03 Describing Data

DMV
suzanne.little@dcu.ie

Today

Recap: Data Sources

What is metadata?

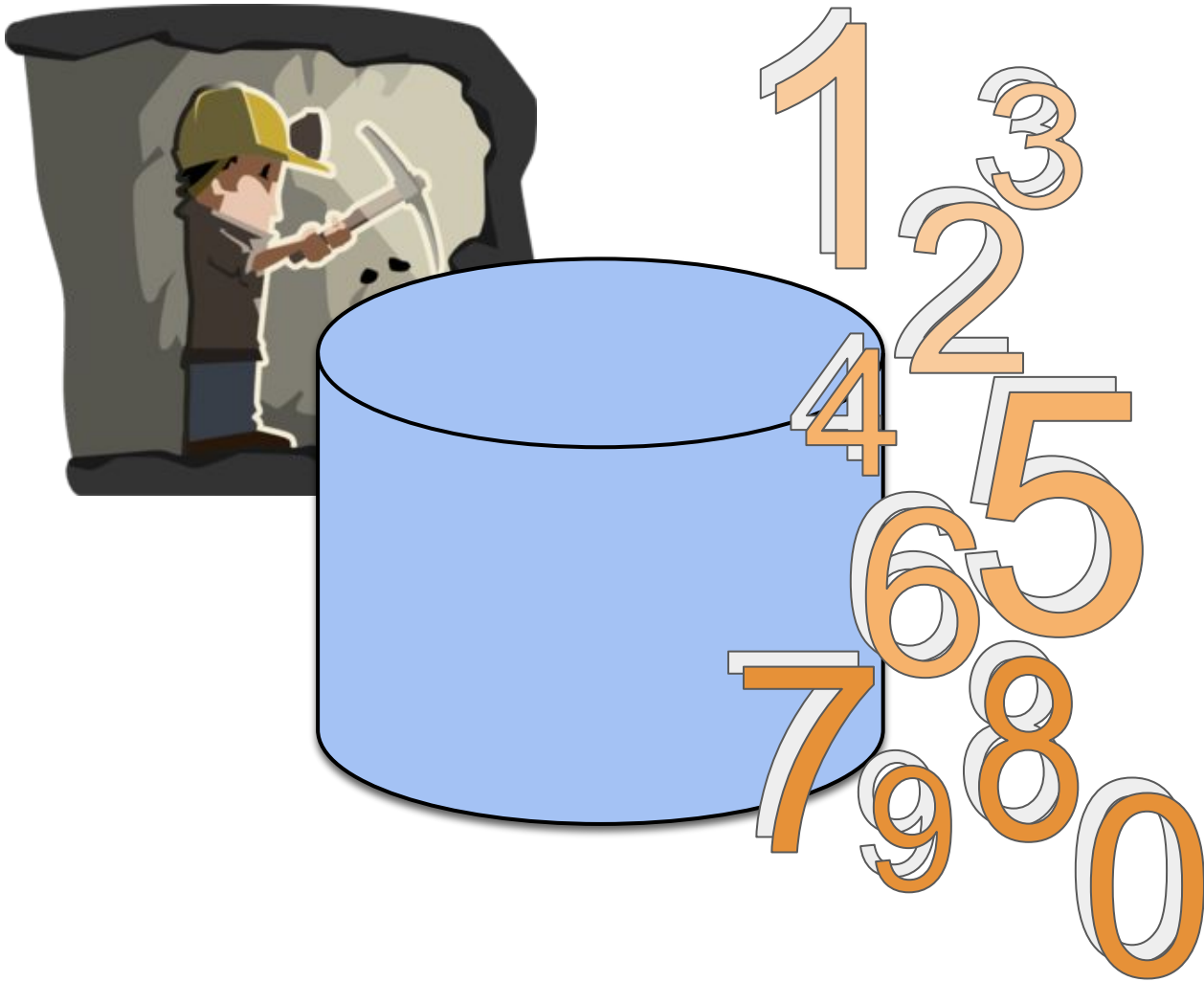
Metadata exercise

Assignment

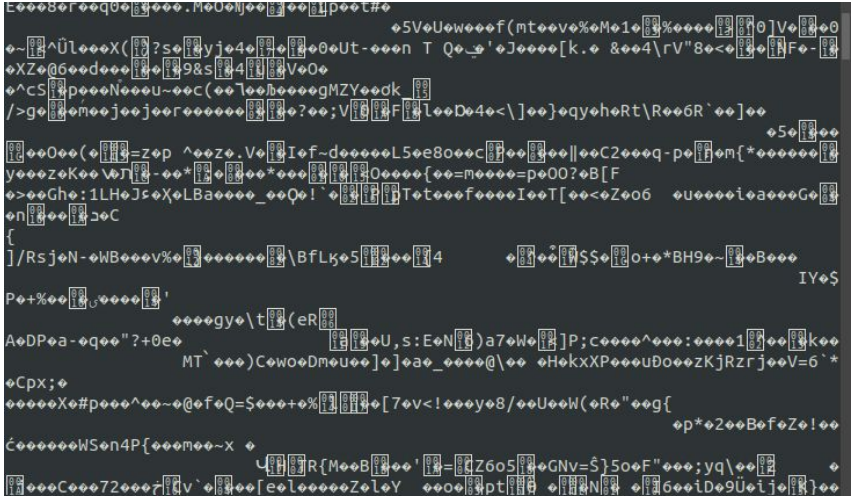
Big data & an example

Data sources

- Files
- Databases
- “The Internet”
- Open Data



Data sources: Files



- Text or Binary
- Open or Proprietary
- Tabulated data - CSV, TSV, DB
- Other text data
  - JSON -- JavaScript Object Notation
  - XHTML -- web pages
  - KML (https://developers.google.com/kml/documentation/kml_tut?csw=1)
  - many other XML-based! (YAML …)
  - Specialist data formats (GDP, ASX etc.) -- may be proprietary
- List of file formats … (https://en.wikipedia.org/wiki/List_of_file_formats)

10 common data science file formats

1. CSV (& TSV & TXT)
2. JSON
3. XLS or XLSX
4. SQL
5. PDF
6. HTML
7. DOC or DOCX
8. HDF5
9. ZIP (or GZ or TGZ)
10. XML

How many do you know how to process?

Data sources: Databases

More to come!

- Traditional relational db: Oracle, MySQL, Postgres, etc.
  - Tables (“relations”) of rows and columns
  - Unique key per row
  - Links between rows (“foreign key”)
  - Optimise structures (the database schema)
  - Stored procedures (queries) to speed up responses
  - Most commonly use SQL - Structured Query Language
    - `SELECT CustomerName,City FROM Customers;`
    - `SELECT CustomerName,Age FROM Customers WHERE City='Dublin';`
- In memory databases: SAP Hana (http://hana.sap.com/abouthana.html)
- NoSQL, document, column, graph, etc.

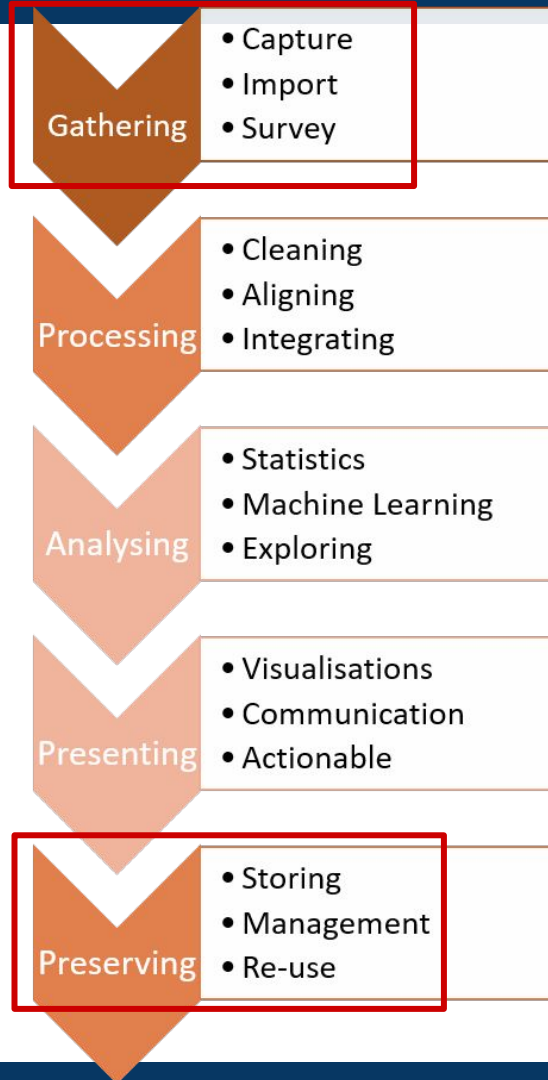
Data sources: the Internet

- Crawlers or spiders
  - Scraping data from semi-structured sources
    - Parse HTML
    - Match Patterns to extract data
    - Identify links (repeat)
- URL
  - Files and databases on the web
  - Many libraries and apps will accept either a local path or url
- How many file formats?

Exercise & Information on loop

Data

- Data is collected information
- Where does data come from?
  - Files
  - The Internet
  - Databases



Metadata

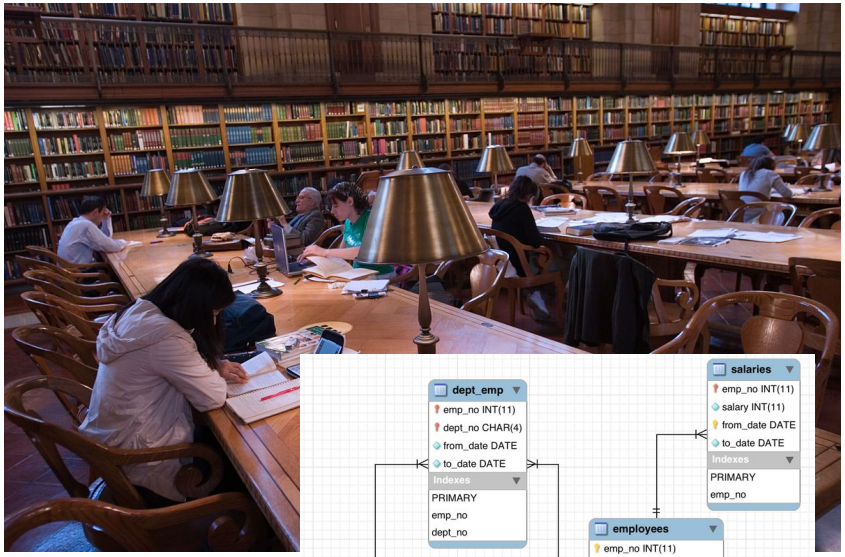
What is metadata?

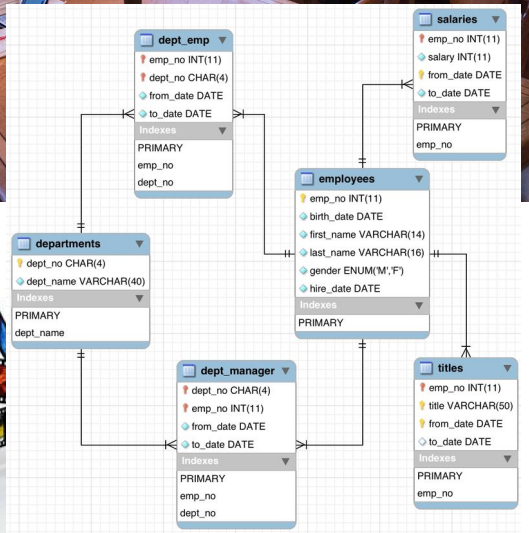
“data about data”

“inferior type of cataloging”

Why is it useful?

"*Information on the organization of the data, data domains, and the relationship between them*"
(Baeza-Yates)





Metadata

Campo dei Miracoli
La Torre di Pisa
Field of Miracles, Pisa, Italy
Leaning Tower
Pisa, Italy
Sunny day in Pisa
My holidays
Old building

ExifVersion 0220
CompressedBitsPerPixel (5, 1)
DateTimeOriginal 2006:10:01 17:50:11
DateTimeDigitized 2006:10:01 17:50:11
MaxApertureValue (107, 32)
MeteringMode 5
Flash 80
FocalLength (7272, 1000)
ApertureValue (128, 32)
FocalPlaneXResolution (3264000, 286)
Make Canon
Model Canon PowerShot S80
Orientation 1
YCbCrPositioning 1
SensingMethod 2
XResolution (180, 1)
YResolution (180, 1)
ExposureTime (1, 640)
ColorSpace 1
FNumber (40, 10)
DateTime 2006:10:01 17:50:11
ExifImageWidth 3264
FocalPlaneYResolution (2448000, 214)
ExifImageHeight 2448

https://readexifdata.com/

Metadata

Why is it useful?

Find, Locate, Identify, Select, Obtain, Navigate

Use data

Rights and data management

Doctorow on Meta-utopia

- People lie
- People are lazy
- People are stupid
- People delude themselves
- Schemas aren't neutral
- Metrics distort or limit
- There's more than one way!

http://www.well.com/~doctorow/metacrap.htm

Types of metadata

DESCRIPTIVE metadata

what the information object is about; inherently intrinsic properties

ADMINISTRATIVE metadata

who, what, why, where of the object's creation and management; inherently extrinsic properties

STRUCTURAL metadata

information about the structure, format, and composition of the thing being described; can be intrinsic or extrinsic

Exercise

Describe your favourite book or movie.

What qualities did you use? Title, Author, Year, Genre, Characters ?

How would you use the “metadata” to Find, Locate, Identify, Select, Obtain, Navigate

Exercise: Metadata about your favourite book

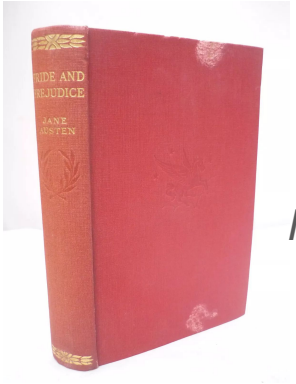
Title: Pride and Prejudice

Author: Jane Austen

Summary: “Pride and Prejudice is an 1813 novel of manners by English author Jane Austen. The novel follows the character development of Elizabeth Bennet, the protagonist of the book, who learns about the repercussions of hasty judgments and comes to appreciate the difference between superficial goodness and actual goodness.”

Description: 1916 red hardback printing

Location: storage box 12



Not my copy!

What is being described?

Two separate dimensions the metadata can be associated with:

- Abstraction hierarchy
- Granularity

An example abstraction hierarchy

WORK - an abstract entity; the distinct intellectual or artistic creation; it has no single material manifestation

EXPRESSION - the multiple realizations of a work in some particular medium or notation, where it can actually be perceived

MANIFESTATION - each of the formats of an expression that have the same appearance; but not necessarily the same implementation

ITEM - a single exemplar of a manifestation; if we distinguish this level it is because otherwise identical manifestations have some differentiation

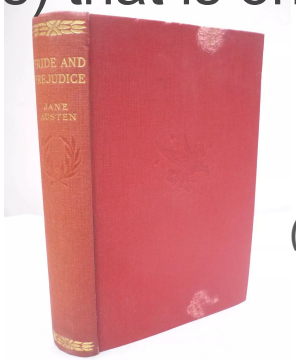
Abstraction hierarchy

WORK: “Pride & Prejudice” by Jane Austen

EXPRESSION: a book (c.f., movie, comic, audio recording, ??)

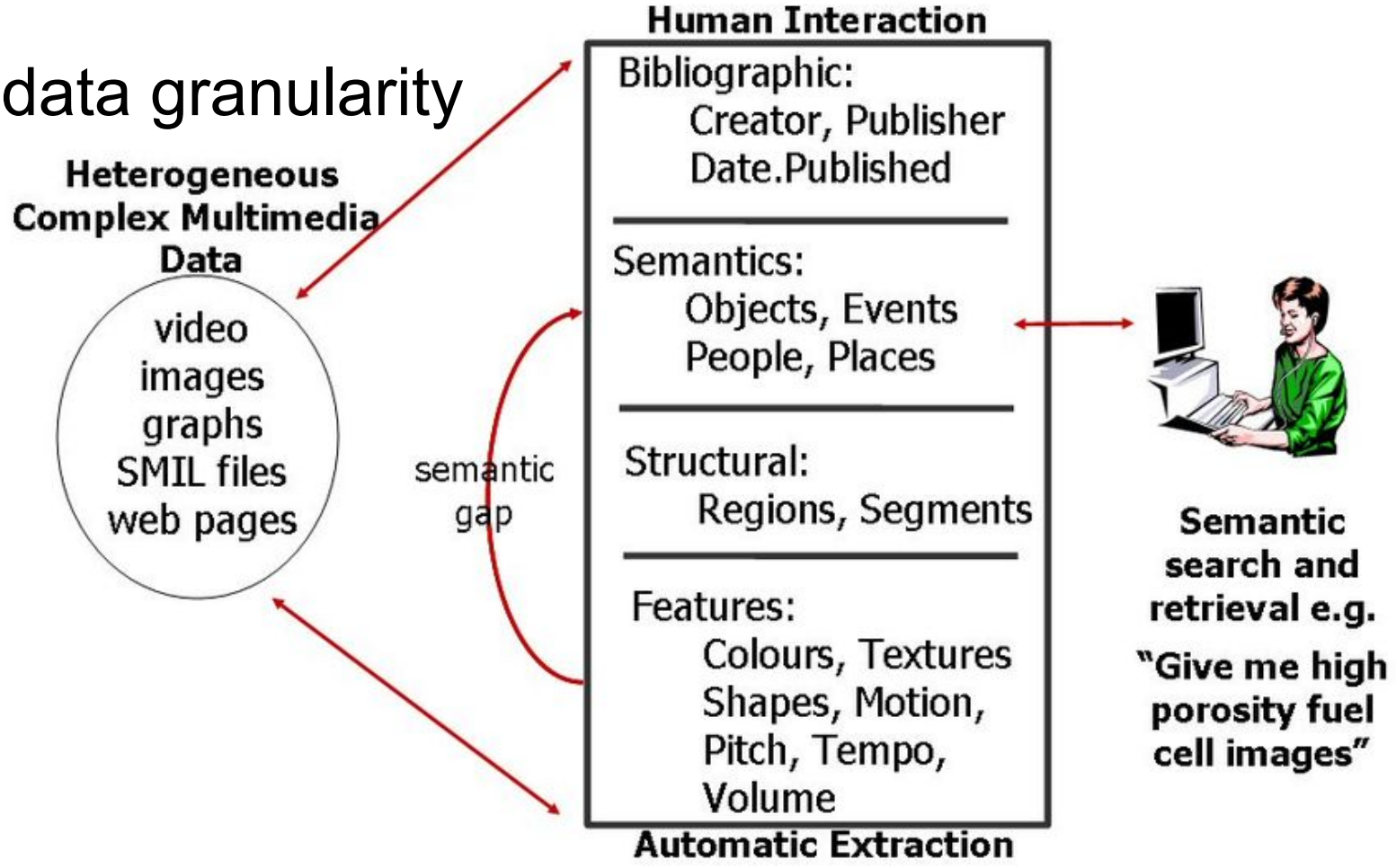
MANIFESTATION: 1916 edition, hardback

ITEM: my copy (with the tea stained pages) that is on my bookshelf



(Not my copy!)

Metadata granularity



So how much metadata do we need?

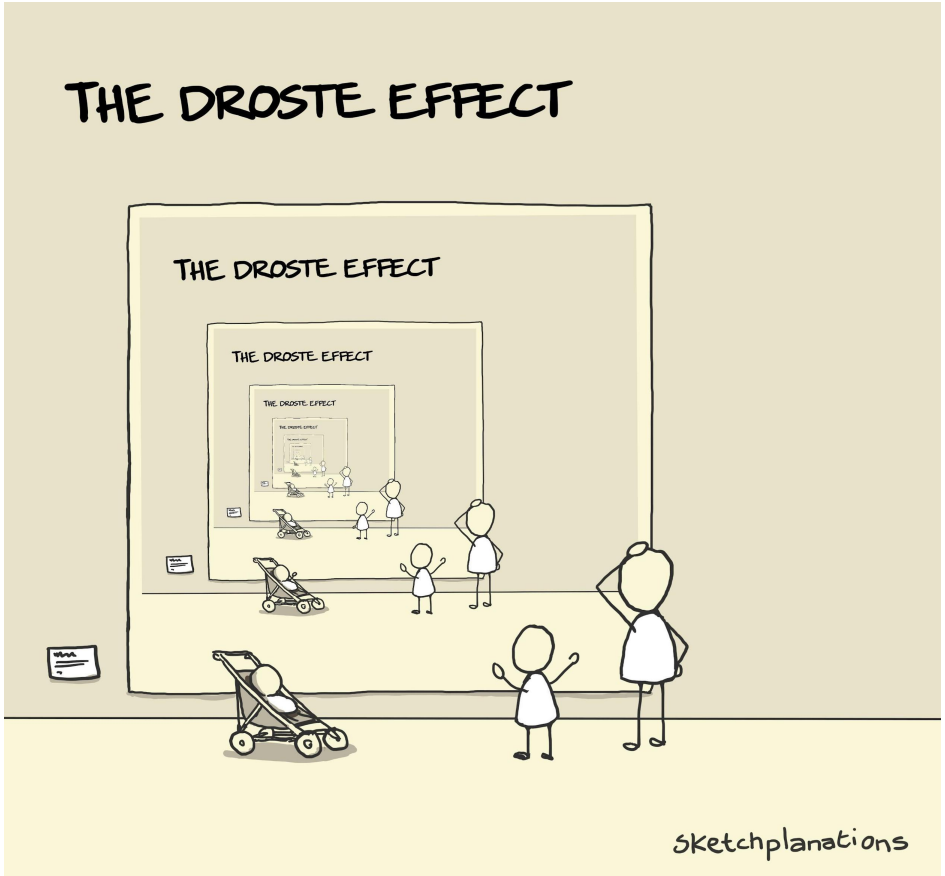
Consider the **tradeoffs** between

organization (adding, duplicate detection, storage) and retrieval (query, search)

Not all documents / resources need the same amount of metadata

Could someone else understand and use your dataset?

When does metadata become a dataset?



Data

Metadata

Meta metadata

Meta meta metadata

???

https://sketchplanations.com/the-droste-effect

Where does metadata come from?

Simple
- EXIF, document headers, time stamps, “ad hoc” labels

Structured
- Adhering to a standard

Professional
- Created by a librarian or curator

Crowd Sourced
- hashtags, comments

Metadata standards

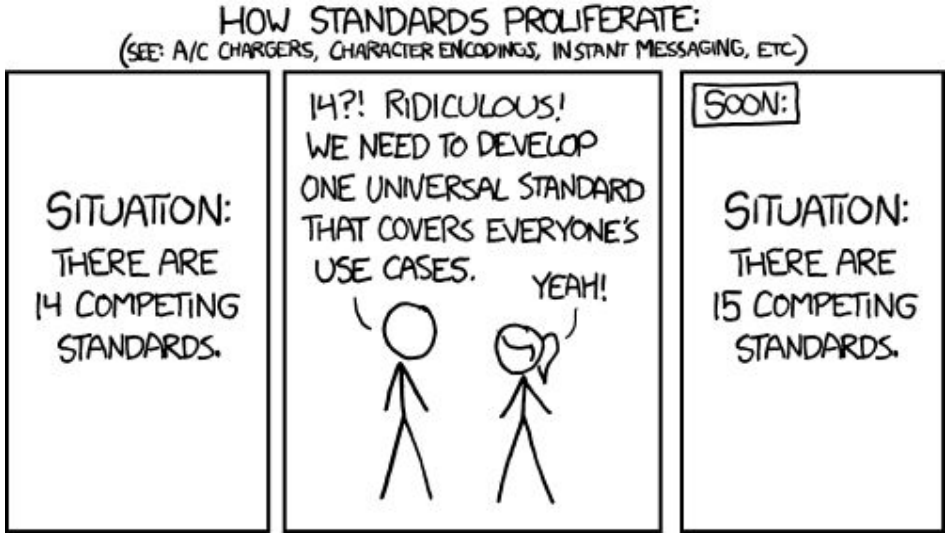
Specify the structure - Syntax or Content?

Many exist, not always compatible

MARC, Dublin Core, MPEG-7

Publishers: ISO, RFC, IEEE, W3C

https://xkcd.com/927/



Example: Dublin Core

Proposed in 1995 as standard set of metadata elements, simple enough to be supplied by document's author rather than professional curator

DC is the set of elements, described abstractly and all optional

Semantics of DC established by international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields

Specifications of how to use it in numerous syntaxes (especially XML and RDF) and languages

Dublin Core

TITLE
IDENTIFIER
SUBJECT
CREATOR - makes the content
CONTRIBUTOR
PUBLISHER
DATE
DESCRIPTION
LANGUAGE
TYPE - nature or genre
RIGHTS
SOURCE - if derived from something
RELATION
COVERAGE
AUDIENCE

Go back to your movie or book description. Can you see what a Dublin Core description would look like?

http://www.dublincore.org/documents/2000/07/16/usageguide/generic/

Problems

“Some information may appear to belong in more than one metadata element”

“There is potential semantic overlap between some elements”

“There will occasionally be some judgment required from the person assigning the metadata”

https://catalog.data.gov/dataset/consumer-complaint-database

Open Data

Most have some sort of metadata

This is data about the dataset

Formal and structured is better

Where else do you use metadata in your daily life?

Metadata in last week's lab

title	documentId	revisionId
Introduction to Jupyter Notebooks	1B6jwll2zgKFOcXszeFNmT0vpaH-07a45chnLKJjHYGE	ALBJ4Lt69L1cCokXaFjP8088EX4yZyAw2IN8B8odoebJf9Rga2kiGxYVeF4k7uc1fkCveXqqIV

If you got to the DataSampler exercise, this is an illustration of what happens when there is no metadata for a dataset (or even useful column headings)!

Today

Open data exercise

What is metadata?

Metadata exercise

<break>

Assignment

Big data & an example

CSC1143 Visualisation Assignment

Visualisation Assignment - loop has specification

Assignment due by Friday Nov 29th 23:59 (end of teaching)

In pairs, create a visualisation that **illustrates a point, answers a question or tells a story (explanatory)**.

Short report (following the template) plus screencast video showing and describing your visualisation

A simple chart on limited data is not sufficient. Remember it is worth 25%. Marking criteria are included in the description.

Use any tool or tools but remember you want to demonstrate your skill.

Assignment Marking Criteria

1. Dataset [30%]:
    a. “big” data
    b. showing either data cleaning **or** transformation **or** integration.
2. Visualisation [50%]:
    a. suitable graph choice;
    b. difficulty level;
    c. good design/style;
    d. use of interactivity **or** animation.
3. Report [20%]:
    a. follows instructions and template;
    b. good abstract;
    c. critique and reflection.

Assignment - good things to know

Present **1 (one)** explanatory graph. No dashboards, please.

Keep your report concise, following the template, and your video screencast brief.

Not allowed to use the MovieLens, Chicago Crime, New York Traffic or derivative datasets.

Critique & reflect is very important.

Objective: Assessing your data processing, graph *selection* and *design* skills.

Planning your visualisation assignment **[OPTIONAL]**

Link in Loop

Submit a brief description (< 500 words) of your idea.

What data? What question or idea? What tools?

I'll try to give individual feedback but may end up summarising for the class.

Submit before Oct 30th (so I can comment in class)