs better. And then something like this happens.

Dedoose is a web application for conducting qualitative and mixed-methods research across data types in a secure environment that enables collaboration. It is available for individual researchers, small groups, or as an enterprise solution. It costs money, but the individual accounts are pretty cheap ($12.95/mo for an individual account). Given the services Dedoose offers and the working needs of researchers that we’ve just been looking at, it’s easy to see why a product like this is so attractive – particularly for the social science researchers we’re trying to focus on. It’s cloud storage that comes with the added attraction of software for coding and analyzing the data you’re storing.
In May of this year, Dedoose suffered a catastrophic system failure that resulted in permanent data loss for some of their customers. This article from The Chronicle of Higher Education presents the crash as a lesson in not trusting third parties (aka: the cloud) with your data.

But Jonathan Rochkind’s rebuttal to this article steers the focus away from “the cloud” as the problem. Rather, digital research data is the problem, and storing confidential and sensitive data securely is a job best left to professionals, not to individual researchers. He argues that research data is much more at risk when stored on individuals’ hard drives and other media than when it is stored in the cloud – although it’s essential to evaluate the cloud service you’re employing for this job. So whether the cloud is provided by a third party or by the institution, the point is that research data is better off in a professionally-managed environment.
We know from the literature that researchers need the file syncing across devices and easy sharing with colleagues that are such attractive features of 3rd party cloud services. And we know that social science researchers, particularly, are concerned about the security of their data, so while they may need this flexibility they also need data security and services for complying with IRB and federal regulations. Researchers at the University of Iowa reported that they would prefer to use university-supplied storage services that enabled them to sync and share and still remain in compliance with IRB rules and other regulations (Averkamp, Gu, & Rogers, 2014). Even researchers who are not working under IRB rules need to follow their country’s and institution’s laws and policies. At the University of Bath, over ¼ of that survey’s respondents used 3rd party cloud storage for something, raising concerns about conflicts between the institution’s policies, EU regulations, and the 3rd parties’ terms of service (Pink, Cope, Jordan, & Jones, 2013).

There’s a sweet spot somewhere that gives researchers the data security they need and the working flexibility they also need. Even if there are already services available to researchers that merge these two
needs, how aware are they of the services and are they willing to use them? This brings us to one of the most intractable problems of academia ---
Trust.
Let’s face it – researchers can be very territorial about their data and can also deeply distrust anyone who wants to “help” them with it.

Are you trusted on your campus to provide data storage services that meet the security and flexibility needs we’ve been talking about? How you answer this question depends on the culture of your campus and on your researchers’ personal experiences with IT AND on the anecdotes they hear from colleagues. It only takes one bad experience to break this trust relationship and it’s a heck of a lot of work to get it back.
The literature offers perspectives, but it isn’t particularly good news for institutionally-sponsored storage. One respondent even reported trusting 3rd party cloud storage over university storage because of privacy concerns! (Averkamp, Gu, & Rogers, 2014, p. 11).

Here are some informative quotes [click and read]
It’s a matter of trust

“...a combination of academic secrecy and misplaced concern about the security of sensitive data on central servers ... contributed to the perpetuation of suboptimal data storage practices both for live and legacy data.”

(Pink, Cope, Jordan, & Jones, 2013, p. 35).
So if you walk away from this session with nothing else, walk away with the knowledge that we can offer services all day long but if we don’t work on that last mile problem of building trust and educating researchers, we won’t be helping anyone.
Let’s start to extrapolate a little from the literature we’ve been looking at and see where this current landscape might take us. We’ll start by looking at a set of threats that seem imminent based on what we’ve been talking about.

Jonathan mentioned researchers “fending for themselves” and I do think this is the biggest threat to confidential data. I’m not at all saying that researchers don’t take the security of their data or the privacy of their research subjects very seriously – I’ve talked to them and they do, very much. I am saying that they are busy people and the management of secure data is not what they’ve been trained to do. So they’re relying on practices they’ve developed themselves, things they’ve learned from mentors and colleagues, and the storage options that are available to them at a given point in time. While an individual researcher’s practices may be very good, collectively this is still a dangerous situation for confidential data.

Remember that researchers are still going to work with their data, whether it’s confidential or not. That means likely engaging in all the practices we’ve been talking about – syncing, copying, and (hopefully) backing up their data, and possibly sharing a de-identified subset with one or more colleagues. This leads to multiple storage devices and media – any of which could be a weak point in data security. It
only takes one lost USB drive to compromise secure data. Research teams add another layer of complexity, because you’re talking about multiple people with access to the data (depending on how the project is structured) – and any one of those people could have a weak password or leave their laptop on the bus.
2014 was in some ways the year of the data breach – high profile theft or exposure of financial and personal data affecting millions of people. Home Depot and Target were only 2 of the affected businesses. For your institution, the question really is, do you want to be the next headline?

These two breaches involved billing and credit card information. Not that these weren’t serious, but confidential research data can include personal information that identifies respondents, records their identity keys (SSN), and links them to potentially damaging information – about their health, opinions, or participation in illegal activities, just to name a few. A breach of this kind of data could be very damaging. And we haven’t talked about sensitive data that’s under export control regulations or private non-disclosure agreements – breaches of these can be very damaging in other ways.

The threats are not isolated to the institution – the institution is really the proxy here for the individual researcher and the research subjects who have trusted their information to the researcher. Consequences of course include bad publicity, but also the researcher’s career, the ability of the institution to receive future grant
funding, and the privacy and possibly the safety of individual research subjects.
This, I think, is one of the best quotes in the available literature, because it points out how critical technical support is for the success of current research projects. And we’re not just talking about currently-active research projects. We’re talking about people’s ideas for research that they might do, and might get funding to do – and if they don’t see the technical support available to shore up their ideas, they may never voice those ideas at all – and that’s both very sad and a loss to research and to the academy.
Across multiple studies, one of the things that became abundantly clear was that, in the absence of enough storage space or support for using networked storage, researchers pieced together their own storage strategies using the media they had access to, and only sometimes did this involve networked storage. As you can imagine, the proliferation of files across multiple devices and media makes it impossible to know simple things like: “Which is the most recent version of my data?” or “Where are all the files I would need to document this data set?” You can make a mess really quick. Data that’s lost in this manner is unavailable for future research or even for replicating current results.
I just want to be clear here that it is every researcher’s professional and ethical responsibility to take good care of their data, particularly if it’s confidential data. And institutions are responsible for creating environments where their researchers can succeed, which includes providing adequate data storage services. Laying all of this information out from the literature, it may look as if I’m questioning researchers’ and institutions’ commitments on these two points, but I’m not. I’m saying that both researchers and institutions are doing the best they can, day-to-day, to get their work done, and in all this busyness, the long-term ramifications for what people are doing every day get lost. The literature I’ve been presenting to you clearly show those gaps. In particular, we’ve seen gaps in trust and in awareness. So the next question is, how can we fill them?
For much of the talk I’ve really been talking to the IT people in the room, although these are issues that librarians need to be aware of too. Now we’re going to bring the two together, since you may ask why the library has a stake in this problem? I approached that a little bit in previous slide about threats to future research RE data management and curation – something the library definitely has a stake in. It’s also important to remember libraries’ traditional role as neutral party; it has built a lot of trust with faculty over time (depends on the institution and department, though).

It can serve as a concierge of sorts to help researchers locate services in what can be a VERY confusing campus environment. At the University of Iowa, respondents identified at least 7 different entities offering some data service (Averkamp, Gu, & Rogers, 2014).

University of Houston researchers reported that it was very hard to know who to go to for data services, because “no one really knows who is offering what service and to whom” (Peters & Dryden, p. 397).

Advocacy can be tricky, because we don’t want to be taking sides with one party
over another. What we can do is represent multiple points of view and keep some of those bigger picture issues visible to all parties – What is good for scholarship? What is responsible and ethical? What benefits the institution AND the researcher? Some of these things we advocate for look really good to researchers AND IT professionals, e.g.: using data management techniques, you can find your files and more efficiently use your storage space! Implementing some data curation practices can keep your files readable over the long term and help you replicate or build on previous work!
Here are some opportunities for all involved parties.

IT professionals need to change the conversation, here taking a cue from Jonathan Rochkind’s blog post cited earlier, and talk not about “cloud vs. local” but about “professionally managed vs. individually managed.” This sounds like a marketing trick but it changes the emphasis from WHERE the data is stored to WHO IS RESPONSIBLE for this data’s storage medium. Emphasizing the responsibility element may help researchers adjust their practices.

IT also needs to offer robust services that meet researchers where they are – so accounting for work practices like syncing across devices, sharing with cross-institutional partners, offering sufficient networked storage that’s easy to use, and offering AND advertising tools for secure data.

IT also needs to remember that a decent amount of fieldwork takes place offline and teach best practices for the circumstances. Offering relevant instruction on encrypting a hard drive or replicating data across multiple devices may go a long way toward repairing any trust gaps with a campus’s researchers.
Libraries, in order to be that neutral ground and advocate, need to be knowledgeable about researchers needs and practices AND the available services from IT. Libraries can also offer opportunities to get researchers and IT professionals together in non-threatening spaces/events.
Researchers have responsibilities here, too, primarily to try to let the past go. This is very difficult if someone has lost time, data, or money to something that happened previously, but dwelling on it won’t help now. Researchers need to educate themselves on the services that are available now and take responsibility for applying the useful ones to their own work. Ask yourself, “According to professional ethics, what is the most responsible way to store and treat my data now?”
Note that the title of this picture is “Goodbye Mattress.”
We’ve heard a lot about the current landscape – both what motivates researchers to choose the storage practices they do and how informed they might/not be about the storage facilities that are available to them and the practices needed to keep confidential data safe. We’ve speculated a little on some ways we might start to fill those gaps. Now, in our remaining time, let’s talk about ways we can work together to say, “Goodbye, mattress.”
Works cited & consulted


Works cited & consulted


Works cited & consulted


