Tidy Data + Data Packaging

LIS 598
Data Curation II

- **Curation**

- **Data**

- **Tidy data**: semantics and principles

- **Data packages**: Research objects, structural representations, BAGs and FAIR data

- **Curation** is the active and on-going maintenance of data throughout a lifecycle of use.

- Curation privileges

    - Efficiency in reusing data through cleaning, normalization, using standards to describe and document.

    - Fitness for purpose - the amount of time and effort needed to prepare data are requisite to the importance of a dataset.

    - Ease of computation

Differentiating data

- Type vs Role distinctions
  - Methods of production: observational, instrumental, experimental, hermeneutic
  - Form: text, video, image, tabular…
  - Logic: quantitative, qualitative
  - Value containment: **Ordinal** (first, second, …), **categorical**(small, medium, large), **continuous** (fractional numbers) … string or integer

Data Definition

(at arms length)

- Data are information objects playing the role of evidence.
    - The object may be a complex or simple **type**
    - The **role** may be temporary and poorly understood or enduring and well documented
    - The **evidence** may be contested or mundane

Tidy Data

Tidy Data

- Simple premise for structuring datasets to facilitate analysis:
    - Each variable is a column
    - Each observation is a row
    - Each type of **observational unit** is a table (because we should be able to join tables in a relational schema)

Tidy Data: Semantics

- Datasets are made of values (numbers or strings)
- Values are organized by variables (columns) and observations (rows).
- Variables are the attributes of a unit of observation.
- Observations are all values measured on the same unit across attributes.

Tidy Data: Semantics

- Imputing missing values: We want to be able to generally represent when something was not recorded (an error, a missed sample, etc) and when something was infeasible to record (e.g. pregnant men).

- Imputing missing values is domain specific. In general we want to avoid introducing different types of values to our datasets. (e.g. using `NA` or `/` in a dataset of numerical values; or risk confusing by entering `0` where the value isn't zero but absent)

- Even if there is strong agreement on how to represent missing or null values the approach taken should be documented.

Tidy Data: Principles

- “Data curation privileges ease of computation”
  - It is easier to describe **functional relationships** between variables (e.g., z is a linear combination of x and y, density is the ratio of weight to volume) than between rows
  - It is easier to make comparisons **between groups of observations** (e.g., average of group a vs. average of group b) than between groups of columns.

Data Packages

Data Packages

- Research Objects: Why read a paper from 2010 about the future of data packaging?

- Because providing access to complex aggregations of information objects is still very hard (i.e. it could have been written in 2019!)

- Research objects are supposed to structurally package **data** + **methods** (software / instrument / etc) + **people** (institutions + individuals)

Data Packages

- Let’s forget about the methods and people for a moment… What about structurally representing what we know about a dataset?

	Order of Birth	Weight	Sex
Alvin	First	4.8	M
Simon	Second	4.8	M
Theodore	Third	4.8	M

- That is more than just descriptive metadata… we want that kind of information to “travel” with our data so that it is easily accessible and reusable

Data Packages

- Data packages are emerging through different “specifications”
- These create a kind of template or set of rules for what information (structured and unstructured metadata), what kinds of encodings (e.g. XML) and what type of wrapper (e.g. a zip file) should be used to “package” a dataset
- The good news: There are no right or wrong answers
- The bad news: There are no right or wrong answers

Data Packages

- These create a kind of template or set of rules for what information (structured and unstructured metadata), what kinds of encodings (e.g. XML) and what **type of wrapper** (e.g. a zip file) should be used to “package” a dataset

    - **BAGit** is an archivist Zip file.

    - Bag = Payload (content) and Tags (metadata files).

    - Every bag must contain a Mainfest File that creates a map between all payloads and all tags with a corresponding checksum (integrity)

Data Packages: TLDR

- Almost all data packaging specifications are extensions of BAGit
- They are attempting to describe all of the necessary payload and tag files that SHOULD accompany a dataset
- Why is this important to curation? Bags are like a cargo container - it doesn’t make what’s inside necessarily better, but it makes the process of transferring, storing, and discovering much easier
- Another way to say this is that data become more **F.A.I.R** when packaged.