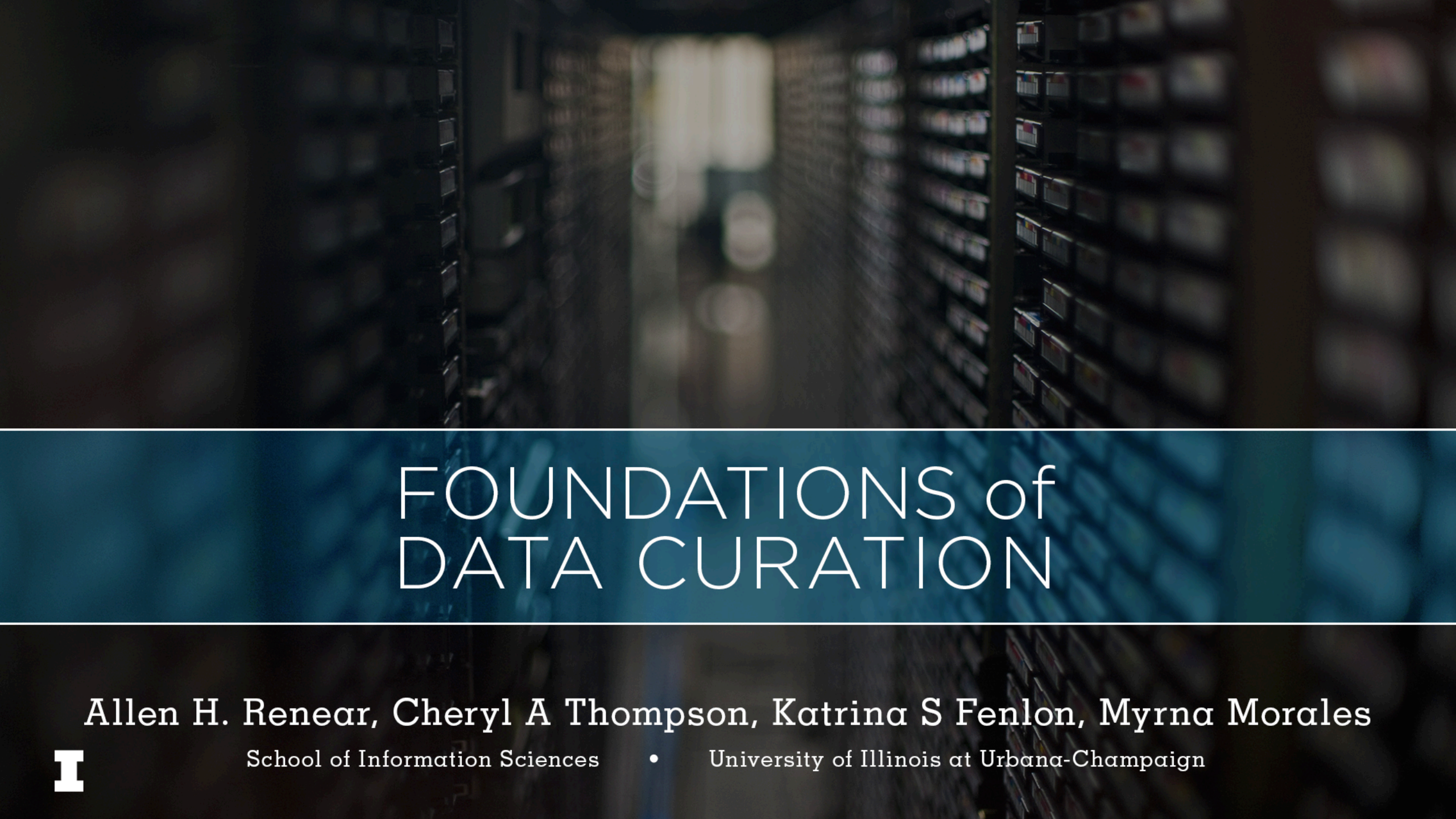
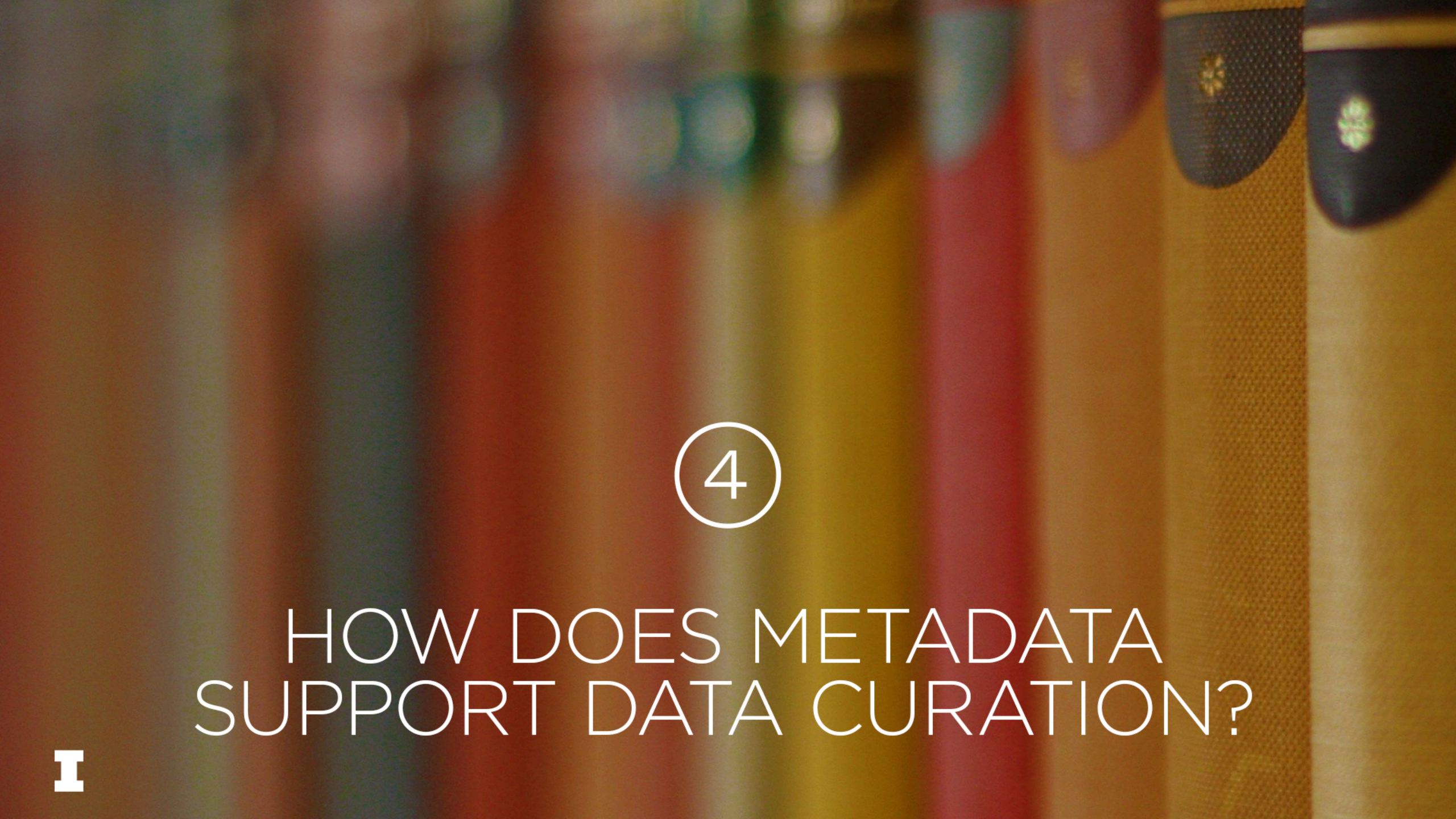


FOUNDATIONS of DATA CURATION

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales

School of Information Sciences • University of Illinois at Urbana-Champaign


METADATA


4
HOW DOES METADATA
SUPPORT DATA CURATION?

How Does Metadata Support Data Curation?

Relating metadata to curation *objectives*, *activities*, *actions*

Recall the data curation objective

Data curation is concerned with all aspects of the management of data

in order to efficiently and reliably support the analysis of data,
and enable reuse over time

Recall: Areas of curatorial activities

Collection:	Support the collection and acquisition of data
Organization:	Employ an appropriate data model and use appropriate standards
Storage:	Support reliable and effective storage
Preservation:	Ensure that data will be understandable and useable in the future
Discoverability:	Support the ability to search for and locate relevant data
Access:	Support the ability to retrieve and distribute data
Workflow:	Support the ability to systematize data workflows
Identification:	Support the ability to identify, authenticate, and validate data
Integration:	Support integration of data from different sources using different data models
Reformatting:	Support reformatting for use by different tools or to match new format standards
Reproducibility:	Support ability to reproduce results, ensuring scientific validity
Sharing:	Support sharing data between researchers, teams, and institutions.
Communication:	Support representation, publishing, and visualizations that provide insight
Provenance:	Support identifying what inputs and calculations are responsible for data values
Modification:	Support management of corrections and updates
Compliance:	Ensure compliance to legal, regulatory, and local policy requirements
Security:	Ensure that data is secure from tampering or inappropriate access and distribution

Recall: Methods of curatorial action

Analysis
To determine needs, and develop relevant data models and *metadata*, and reformat, correct, or update data.

➔ **Documentation**
To record essential information (typically via *metadata*)

System design and implementation
To support all data curatorial activities
To support the generation and use of data documentation and processing documentation

Policy
To specify objectives, procedures, practices, and formats.

Process
To ensure success and efficiency by managing the development of appropriate organizational units and roles, providing training, advocating for change, and managing curatorial activities.

➔ Metadata is rigorously defined machine-processable documentation

Areas of curatorial activities (Part I: the core)

Area	Description	Supported by metadata that documents. . .
Collection	Support the collection and acquisition of data	Method, location, time, instruments, settings, calibration. . .
Organization	Employ an appropriate data model and use appropriate standards	Schemas and schema documentation for semantics, syntax, and encoding.
Storage	Support reliable and effective storage	Authoritative and alternative copies, physical locations, redundancy, compression, reduction, backups
Preservation	Ensure that data will be understandable and useable in the future	[see Organization, Identification, Storage]
Discoverabilty	Support the ability to search for and locate relevant data	Topic, coverage, formats, availability, currency.
Access	Support the ability to retrieve and distribute data	[see Organization, Security], licensing, owner, location.
Workflow	Support the ability to systematize data workflows	[See Organization], scripts, processes, transformations, inputs.
Identification	Support the ability to identify, authenticate, and validate data	[See Organization], identifiers, version data, integrity checks, authentication.
Integration	Support integration of data from different sources using different data models	[See Organization, Identification, Access, Discoverability]

Areas Of Curatorial Activities (Part II: Partial Dependencies)

Activity	Description	Metadata documenting . . .
Modification	Support management of corrections and updates	[See Organization, Workflow, Provenance, Identification]
Reformatting	Support reformatting for use by different tools or to match new format standards	[See Organization, Identification, Workflow]
Provenance	Support identifying what inputs and calculations are responsible for data values	[See Workflow, Identification, Reproducibility]
Reproducibility	Support ability to reproduce results, ensuring scientific validity	[See Organization, Workflow, Provenance, Identification]
Preservation	Ensure that data will be understandable and useable in the future	[see Organization, Identification, Storage]
Compliance	Ensure compliance to legal, regulatory, and local policy requirements	[see Organization, Provenance, Workflow, Discoverability]. Certification.
Security	Ensure that data is secure from tampering or inappropriate access and distribution	[see Organization, Provenance, Workflow, Discoverability]. Encryption, Certification.
Communication	Support representation, publishing, and visualizations that provide insight	[See Organization, Identification, Reproducibility, Compliance]
Sharing	Support sharing data between researchers, teams, and institutions.	[See Discoverability, Organization, Workflow, Provenance, Identification]

Frictions In Creating And Managing Metadata

Metadata Friction Categories	Specific frictions
Standardization	• Unfinished and multiple metadata versions • Proliferation of metadata standards
Temporal	• Metadata in the life cycle • Timeframe for metadata use
Data Sharing	• Audience for metadata • Individual vs. group knowledge
Human Support	• Metadata is nobody's job • Proliferation of metadata tools

Metadata frictions specific to the "problems and conflicts that must be addressed to make metadata useful"

Adapted from Mayernik, M. S., Batcheller, A. L., & Borgman, C. L. (2011). How institutional factors influence the creation of scientific metadata. In *Proceedings of the 2011 iConference* (pp. 417–425). Retrieved from http://dl.acm.org/citation.cfm?id=1940818

FOUNDATIONS OF DATA CURATION (IS531)

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales
School of Information Sciences
University of Illinois at Urbana-Champaign

Includes material adapted from work by Carole Palmer, Melissa Cragin, David Dubin, Karen Wickett, Bertram Ludaescher, Ruth Duerr and Simone Sacchi.

Comments and corrections to: renear@illinois.edu.