Introduction to Data Curation

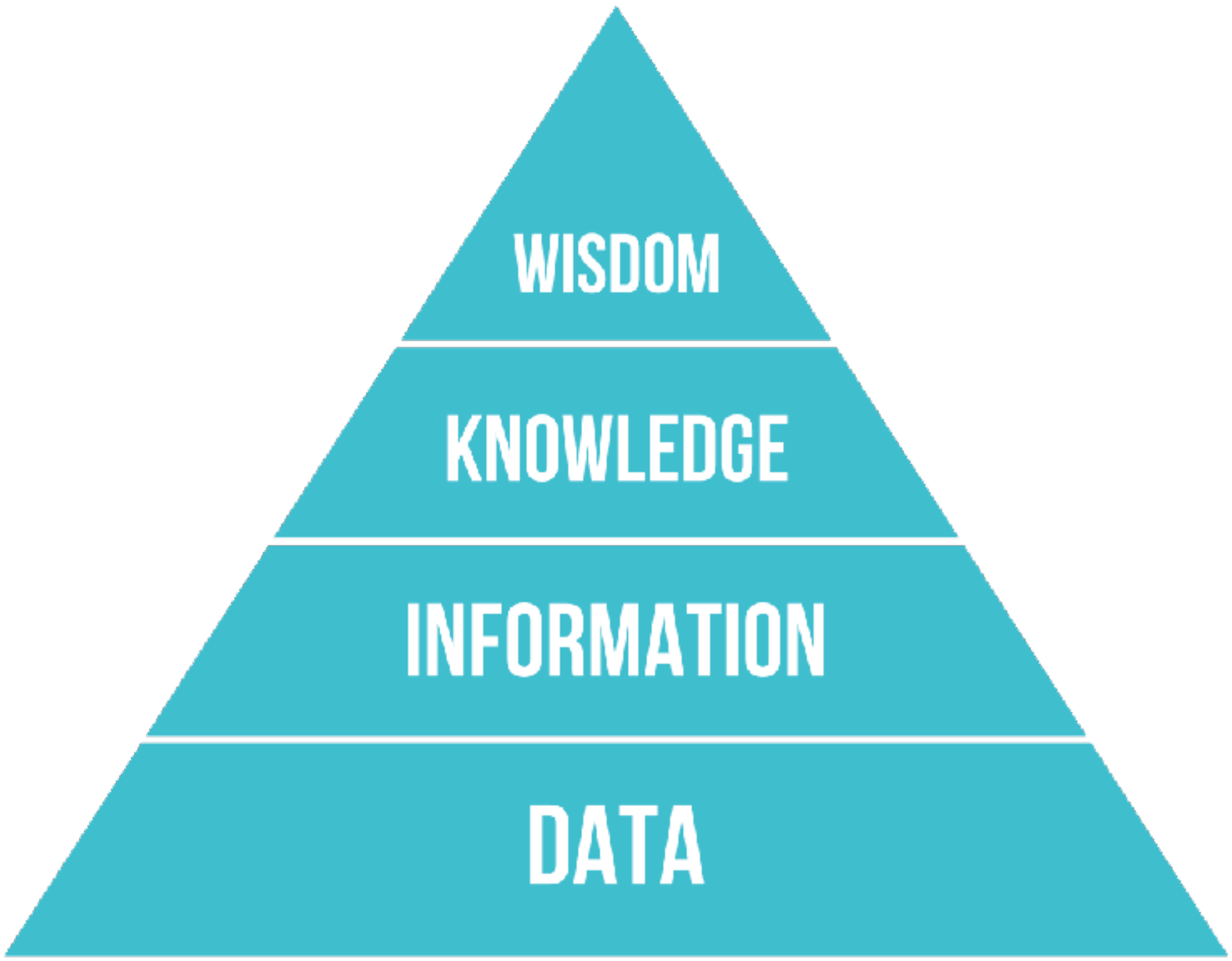
01.23.2018

- Data
- Curation
- Data Curation
- Data Quality and Data Cleaning
- Metadata and forms of Documentation

Declarative	Procedural
I know **that** something is the case…	I know **how** to do something…
I know that longitude is the geographic coordinate that specifies the east-west position of a point on the Earth's surface.	I know how to locate a point on a map given the Longitude and Latitude coordinates…
Often precedes and informs procedural knowledge	Often follows from declarative knowledge (but not always)
I know that XML is an encoding standard for structuring data…	I know how (and when it is appropriate) to create well-formed XML encodings of free text data.

For this lecture, let's assume that….

Data Curation is the active and ongoing management of **data** throughout a lifecycle of use, including its reuse in unanticipated contexts.


WISDOM
KNOWLEDGE
INFORMATION
DATA

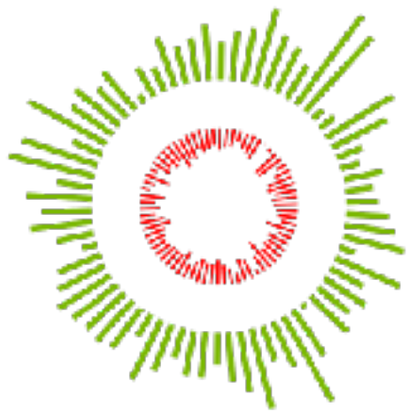
Research Data

"The data, records, files or other evidence, irrespective of their content or form (e.g. in print, digital, physical or other forms), that comprise research observations, findings or outcomes, including primary materials and analysed data."

AUSTRALIAN NATIONAL DATA SERVICE

Open Data

"Open data is data that can be freely used, shared and built-on by anyone, anywhere, for any purpose."



For this lecture, let's assume that….

Data are various **types** of information objects playing the **role** of evidence.

Evidence of…
Infection (Public Health)
Patterns of Consumer Behavior (Business)
Weather (Atmosphere)

Types vs Roles

Type vs Role
distinctions

Type:
Donald Trump is a person.

Role:
Donald Trump is POTUS

Type vs Role
distinctions

Type:
XML

Role:
XML as bibliographic data

Data are various **types** of information objects playing the **role** of evidence.

Types of Data
(by file format)

XML

Databases

Flat Files

EDI

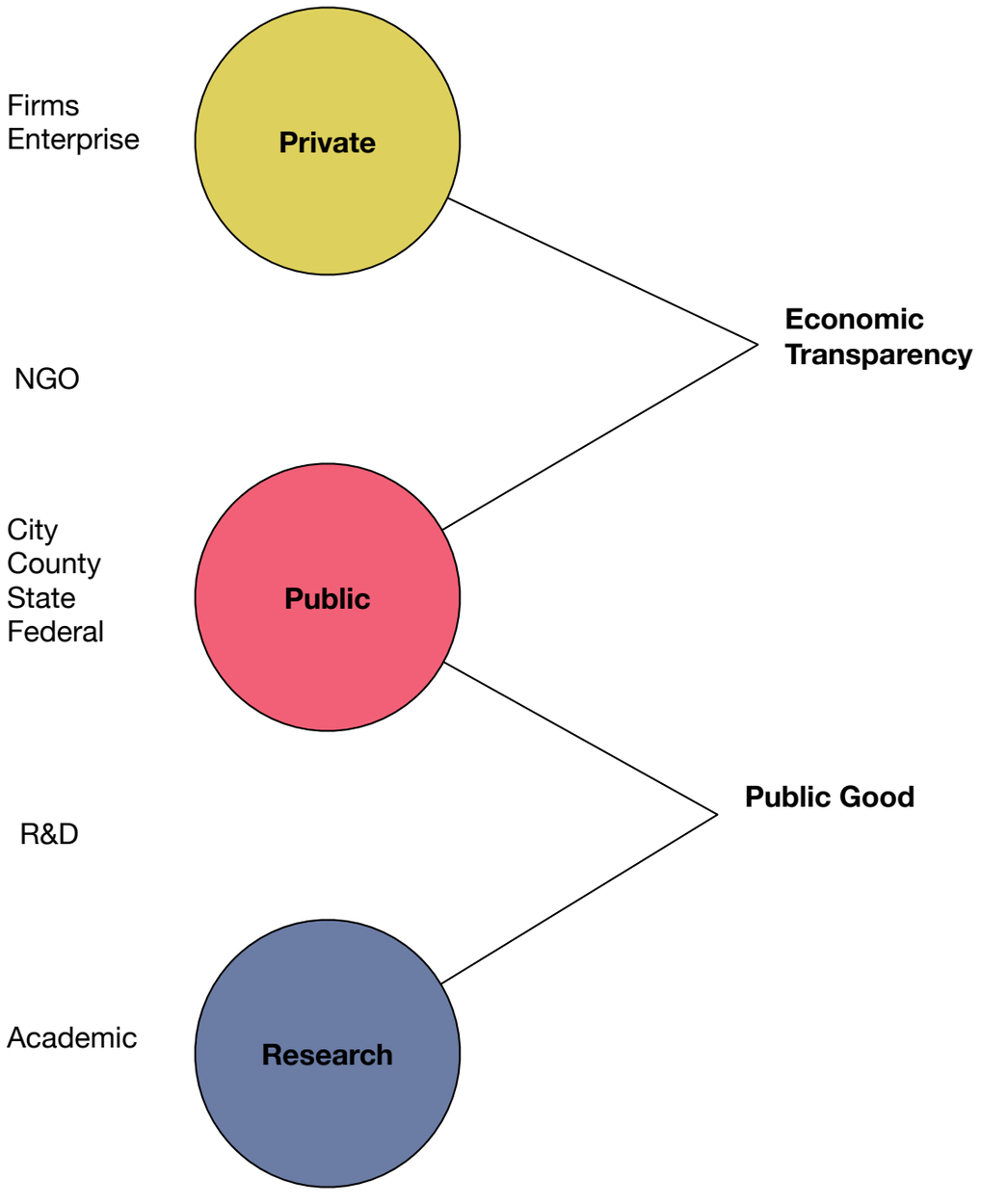
Excel

XBRL

JSON

Web Services

Types of Data
(by sector)



Roles

OPEN ACCESS Freely available online

PLOS ONE

The Role of Gender in Scholarly Authorship

Jevin D. West[1]*, Jennifer Jacquet[2], Molly M. King[3], Shelley J. Correll[4], Carl T. Bergstrom[1,4]

1 Department of Biology, University of Washington, Seattle, Washington, United States of America, 2 Environmental Studies, New York University, New York, New York, United States of America, 3 Department of Sociology, Stanford University, Stanford, California, United States of America, 4 Santa Fe Institute, Santa Fe, New Mexico, United States of America

Abstract

Gender disparities appear to be decreasing in academia according to a number of metrics, such as grant funding, hiring, acceptance at scholarly journals, and productivity, and it might be tempting to think that gender inequity will soon be a problem of the past. However, a large-scale analysis based on over eight million papers across the natural sciences, social sciences, and humanities reveals a number of understated and persistent ways in which gender inequities remain. For instance, even where raw publication counts seem to be equal between genders, close inspection reveals that, in certain fields, men predominate in the prestigious first and last author positions. Moreover, women are significantly underrepresented as authors of single-authored papers. Academics should be aware of the subtle ways that gender disparities can occur in scholarly authorship.

Citation: West JD, Jacquet J, King MM, Correll SJ, Bergstrom CT (2013) The Role of Gender in Scholarly Authorship. PLoS ONE 8(7): e66212. doi:10.1371/journal.pone.0066212

Editor: Lilach Hadany, Tel Aviv University, Israel

Received January 25, 2013; **Accepted** May 7, 2013; **Published** July 22, 2013

Funding: This work was supported in part by NSF grant SBE-0915005 to CTB, NSF Graduate Research Fellowship grant DGE-1147470 to MMK, and a generous gift from JSTOR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jevinw@u.washington.edu

Introduction

Gender inequities and gender biases persist in higher education. After decades of high female enrollment in most PhD fields, women represent one-quarter of full professors and earn on average 80% of the salary of men in comparable positions [1]. A recent report [2] surveyed 1000 faculty across six science and engineering disciplines and found men publish significantly more in chemistry and mathematics, while women publish more in electrical engineering (here were no significant differences found in biology, civil engineering, and physics). A recent experiment tested the role of gender in hiring by asking 127 science faculty to evaluate potential lab manager applications and found faculty gave identical applications higher scores if the applicant had a male name [3]. Another recent analysis of commissioned articles in two prestigious journals published in 2010 and 2011 showed that women scientists are underrepresented; for instance, women wrote just 3.8% of earth and environmental sciences articles for Nature News & Views, although they represent 20% of the scientists in this discipline [4]. With the use of alphabetical authorship listings declining over time [5], and given the complexity of evaluating intellectual contributions [6] in increasingly collaborative efforts, understanding patterns of authorship order becomes increasingly important.

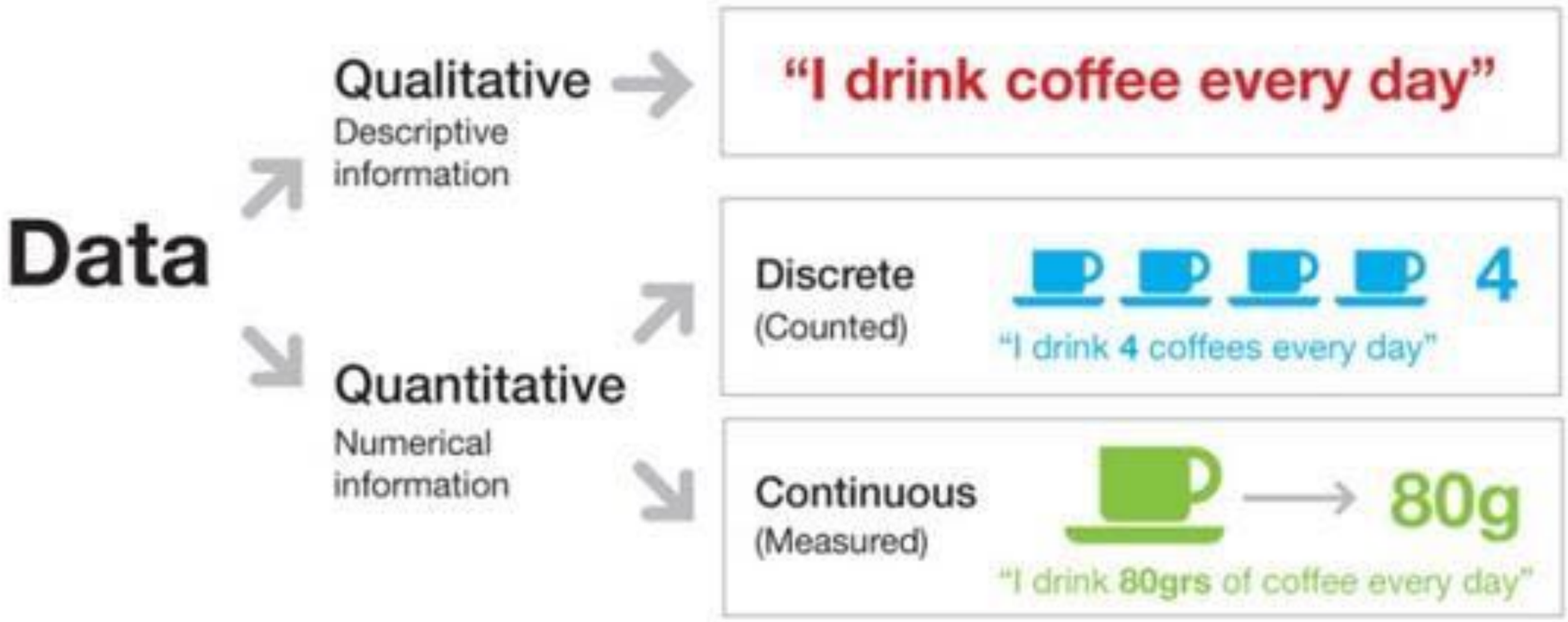
Here we use the JSTOR corpus—a body of academic papers from a range of scholarly disciplines spanning five centuries—to examine trends in the gender composition of academic authorship through time. We pay particular attention to authorship order, given that first and sometimes last author publications are at least as important as raw publication counts for hiring, promotion, and tenure, particularly in scientific fields [7]. Studies of authorship in the medical literature reveal, for instance, that women have been historically underrepresented in the prestige positions of first and last author, and that while discrepancies have recently declined in the first author position, women remain underrepresented as last authors [8,9,10,11]. To view authorship patterns in their disciplinarily context, we use a network-based community detection approach to categorize hierarchically each paper in our study corpus. This yields a hierarchical classification of all papers in our study and allows us to study and compare patterns of gender representation in individual fields of any size and scale.

Methods

The JSTOR corpus

The JSTOR corpus (http://www.jstor.org) is a digital archiver of published scholarly research that spans the sciences and humanities from 1545 to the present day. At the time of this analysis, the JSTOR corpus comprised 8.3 million documents ranging from 1545 until early 2011, including 4.2 million research articles. Approximately 1.8 million of these documents (91% of which are research articles) cite or are cited by other documents in the JSTOR corpus and thus are amenable to network analysis. We call this group the "JSTOR network dataset". Moreover 94% of these 1.8 million articles are part of a single giant component of the citation network, such that any of these articles can be reached from any other by following citation trails forwards and backwards. We restrict our analysis to the JSTOR network dataset because this is the portion of the JSTOR corpus that we can hierarchically categorize using citation information. For a list of the main fields available in JSTOR dataset, see Table 1. The gender composition of the identified authors in the network dataset (21.9% female) is close to that of the identified authors in the entire corpus (20.8% percent).

Roles

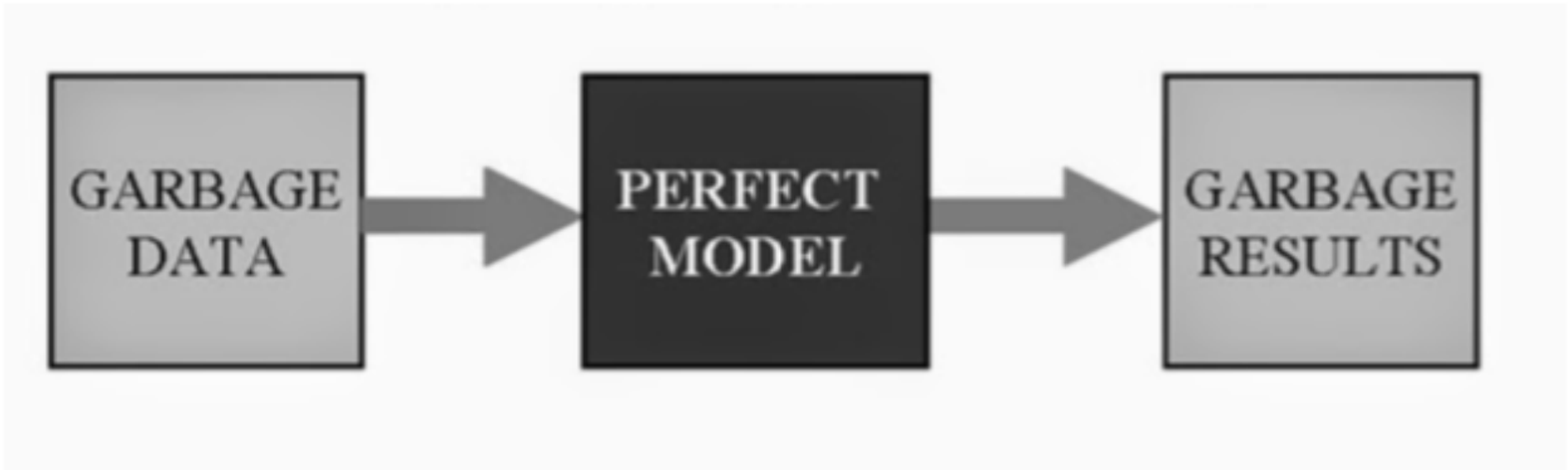

Data
Qualitative
Descriptive information
"I drink coffee every day"
Quantitative
Numerical information
Discrete
(Counted)
4
"I drink 4 coffees every day"
Continuous
(Measured)
80g
"I drink 80grs of coffee every day"

Curation

Data Curation is the active and ongoing management of data throughout **a lifecycle of use**, including its reuse in unanticipated contexts.

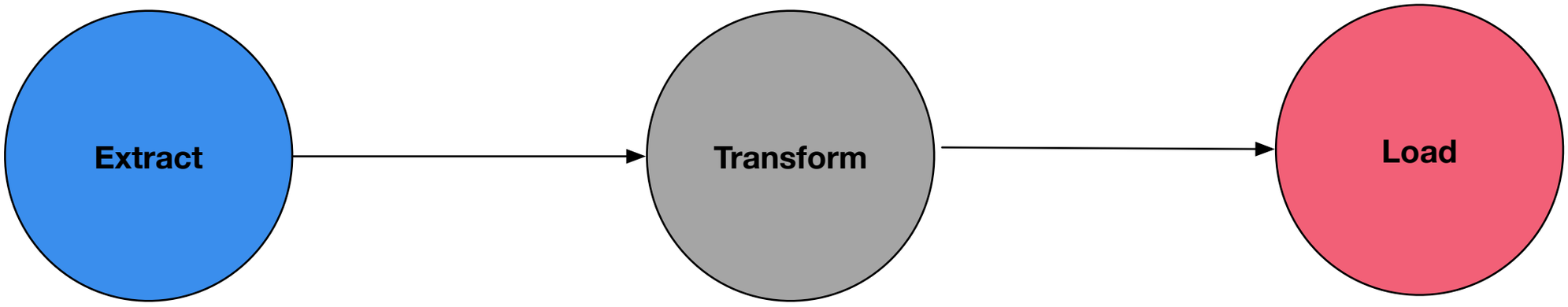
Data Curation

Old computer science saying ”Garbage in = Garbage out”

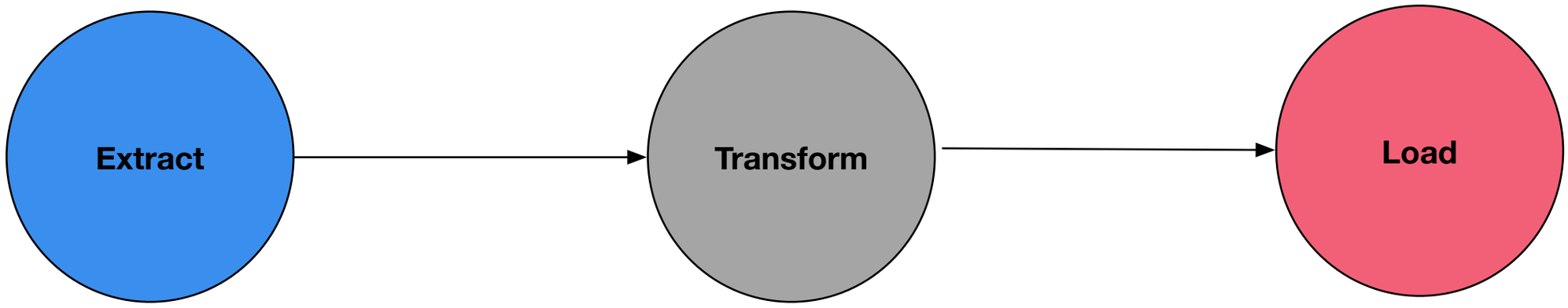


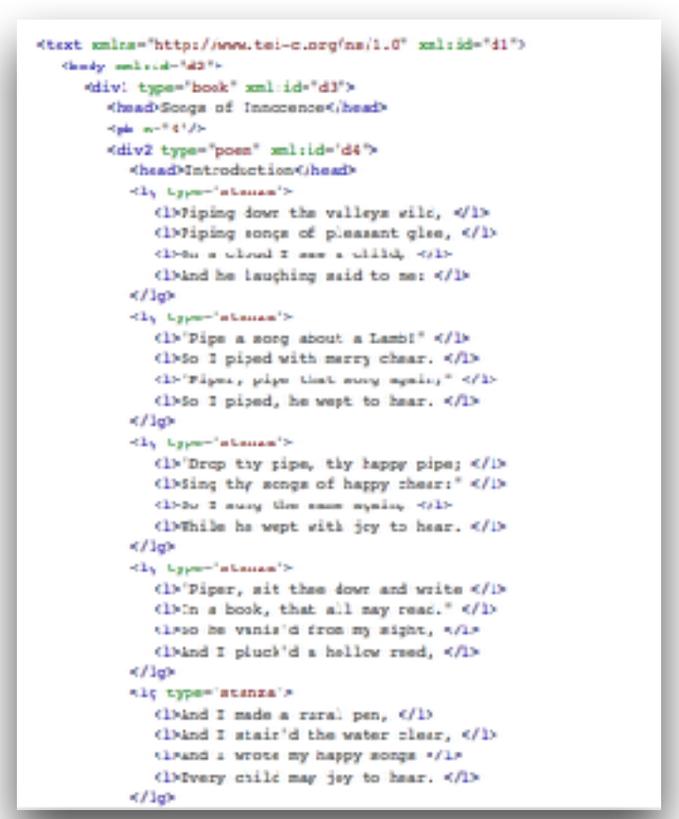
Data curation says “Quality in = Quality out”

ETL: Simplest form of Curation Workflow



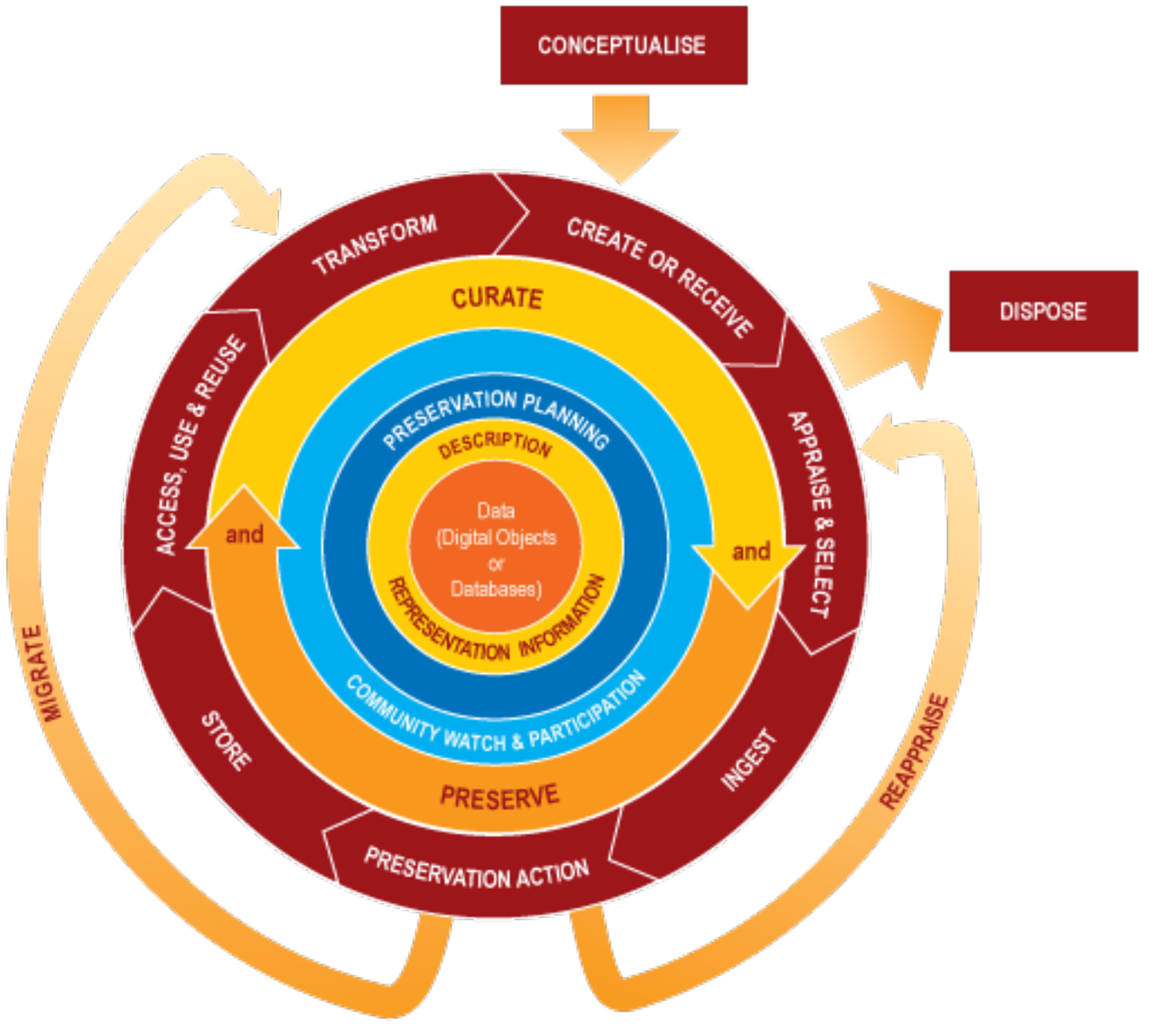
ETL: Simplest form of Curation Workflow



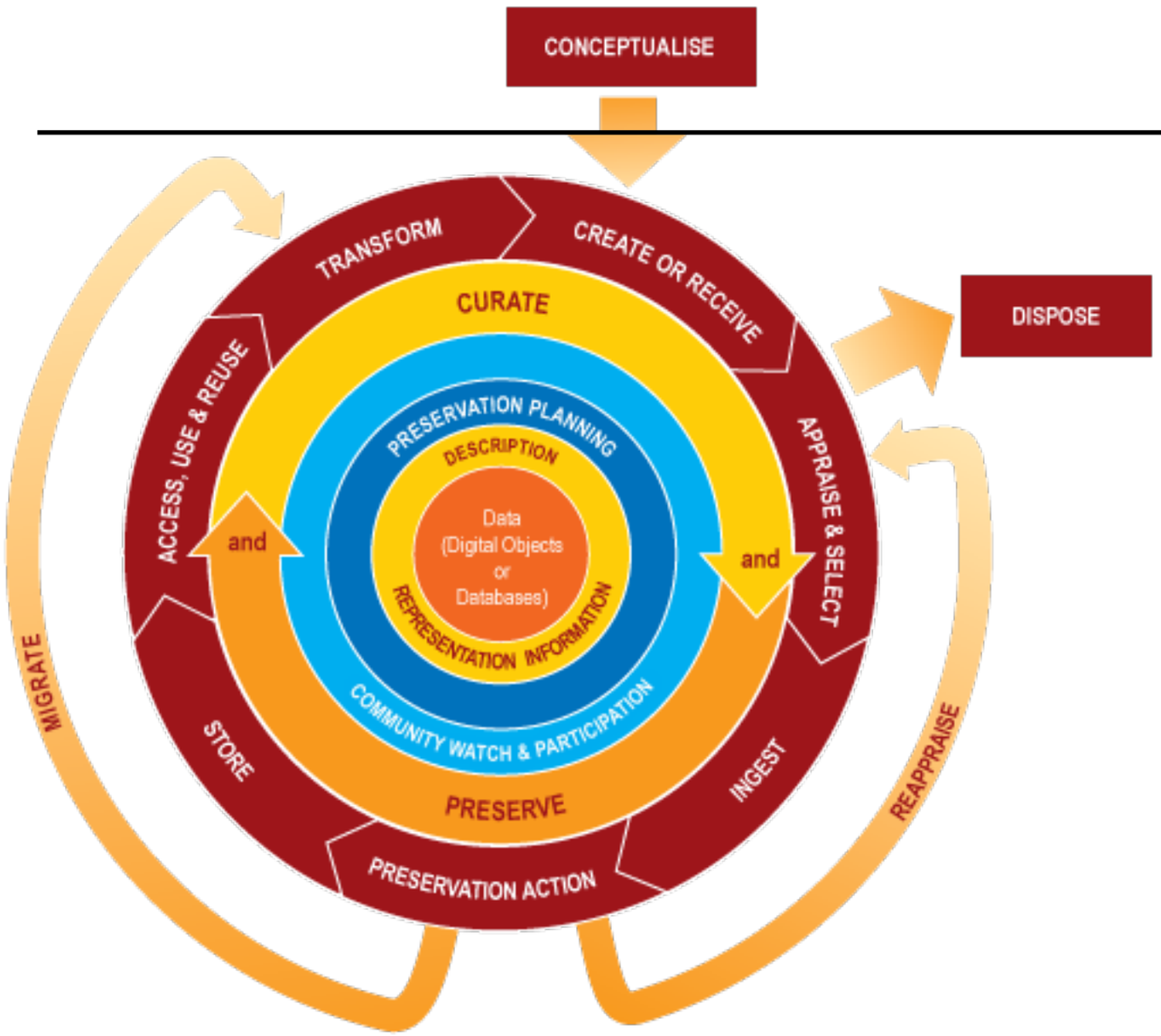


```
<text xmlns="http://www.tei-c.org/ns/1.0" xml:id="d1">
  <body xml:id="d2">
    <div1 type="book" xml:id="d3">
      <head>Songs of Innocence</head>
      <pb n="4"/>
      <div2 type="poem" xml:id="d4">
        <head>Introduction</head>
        <lg type="stanza">
          <l>Piping down the valleys wild, </l>
          <l>Piping songs of pleasant glee, </l>
          <l>On a cloud I saw a child, </l>
          <l>And he laughing said to me: </l>
        </lg>
        <lg type="stanza">
          <l>"Pipe a song about a Lamb!" </l>
          <l>So I piped with merry chear. </l>
          <l>"Piper, pipe that song again," </l>
          <l>So I piped, he wept to hear. </l>
        </lg>
        <lg type="stanza">
          <l>"Drop thy pipe, thy happy pipe; </l>
          <l>Sing thy songs of happy chear:" </l>
          <l>So I sung the same again, </l>
          <l>While he wept with joy to hear. </l>
        </lg>
        <lg type="stanza">
          <l>"Piper, sit thee down and write </l>
          <l>In a book, that all may read." </l>
          <l>So he vanish'd from my sight, </l>
          <l>And I pluck'd a hollow reed, </l>
        </lg>
        <lg type="stanza">
          <l>And I made a rural pen, </l>
          <l>And I stain'd the water clear, </l>
          <l>And I wrote my happy songs </l>
          <l>Every child may joy to hear. </l>
        </lg>
```

Data Curation Lifecycle



Data Curation Lifecycle
(more complex workflow)



Upstream

Downstream

Curation

Data Curation is the active and ongoing management of data throughout a lifecycle of use, including its **reuse in unanticipated contexts**.

Data Quality

*"…the degree to which a set of **characteristics** of data fulfills **stated requirements**."*

*Examples of **characteristics** are:*
completeness, validity, accuracy, consistency, availability and timeliness

ISO 8000 & ISO 9000

- Normalization
    - Literally - making data conform to a normal schema
    - Figuratively - transforming data structures, organizing variables (columns) and attributes (rows), and editing values so that they are consistent, interpretable, and match best practices in a field.

Database Normalization
(structure)

#	Customer	Order	Item	Delivery Address
1	Linda Porch	01366	Cosa-1	520 Alpha St.
2	Elliott Roof	01377	Ding-1	205 Beta Dr.
3	Kevin Chair	01334	Coisa-1	052 Theta Circle.
4	Todd Window	01355	Veshch-2	502 Gamma Avenue.
5	Diane Door	01353	Koto-2	250 Delta Rd.

Customer	Delivery Address
Linda Porch	520 Alpha St.
Elliott Roof	205 Beta Dr.
Kevin Chair	052 Theta Circle.
Todd Window	502 Gamma Avenue.
Diane Door	250 Delta Rd.

Customer	Order
Linda Porch	01366
Elliott Roof	01377
Kevin Chair	01334
Todd Window	01355
Diane Door	01353

Data Normalization
(values)

#	Customer	Order	Item	Delivery Address
1	Linda A. Porch	01366	Cosa 1	520 Alpha St.
2	Elliott Roof	1377	Ding-1	205 Beta Drive

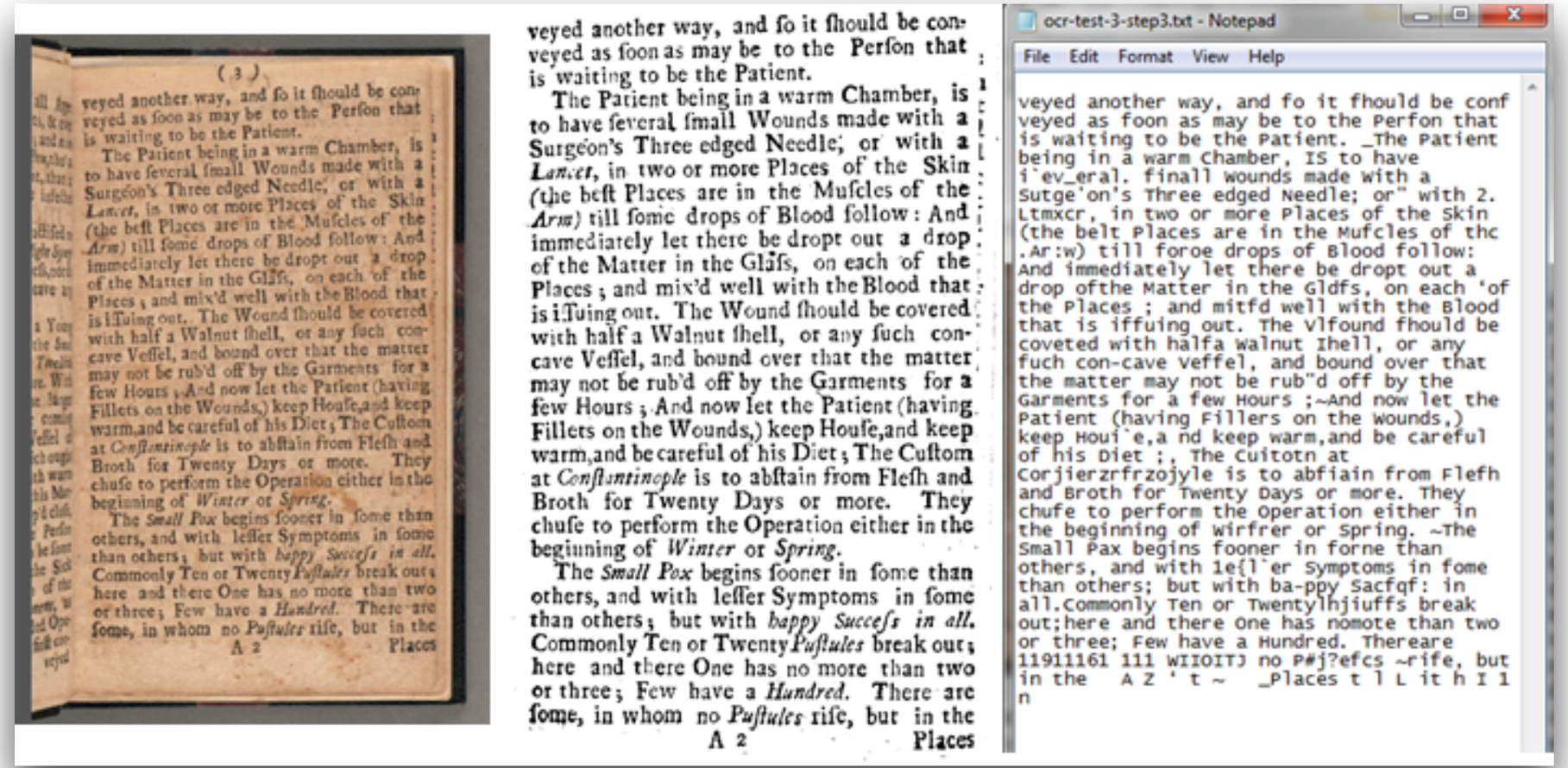
Parsing and standardization — aligning rows and columns, formatting of values into consistent layouts and convention of local standards (for example, postal authority standards for address data). Think of this as the Data Dictionary has to match the Data Value.

Cleaning (aka wrangling / scrubbing)— Modification of data values to meet domain restrictions (e.g. all blank values are to be titled NA), making sure that the **value constraints** of a data dictionary are met in the dataset.

Matching — Identification, linking or merging related entries within or across sets of data (e.g. Item numbers across datasets)

Profiling — Analysis of data to capture statistics that provide insight into the quality of the data and aid in the identification of data quality issues (e.g. We will use a “clustering” feature in Open Refine to group different values together.

Enrichment — Enhancing the value of internally held data. Oftentimes this means adding information (e.g. geographic coordinates; zip codes; etc.)



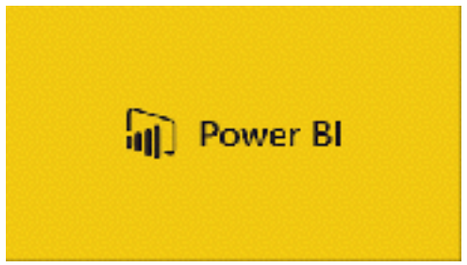
https://smallpoxinoculation.wordpress.com/ocr/

Text Normalization (unstructured data)

- Spelling (e.g. theatre or theater; organise vs organize)
- Vocabulary (e.g
- Punctuation (e.g. on-line vs online)
- Chunking (pargraphs, sentences, stanzas, acts, scenes, etc.)
- Markup (what schema did our XML use to encode a text?)

Data Normalization

without writing any code.

	Open Refine	FLOSS	Data Cleaning	Looks like a spreadsheet acts like a database
	R	FLOSS	Statistics + Data Cleaning	Commandline power
	Microsoft	Proprietary	Analytics + VIZ	Plug + Play
	Tableau	Public	VIZ	Plug + Play

PARKING

Can I park here?

R7-1 NO PARKING

R7-2 NO PARKING

What about here…?

https://data.seattle.gov/Transportation/SDOT-Street-Signs/atig-uucb#revert

Data Curation is the active and ongoing management of data throughout a lifecycle of use, including its **reuse in unanticipated contexts**.

The reuse of data creates **friction**…
Between person who originally produced the data…
And person trying to understand and use data…

Metadata is a kind of **lubricant** that reduces friction between data producers and data users

Metadata is most simply a set of **standardized attribute-value** pairs that provide **contextual information** about an object or artifact:

Attribute	**Value**
Title	Hitchhiker’s Guide to the Galaxy
Creator	Douglas Adams

Metadata is most simply a set of **standardized attribute-value** pairs that provide **contextual information** about an object or artifact:

Attributes - properties, features, or characteristics of instances (and by inheritance, classes)

Attribute	**Value**
Name	Masha
Eye Color	Blue

Expressivity vs. **Tractability**
(inverse relationship)

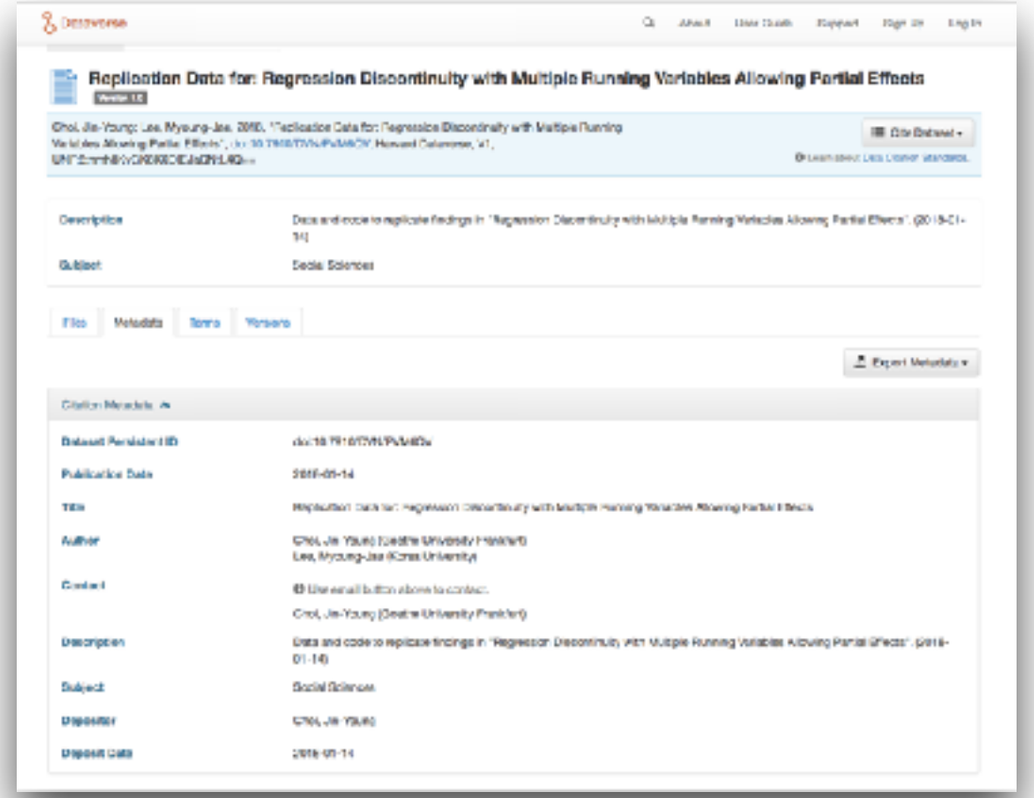
The *more expressive* we make our metadata, the *less tractable* it is in terms of generating, managing, and computing for reuse.

The challenge of metadata and documentation for data curation is balancing expressivity and tractability.

Structured vs. **Unstructured** Metadata

```
<metadata xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns="http://dublincore.org/documents/dcmi-terms/">
<dcterms:title>
Replication Data for: Regression Discontinuity with Multiple Running Variables Allowing Partial Effects
</dcterms:title>
<dcterms:identifier>http://dx.doi.org/10.7910/DVN/PVH6QY</dcterms:identifier>
<dcterms:creator>Choi, Jin-Young</dcterms:creator>
<dcterms:creator>Lee, Myoung-Jae</dcterms:creator>
<dcterms:publisher>Harvard Dataverse</dcterms:publisher>
<dcterms:issued>2018-01-14</dcterms:issued>
<dcterms:modified>2018-01-14T17:01:39Z</dcterms:modified>
<dcterms:description>
Data and code to replicate findings in "Regression Discontinuity with Multiple Running Variables Allowing Partial
Effects".
</dcterms:description>
<dcterms:subject>Social Sciences</dcterms:subject>
<dcterms:contributor>Choi, Jin-Young</dcterms:contributor>
<dcterms:dateSubmitted>2018-01-14</dcterms:dateSubmitted>
<dcterms:license>CC0</dcterms:license>
<dcterms:rights>CC0 Waiver</dcterms:rights>
</metadata>
```

Machine Readable



Human Readable

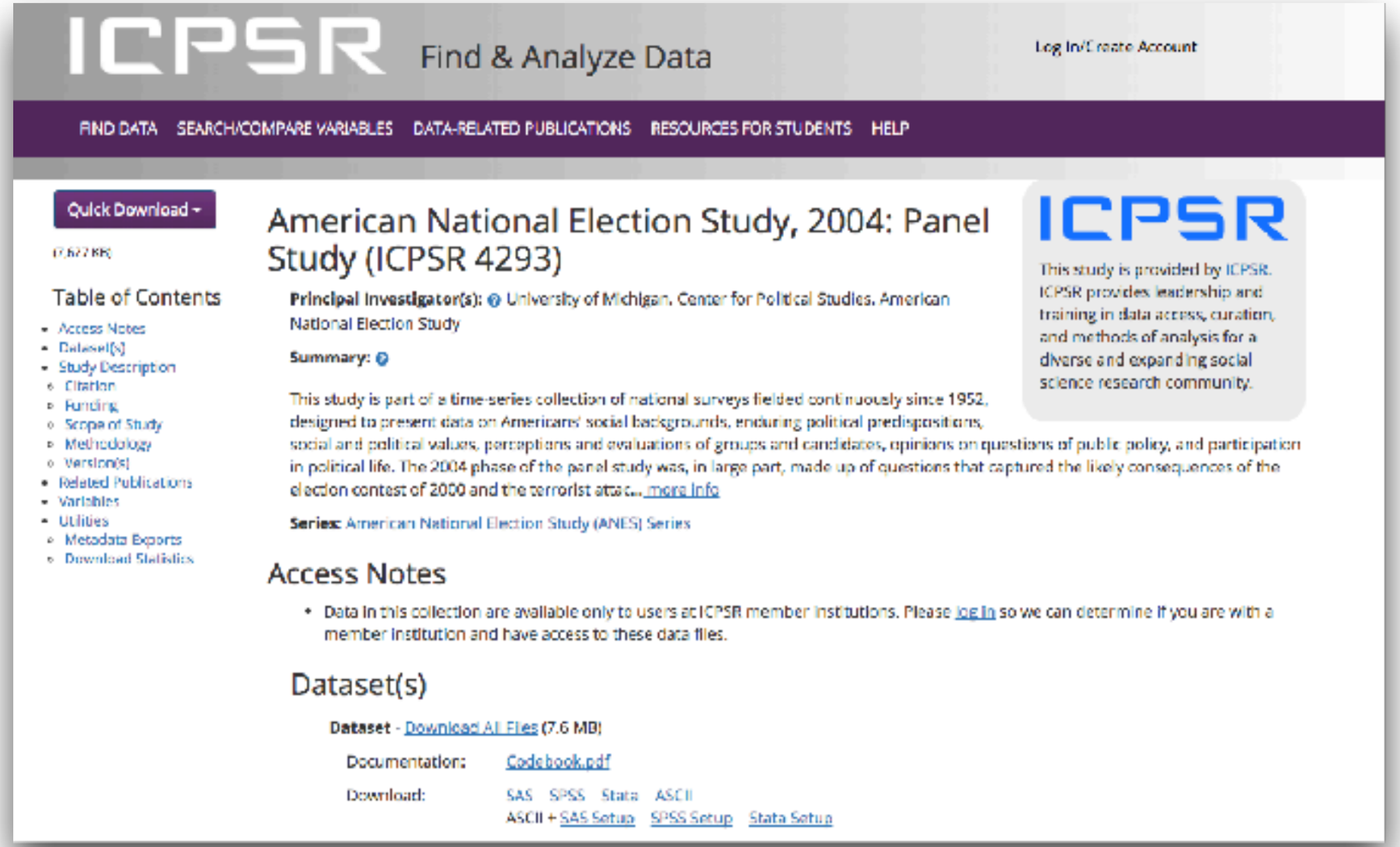
Three basic forms of structured metadata in data curation

Descriptive Metadata: Tells us about objects, their creation, and the context in which they were created (Title, Author, Date)

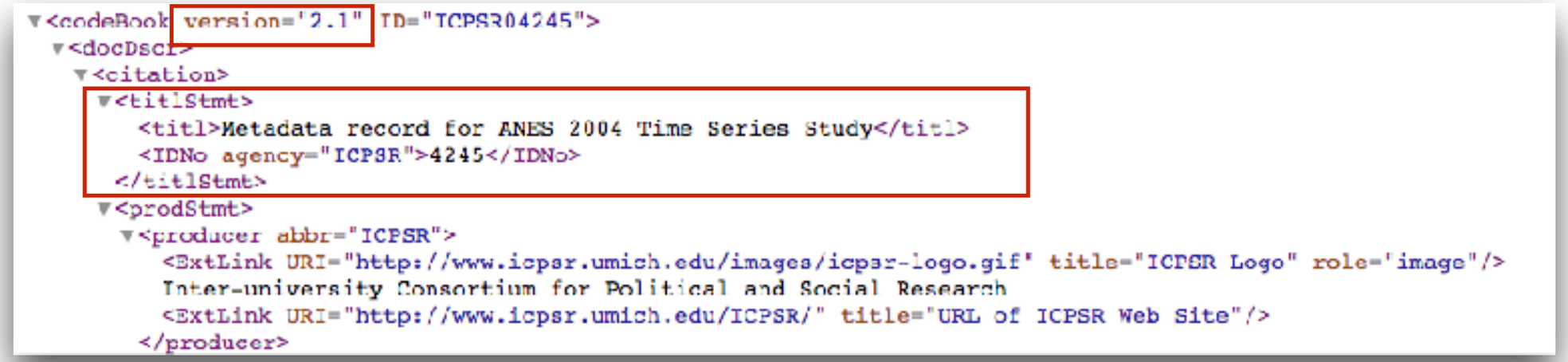
Technical Metadata: Tells us about the context of the data collection (Instrument, Computer, Algorithm)

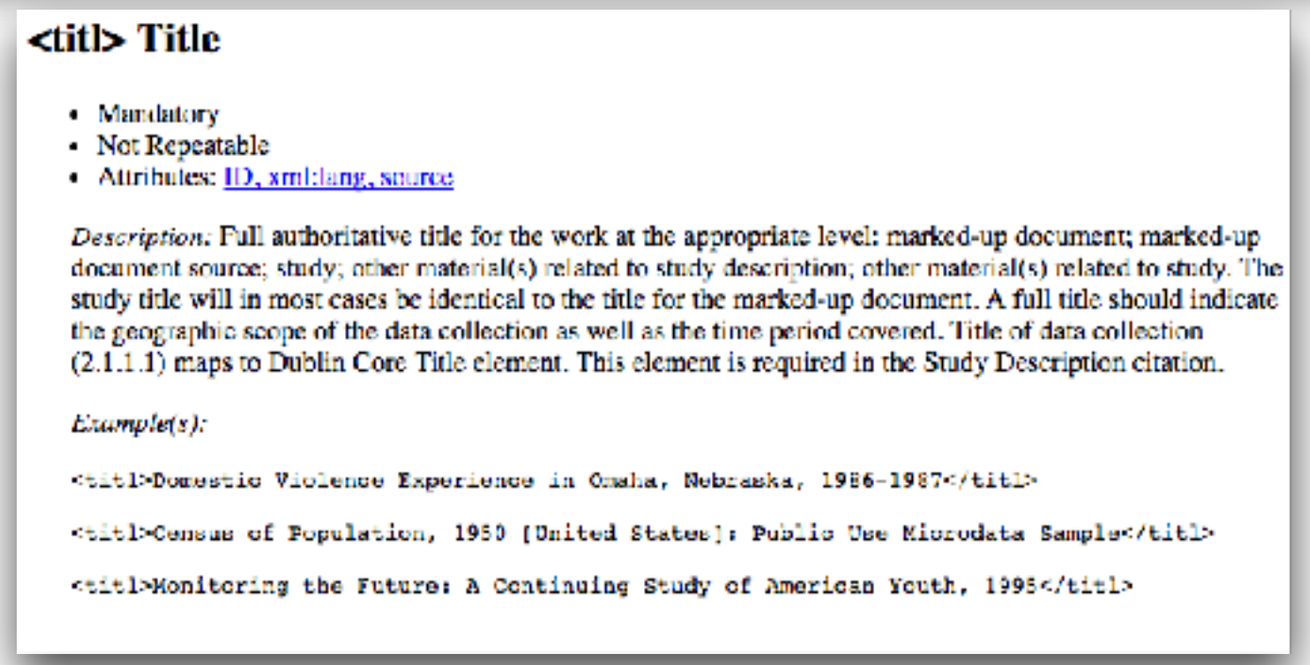
Administrative Metadata: Tell us about the management of that data (Rights statements, Provenance, etc.)

Descriptive Metadata: Tells us about objects, their creation, and the context in which they were created (Title, Author, Date)



Descriptive Metadata: Tells us about objects, their creation, and the context in which they were created (Title, Author, Date)





<titl> Title

- Mandatory
- Not Repeatable
- Attributes: ID, xml:lang, source

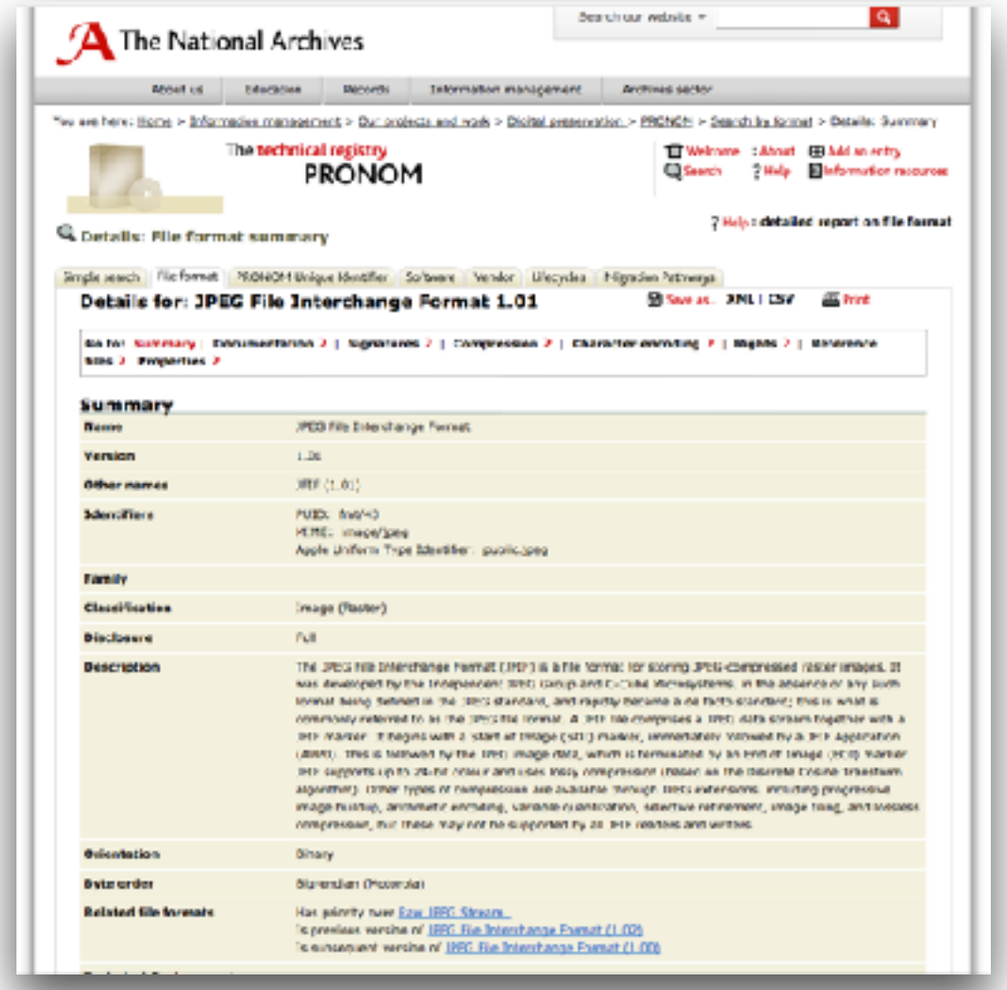
Description: Full authoritative title for the work at the appropriate level: marked-up document; marked-up document source; study; other material(s) related to study description; other material(s) related to study. The study title will in most cases be identical to the title for the marked-up document. A full title should indicate the geographic scope of the data collection as well as the time period covered. Title of data collection (2.1.1.1) maps to Dublin Core Title element. This element is required in the Study Description citation.

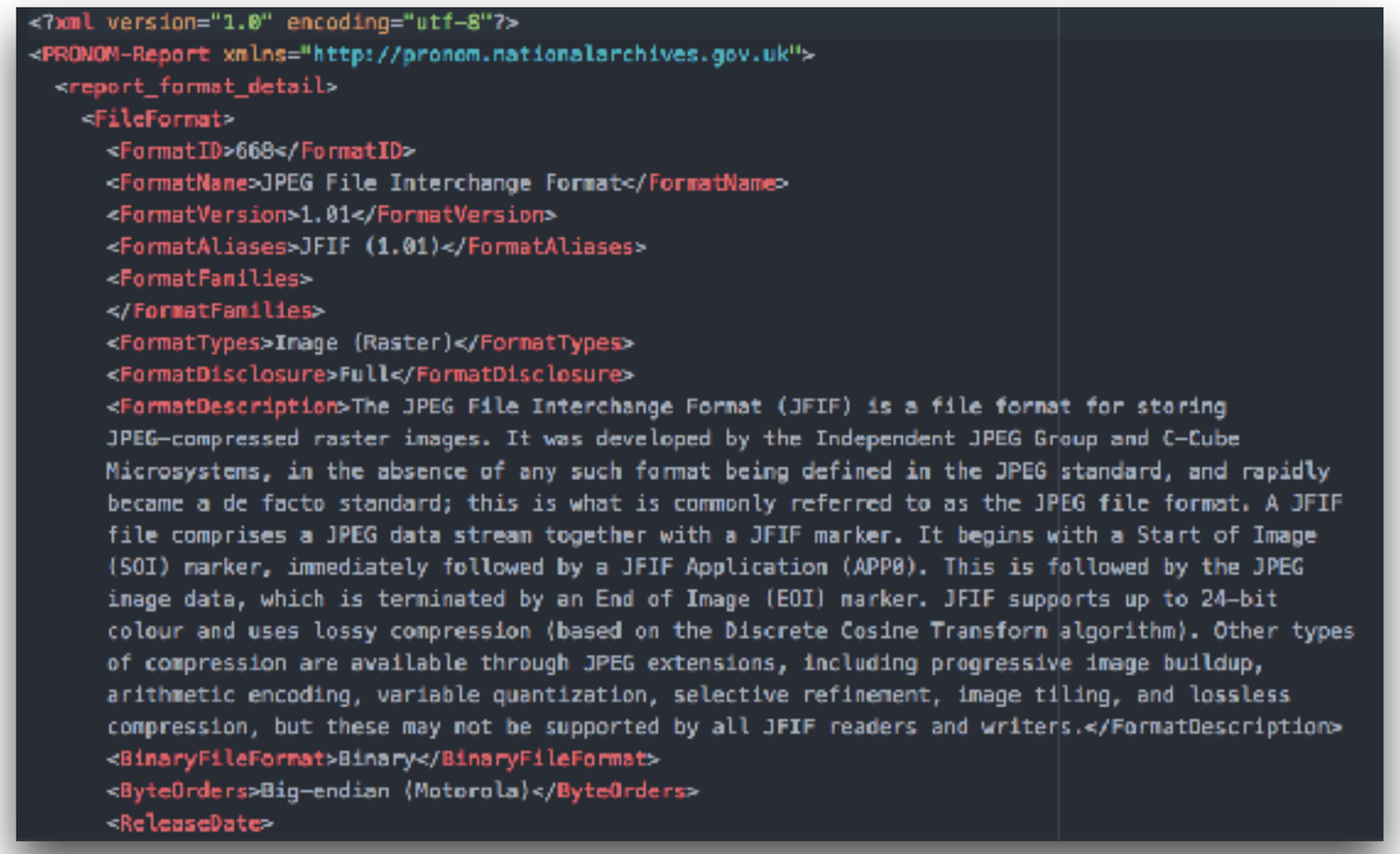
Example(s):

```
<titl>Domestic Violence Experience in Omaha, Nebraska, 1986-1987</titl>

<titl>Census of Population, 1950 [United States]: Public Use Microdata Sample</titl>

<titl>Monitoring the Future: A Continuing Study of American Youth, 1995</titl>
```

https://www.ddialliance.org/Specification/DDI-Codebook/2.1/DTD/Documentation/titlType.html

Technical Metadata: Tells us about the context of the data collection (Instrument, Computer, Algorithm)





```
<?xml version="1.0" encoding="utf-8"?>
<PRONOM-Report xmlns="http://pronom.nationalarchives.gov.uk">
  <report_format_detail>
    <FileFormat>
      <FormatID>668</FormatID>
      <FormatName>JPEG File Interchange Format</FormatName>
      <FormatVersion>1.01</FormatVersion>
      <FormatAliases>JFIF (1.01)</FormatAliases>
      <FormatFamilies>
      </FormatFamilies>
      <FormatTypes>Image (Raster)</FormatTypes>
      <FormatDisclosure>Full</FormatDisclosure>
      <FormatDescription>The JPEG File Interchange Format (JFIF) is a file format for storing
      JPEG-compressed raster images. It was developed by the Independent JPEG Group and C-Cube
      Microsystems, in the absence of any such format being defined in the JPEG standard, and rapidly
      became a de facto standard; this is what is commonly referred to as the JPEG file format. A JFIF
      file comprises a JPEG data stream together with a JFIF marker. It begins with a Start of Image
      (SOI) marker, immediately followed by a JFIF Application (APP0). This is followed by the JPEG
      image data, which is terminated by an End of Image (EOI) marker. JFIF supports up to 24-bit
      colour and uses lossy compression (based on the Discrete Cosine Transform algorithm). Other types
      of compression are available through JPEG extensions, including progressive image buildup,
      arithmetic encoding, variable quantization, selective refinement, image tiling, and lossless
      compression, but these may not be supported by all JFIF readers and writers.</FormatDescription>
      <BinaryFileFormat>Binary</BinaryFileFormat>
      <ByteOrders>Big-endian (Motorola)</ByteOrders>
      <ReleaseDate>
```

http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=668#

Administrative Metadata: Tell us about the management of that data (Rights statements, Provenance, etc.)

PREMIS PRESERVATION METADATA MAINTENANCE ACTIVITY

```
<premis:object>
  <--!other metadata-->
    <premis:signatureInformation>
      <premis:signatureInformationEncoding>BASE 64</premis:signatureInformationEncoding>
      <premis:signer>Susan Thomas</premis:signer>
      <premis:signatureMethod>DSA-SHA1</premis:signatureMethod>
      <premis:signatureValue>qUADDMHZkyebvRdLs+6Dv7RvgMLRlUaDB4Q9yn9XoJA79a2882ffTg==
      </premis:signatureValue>
      <premis:signatureValidationRules>Add reference to repository documentation detailing signature
      validation rules</premis:signatureValidationRules>
      <premis:signatureProperties>2006-11-01T10:15:16</premis:signatureProperties>
    </premis:signatureInformation>
  <--!other metadata-->
</premis:object>
```

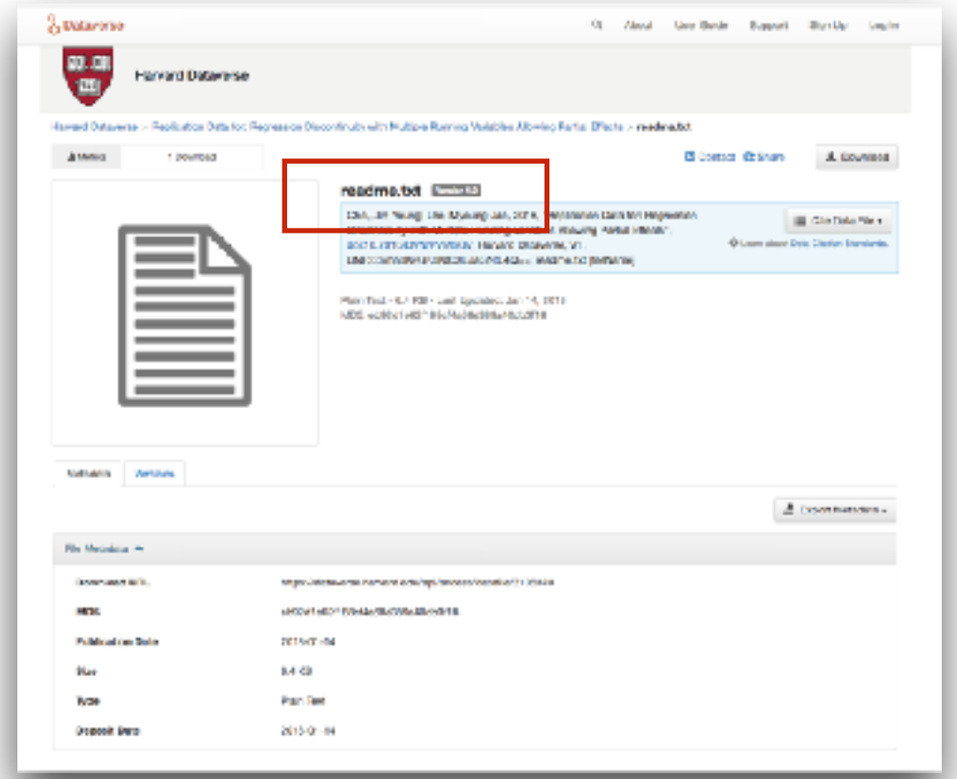
Unstructured Metadata or **Documentation**
(human readable)

README.txt - provides narrative explanation of what a dataset contains, how it was produced, and how it can or should be used.

Data Dictionary - defines the variables (and constraints on the values of those variables) in a dataset

CodeBook - defines what codes were created to analyze, or summarize a dataset

readMe.txt




```
There are two folders to replicate the empirical results of the paper: STATA folder and
GAUSS folder.
The STATA folder provides the graphic outputs in *.gph files, and the GAUSS folder provides
the table outputs in *.txt files.

Even if the user is unfamiliar with GAUSS, he/she can still obtain at least parts of the
table outputs by running the STATA program:
specifically, the estimates of the tables in the paper, and the t-values computed with the
usual OLS asymptotic variance estimator,
but not the confidence intervals (CI's) computed with bootstrap in the paper.

The details of the STATA and GAUSS folders are as follows.




---------- STATA FOLDER DESCRIPTION----------


The enclosed STATA program "Election 26AUG2017 Stata.do" produces Table 1, all estimates in
Tables 2 and 3, and Figures 2 and 3.
The *.log file is the saved result corresponding to the .do file and it includes Tables 1,
2, and 3.
And the *.gph files are figure outputs also generated with the .do file.


What the STATA program does not produce is the confidence intervals (CI) based on bootstrap
in Tables 2 and 3;
instead of the CI's, the STATA program provides the usual t-values based on the OLS
asymptotic variance estimator for all OLS-based estimates.
Because of this, the OLS CI's in the paper differ somewhat from those in the STATA output
file.


The STATA program does not provide any t-value for the "boundary-weighting (BW)" estimator
in Tables 2 and 3,
because BW is a complicated estimator, not based on OLS.

If the reader desires to generate bootstrap CI's, he/she may use the bootstrap option for
OLS provided by STATA.


In the STATA program, "mf" appears, which stands for "multiplicative factor" in selecting
the bandwidth

   h=mf*SD(S)*N^(-1/6)   where S is the running variable in use.


The "mf" value is typically about 0.5-2.5, and it was already chosen with Cross-Validation
(CV) using a GAUSS program.
The STATA file uses the pre-selected value of "mf" without redoing the CV procedure.


The reason for not providing the bootstrap CI's and not doing the CV procedure in the STATA
program is that
these procedures require a sophisticated programming with STATA,
which the authors could not do, as they are not regular users of STATA.




---------- GAUSS FOLDER DESCRIPTION----------


In the GAUSS folder, all files are written in GAUSS, which is a programming language from
Aptech Systems Inc.
GAUSS files can be opened with any text file editor (e.g., notepad or wordpad).
In our paper, empirical parts were done with GAUSS, except for Figures 2 and 3.
```

Data Dictionary

Department	Dataset Name	Field Name	Field Alias	Field Type	API Key	Field Definition	Field Type Flag
Rent Arbitration Board	Eviction Notices	City		text	city	The city where the eviction notice was issued. In this dataset, always San Francisco.	
Rent Arbitration Board	Eviction Notices	State		text	state	The state where the eviction notice was issued. In this dataset, always CA.	
Rent Arbitration Board	Eviction Notices	Eviction Notice Source Zipcode		text	zip	The zip code where the eviction notice was issued.	
Rent Arbitration Board	Eviction Notices	File Date		timestamp	file_date	The date on which the eviction notice was filed with the Rent Board of Arbitration.	
Rent Arbitration Board	Eviction Notices	Non Payment		boolean	non_payment	This field is checked (true) if the landlord indicated non-payment of rent as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Breach		boolean	breach	This field is checked (true) if the landlord indicated breach of lease as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Nuisance		boolean	nuisance	This field is checked (true) if the landlord indicated nuisance as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Illegal Use		boolean	illegal_use	This field is checked (true) if the landlord indicated an illegal use of the rental unit as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Failure to Sign Renewal		boolean	failure_to_sign_renewal	This field is checked (true) if the landlord indicated failure to sign lease renewal as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Access Denial		boolean	access_denial	This field is checked (true) if the landlord indicated unlawful denial of access to unit as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Unapproved Subtenant		boolean	unapproved_subtenant	This field is checked (true) if the landlord indicated the tenant had an unapproved subtenant as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Owner Move In		boolean	owner_move_in	This field is checked (true) if the landlord indicated an owner move in as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Demolition		boolean	demolition	This field is checked (true) if the landlord indicated demolition of property as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Capital Improvement		boolean	capital_improvement	This field is checked (true) if the landlord indicated a capital improvement as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Substantial Rehab		boolean	substantial_rehab	This field is checked (true) if the landlord indicated substantial rehabilitation as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Ellis Act WithDrawal		boolean	ellis_act_withdrawal	This field is checked (true) if the landlord indicated an Ellis Act withdrawal (going out of business) as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Condo Conversion		boolean	condo_conversion	This field is checked (true) if the landlord indicated a condo conversion as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Roommate Same Unit		boolean	roommate_same_unit	This field is checked (true) if the landlord indicated if they were evicting a roommate in their unit as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Other Cause		boolean	other_cause	This field is checked (true) if some other cause not covered by the admin code was indicated by the landlord. These are not enforceable gr	
Rent Arbitration Board	Eviction Notices	Late Payments		boolean	late_payments	This field is checked (true) if the landlord indicated habitual late payment of rent as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Lead Remediation		boolean	lead_remediation	This field is checked (true) if the landlord indicated lead remediation as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Development		boolean	development	This field is checked (true) if the landlord indicated a development agreement as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Good Samaritan Ends		boolean	good_samaritan_ends	This field is checked (true) if the landlord indicated the period of good samaritan laws coming to an end as a grounds for eviction.	
Rent Arbitration Board	Eviction Notices	Constraints Date		timestamp	constraints_date	In the case of certain just cause evictions like Ellis and Owner Move In, constraints are placed on the property and recorded by the the City	
Rent Arbitration Board	Eviction Notices	Supervisor District	Supervisor D	number	supervisor_district	There are 11 members of the Board of Supervisors in San Francisco, each representing a geographic district. These are numbered 1 throug	
Rent Arbitration Board	Eviction Notices	Neighborhoods - Analysis Boundaries	Neighborhoo	text	neighborhood	The Department of Public Health and the Mayor's Office of Housing and Community Development, with support from the Planning Dep	
Rent Arbitration Board	Eviction Notices	Location	Geometry	geometry: p	client_location	Contains the geometry of the record in Well Known Text (WKT) format.	

https://data.sfgov.org/City-Management-and-Ethics/-alpha-Master-data-dictionary/wn8x-uk7i#

Codebook

CODEBOOK FOR ICPSR 9028

UNIFORM CRIME REPORTING PROGRAM DATA [UNITED STATES]

PART 1: OFFENSES KNOWN AND CLEARANCES BY ARREST, 1980

PLEASE NOTE: The "M" between the code and the code label indicates the code has been designated as a missing value.

NAME	VARIABLE LABEL	BEG COL	END COL	FMT
V1	ID CODE	1	1	F1
	1 Offenses known			
V2	NUMERIC STATE CODE	2	3	F2
	1 Alabama			
	2 Arizona			
	3 Arkansas			
	4 California			
	5 Colorado			
	6 Connecticut			
	7 Delaware			
	8 District of Columbia			
	9 Florida			
	10 Georgia			
	11 Idaho			
	12 Illinois			
	13 Indiana			
	14 Iowa			
	15 Kansas			
	16 Kentucky			
	17 Louisiana			
V3	DIVISION	13	13	F1
	0 Possessions			
	1 New England States			
	2 Middle Atlantic States			
	3 East North Central States			
	4 West North Central States			
	5 South Atlantic States			
	6 East South Central States			
	7 West South Central States			
	8 Mountain States			
	9 Pacific States			
V6	YEAR	14	17	F4
V7	CITY SEQUENCE NUMBER	18	22	F5
V8	CORE CITY INDICATION	23	23	A1
	N No, not core city of MSA			
	Y Yes, core city of MSA			
V9	COVERED BY CODE	24	30	A7
V10	LAST UPDATE	31	38	F8
V11	FIELD OFFICE	39	42	F4
V12	NUMBER OF MONTHS REPORTED	43	44	F2
	0 No months reported			
	1 Jan last reported			
	2 Feb last reported			
	3 March last reported			
	4 April last reported			
	5 May last reported			
	6 June last reported			

https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/09028

- Metadata helps reduce friction between data producers and data users

- Comes in two forms: Structured and Unstructured

  - Structured metadata uses an encoding, and a formally defined schema to make metadata **Machine Readable**

  - Unstructured Metadata is meant to provide contextual information that is **Human Readable**

Data Curation is the active and ongoing management of data throughout a lifecycle of use, including its reuse in unanticipated contexts.

Data are various **types** of information objects playing the **role** of evidence.

“…the degree to which a set of characteristics of data fulfills stated requirements.”