

Data in the Transformative Research Library

Kristina M. Spurgin

Library Data Strategist, UNC Chapel Hill Libraries - @kspurgin

2018-10-16

University of Notre Dame Hesburgh Libraries

Introduction

- Who am I?
- Notes and disclaimers
- Presentation + supplementary materials at: https://is.gd/20181016_ndhl

Speaker notes

- Library Data Strategist at UNC Chapel Hill Libraries
- It is impossible to comprehensively cover this topic in 45 minutes.
- I have chosen to cover some areas I feel are particularly important, or where there are developments I believe are particularly interesting
- Though the topics are presented in a more or less linear fashion, they are truly overlapping and inter-related
- This presentation is available online with speaker notes that include details mentioned and citations

Orthogonal theme: Data Literacy

Information that directly impact people's lives is increasingly accessible but civil society is falling behind in making effective use of it. --School of Data

Universities have a responsibility to ensure that today's students learn to engage with and use data to understand the world and inform decision-making.

Research libraries can be key partners in making this happen.

Speaker notes

This is a theme that I wanted to nod to in almost every slide in this presentation, so I wanted to make this statement up front so I can point back to it.

Data literacy is an important aspect of being an educated person today. The ability to understand and grapple with data cannot be left to the tech nerds and 'business intelligence specialists.' The news cycle provides plenty of dystopian examples of what these folks will do with data when there is no good data-informed oversight or critical response.

Whether universities are taking seriously the responsibility to build a data literate populace is out of scope of this talk, but this is a topic that research libraries can champion within their organizations and we are well-located to contribute to this goal.

As the rest of this presentation will show, libraries are already a hub of many of essential data literacy skills. We already partner with faculty to enrich learning experiences with our collections and expertise. We can expand these partnerships to embed data literacy in classrooms across disciplines.

More info/references

- [School of Data](#) - "School of Data is a global network committed to advancing data literacy in civil society. Information that directly impact people's lives is increasingly accessible but civil society is falling behind in making effective use of it. Through our global network of data literacy practitioners and trainers, School of Data seeks to address this data skills gaps in order to amplify the messages of civil society through the use of data. We level the playing field by ensuring that civil society organisations and newsrooms have the knowledge, resources and tools they need to participate fully in the information age... School of Data is a network of data literacy practitioners composed of organisations and individuals. Together, we implement an array of data literacy programmes in our respective countries and regions. Members of School of Data network work to support civil society organizations (CSOs), journalists, and citizens to engage with and use data in their efforts to create better, more equitable and more sustainable societies. Over the past four years, School of Data has succeeded in developing and sustaining a thriving network of data literacy practitioners across Europe, Latin America, Asia and Africa."

Context: The transformative research library

- Libraries are in transformation
- Working in a transformative library transforms you
- A transformative research library transforms the world

Speaker notes

In my research, I found that being a transformative research library is a core part of your library's mission. I did not find a clear outline what that means to you in practice, so I thought about what it means to me in order to scope and bound this talk. What I came up with was:

First, research libraries themselves are in a transformation process. Many libraries are expanding our missions from a focus on collecting and providing access to research outputs, to providing support for the entire research lifecycle. This expanded mission calls for new services (where collections themselves are understood to be services, and we broaden our definition of what collections are). These changes are happening in an environment where many research libraries face ever-decreasing resources. This raises difficult questions about what we will stop doing and stop providing in order to rise to our new missions, given that assuming we can perpetually "do more with less" is unrealistic and a recipe for failure and burnout

In this environment, library staff are required to:

- iteratively rethink why and how work is done, in alignment with overall library mission and goals
- extend their existing skills and expertise into new domains
- gain new skills and expertise
- build relationships with colleagues and partners inside and out of the libraries in new and creative ways

The transformative research library changes the world by:

- playing a core role in producing a data-literate society
- helping researchers find, access, use, and create new knowledge
- creating collections that can be aggregated and used programmatically in novel ways
- knocking down barriers to access

Which doesn't actually narrow it down that much, but oh well...

More info/references For an inspiring vision of the transformative future of libraries, see the [MIT Future of Libraries Task Force Preliminary Report](#). [TODO: zcite]

Context: Data

- Research data
- Collections as data
- Library data
- Patron data
 - (Library data)
 - Vendors and third party applications
 - All the data about an individual

Speaker notes

I'm going to talk about four general categories of data in research libraries today.

(READ CATEGORIES)

I show you this as an overview of what's coming in the rest of this presentation. I'll talk more about how I define each of these as I get to the relevant sections.

Research Data

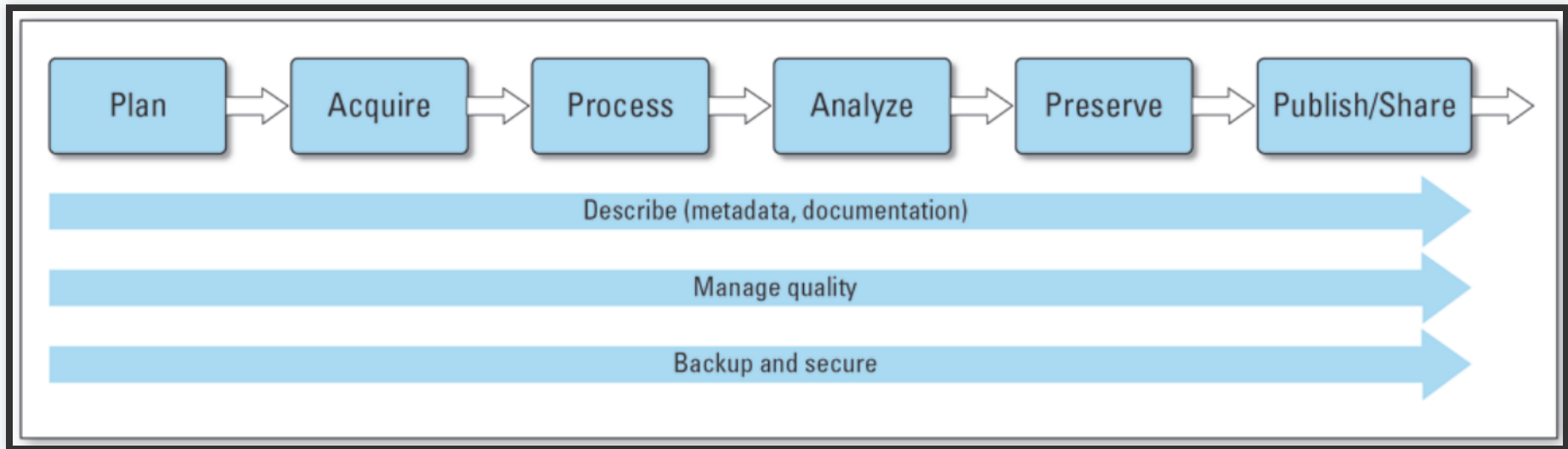


Figure 1: USGS Science Data Lifecycle Model. Boxes indicate the main Model elements, and the shaded arrows below represent cross-cutting elements.¹

Speaker notes

Research data is data in any format gathered, created, and/or used in the process of research. This includes numeric data, textual data, audio and visual data, sensor data, etc.

Overall, this has typically been library-exogenous data, created independent of the library and becoming our concern only when researchers need help finding datasets to use, or a place to store/publish their own research data.

The traditional research library mission was to collect the published/shared products of research. Today, research libraries are expanding their missions to include support for the entire research lifecycle. This future vision includes much deeper library involvement with research data. Let's look at some of the aspects of this:

More info/references

1. Faundeen, John L., Thomas E. Burley, Jennifer A. Carlino, David L. Govoni, Heather S. Henkel, Sally L. Holl, Vivian B. Hutchison, et al. 2014. "The United States Geological Survey Science Data Lifecycle Model." Report 2013–1265. Open-File Report. Reston, VA. USGS Publications Warehouse. <https://doi.org/10.3133/ofr20131265>.

Data management plans (DMPs)

- Required by an increasing number of funders ([src](#))
- Calls for decisions about:
 - metadata
 - organizing data
 - selecting file formats
 - supporting sharing and reuse of data
 - data archiving and preservation
 - rights, licensing, open access considerations

These are not new concepts or skills for libraries!

Speaker notes

A growing number of funders require researchers applying for funding to file a data management plan. Different funders have different requirements.

The library is a place where this expertise already exists. It is a much smaller leap for librarians to extend their existing expertise in these areas to apply to data, than it is for researchers to learn all these skills from scratch.

Library services and tools related to DMP

- [DMPTool](#)
- Online resources and guides ([NYU](#), [MIT](#), [Minnesota](#))
- Workshops, trainings, one-on-one consultations

Speaker notes

I won't go into detail about this stuff because, from your website and workshops calendar, it looks like you already know about these things and can talk to folks in Navari Family Center for Digital Scholarship for details.

Quickly, if you don't know:

- **DMPTool** is an open-source application (created in part by libraries) that researchers can use to create DMPs meeting specific funders' requirements
- I have included some links to a few of the most extensive and linked-to library **Resources/guides** on DMPs (and other research data topics) that I know of

More info/references

- DMPTool's original contributing institutions in 2011 included:
 - California Digital Library
 - UCLA Libray
 - UC San Diego Libraries
 - University of Illinois, Urbana-Champaign Library
 - University of Virginia Library

Managing, processing and analyzing research data

Training and consultation in:

- Data cleaning and remediation
- R, Python, or other languages for manipulating and analyzing data
- Data visualization
- GIS data and mapping
- Corpus linguistics tools and methods + Data mining
- Creating transparent, reproducible research using [Jupyter Notebooks](#) or other tools
- Distributing/sharing and version controlling data ([Dat Project](#))
- Principles of [frictionless data](#)

Speaker notes

Many libraries, this one among them, provide **training and consultation** on a wide range of techniques and tools for working with data:

I see this as an extension of libraries' long history of collaborating with researchers to provide text encoding, analysis, and custom web interfaces to explore digital research projects mainly in the digital humanities.

The Jupyter Notebooks for metadata mapping documentation are publicly viewable on Github. Look for the .ipynb files in [this directory](#) and eventually other sections of the [TRLN data documentation repository](#).

More info/references "The **Jupyter Notebook** is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more." ([src](#))

"**Dat** is a data distribution tool with a version control feature for tracking changes and publishing data sets. It is primarily used for data-driven science, but it can be used to keep track of changes in any data set. As a distributed revision control system it is aimed at speed, simplicity, security, and support for distributed, non-linear workflows." ([src](#))

Frictionless data: " we have learned that there is too much friction in working with data. The frictions we seek to remove—in getting, sharing, and validating data—stop people from truly benefiting from the wealth of data being opened up every day. This kills the cycle of find/improve/share that makes for a dynamic and productive data ecosystem." Focused – Web-oriented – Distributed – Open – Built around existing software – Simple

See also [csv,conf](#)

Data discovery

Our users need data to:

- conduct research
- complete coursework
- meet personal information needs

Libraries are exploring ways to help users to find the data they need

Data catalogs

*A data catalog is an **aggregation of metadata and corresponding links to data**. The catalogs are used to bring together related data that may be hosted in different repositories to make it easier for researchers to find data. Current catalogs range from aggregating research data from an institution to from an entire field. --[National Network of Libraries of Medicine Data Thesaurus](#)*

- [Columbia University Libraries Digital Social Science Center Data Catalog](#)
- [Data Catalog Collaboration Project \(DCCP\)](#) (NYU, UPitt, Duke, UMB, UVA, UNC, Wayne State)

More info/references A relatively new trend on my radar, seeming to be coming primarily out of the Health Sciences, is data catalogs.

(READ DEFINITION)

This is different than the hand-curated catalog lists of available data sets maintained by Columbia University Libraries linked to [here](#).

"The Data Catalog Collaboration Project (DCCP) helps researchers make their own data discoverable, and locate usable biomedical data that is not readily accessible elsewhere online. The DCCP is a collaboration of academic libraries working to highlight institutional biomedical research data using an open source catalog."

"[DCCP] metadata has been mapped to the Data Tag Suite (DATS) developed by NIH bioCADDIE to ensure that it can be indexed in national discovery systems like DataMed."

DCCP is a relatively new project with catalogs still rather small. Process of creating descriptions is labor intensive. At UNC, it has involved conducting interview with each dataset creator.

I have questions about:

- creating more siloes
- sustainability in terms of level of effort
- sustainability in terms of what happens when researcher who has the data leaves an institution
- how to best facilitate access after discovery?

Repositories and data

- Institutional repository seems a natural fit
- And the data is now discoverable, right??
- Disciplinary/subject repository
- Data-specific repositories¹

Speaker notes

Some funders require that research data be made available in an open access repository. ([src](#))

Many research libraries are responsible for their university's insitutional respository (IR).

This would seem a natural place to encourage affiliated researchers to deposit their research data sets, and it looks like Notre Dame allows researchers to do that, which is great.

However, IR design often prioritizes ingest, preservation, and access over discovery functions and user experience. Further, "each individual repository is of limited value for research"² because it's an institution-specific silo.

Even if it works well to store data in the IR, it's a good idea to think about how to improve the discoverability of this data. More on this in a few...

At UNC, we've historically received feedback from some researchers that no one is going to come to UNC's IR to find datasets. It exists outside the disciplinary data ecosystems where such data will be best described, discovered, and used.

Some disciplines have trusted repositories already in place. Also, there are dedicated data repository tools.

There are pros and cons to all of these approaches, but the big takeaway for me here is the importance of metadata and interoperability.

More info/references

1. Dataverse Project. "A Comparative Review of Various Data Repositories." Blog. Dataverse Project Blog, July 25, 2017. <https://dataverse.org/blog/comparative-review-various-data-repositories>.
2. Confederation of Open Access Repositories (COAR). Working Group 2: Repository Interoperability. "The Case for Interoperability for Open Access Repositories," July 2011. <https://www.coar-repositories.org/files/A-Case-for-Interoperability-Final-Version.pdf>.

Responsibilities in larger data discovery ecosystem

If we are building institutional or consortial data repositories or catalogs:

- Support harvesting and aggregation of your metadata
 - OAI-PMH, [ResourceSync](#), or an API that supports metadata harvesting
- Ensure metadata is interoperable
 - Use standard data description schemata ([DDI](#), [ABCD](#), [DATS](#), [etc.](#))
 - Share your metadata application profiles
- Register your collections with appropriate external resources
 - [Registry of Research Data Repositories](#)
 - [DataMed](#)

True interoperability is extremely complex and difficult to achieve. However this slide shows some basic best practices that will get us closer to being able to effectively aggregate research data for discovery.

When these responsibilities are met, it's possible to do cool things like... (next slide)

More info/references "The **Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)** is a low-barrier mechanism for repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest that metadata." ([src](#))

--

"ResourceSync is a self-describing set of capabilities designed to keep content in sync between a provider and consumer of that content. The capabilities of a ResourceSync endpoint can be adapted to meet specific community requirements as it extends the Sitemaps protocol used by Google and other search engines.

The project team has been motivated to leverage ResourceSync as an alternative, or next-generation, approach to harvesting repository metadata by aggregators. ResourceSync is attractive because it utilizes native qualities of the web to solve the problem of keeping web-published resources in sync as inevitable changes occur. Nothing special is required beyond publication of a sitemap and change lists, leveraging timestamps to indicate that changes have been published and when they occurred. We anticipated that it'd be an improvement over the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)." ([src](#))

--

"DDI [Data Documentation Initiative] encourages comprehensive description of data for discovery and analysis and supports effective data sharing. Because DDI is a structured standard, it facilitates machine-actionability and interoperability and it can actually be used to drive systems. Another feature of DDI is its focus on metadata reuse;

“enter once, use often” means you can reuse metadata over the course of the data life cycle to avoid costly duplication of effort." ([src](#))

--

"The **Access to Biological Collections Data (ABCD) Schema** is an evolving comprehensive standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data). The ABCD Schema attempts to be comprehensive and highly structured, supporting data from a wide variety of databases. It is compatible with several existing data standards. Parallel structures exist so that either (or both) atomised data and free-text can be accommodated." ([src](#))

--

"**Data Tag Suite (DATS) model** to support the DataMed data discovery index. DataMed’s goal is to be for data what PubMed has been for the scientific literature. DATS has a core set of elements, which are generic and applicable to any type of dataset, and an extended set that can accommodate more specialized data types. DATS is a platform-independent model also available as an annotated serialization in schema.org, which in turn is widely used by major search engines like Google, Microsoft, Yahoo and Yandex." ([src](#))

"DataMed is a prototype biomedical data search engine. Its goal is to discover data sets across data repositories or data aggregators." ([src](#))

Aggregation of metadata from data repositories in discovery tools

- Triangle Research Libraries Network (TRLN) shared catalog
 - One shared index and [union catalog](#) of Duke, NCCU, NCSU, and UNC holdings
 - Individual institutional catalogs for Duke, UNC, and NCSU
- External feeds of metadata from two data repositories mapped into catalog
 - [UNC Odum Institute Archive Dataverse](#)
 - ~2895 dataset records
 - unrestricted sets only
 - appear for all institutions
 - OAI-PMH harvest
 - [Inter-university Consortium for Political and Social Research \(ICPSR\)](#)
 - ~10,696 study records
 - appear for Duke, NCSU, and UNC only
 - regular data set refresh (.tar file)

Speaker notes

UNC Chapel Hill Libraries is a member Triangle Research Libraries Network (TRLN). A major ongoing TRLN initiative is our consortial shared catalog, which is used by 3 of the 4 institutions as our primary catalog-level discovery tool (as opposed to journal contents/full text search of e-resources level discovery tool such as EDS or Summon).


The shared catalog contains data not only from our respective integrated library systems, but also from selected digital collections, Encoded Archival Description records, enhanced indexable content (for tables of contents and book summaries) from Syndetics Solutions, and other sources.

One of the ways we have increased the discoverability of research data sets across our institutions is by mapping metadata from two external data set repositories into our shared catalog.



More info/references "The Odum Institute Archive Dataverse contains social science data curated and archived by the Odum Institute Data Archive at the University of North Carolina at Chapel Hill. Some key collections include the primary holdings of the Louis Harris Data Center, the National Network of State Polls, and other Southern-focused public opinion data."

ICPSR: "An international consortium of more than 700 academic institutions and research organizations...ICPSR maintains a data archive of more than 500,000 files of research in the social sciences. It hosts 16 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields."

UNC Odum Institute Archive Dataverse record

**Harris 2013 Listening to Mothers III, study no. 42389/42390**


Author: [Harris Interactive, Inc.](#)


Format:  Internet resource;  Statistical Dataset


Online Access: [Open Access resource](#)


Summary: [Childbirth Connection's](#) national Listening to Mothers surveys describe women's childbearing experiences from before pregnancy through the postpartum period, and their views about these matters.


Listening to Mothers III is the third of a series of national surveys that explore women's experiences from before pregnancy thr... ([see more](#))

 Email

 Print

 Text Message

 Export to ...

 delicious

DetailsSubjectsSummary

Authors

- [Harris Interactive, Inc.](#)

Notes

- Citation requirements:

Research publications, news or magazine articles and radio or television broadcasts employing statistical summaries of the Odum Institute data should give an appropriate citation to the Institute as the source of the data.
- Data Disclaimer:

Files are offered "as is" with no warranty or claim of fitness for any purpose. In no event shall the University be liable for any actual, incidental or consequential damages arising from use of these files.
- Citation: Harris Interactive, Inc., 2013, "Harris 2013 Listening to Mothers III, study no. 42389/42390", <http://dx.doi.org/10.15139/S3/11925> UNF:5:oT9l46dcJ/RDqHmpHWMdJg== Odum Institute for Research in Social Science [Distributor] V2 [Version]
- Date of Collection: 2012 - 2013
- Files are offered "as is" with no warranty or claim of fitness for any purpose. In no event shall the University be liable for any actual, incidental or consequential damages arising from use of these files.
- Files are publicly available. Institute staff **are not** available to assist extramural clients with accessing and using these files.
- Research publications, news or magazine articles and radio or television broadcasts employing statistical summaries of the Odum Institute data should give an appropriate citation to the Institute as the source of the data.

DataverseOdumCollectiondoi1015139S311925

Figure 2: UNC Odum Institute Archive Dataverse record appearing in the TRLN shared catalog ([link](#))

Speaker notes

I know you probably cannot see these records well and this presentation is not the place to look at them in detail.

BUT I wanted to show them to you so you can notice that the overall shape of the record is somewhat different because of the differences in the underlying metadata that we map into our own catalog.

ICPSR record in TRLN Shared Catalog

The screenshot displays the ICPSR record for the "National Corrections Reporting Program, 1991-2015: Selected Variables". The record is presented in a structured format with a top header, a main content area, and a detailed summary section.

Header: The title "National Corrections Reporting Program, 1991-2015: Selected Variables" is prominently displayed. To the right of the title are icons for "Add", "Email", "Print", "Text Message", "Export to ..." (with a dropdown arrow), and "delicious".

Main Content Area: This section provides key metadata about the dataset:

- United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics**
- Series:** [National Corrections Reporting Program \(NCRP\) Series](#)
- Format:** [Internet resource:](#) [Statistical Dataset](#)
- Published:** Ann Arbor, MI: Inter-university Consortium for Political and Social Research 2018
- Language:** English
- Online Access:** [Online \(Duke only\)](#), [Online \(NCSU only\)](#), [Online \(UNC only\)](#)

Summary: The summary text states: "The National Corrections Reporting Program (NCRP) compiles offender-level data on admissions and releases from state and federal prisons and post-confinement community supervision. The data are used to monitor the nation's correctional population and address specific policy questions related to recidivism, prisoner reentry, and trends in demographic characteristics of the incarcerated and communit... ([see more](#))".

Details Section: Below the main content, there are three tabs: "Details", "Subjects", and "Summary". The "Details" tab is selected, showing the following information:

- Authors:**
 - United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics
 - [Inter-university Consortium for Political and Social Research](#)
 - United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics
- Series:**
 - [National Corrections Reporting Program \(NCRP\) Series](#)
 - [ICPSR 36862](#)
 - [National Corrections Reporting Program \(NCRP\) Series 36862](#)
- Item Description:**
 - Published: Ann Arbor, MI: Inter-university Consortium for Political and Social Research 2018
 - Edition: ICPSR ed.
- Notes:**
 - Contents: 4 data files
 - Contents: Stata system file(s), rda Data file(s), sav Data file(s), stc Data file(s), tsv Data file(s), Data file(s), and electronic documentation.
 - Geographic Coverage: United States
 - Geographic Unit(s): State
 - Title from ICPSR DDI metadata of 2018-03-02
 - Data Source: administrative records data
 - Data Source: mail questionnaire
 - Time Period: 1991-01-01 to 2015-12-31
 - Universe: All persons admitted to state prison, released from state prison, or in state prison at year end from 1991 to 2015 in the United States.
 - Access Restrictions: AVAILABLE. This study is freely available to the general public.

Figure 3: ICPSR record appearing in the TRLN shared catalog ([link](#))

Collections as data

"Aims to encourage computational use of digitized and born digital collections. By conceiving of, packaging, and making collections available as data, cultural heritage institutions work to expand the set of possible opportunities for engaging with collections."--[Santa Barbara Statement on Collections as Data](#)

Speaker notes

Collections as data is an interesting emerging area focused on how libraries (and other cultural heritage institutions) can transform their collections (or, typically, slices of/selections from the collections) into data that can be used programmatically/computationally by researchers.

I see three main categories of collections as data initiatives and will talk about them each briefly.

More info/references For much more on this topic, see the [Collections as data - projects, initiatives, readings, tools, datasets group Zotero library](#) - "Ongoing collection of projects, readings, initiatives, tools, and datasets that are in some way or another related to collections as data. This group is an open resource, welcoming contributions from anyone who has a resource to share."

Digitizing texts – beyond page images

*"Libraries should move beyond the creation of digital images of original sources. Digital materials should allow scholars to do interesting and amazing things with our unique collections beyond what is possible with their physical incarnation rather than trying to replicate the experience of the original."--
Zarafonetis, Michael, and Sarah M. Horowitz. "Beyond Penn's Treaty."*

What if users could leverage our collections for:

Text mining and analysis - Topic modeling - Network modeling - Machine learning -
Feature and named entity extraction - Other natural language processing tests

To some extent, this is not at all a new idea. Libraries have been engaged in this for a long time. For example, UNC Chapel Hill Libraries' Documenting the American South was transcribing and encoding in TEI/XML slave narratives, other first person narratives, and additional primary source literature in 2004!

What seems new is the scale we are aiming for, some of the newer tools available for getting this work done, and what seems like a changing approach to quality. UNC's DocSouth project was hand-encoded and extremely close attention to quality.

While some of the projects currently described on the Collections as Data site are similar, a number of them also seem to be willing to accept a lot more messiness, with the hope of FIRST getting the data out there; and SECOND accepting an iterative quality improvement process, perhaps leveraging the fact that smart people crunching the data could help identify quality problems and means of repairing or mitigating them.

- MIT - electronic theses and dissertations (this is a use case that arose at UNC just last week!)
- A number of historic newspapers projects
- journals and letters written by Quaker travelers in the late eighteenth and early nineteenth centuries

Very important: leverage and re-negotiate existing resource agreements – researchers want to be able to do things with vendor-provided collections too

More info/references Zarafonetis, Michael, and Sarah M. Horowitz. "Beyond Penn's Treaty." Collections as Data Facets. Accessed October 15, 2018. <https://collectionsasdata.github.io/facet11/>.

Making analog tabular data computationally actionable

UNDER REVIEW

Southern Weather Discovery

EXPLORE PROJECT ▾

Barometer		
Un-corrected Reading	Att. Ther.	Pressure at Sea Level
994,7	+0,3	995,0
1001,4	0	1001,4
1011,1	-0,3	1010,8
1017,0	-0,4	1016,6

TASK

TUTORIAL

Step 2.

Enter the values for the *first* column.

Please make sure you start at the top of the column and end at the bottom.

If there are missing values for any rows, please use a forward slash (/) or asterisk (*).

You can separate the rows with a comma or put them on different lines.

unclear

NEED SOME HELP WITH THIS TASK?

Back

Next →

Figure 4: Interface for transcribing old weather data from ocean voyages via the Southern Weather Discovery project on Zooniverse

Speaker notes

As a spreadsheet nerd and wannabe science nerd, this category is particularly exciting to me. It's also particularly tricky in that it's very hard to OCR this data in the proper tabular format.

The image here is from a Zooniverse project under review, which uses crowdsourcing to transcribe old climate data. Imagine if all the old data in logbooks were made searchable, crunchable! What might we learn?

More info/references Some examples of projects in this category:

- [Hopkins Marine Station CalCOFI hydrobiological survey of Monterey Bay, CA: 1951 - 1974](#)

" Description: In 1951, the Hopkins Marine Station of Stanford University became a partner in the California Cooperative Oceanic Fisheries Investigations (CalCOFI) program in order to collect oceanographic data in and near Monterey Bay. The aim of the program was to conduct joint fisheries-oceanographic cruises that would help researchers understand what contributed to observed fluctuations in the California sardine fishery. Hopkins conducted weekly sampling (more or less) continuously from March 1951 through June 1974. The raw and aggregated data for most of these cruises currently reside in analog form (handwritten data logs, annual reports, etc.) in the library at the Hopkins Marine Station. The dataset includes variables such as temperature, salinity, oxygen, phosphate, silicate, phytoplankton and zooplankton community structure and abundance, meteorological conditions, fish and marine mammal counts, and more. The collection includes forty-four 3-ring or loose-bound notebooks, twenty-two small, bound notebooks, minutes from annual meetings, annual data reports, and other ephemera. The Hopkins CalCOFI collection is large, completely analog, and very heterogeneous. We are in the early phases of planning a curation strategy, but our general objectives for the dataset are to digitize it, add metadata, convert sampling data to actionable formats, and make it all public. "

- [American Philosophical Society Library data](#) - historic prison data – a post office book kept by Benjamin Franklin during his tenure as Postmaster of Philadelphia – a record of indentured individuals arriving in Philadelphia during the years of 1771-1773.

Catalogs as data sets

Museums

- [Museum of Modern Art \(MoMA\)](#) - Artists (15,651 records) and Artworks (135,423 records) - CSV and JSON - updated monthly
- [Carnegie Museum of Art Collection Data](#) - data on 28,269 museum objects and 59,031 items in Teenie Harris Archive - CSV and JSON

Libraries

- [University of Pennsylvania Libraries](#) – Open bibliographic records (2 files - created by Penn, derived from other sources – OPENN (high-resolution archival images of manuscripts and cultural heritage material, with machine-readable descriptive and technical metadata.)
- [Harvard Library bibliographic dataset](#) - Over 12 million bibliographic records, many from OCLC and LC
- [Library of Congress Meme Generator and GIPHY data set metadata downloads](#) released last week

Speaker notes

Some museums and libraries are releasing dumps of metadata, on the premise that it might be usable for research and other purposes.

As a former instructor of library cataloging, I greatly appreciate the educational benefits of having such data sets freely available!

Challenges

- Skills
- Scale
- Quality
- Rights
- Ethics

Speaker notes

Work in this area is very exciting to me, and becomes even more so when I think about the opportunities that will be afforded as we develop practical ways for libraries to work with linked data at scale.

However this work has somewhat daunting challenges, including:

- NEED FOR NEW SKILLS
 - text mining
 - creating and preparing corpora
 - database applications
 - data manipulation software or programming languages
 - large-scale file management
 - cloud/distributed computing
- SCALE OF THE DATA FOR LARGE COLLECTIONS (requiring cloud/distributed computing)
- DATA QUALITY
 - There's no better way to find out all the things that are wrong with your data than to try to use it to do something other than its initial intended purpose. The promise of collections as data is marred in large part by the fact that so many libraries have for a long time accepted "good enough" metadata that was only good enough for its use in a traditional library catalog.
 - OCR text quality is very poor to impossible for many older printed materials and handwritten materials
 - Crowdsourcing is one model for solving problems in large datasets. See library projects such as [LC Labs' Beyond Words](#) and [Biodiversity Heritage Library's Science Gossip project](#) using the Zooniverse platform
- UNDERSTANDING RIGHTS ISSUES
 - Under what licenses do you release collections as data? It's interesting to observe the variations in how different libraries release their catalog data:
 - Release only bib records originally created by your institution? ([UMich](#))
 - Or include the whole catalog (including vendor and OCLC records(released under Open Data

Commons ODC-BY)) (Harvard, [Columbia Univ Libraries](#))

- Or split the two into separate files, released under separate licenses? ([UPenn](#))

- ETHICS, ETC.

- Do our best to ensure no unintended consequences/conclusions drawn from data once it can be analyzed at large scale. However this is impossible to truly predict. How do we minimize harm?
- Acknowledge that algorithms are biased and tend to reinforce existing structures and hierarchies of privilege.
- Cultivate awareness of what collections, populations, voices are missing and work to represent them

Lowering barriers to use

"Collections as data stewards aim to lower barriers to use. A range of accessible instructional materials and documentation should be developed to support collections as data use. These materials should be scoped to varying levels of technical expertise. Materials should also be scoped to a range of disciplinary, professional, creative, artistic, and educational contexts. Furthermore the community should be motivated and encouraged to build and share tools and infrastructure to facilitate use of collections as data."--[Santa Barbara Statement on Collections as Data](#)

Speaker notes

And here is that link back to data literacy. Releasing collections as data is how libraries provide the raw materials for people to gain and hone these skills.

Library data

Speaker notes

My daily work centers on a subset of "library data" so I'd claim some decent level of expertise in this category. By "library data," I mean data the library creates, compiles, gathers, or uses in the process of carrying out the work of the library. It includes data from external sources (discovery service knowledge bases, partner institutions, etc) that get used in the library's work.

Everyone working in a library interacts with and/or contributes to library data, but not everyone interacts with it or thinks about it **as** data. When I talk about "library data," it involves doing stuff with that data at scale. Not necessarily "big data" scale, but not manually, one record at a time. For example: when cataloging an online database, I am thinking of a MARC record as a description of that specific resource. This isn't a data-centric view. On the other hand, I am working with bib records as data when I extract from our ILS all the MARC records coded as online databases and analyze the fixed field coding patterns in order to make decisions about transforming the records so they scope properly as databases in our online catalog.

"A surprising takeaway for us has been that one of the primary users of our public data has been the museum itself. Easy access to our own data has enabled internal projects to be built on top of the published data, both because it's in an easy-to-use form, but also because of the permissive license." –Carnegie Museum

<https://collectionsasdata.github.io/facet2/>

Trends I think I see

- more positions requiring data-oriented skills
- more positions with "[meta]data strategy" or "systems strategy" in the title

Speaker notes

Overall, I think there is recognition that we need to be doing more library work at scale. I saw two job postings last week for metadata positions focused on large scale batch metadata work.

I think there is also recognition that our systems and data landscapes and the data and metadata that flow through them are becoming increasingly complex and interconnected. And that there is a need within our organizations to have someone responsible for uncovering and documenting this knowledge.

One example from UNC that catalyzed my thinking on this:

- I knew how the bib records and the EAD data got linked/merged in the public catalog data layer
- Someone else knew how to create and mount the EADs into our existing system
- Someone else was working on a new system to manage our EADs
- Someone else (catalog front end developer) knew the details of how requesting archival collections via Aeon works
- Someone else had implemented a feature so that there was a request button displayed on each rendered EAD
- NO ONE REALIZED that when the storage location of the EADs changed, the request buttons on the rendered EADs would all break because actually all the things I just mentioned were dependent on each other in ways no one taking care of things on the ground would be aware of.

My colleagues at Duke University Libraries, Jaquie Samples and Dennis Christman (along with many others) developed a data flow documentation—essentially a map of dependencies like the one above. Developing a similar document for UNC Libraries is one of my goals.

Some neat projects we have done recently

- leverage HathiFiles to semi-automate HSL weeding project
- support of Materials Review project
- moving toward one extractor to rule them all...
- metadata-first IA digitization->HT ingest workflow

HSL weeding project Our Health Sciences Library was starting a weeding project which included decision-making about a fair amount of older material. Their initial idea was to have student workers search HathiTrust for each title published before a certain year to see if the full text was freely available there. Thankfully they brought this project to us instead. We extracted from our ILS the relevant bibliographic information on the HSL collection being weeded. We compared that data against the [HathiFiles](#). That comparison was used to produce a report of:

- clear matches where the full text of the HSL volume was definitely freely available via HathiTrust
- candidates for manual checking – this was a report of ambiguous matches that a human should look at

Materials Review project Library Data Strategy & Services staff assisted the Materials Review team to develop spreadsheet reports that selectors and faculty could use to make decisions about materials cancellations. Creating these reports involved retrieving data from multiple sources, manipulating it so it could be merged, and outputting a useful final product. Data sources included:

- bibliographic records from ILS
- order records from ILS (cost)
- reports from SerialsSolutions on our tracked holdings, including print and online ISSN for each title (when available), packages of which the title is part, dates of holdings for the title in each package, OA status
- usage statistics

One Extractor project We had an "extract the catalog" script for our existing consortial online catalog. Then, we began developing a new version of that catalog, which required the data to be extracted in a slightly different way. In addition, we needed to extract records from our catalog in a slightly different way in order to prepare metadata for batches of titles to be ingested into HathiTrust. At some point, we also needed to extract our whole catalog—in a slightly different way—so that Google could analyze whether our percentage of unique materials made it worth partnering with us for a large scale digitization project.

The extraction process overall is always the same. What changes is stuff like:

- need to extract one bib record with data from all attached item and holdings records; versus
- need to extract one copy of the bib record for each attached item record, with data from that item record; and
- what fields/subfields data from the attached non-bib records are mapped to
- what field/subfield bib record number is mapped to

Instead of developing a complete extract script for each of these needs, LDSS staff have created an general extraction script that can be configured to output the extracted data in various ways.

Metadata-first IA -> HT digitization to preservation workflow We are on our second Internet Archive (IA) Scribe contract. As part of this contract, IA provides a Table Top Scribe book scanner at our institution, along with Scribe operations staff to do the digitization.

IA is a great resource for bulk digitization and making digitized works accessible, however it is not a true preservation platform. UNC also ingests all Scribe-digitized materials into HathiTrust, which is a preservation platform. UNC librarians also want the URLs for the digitized versions added to the bib records of the print items that were scanned.

LDSS is responsible for adding these URLs back into our catalog records, and the preparation of metadata for HathiTrust ingests.

In our first Scribe contract, the workflow was roughly:

- selectors choose items for digitization
- staff and student workers:
 - assess whether item meets physical requirements for Scribe scanning
 - verify item has not already been digitized by another institution and added to a trusted repository
- If the item passed the previous two tests, the worker then:
 - Found bib record in online catalog or ILS and pasted that into IA Scribe metadata spreadsheet
 - Copy/pasted various metadata elements (author, title, language, publisher, etc) from catalog into IA Scribe metadata spreadsheet
- Items and spreadsheet were passed to Scribe operator who conducted digitization.
 - Part of this process involved IA's software using the bib record number from the spreadsheet to grab that bib record from our catalog via a Z39.50 request – the bib record is stored as MARC-XML in the IA data package ([example](#)). The bib record data is also mapped into data for display in the IA record page for the item ([example](#)).
- Monthly, LDSS pulled report of newly digitized items from IA. Report included bib record number submitted for item, URL of digitized item, and any volume/issue designation submitted for item.
- Quarterly, LDSS used similar report from IA to compile a HathiTrust ingest

A number of metadata problems tended to emerge and complicate or block LDSS work, leading to delays in catalog link addition and HathiTrust ingest. These included:

- Inconsistency in bib record number format submitted at digitization time (with vs. without check digit) leading to extra work to normalize bib numbers. Our processes hinge upon a standard form of bib number.
- IA has basically no metadata quality requirements, while HathiTrust has fairly stringent metadata quality requirements. At the point of submitting ingests, HathiTrust would sometimes reject hundreds of items for issues such as bib record lacking OCLC number or date data, or there being no volume/issue for a serial or multi-volume monograph.
- Missing volume/issue data caused delays and tedious manual work to add URLs to bib records for serials and multi-volume sets. LDSS staff would have to try to figure out which item each digitized object represented, add that item's volume/issue info to the IA data, and add it to the \$3 of the 856 field containing the link to that item in the bib record. We believe users should know from looking at the record where they are going to be taken when they click a link. They shouldn't have to click on multiple links to get to the volume that's of interest.

For these reasons, in planning for our second Scribe contract, I advocated firmly for a metadata-first approach. This centers upon a newly designed spreadsheet for preparing the metadata to be passed along to the Scribe operator with the items to be digitized.

When a worker begins work on a new batch of items to be Scribed, they make a copy of our Scribe spreadsheet template. This spreadsheet is set up to connect to our ILS (via III Sierra's bib and item record APIs).

Now, when a physical item has been selected and has passed the first couple of checks, all the worker has to do is find the matching bib **and item** records in the ILS and paste the bib and item record numbers into the spreadsheet.

The spreadsheet:

- normalizes the record numbers, ensuring they are passed on to Scribe operator without check digit
- pulls in the title, author, language, etc metadata that used to need to be manually copy/pasted (which left room for errors)
- consistently pulls in volume/issue information for every item
- runs several metadata quality checks for issues that will block HathiTrust ingest
- if metadata issues are found, uses the item location code to show where to route the item for metadata updates

One person in our Special Collections Technical Services and one person in our main Technical Services have been assigned as main contacts for IA/HT metadata fixes. When the spreadsheet flags a metadata problem for an item, the worker physically places the item on the designated shelf where the Tech Services folks will check for them. The metadata fixes typically are extremely quick, and then the Tech Services staff route the item back to the worker who brought it to them. Often the item proceeds for digitization with the original batch of which it was part.

The effect of this is:

- metadata issues are taken care of at the point in the process where it makes the most sense—when the physical item is in hand and can be referred to—rather than discovered later in bulk by LDSS, a team who does not usually handle physical materials at all
- metadata issues are fixed before we push metadata out into the larger metadata ecosystem. The metadata in IA and HT is correct from the start, with no extra work after the fact
- No manual copy/paste errors and no missing volume/item data

On the table: Data warehousing

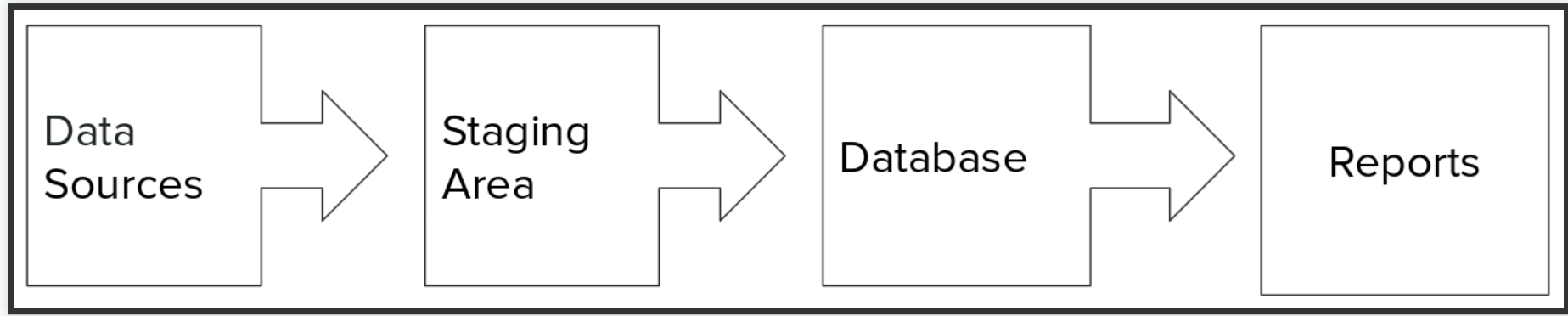


Figure 5: Conceptual flow of data warehousing¹

Data sources could include:

- ILS (bib, order, circ, financials, patron...)
- Discovery knowledge base data
- Usage statistics (of repository, digital collections, vendor-hosted resources)
- E-resource entitlement lists
- Web analytics
- Search logs
- Interlibrary borrowing and document delivery system data
- and more...

As libraries face greater-than-ever resource pressures, we see assessment and analytics as a very convincing way to tell stories about our value.

I am in no way an assessment librarian, and am by nature a critic of overreliance on quantitative measures and anything that suddenly becomes a buzzword, like "metrics." BUT, I do see the value of being able to generate stats, reports, visualizations, and dashboards with minimal friction, and I have intimate understanding of how our complex, siloed systems that generate and store library data create a lot of friction and drag.

One approach to making this work easier is data warehousing. There seems to be a trend in this direction. I keep seeing programs on the topic at Innovative Users Group meetings, and there was a racous breakout session on the topic at Code4Lib 2018.

A couple of years ago I had floated the idea of data warehousing at UNC, primarily to support technical services workflows not supported by our ILS and other tools, such as:

- semi-automated reconciliation of vendor-provided MARC record sets against entitlement lists for e-resource collections
- experiments with leveraging open linked data for authority control work
- ability to run (and schedule to run) more flexible and sophisticated reports than we can within our ILS, with applications like: finding records with invalid MARC

Now that there's movement/interest from the assessment side, it seems this might make it onto our real projects list when certain key positions are finally filled.

1. Yoose, Becky. "Wrangling Library Patron Data." presented at the Privacy in Libraries, a LITA webinar series, April 11, 2018. https://docs.google.com/presentation/d/1_W-3I9CSz6Uu5pFnKsc2USMGA4kOxzx25XiUj_e57bE/edit#slide=id.p.

Patron data

- patron data as library data
- vendor and third party applications collection/use of patron data
- patron data as the patron's individual personal information environment

Speaker notes

Finally, we have patron data, which I break into three subcategories. In part, patron data is a subcategory of library data. This includes the obvious such as borrower status, check-out history (maybe), current items checked out, and fines. But we also have patron data in our search and reference chat logs, web analytics, studies we do of how our spaces are used, etc.

Aside from this, the vendors and third party applications we use to provide content to our patrons certainly also collect data about our patrons. Do we know what they collect and what they do with it? What are our responsibilities here?

Lastly, there's the fact of each of our patrons as individuals with their own personal data—about them, created by them, belonging to them, in or out of their control. What connection do research libraries have to this?

Shout outs

Wrangling Library Patron Data - Becky Yoose, LITA Webinar 2018-04-11

"Ethics in Research Use of Library Patron Data: Glossary and Explainer." Digital Library Federation, Ethics Subgroup, October 2, 2018. <https://doi.org/10.17605/OSF.IO/XFKZ6>.

Salo, Dorothea. "We, Surveilled and Afraid, in a World We Never Made." Speaker Deck, October 11, 2018. <https://speakerdeck.com/dsalo/we-surveilled-and-afraid-in-a-world-we-never-made>.

Keep an eye out for report out of: "National Web Privacy Forum - MSU Library | Montana State University," September 12, 2018.
<https://www.lib.montana.edu/privacy-forum/>.

Speaker notes

I am just touching the surface of this topic and would point you the resources shown here, from which I have heavily cribbed.

Patron data as library data

We protect each library user's right to privacy and confidentiality with respect to information sought or received and resources consulted, borrowed, acquired or transmitted. – ALA Code of Ethics¹

Speaker notes

This presentation takes a little turn toward the dark side here, because I'm afraid we really are falling down on our professional code of ethics here, given the reality of the "surveillance capitalist" world in which we are embedded.

On the 11th of this month Dorothea Salo gave a keynote talk at the Minnesota Library Association Annual Conference in which she pulled no punches pointing out the ways in which libraries are being complicit collaborators in this surveillance panopticon, including:

- serving up insecure (http instead of https) library websites
- having ad trackers (DoubleClick, Ad Nexus) installed in library websites and apps
- using Google analytics, given Google's privacy track record
- Enthusiastically buying in to "learning analytics" that rely on student surveillance (including what e-resources they use and what books they check out) to prove our value within our organizations

There are no easy answers here, though I agree with Salo that we need to remember that "No." can be an answer and a complete sentence. As both Salo and Yoose point out, simply asking "Is this ethical?" and "Why are we doing this?" can create a pause in which we realize we can do better.

More info/references

1. American Library Association. "ALA Code of Ethics," Adopted 1939, last amended January 22, 2008.
<http://www.ala.org/tools/ethics>.

What patron data do we even have?

"Expect any data you collect and store to be used for purposes you didn't intend—and maybe wouldn't approve of."—Dorothea Salo¹

- What data are we collecting?
- Why are we collecting it? Is there an actual solid business need for it?
- Where is this data stored?
- Who has access to this data? Audit regularly!

Speaker notes

1. Salo, Dorothea. "We, Surveilled and Afraid, in a World We Never Made." Speaker Deck, October 11, 2018.
<https://speakerdeck.com/dsalo/we-surveilled-and-afraid-in-a-world-we-never-made>.

Data warehousing, again

- Extract -> Transform -> Load (ETL)
- Transform is a magical patron data protecting step.
- (But not **that** magical)

Speaker notes

In transformation step you can:

- remove data fields altogether
- obfuscate data
- aggregate data
- de-identify data

But don't be too confident:

- De-identification methods do not provide adequate privacy protection for outliers in a service population, or a small overall service population or subset. Nor does it protect against identifiable patterns in the data (e.g. AOL search logs used to reconstruct specific identities belonging to distinct persons) – or what identifiable data may emerge if your data set is matched up with another data set(eg NYC Taxicab data set + images from Google image search + other external data = identifying individual taxi passengers) (Yoose, Wrangling..., slide 27, 38)

What our vendors and third party applications do with our patrons' data

- Start on this early with each new agreement
- If you haven't been on it from the start, consider working to add addendums to existing contracts/licenses, that address:
 - basic data standards we expect to be followed (HIPPA, COPPA, ALA Library Bill of Rights, etc.)
 - expected data disclosure and confidentiality practices
 - vendor liability for data breaches/leaks
- AND really thinking ahead:
 - Can we take our data (and our patrons' data) with us if we move to a different product
 - Is the system even able to truly delete your data?

Speaker notes

Again, this is one of those "Do we even have any idea?" questions. And it's a long haul and a lot of work, but it is the right thing to do.

Helping our patrons manage their own data and privacy

Instruction and workshops to think about if we're not doing them already:

- [Surveillance Self-Defense \(EFF\)](#)
- [Data Detox Kit](#) from [Tactical Technology Collective](#)

Speaker notes

This again brings it back to that orthogonal topic of helping our users become more deeply data literate, which is where I will stop talking and see what you have to say...

Thank you



Figure 6: Any questions?

Presentation + supplementary materials: https://is.gd/20181016_ndhl