

Hosted by:

# DATA CURATION NETWORK



Journal of  
eScience  
Librarianship

# Data Soup



# Today's Menu

Welcome to Data Soup

---

## Presentations

- ★ [Creating Guidance for Canadian Dataverse Curators: Portage Network's Dataverse Curation Guide](#)

*Alexandra Cooper, Michael Steeleworthy, Ève Paquette-Bigras, Erin Clary, Erin MacPherson, Louise Gillis, and Jason Brodeur*

- ★ [Active Curation of Large Longitudinal Surveys: A Case Study](#)

*Inna Kouper, Karen L. Tucker, Kevin Tharp, Mary Ellen van Booven, and Ashley Clark*

# Today's Menu

## Presentations cont'd

- ★ [Data Curation through Catalogs: A Repository-Independent Model for Data Discovery](#)

*Helenmary Sheridan, Anthony J. Dellureficio, Melissa A. Ratajeski, Sara Mannheimer, and Terrie R. Wheeler*

## Q & A Discussion

---

DATA  
**CURATION**  
**NETWORK**

# DATA CURATION NETWORK



UNIVERSITY OF MINNESOTA

**I** ILLINOIS



Cornell University



Duke  
UNIVERSITY



UC SANTA BARBARA

[datacurationnetwork.org](http://datacurationnetwork.org)

## Mission

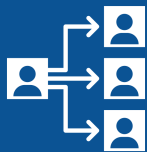
---

Trusted, community-led  
network of curators  
advancing open research  
by making data

**Ethical. Reusable. Better.**



ALFRED P. SLOAN  
FOUNDATION



## DCN Curation



Curate data as a cross-institutional network of nearly 50 individual experts



## DCN Education



Offer professional development opportunities for an emerging data curator professional community



## DCN Primers



Create and openly share data curation best practices



## DCN Interest Groups



Informal research teams addressing a specific topic: big data, human subjects, racial justice...

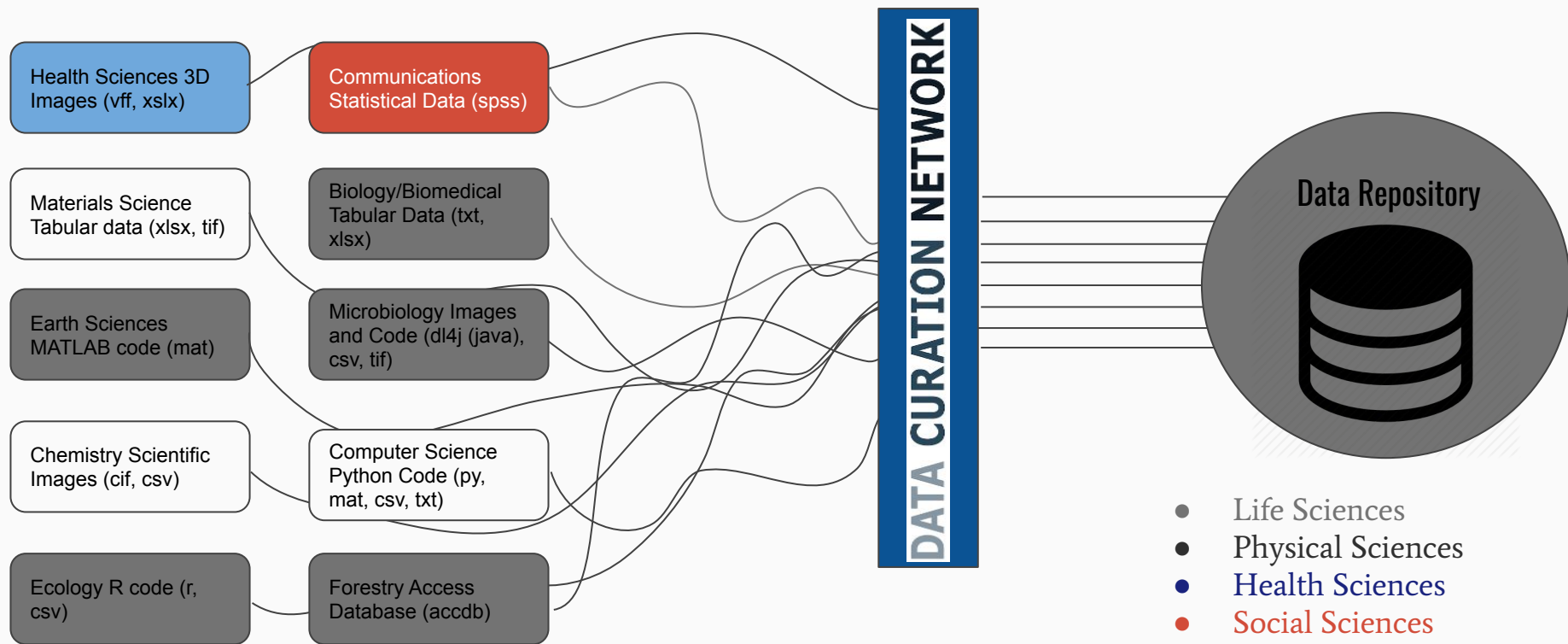


## DCN Community

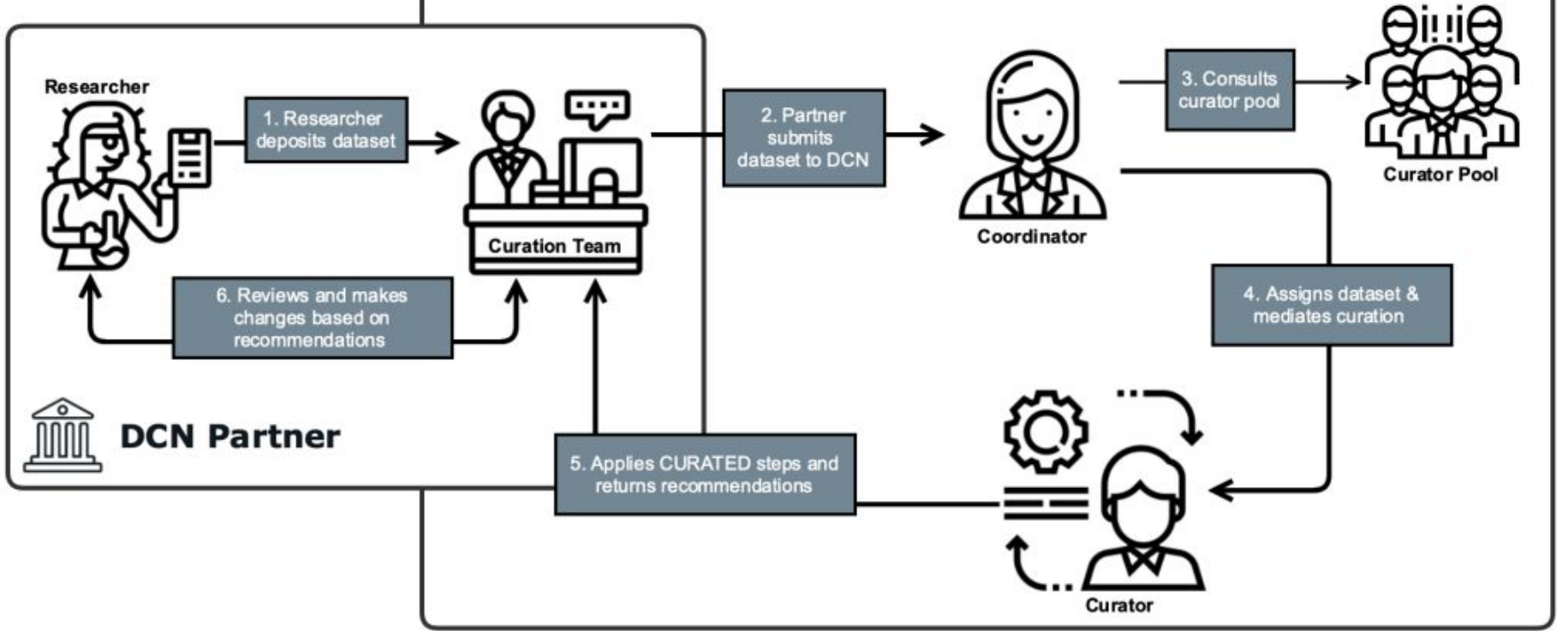


Build community at annual events for discussion, training, and networking.

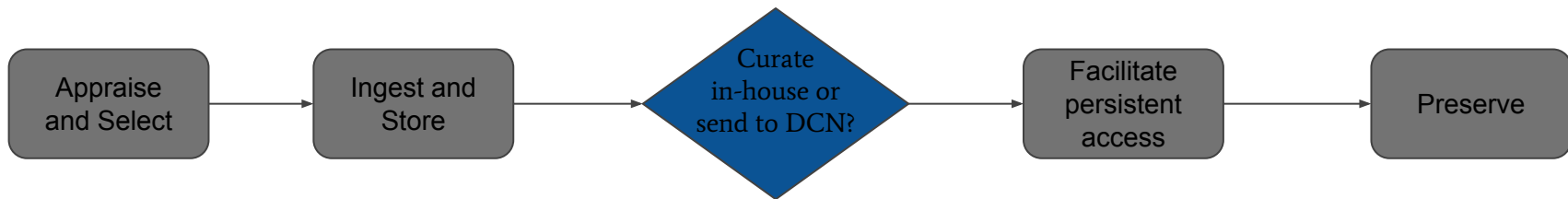
# Curation at Scale



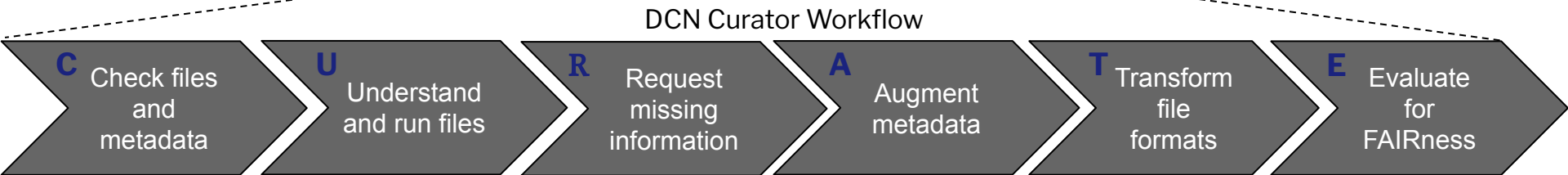
# DATA CURATION NETWORK







## DATA CURATION NETWORK

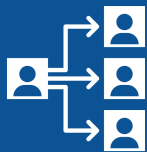


\*CURATE(D) = Document curation process throughout

Ethical. Reusable. Better.

**DATA CURATION NETWORK**

[datacurationnetwork.org](http://datacurationnetwork.org)



## DCN Curation



Curate data as a cross-institutional network of nearly 50 individual experts



## DCN Education



Offer professional development opportunities for an emerging data curator professional community



## DCN Primers



Create and openly share data curation best practices



## DCN Interest Groups



Informal research teams addressing a specific topic: big data, human subjects, racial justice...



## DCN Community



Build community at annual events for discussion, training, and networking.

# The CURATE(D) Steps

- C** **Check** files and read documentation.
- U** **Understand** the data (or try to), if not...
- R** **Request** missing information or changes.
- A** **Augment** metadata for findability.
- T** **Transform** file formats for reuse.
- E** **Evaluate** for FAIRness.
- (D)** **Document** your curation activities



<https://datacurationnetwork.org/resources/workflows/>

# Data Curation Primers

- The project began as a capstone to our Specialized Data Curation Workshops that were generously funded by the Institute of Museum and Library Services (IMLS).
- 27 primers released so far on Github!

Ivey, Susan; Koshoffer, Amy; Sneff, Gretchen; Wang, Huajin. (2019). Confocal Microscopy Images Data Curation Primer. [Data Curation Network GitHub Repository](#).



Grant # RE-85-18-004018

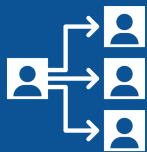
The screenshot shows the title page of a primer from the Data Curation Network. The title is "Confocal Microscopy Data: A Primer for Curators". Below the title is an "Executive summary" section containing a table with two columns: "Topic" and "Description".

Topic	Description
File Extensions	.ism (Zeiss - proprietary) <sup>1</sup>
	.dzi (Zeiss - proprietary)
	.nd2 (Nikon NIS-Elements 2 - proprietary)
	.lif (Leica - proprietary)
	.oib (Olympus - proprietary)
MIME Type	.zip
	.tiff (open source - recommended for archiving) <sup>2</sup>
Image Type	Image/tiff

Free and open!

**Peer-reviewed concise resources** to assist the data curator with recommendations for curating specific formats or curation topics!

**DATA CURATION NETWORK**



## DCN Curation



Curate data as a cross-institutional network of nearly 50 individual experts



## DCN Education



Offer professional development opportunities for an emerging data curator professional community



## DCN Primers



Create and openly share data curation best practices



## DCN Interest Groups



Informal research teams addressing a specific topic: big data, human subjects, racial justice...



## DCN Community



Build community at annual events for discussion, training, and networking.

Please contact us:

---

[dcn-team@googlegroups.com](mailto:dcn-team@googlegroups.com)

<http://datacurationnetwork.org>

Ethical. Reusable. Better.

Thank You!

---

**DATA  
CURATION  
NETWORK**

# Creating Guidance for Canadian Dataverse Curators: Portage Network's Dataverse Curation Guide

*Alexandra Cooper, Michael Steeleworthy (Presenter), Ève Paquette-Bigras, Erin Clary,  
Erin MacPherson, Louise Gillis, Jason Brodeur*

*Journal of eScience Librarianship. 2021. 10(3): e1201. <https://doi.org/10.7191/jeslib.2021.1201>*





CARL PORTAGE

NDRIO RDM

ALLIANCE RDM



# The need – Le besoin

## Provide bilingual curation advice that:

- Is tailored for the Dataverse platform
- Is adaptable to various service models
- Promotes and encourages curation consistency within Canadian Dataverses and other repositories
- Is intended for data curators at all levels of experience, at institutions of all sizes

## Offrir des conseils bilingues sur la curation qui :

- S'adaptent à la plateforme Dataverse
- Sont adaptables à divers modèles de services
- Favorisent et encouragent la cohérence de la curation au sein des dépôts Dataverse canadiens et des autres dépôts
- S'adressent aux curateurs de données de tous niveaux d'expérience, dans des établissements de toutes tailles

# Modelling our work on the DCN Curation Framework



The WG adapted the Data Curation Network's CURATE(D) Framework to fit the needs of Canadian Dataverse curators, in both official languages

//

Le GT a adapté le cadre CURATE(D) du Data Curation Network pour répondre aux besoins des curateurs de Dataverse canadiens dans les deux langues officielles.

Aligner notre travail sur le cadre de curation du DCN

# CURATE(D) → CURATION

<b>C</b>	Check	Consulter
<b>U</b>	Understand	Un peu plus en profondeur
<b>R</b>	Recommend improvements	Recommander
<b>A</b>	Augment	Améliorer
<b>T</b>	Transform	Transposer
<b>I</b>	Include persistent IDs and a reuse licence/agreement	Inclure ID pérennes et les licences/ententes de réutilisation
<b>O</b>	Optimize for FAIRness	Optimiser selon les principes FAIR
<b>N</b>	Note down curation activities	Noter les actions réalisées

# How it works – Fonctionnement du guide

## Service Scenarios

1. Unmediated curation
2. Semi-mediated curation
3. Mediated curation

## Levels of Curation

1. Minimum requirements to make data findable
2. Enhance discoverability and ensure usability
3. Prepare the dataset for reproducibility and preservation.

## Scénarios de services

1. Curation non médiatisée
2. Curation semi-médiatisée
3. Curation médiatisée

## Niveaux de curation

1. Exigences minimales pour rendre les données trouvables
2. Améliorer la facilité de découverte et assurer la facilité d'utilisation
3. Préparer l'ensemble de données pour la reproductibilité et la préservation

# Step 1 Screenshot – Capture d'écran Étape 1

## Check / Consulter

At the **Check** step, confirm that all data and metadata components required by the system to successfully publish the deposit are present. If possible, identify any characteristics that may require special consideration (e.g., data with disclosure risk, or data obtained from a third-party source).

### Level 1

Yes	No	Some issues	N/A	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	The researcher has confirmed that the dataset is free of any licensing and intellectual property issues.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	The researcher has confirmed that the dataset is free of any sensitive information (i.e., information that must be safeguarded against unwarranted access or disclosure).
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Supporting documentation is included. For example, a codebook, data dictionary, methodology, Readme file, etc.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	All files described in the documentation are included in the dataset.

# Adopt the Guide – Adopter le guide

Your curation service and your level of curation will not map perfectly. There are dependencies:

- Type of data you are working with
- Institutional policies and strategic aims
- Expertise and ability
- Resource capacity
- Resource demand

The Curation Guide provides advice and step-by-step instructions on how to curate datasets based on the RDM service your institution is providing.

Votre service de curation et votre niveau de curation ne s'arriment pas parfaitement. Il y a des dépendances :

- Type de données avec lesquelles vous travaillez
- Politiques et objectifs stratégiques de l'établissement
- Expertise et capacité
- Capacité des ressources
- Demande de ressources

Le guide de curation fournit des conseils et des instructions étape par étape sur la façon d'organiser les ensembles de données en fonction du service de GDR fourni par votre établissement.

# Language Challenges – Défis liés à la langue

- Extra Work for our bilingual colleagues
- The CURATED → CURATION Acronym
  - Finding and deploying a bilingual acronym
  - The acronym can affect the concepts we wish to prioritize, or how we talk about them
- Fully bilingual resources are difficult to find
- Precise, technical vocabulary can be difficult for translation
- Travail supplémentaire pour nos collègues bilingues
- L'acronyme CURATED → CURATION
  - Trouver et diffuser un acronyme bilingue
  - L'acronyme peut affecter les concepts que nous souhaitons prioriser, ou la façon dont nous en parlons.
- Les ressources entièrement bilingues sont difficiles à trouver
- Le vocabulaire précis et technique peut être difficile à traduire

# Next Steps – Prochaines étapes

- Adapt based on feedback as Guide is used in practice
- Find more French exemplars and resources
- Add more templates for correspondence
- Create a web-based resource. Easier to:
  - Navigate
  - Adapt locally
  - Update
  - Contribute new content
- Workshop: curate data using the Guide
- Adapter en fonction de la rétroaction lors de l'utilisation pratique du guide
- Trouver plus d'exemples et de ressources en français
- Ajouter d'autres modèles de correspondance
- Créer une ressource en ligne. Plus facile à :
  - Naviguer
  - Adapter localement
  - Mettre à jour
  - Contribuer en ajoutant du contenu
- Atelier : organiser des données avec le Guide



# Dataverse Curation Guide Working Group

//

Groupe de  
travail sur le  
guide de  
curation de  
Dataverse

- Jay Brodeur, McMaster University
- Erin Clary, Digital Research Alliance of Canada
- Alexandra Cooper (co-chair), Queen's University
- Louise Gillis, Dalhousie University
- Erin MacPherson, Dalhousie University
- Ève Paquette-Bigras, Université de Montréal
- Michael Steeleworthy (co-chair), Wilfrid Laurier University
- Lee Wilson, Digital Research Alliance of Canada

Thank you to / merci à

Meghan Goodchild, Queen's University Library and Scholars Portal

Contact us via Erin Clary, Alliance RDM,  
[erin.clary@engagedri.ca](mailto:erin.clary@engagedri.ca)

# Active Curation of Large Longitudinal Surveys

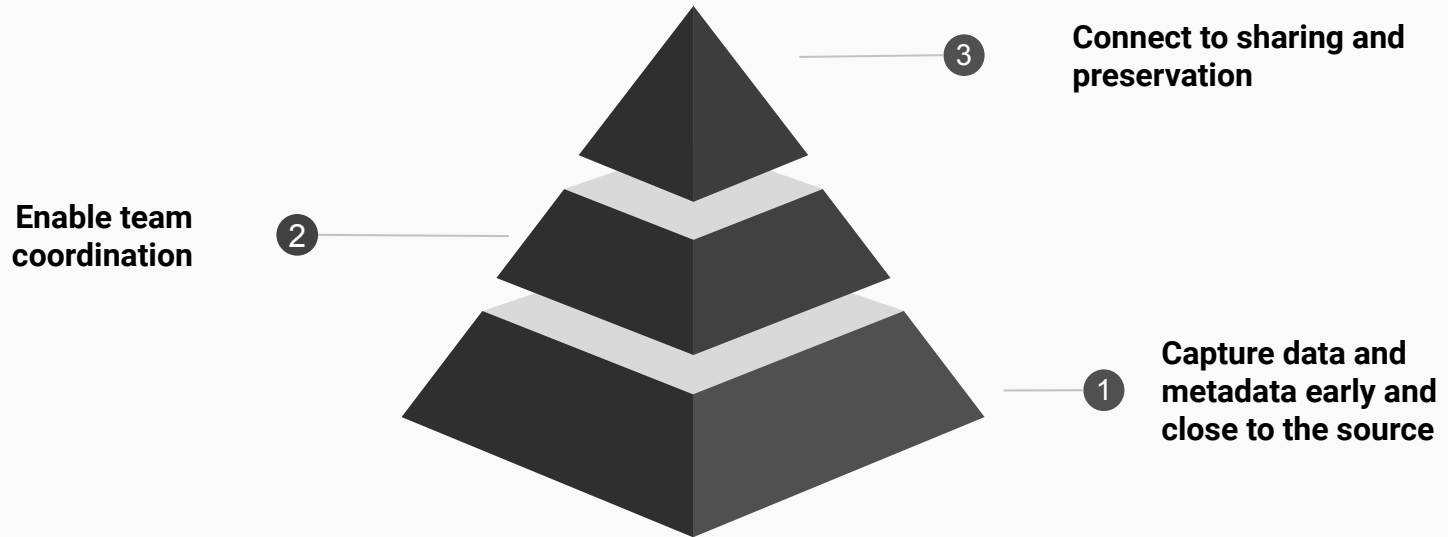
A Case Study

Inna Kouper (presenter),

Karen L. Tucker, Kevin Tharp, Mary Ellen van Booven, Ashley Clark

*Center for Survey Research, Indiana University*

# Active Curation



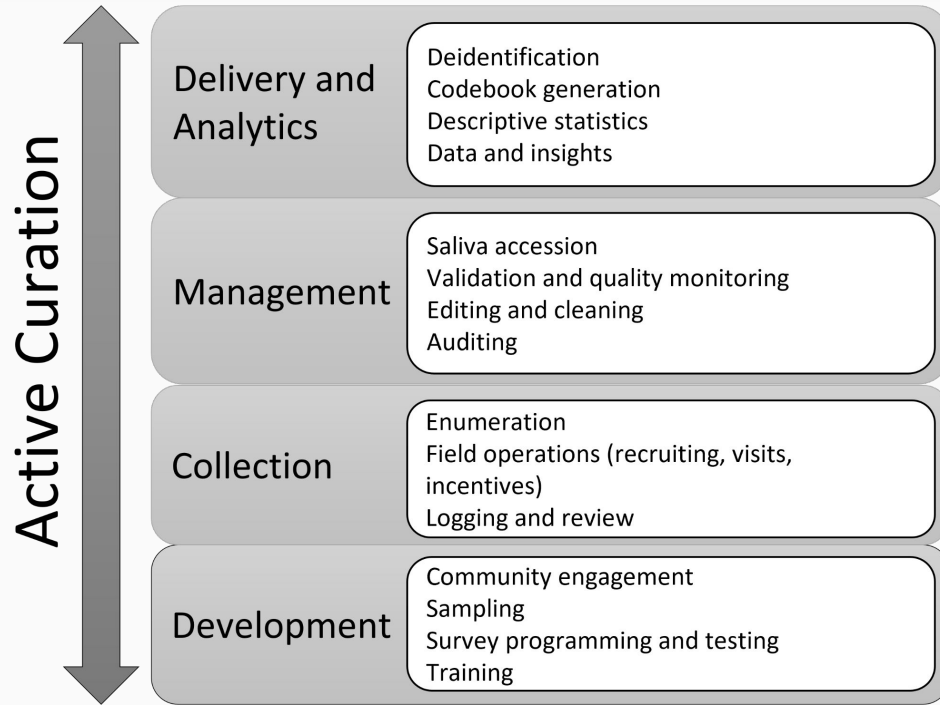
# The Survey: P2P Health Interview Study, Indiana University Precision Health Initiative

A representative survey of ~2,000 residents of one state

100s of questions + biometric information (height, weight, blood pressure) + saliva samples

Team includes researchers (the science team) and data managers (the data team)

# Active Curation Activities



# Curation Objects

## Development

- Sample with individual cases
- Survey instrument
- Pilot data
- Software and documentation

## Collection

- Study information materials
- Cases (information about participants)
- Survey outcomes and paradata
- Supplementary survey materials

## Management

- Survey data
- Saliva samples
- Survey cost and other indicators
- Field interviewer reports

## Delivery and Analytics

- Integration workflows
- Codebook
- Anonymized data
- Derived data

# Recommendations

Develop a consistent approach to working with active (“live”) data

Design curation for current and future data work

Consider working with humans as part of curation

Develop and adopt standards for active curation

# Data Curation through Catalogs: A Repository-Independent Model for Data Discovery

Helenmary Sheridan (presenter), Anthony J. Dellureficio, Melissa A. Ratajeski,  
Sara Mannheimer, and Terrie R. Wheeler



# Why a Catalog?

- Resources: catalogs may be cheaper in staff time and storage costs than a repository
- Scope: catalogs can describe and point to data that would not be included in a repository because of access protocols, researcher requirements, or sensitivity
- Security: can specialize in data governance by integrating with secure data enclaves

# Why a Catalog? (cont.)

Better co-location: records can point to data, code, registration, protocol, and more

Efficiency: there are great repositories out there that researchers are already using. Let's make their submissions more findable!

DCN C-U-R-A-T-E-D Step	Data catalog activity
C: Check data files and read documentation	Somewhat similar. Catalogers read documentation and examine data files to create a high-quality metadata record; however, they do not check files for completion, quality, or file integrity
U: Understand the data	Somewhat similar. Catalogers try to understand the data enough to describe it, but do not comment on the data files unless also offering advice prior to submission to a repository
R: Request missing information or changes	Similar. Catalogers ask for more information to create a metadata record, and may suggest that the authors create documentation
A: Augment with metadata for findability	Very similar. Catalogers create descriptive metadata, incorporating author-supplied terms when possible, and source metadata from controlled vocabularies for interoperability
T: Transform file formats	Does not apply, although catalogers can make recommendations for the data stored elsewhere
E: Evaluate and rate for FAIRness	Does not apply; although catalog staff may have their own checklist for acceptable metadata records, they do not control the data themselves
D: Document throughout curation activities	Somewhat similar. Since no actual datasets are changing hands, submission agreements and chain-of-custody documentation are unnecessary, but institutions may have their own cataloging workflow requirements. The open-source code developed by NYU keeps a basic log of editing dates made to records, and catalogers have the option of adding detailed edit notes

# Who We Are: the Data Discovery Collaboration

**NYU Langone**

**University of  
Pittsburgh  
HSL**

**Memorial  
Sloan  
Kettering**

**Northwestern  
University  
Galter**

**Weill Cornell  
Medicine**

**Hofstra/  
Northwell**

**Montana  
State  
University**

**University of  
Maryland  
HSHSL**

# Evolution of the DDC

**2017**

Formation of the Data Catalog Collaboration Project, a network of health sciences libraries using data catalog software produced at NYU

**2018-2019**

Increasing interest in methods and philosophy from organizations without catalogs (and no interest in supporting one)

**2020+**

Reconfiguration into the Data Discovery Collaboration, a platform-agnostic organization connecting individuals and institutions who are working to increase the discoverability of data

# Some current areas of development

- Representing terms and concepts of interest to basic science researchers, like study organisms: how to model a particular cell line from the kidney tissue of a Sprague-Dewley (Brown Norway) rat?
- Integration of/reference to biomedical registries and taxonomies like SciCrunch or NCBI Taxonomy
- Supporting non-dataset data products like computational models and software code
- Author disambiguation tools for API-assisted metadata ingestion

# Thank you!

Helenmary Sheridan, University of Pittsburgh Health Sciences Library System

Anthony J. Dellureficio, Memorial Sloan Kettering Cancer Center

Melissa A. Ratajeski, University of Pittsburgh Health Sciences Library System

Sara Mannheimer, Montana State University

Terrie R. Wheeler, Weill Cornell Medicine

Data Discovery Collaboration: Contact Nicole Contaxis, [nicole.contaxis@nyulangone.org](mailto:nicole.contaxis@nyulangone.org)

Hosted by:

**DATA  
CURATION  
NETWORK**



Journal of  
eScience  
Librarianship

Discussion  
time!