

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

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.

Context: The transformative research library

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

Context: Data

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

Research Data

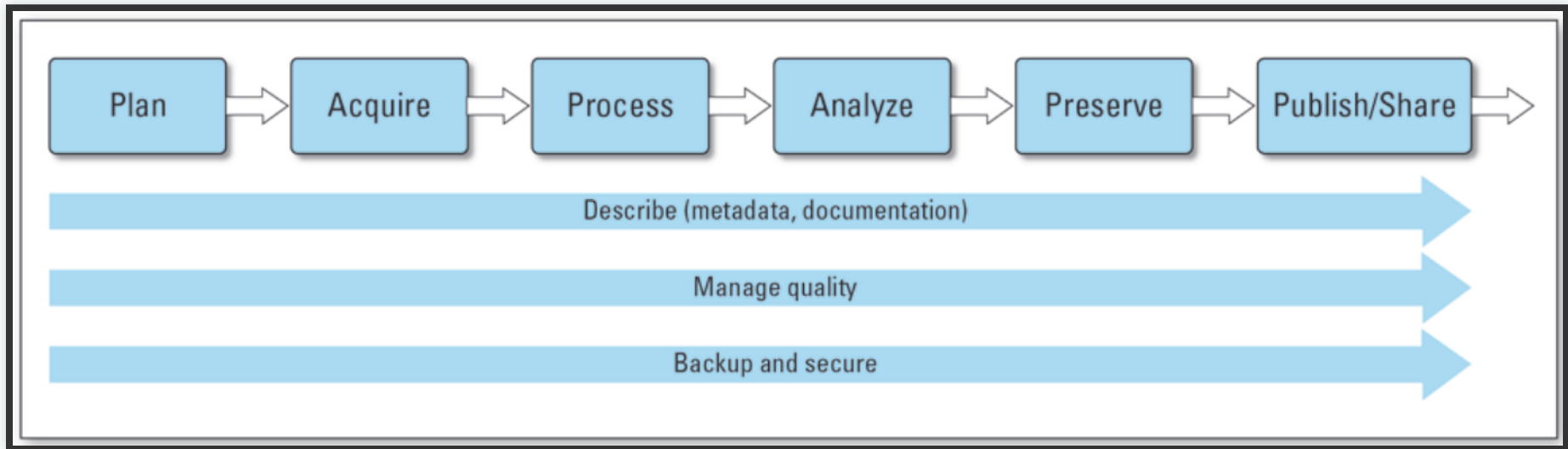


Figure 1: USGS Science Data Lifecycle Model. Boxes indicate the main Model elements, and the shaded arrows below represent cross-cutting elements.¹

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!

Library services and tools related to DMP

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

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](#)

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)

Repositories and data

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

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](#)



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)

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](#))

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 section for basic information and a bottom section for detailed metadata.

Top Section:

- Add** (icon)
- National Corrections Reporting Program, 1991-2015: Selected Variables**
- 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 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](#))

Right Side Panel:

- Email
- Print
- Text Message
- Export to ... (dropdown menu)
- delicious

Bottom Section (Metadata):

- Details** | **Subjects** | **Summary**
- 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](#)

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

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

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

Challenges

- Skills
- Scale
- Quality
- Rights
- Ethics

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](#)

Library data

Trends I think I see

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

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

On the table: Data warehousing

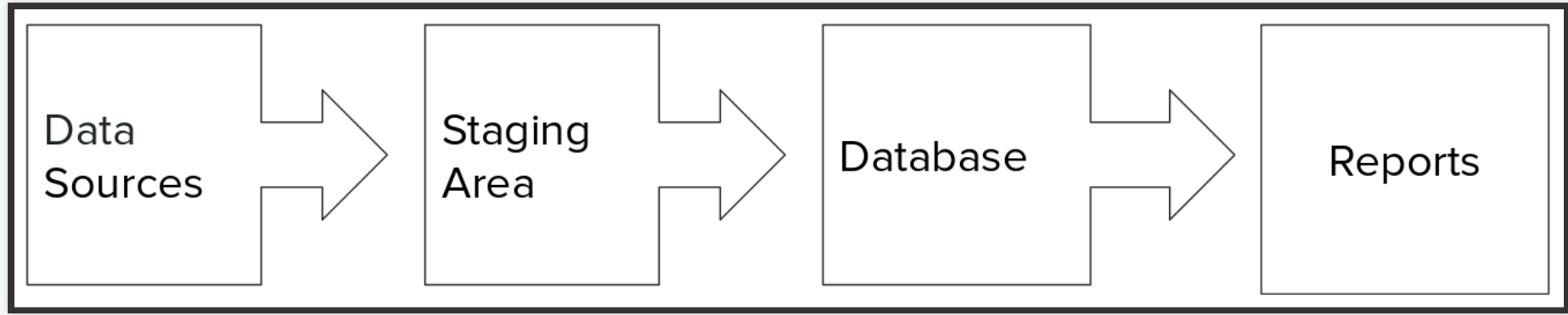


Figure 5: Conceptual flow of data warehousing¹

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 raucous 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

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/>.

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

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!

Data warehousing, again

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

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?

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](#)

Thank you



Figure 6: Any questions?

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