OAK RIDGE National Laboratory

Towards a Big-Data Toolkit: Ensuring Data Governance & Ethical Considerations Are Applied to Large Datasets

DOE Data Days

2022-05-27

Alex May, Olga Kuchar, Katie Knight, Rohit Srivastava

The Fundamental Question:

Do the Data Curation Network's *best practices* and *human-intensive workflows* **scale,** and if not, what are the **ethical implications?**

The Issues:

Defining Big Data	No consistent definition: what is it, exactly?	How “Big" is Big Data?	What is the data lifecycle for Big Data?
Curation	Which data should be stored, and how?	What should get tagged: the entity creating the data, or the actual data?	How do workflows interact with data?

DCN Curation Checklist à Does this Work for 600,000 Images?

C	**Check** files/code and read documentation (risk mitigation, file inventory, appraisal/selection)
U	**Understand** the data (or try to), if not… (run files/code, QA/QC issues, readmes)
R	**Request** missing information or changes (tracking provenance of any changes and why)
A	**Augment** metadata for findability (DOIs, metadata standards, discoverability)
T	**Transform** file formats for reuse (data preservation, conversion tools, data visualization)
E	**Evaluate** for FAIRness (transparent usage licenses, responsibility standards, metrics for tracking use)
D	**Document** all curation activities throughout the process

Data Curation Network, 2022, p. 2

Data Governance and Ethical Considerations

Schwarz et al., 2021

Corrupted binaries and files will **impact reproducibility** and **file format migration**

Personal Identifiable Information (PII) –, e.g., faces in a dataset of 600,000 images and embedded location metadata

Incomplete information for understanding the dataset itself

Increasing use of ML and/or AI makes it possible to identify individuals by combing deidentified datasets

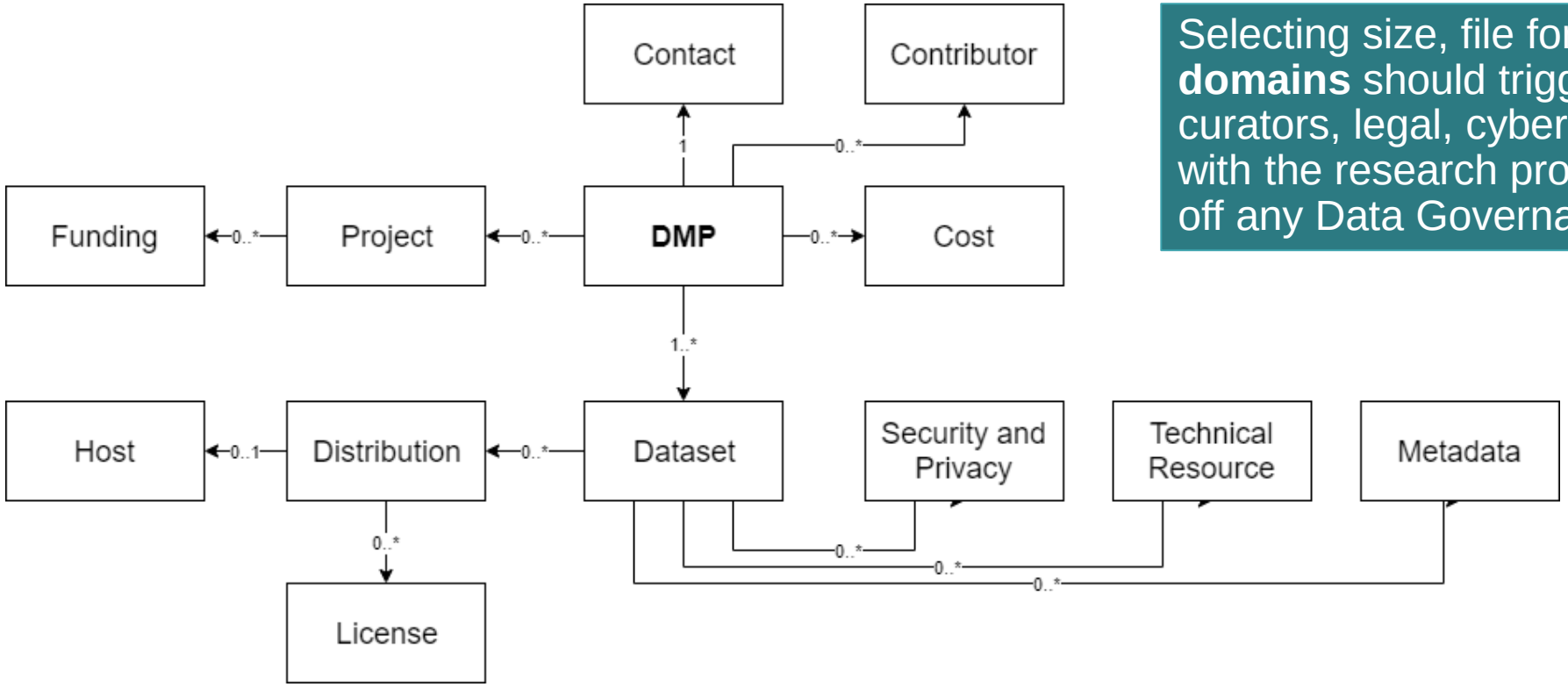
Need for Data Management Policies

- Enable Open Science
- Comply with Federal & International Law
- Risk Management
- Sensitive Data Compliance
- Add Data Value
- Adhere to the Prime Contract
- Long-term stewardship

Adapted from: Expanding the Table: The Role of Library Data Professionals in Data Governance *Speakers:* Abigail Goben, Kristin

Towards the Big Data Toolkit

Machine Actionable Data Management Plans



Selecting size, file formats, and **even domains** should trigger workflows that alert curators, legal, cyber, etc… to get involved with the research process early and head off any Data Governance or Ethical Issues

Adapt the RDA-DMP-Common-Standard to the specific needs of the Labs

Towards the Big Data Toolkit

Don't focus solely on file formats

Create **automated workflows** that are **domain specific**

Create **domain-specific primers**

Start creating a **domain repository**

e.g., BioPortal2, an open, automatically updated repository of versioned biomedical ontologies stored in various formats accessible via Web browsers

Towards the Big Data Toolkit

Invest in a Lab-Wide Data Catalog

For **preservation**, know the:

- Frequency of accession
- What factors contribute to dataset discovery

Adopt **machine-actionable access controls** that can:

- Align with data governance and policies
- Automatically scan metadata, provide file-level sna

Towards the Big Data Toolkit

Sciences & Humanities Can Solve the Problem Together

Look to Digital Humanities work with **AI and ML**

AI for Libraries interest group

Stanford and Library of Congress work in classifying images

Repurpose tools like the MIT's Sonification Toolkit

e.g., quickly identify bad binaries

Look into ARCHANGEL **blockchains**

Data can be added but not overwritten, amended or deleted

Offers a **digital finge** archival materials

Verify authenticity

Towards the Big Data Toolkit

Evaluate Organization-Wide Solutions

Our big data is not "special"

Avoid technical debt

Evaluate and find turn-key solutions

How is this already handled in industry, other domains?

e.g., Google DataFlow, Apache Beam, HubZero

While all products may not work perfectly, they may point to potential solutions

Towards the Big Data Toolkit

Get Creative!

Create virtual, web-based environments for curators	
Create a repository of curation tools	e.g., Re3Data
	No more spreadsheets!
Use HPC to curate data	Harness big computers for big data
Consider Digital Object solutions	BDBag
	FAIR Digital Objects
	RO-Crate

The Big Data Toolkit is, in fact, a Constellation of Services

- **Next Steps**
- Work towards a MaDMP
- Evaluate lab-wide data catalog solution
- Participate in DCN's Big Data Interest Group
- Continue to learn from curating Big Data datasets in order to build the toolkit