Data Curation 2018

November 5th, 2018
Charleston, South Carolina

Part One

Introduction to Data Curation What is Data? What is Data Curation?


CONUNDRUM

Pictures, Pictures, Pictures

Data Curation

Data Curation is “applying the archival principles of library and information sciences to a wide variety of data objects from all disciplines and prepare them for ingest, access, and long term preservation within an environment that facilitates discovery and access while not diminishing their content, authenticity and value” - Lisa R. Johnson

Research Data

Instrument measurements

- Experimental observations
- Still images, video and audio
- Text documents, spreadsheets, databases
- Quantitative data (e.g. household survey data)
- Survey results & interview transcripts
- Simulation data, models & software
- Slides, artefacts, specimens, samples
- Sketches, diaries, lab notebooks ...

Research Data Classes

Observational - (Climate readings, field notes, etc.) - Historical, cannot be reproduced in any way, may need indefinite archiving.

Computational - May require information about hardware, and perhaps the complete software, but not necessarily the results

Experimental Data - Data gained from running experiments. May or may not be reproducible.

Data Curation: Then and Now

Then:

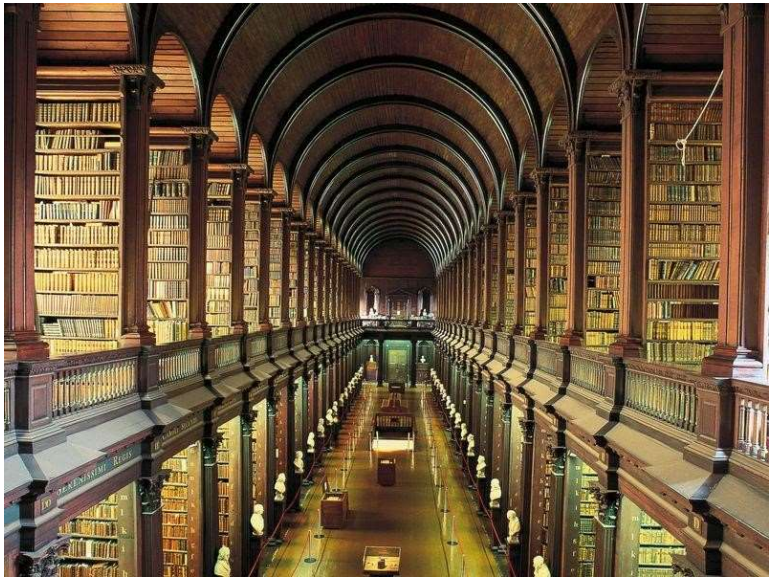
Journal articles primary means of scholarly communication

Data behind the articles accessible, but difficult

Articles contained in structured metadata (journals)

Cited according to rigid standards

No real data standards other than IRB rules.



Data Curation: Then and Now

Now:

Technology makes data easily shareable

Data now required to be made available by

Grant writing bodies - https://www.nsf.gov/bfa/dias/policy/dmp.jsp

Data Curation: Then and Now

Publishers - https://www.elsevier.com/authors/author-services/research-data/data-statement https://authorservices.taylorandfrancis.com/understanding-our-data-sharing-policies/

Data Curation: Then and Now

Now:

Data Curation: Then and Now

New kinds of collaboration

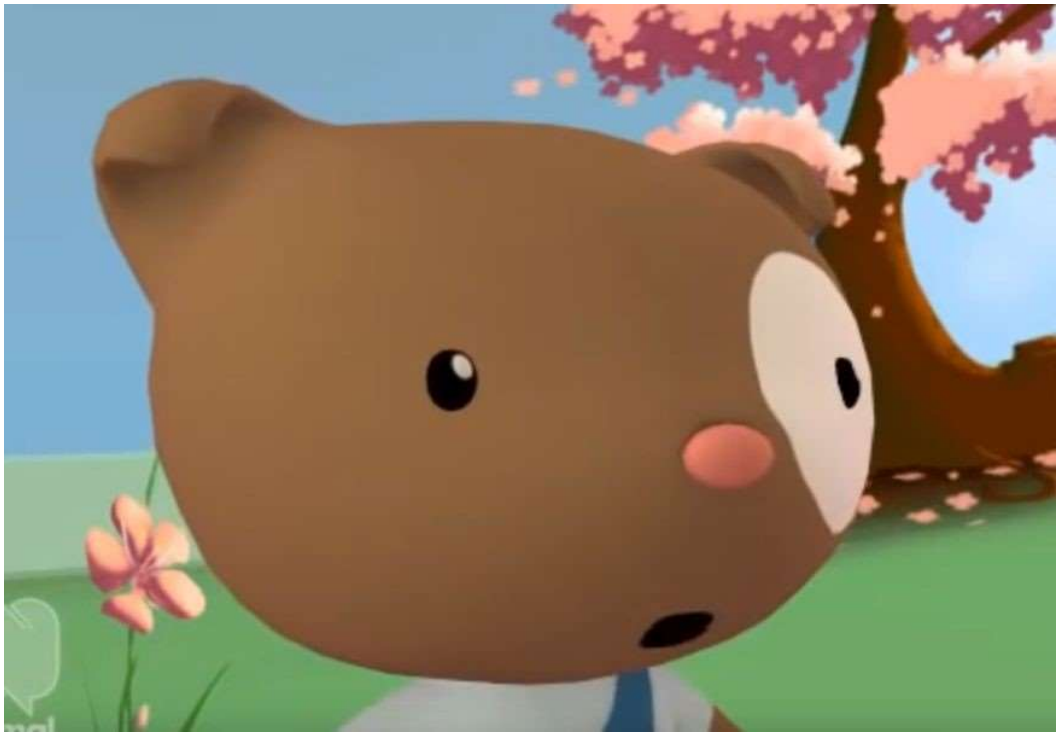
Data must be standardized, catalogued, distributed

Benefits everybody, including the researcher

Data Sharing & Management Snafu in Three Acts

https://youtu.be/66oNv_DJuPc

What's Wrong With This Dog?



Data Curation ‘tions

Preparation

Evaluation

Communication

Clarification

Standardization

Documentation

Purpose of the Curator

Proof - Proofread data

Question - Ask researchers about what's missing, Descriptions or pieces

Describe - Apply or improve labels and metadata

Translate - File formats

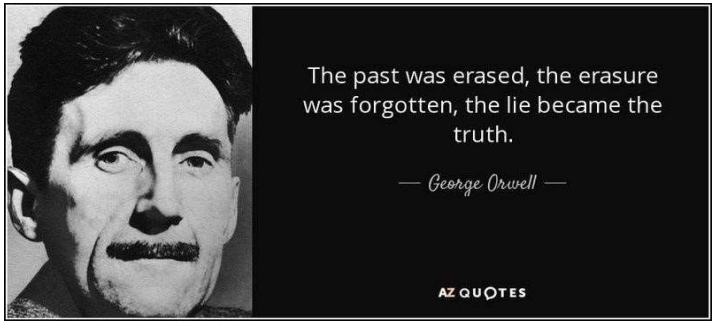
Clarify - Intelligibility to wide audience

Data Curation Thesaurus

Thesaurus.

Reasons to Curate

- Easier for fellow scholars and future collaborators to understand
- More likely to be trusted
- The research they represent are more likely to be reproducible
- More likely to be properly cited
- Represent potential cost-savings



Data is Being Lost

Data decay - [The Quest to Save Millions of Climate Records]

Deliberate scrubbing - [http://datarefuge.org]

[https://freegovinfo.info/node/13099]

Obsolete Media - [Rescuing the Computer Code that Took Us to the Moon]

Other Data Issues

- Open Data - https://cos.io/
  - https://osf.io/grhz7/
- Citizen Science
- https://www.nature.com/articles/d41586-018-07106-5
- https://www.citizenscience.gov/#

Data Curation and Libraries

“How Important are Data Curation Activities to Researchers?”

Johnston, L.R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R. and Stewart, C., 2018. How Important is Data Curation? Gaps and Opportunities for Academic Libraries. *Journal of Librarianship and Scholarly Communication*, 6(1), p.eP2198. DOI: http://doi.org/10.7710/2162-3309.2198

“How Important are Data Curation Activities to Researchers?”

- Focus Groups
- Card Swapping
- Surveys

Johnston, L.R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R. and Stewart, C., 2018. How Important is Data Curation? Gaps and Opportunities for Academic Libraries. Journal of Librarianship and Scholarly Communication, 6(1), p.eP2198. DOI:

http://doi.org/10.7710/2162-3309.2198

“How Important are Data Curation Activities to Researchers?”

- 47 different activities http://bit.ly/DCNcurationActivities
- How important are these activities?
- How satisfied are you with how they are being done?
- Results: “No single data curation activity was happening in ways that satisfied the majority of our participants”

Johnston, L.R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R. and Stewart, C., 2018. How Important is Data Curation? Gaps and Opportunities for Academic Libraries. Journal of Librarianship and Scholarly Communication, 6(1), p.eP2198. DOI:
http://doi.org/10.7710/2162-3309.2198

“How Important are Data Curation Activities to Researchers?”

- “Our study found gaps in support for data curation activities that are very important but that are either not happening or not happening in a satisfactory way…”
- “These may be areas of opportunity for libraries to invest in new services and/or heavily promote services that may already exist but are not reaching the researchers who value them”

Johnston, L.R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R. and Stewart, C., 2018. How Important is Data Curation? Gaps and Opportunities for Academic Libraries. Journal of Librarianship and Scholarly Communication, 6(1), p.eP2198. DOI:

http://doi.org/10.7710/2162-3309.2198

Opportunities for Libraries

- Creating adequate documentation
- Providing secure storage
- Performing quality assurance for data
- Creating or applying metadata
- Visualizing data

Johnston, L.R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R. and Stewart, C., 2018. How Important is Data Curation? Gaps and Opportunities for Academic Libraries. Journal of Librarianship and Scholarly Communication, 6(1), p.eP2198. DOI:
http://doi.org/10.7710/2162-3309.2198

If Libraries Don't…..

Others will: Springer Research Data Support - https://www.springernature.com/gp/authors/research-data-policy

Activity #1: Make the Case

F A I R

Data Principles

FAIR Data Principle Guidelines

Created at an International Loretz Center workshop in 2014 by a group representing academia, industry, grant funding bodies, and publishers

Published in *Nature* in 2016

No real binding power, but best practices to follow

Not a standard or specification, but rather guidelines to assist curators and data producers in ascertaining whether their practices are optimizing the usefulness of their data

The F A I R Data Principles

Findable

Accessible

Interoperable

Reusable

F A I R: Findable

F1. Assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data it describes

F4. Data are registered or indexed in a searchable resource

F A I R: Findable

DOI: Digital Object Identifier

Providers:

DataCite - https://www.datacite.org/dois.html

OSF (Open Science Framework): http://help.osf.io/m/sharing/l/524208-create-dois

F A I R: Findable

Google Dataset Search

F A I R: Accessible

A1. Data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

Repositories

R3Data Repository Search Engine: https://www.re3data.org/search/results/?term=

Dryad: https://youtu.be/RP33cl8tL28

ICPSR: https://deposit.icpsr.umich.edu/deposit/home

Zenofo: https://zenodo.org/

Open Source Repository Software - http://oad.simmons.edu/oadwiki/Free_and_open-source_repository_software

F A I R: Interoperable

I1. Data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. Metadata use vocabularies that follow FAIR principles

I3. Metadata include qualified references to other data

F A I R: Interoperable

W3C Best Practices for Data on the Web - https://www.w3.org/TR/dwbp/

Metadata: Digital Curation Centre: http://www.dcc.ac.uk/

1. Consider what information is needed for the data to be read and interpreted in the future.

2. Understand your funder requirements for data documentation and metadata.

3. Consult available metadata standards in your field:
http://www.dcc.ac.uk/resources/subject-areas/general-research-data

Metadata: Digital Curation Centre: http://www.dcc.ac.uk/

4. Describe data and datasets created in your research lifecycle. Assign or capture:

   Descriptive - Creator, Author, Title(s), File Name, File Location, File Size

   Technical - Format, Compression, Software, Hardware, OS Used

   Administrative - Creation, Updates, Migration, etc.

F A I R: Reusable

R1. Data are richly described with a plurality of accurate and relevant attributes

- R1.1. Data are released with a clear and accessible data usage license
- R1.2. Data are associated with detailed provenance
- R1.3. Metadata meet domain-relevant community standards

F A I R

Licensing: Creative Commons - https://creativecommons.org/share-your-work/

Attribution (by) This is a requirement of all CC licenses in that it establishes that any use of your work must contain an attribution to its source. This does does indicate that you, as original creator, necessarily endorse or support the subsequent work, however.

ShareAlike (sa) This gives others permission to use, copy, distribute, and modify your work as long as they distribute the new work with the same terms.

NonCommercial (nc) This gives others permission to use, copy, distribute, and modify your work as long as they do not do so for commercial purposes.

NoDerivatives (nd) This gives others permission to use, copy, distribute, original copies of your work, and prohibits modification of your work.

Exercise #2

DataSets in the Wild

https://docs.google.com/document/d/1fRBiWktZfC7CvPxKPt-eikC1BiHLK2Kg2boGxu1VISM/edit?usp=sharing