Policy,
Data Governance &
Licensing

INFX 551
Winter 2018

Information School
UNIVERSITY of WASHINGTON

Course Outline

Week 7

Data	Data Systems	Policy, Privacy, & Ethics
Types & Roles	Repositories	Policy
Documentation	Preservation	Privacy
Standardization	Cost Models	Ethics

Agenda

- Policy and its relation to data curation
- Data Governance
- Data Licensing + IP

Data Policy vs Procedure

- Policy answers who, what, and why questions.

  What format are text files stored in? *All text files are stored as ASCII-encoded plain text as well as PDF*

  Why are text files stored as plain text and PDF? *One serves as a preservation copy (plain text) and the other serves as an access copy (PDF)*

  Who ensures that all text are properly preserved in both formats? *A curator is responsible for all file format conversion.*

- Procedure answers how questions

  How are text transformed? All files converted using the tool Pandoc.

Data management plans create a set of policies for a specific dataset, or a specific project.

Data governance creates a set of policies for all data, or all projects in a school, a lab, a business, etc.

“A *policy* is a deliberate system of principles to guide decisions and achieve rational outcomes.”

–Daniel Beland, 2010 (they even put him in Wikipedia!)

Types of policies:

- Public policy, Social Policy, Economic Policy, Immigration Policy, etc.
- Data policy: Regarding the creation, access, use, and rights of creators and users of digital information objects

Policy as an instrument …

- **Money** (funding agencies create data management plan policies)

- **Prestige** (Journals create data sharing policies for publishing a paper)

- **Freedom** (Regulatory compliance with government data protection policies can (and have) resulted in jail time)

Money

Funder	DMP Policy (last policy updated)	Requirements	Notes
NSF	Dissemination and Sharing (2013)	Data sharing Data management plan	Since Jan 2011, a 2-page data management plan is required for all new proposals.
NIH	Data Sharing Policy (2012)	Data sharing	2-paragraph "data sharing statement" for proposals requesting > $500K and others. Check your solicitation. See our guidelines for preparing the statement.
NEH	Data management plan (2016)	Data management plan for Digital Humanities grants	Certain other NEH solicitations may require data sharing and management plan components.
NASA	Data management plans (2016)	Data management plan Project data accessible though NASA repositories when possible. At minimum, data associated with publications should be accessible, or a broader range of project data.	Fall 2016, a new web portal has been added including a data portal and information on their PubSpace publication repository.
NOAA	Data Sharing Directive (2016)	Data sharing Data management plan	NOAA maintains the NCEI in which data can be archived.
US Dept. of Education	IES Policy Statement (2016)	Data Sharing Plan required for IES "Effectiveness Goal" grants.	Plans not currently required for other "research goal" grant categories beyond Effectiveness (formerly "scale-up") grants.

http://dms.data.jhu.edu/data-management-resources/plan-research/funders-data-sharing-requirement/funder-data-related-mandates-and-public-access-plans/

Prestige

Journal	Policy (Last policy updated)	Requirements	Notes
Nature	Availability of data (2016)	Nature has specific and well documented recommendations for different types of data, materials, and computer code.	Nature journals' data availability policies are compatible with Springer Nature's standardized research data policies.
PLoS	Editorial and Publishing Policies (2016)	Data associated with publications must be publically available with rare exceptions.	Each PLoS journal has its own more specific set of guidelines and policies. These are accessible from the main policy link.
Science	Data Deposition and Availability of Data (2016)	Large data must be deposited in a database with identifier prior to publication and all data and materials must be available to public after publication.	For full data policy explanation see both "data deposition" and "availability of data and materials after publication" headings on the policy page.
PNAS	Materials and Data Availability (2016)	Authors required to make data, protocols and materials available to researchers and disclose where restrictions apply.	Under the Materials and Data Availability section on the editorial policies page there are type-specific data policies listed in addition to the general data sharing policy

http://dms.data.jhu.edu/data-management-resources/plan-research/funders-data-sharing-requirement/academic-journals-and-data-sharing-requirements/

Data Policies touch each of these areas…

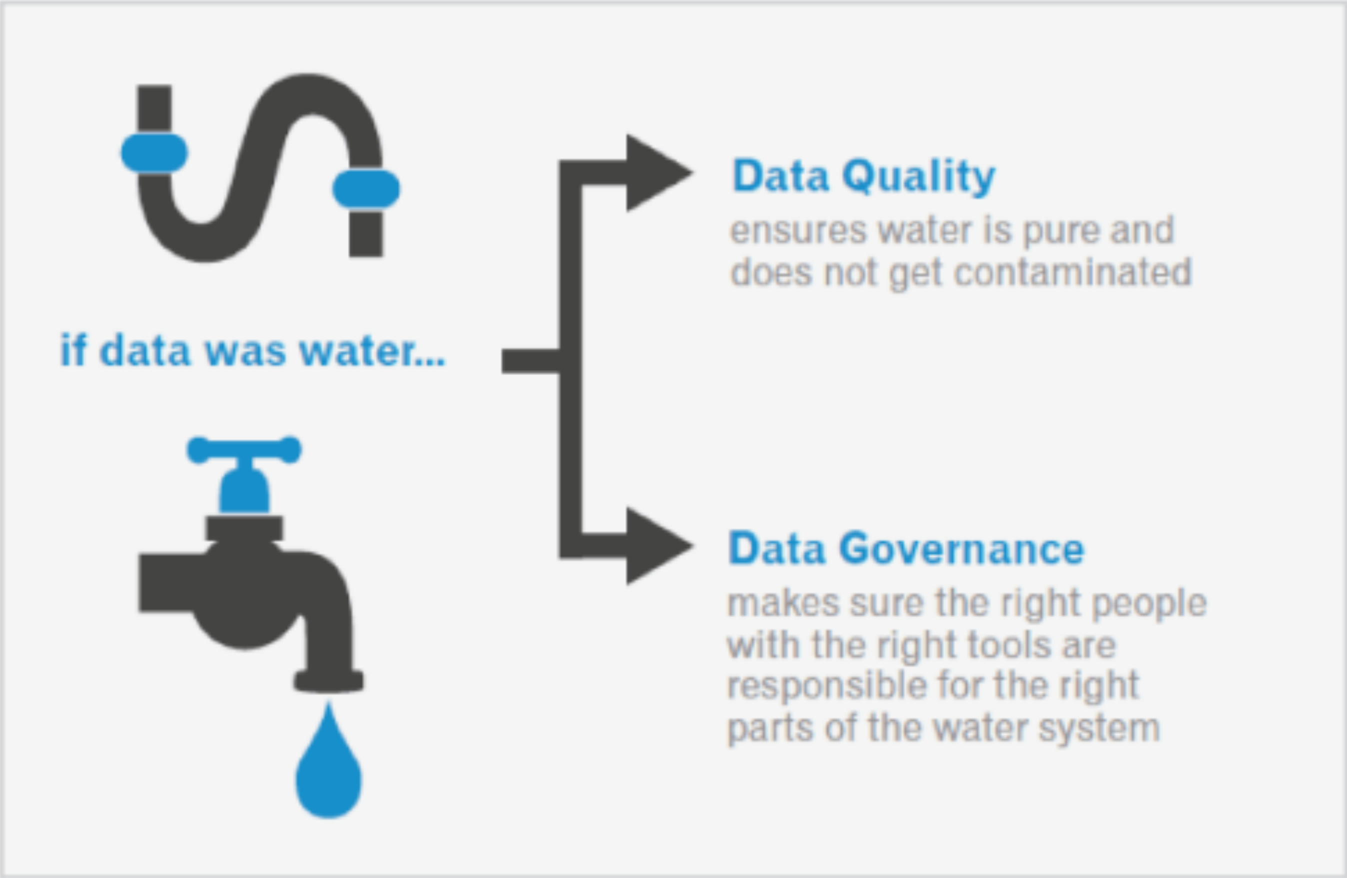
- **Data Curation**: Active and on-going management of data throughout a lifecycle of use (including reuse).
- **Data Management**: The activities involved in planning and handling data from the point of collection to storage in a long term repository.
- **Data Governance**: The policies required to put a curation program and management plan into action.

Data Governance

“…refers to who holds the decision rights and is held accountable for an organization's decision-making about its *data assets*.”

Khatri & Brown, 2010

1. Emphasizes data as an asset
2. Focuses on decision making, and in particular RIGHTS to make decisions



if data was water...

Data Quality
ensures water is pure and does not get contaminated

Data Governance
makes sure the right people with the right tools are responsible for the right parts of the water system

As a curator you will interact with nearly all of the people, systems, and users of an institution's data. Therefore, you will almost certainly have an important role to play in shaping data governance.

Governance Area	Decisions	Potential Decision Makers
Data Quality	What are standards for accuracy, timeliness, completeness and credibility? What metrics will be used to evaluate data quality?	Data creator, subject matter expert (data consumer), data curator
Data Access	How will users gain access to data? What security measures are taken to prevent privacy breaches? **What licenses will be used for data sharing and reuse?**	Data Curator, IT staff
Data Preservation	How often will data be checked for preservation assurance? What are the recommended preservation actions upon detecting data integrity issues? What is the frequency of creating back-ups?	Data Curator, IT staff
Metadata + Documentation	What documentation will be created? Using what standards? How often will documentation be updated?	Data creator, Data Curator

Data Licensing and Intellectual Property.

- Data vs Databases

    - **Content:** Data are often facts about the world (most of the time) and so one cannot claim property rights for factual data (e.g. whitepages, lists of English words, melting point of chemicals)

    - **Container:** Structured data has a logic that is designed and purposefully (intellectually) engineered- One can claim right to copyright protection for structure of data. (e.g. Sports statistics, bibliographic databases, chemical abstracts)

Data Licensing

- Rules for copyright claims and protections are jurisdictional - they vary greatly by country.

- Using, reusing and redistributing data without explicit permission can and will happen. LICENSES are the answer to creating some enforceable standards around who can use what data, when, and for what purposes.

- Licenses are differentiated by:

    - Attribution (Do content creators have to be acknowledged?)

    - Context and Reuse (Do data have to be shared with same license? Can data be commercialized? Recombined?)

4 Open Data Licenses

Public Domain Dedication and License (PDDL): This dedicates the database and its content to the public domain, free for everyone to use as they see fit.

Attribution License (ODC-By): Users are free to use the database and its content in new and different ways, provided they provide attribution to the source of the data and/or the database.

Open Database License (ODC-ODbL): ODbL stipulates that any subsequent use of the database must provide attribution, an unrestricted version of the new product must always be accessible, and any new products made using ODbL material must be distributed using the same terms. It is the most restrictive of all ODC licenses.

Public Domain mark (PDM): It is used to mark works that are in the public domain, and for which there are no known copyright or database restrictions. It is possible to flag factual data as PDM in a database, for example, in order to make it clear it is free to use.

https://data.research.cornell.edu/content/intellectual-property

How to Choose a License:

- Place - Where you are publishing / storing data is most important aspect of IP rights, and therefore licenses

- Funding / Institution - Most funders require open access, most business require Attribution and No Reuse, Most universities are IP rights holders

- Attribution + Context of Reuse - These are often tradeoffs - especially when it comes to data where completely open in public domain is the only REAL way to make data open and easily reusable. (e.g. With an attribution license, you can’t reuse data through an API)