Data Curation and LIS 530

Dr. Nicholas Weber
02.20.2019

Slides, data and docs at:
https://github.com/nniiicc/LIS-530-2019

www.nuforc.org

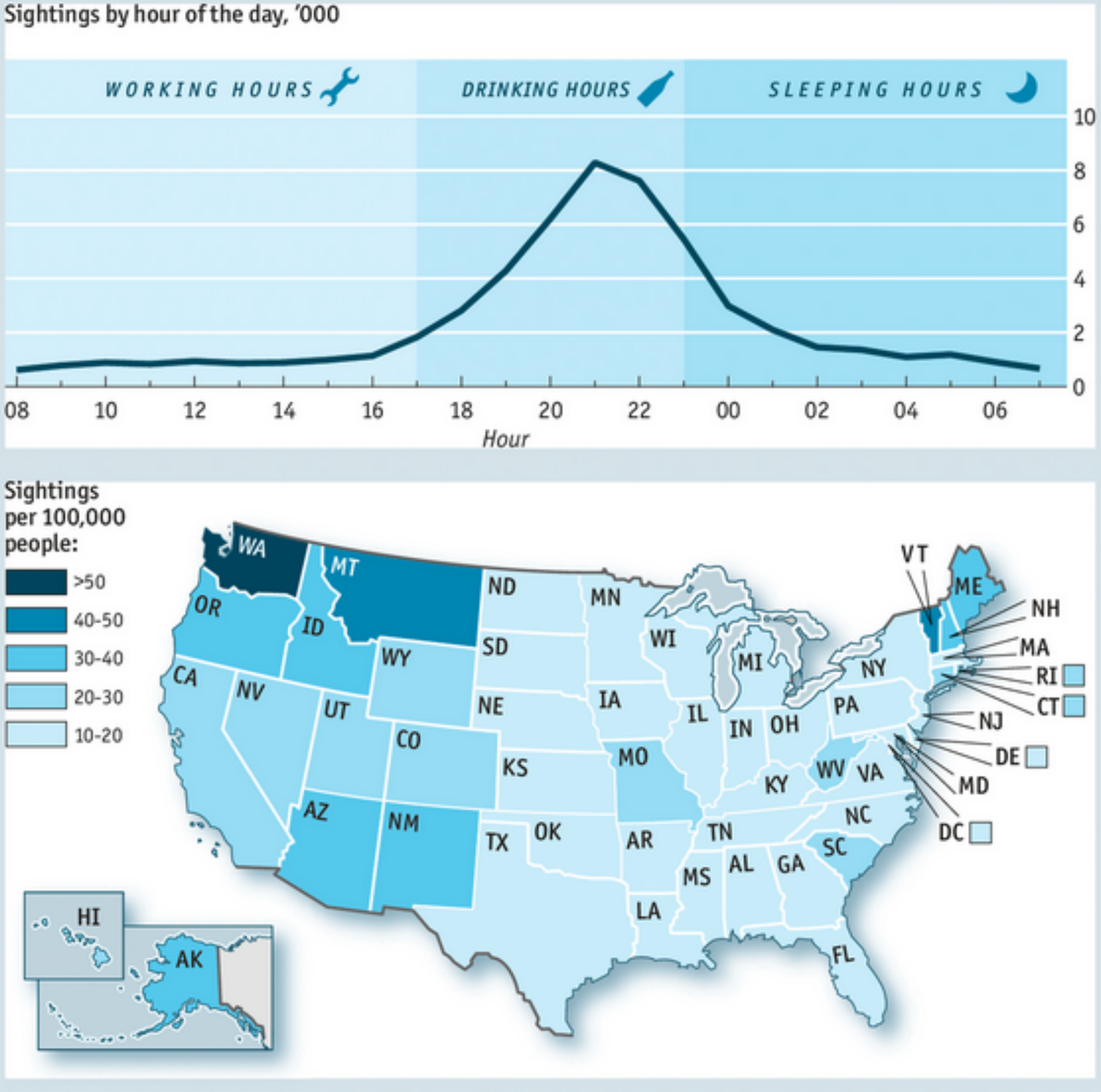
You can look at any dataset, but this is the latest:
http://www.nuforc.org/webreports/ndxe201902.html

UFO Data…

1. What are the data being served by nuforc ?
2. How are data collected?
3. What interesting questions might you ask of this kind of data?
4. What needs to be done to the data in order to answer those questions?

America's UFO sightings

2000-14



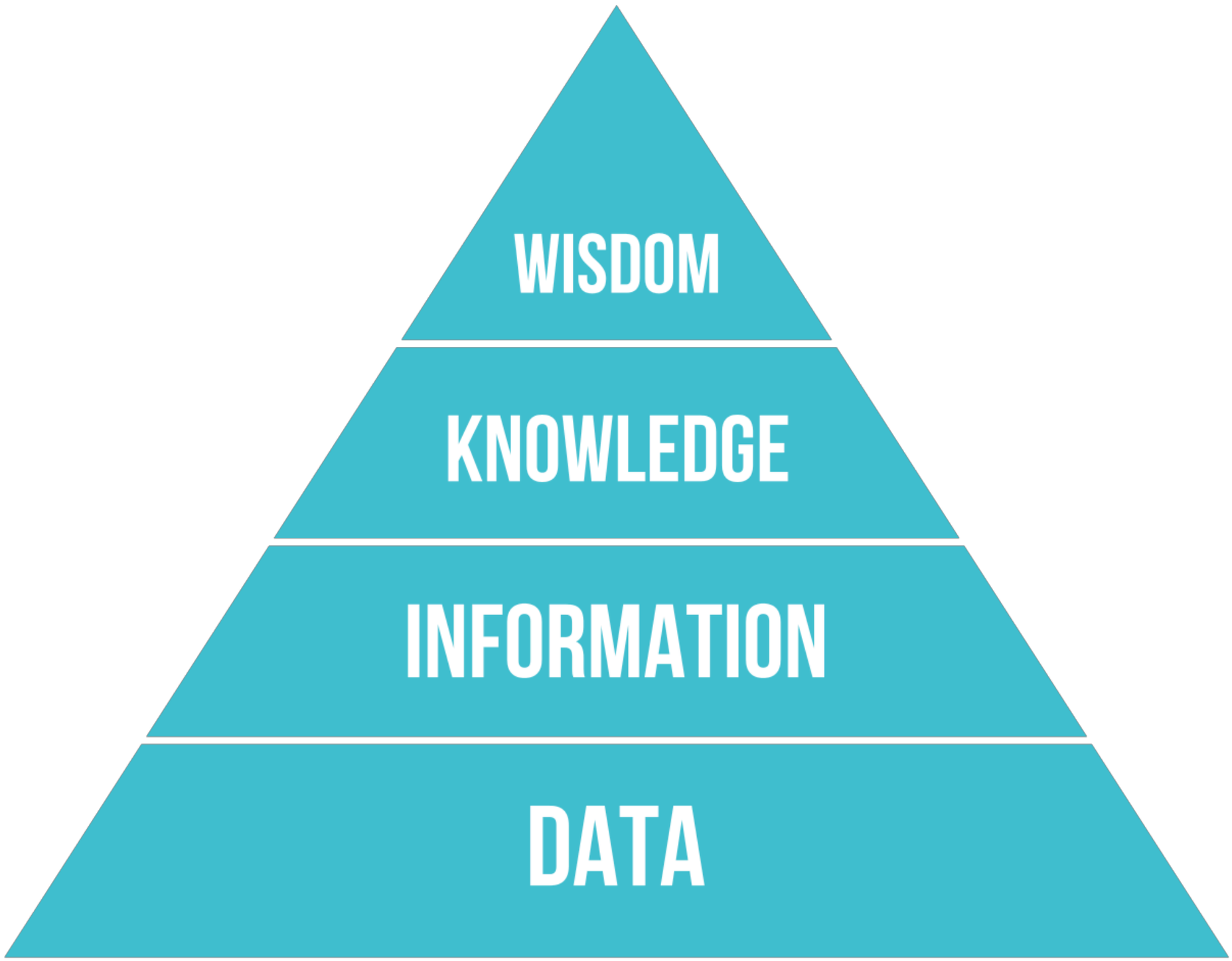
Source: National UFO Reporting Center; www.ufocenter.com

https://www.theatlantic.com/technology/archive/2014/06/the-hard-data-on-ufo-sightings-its-mostly-drunk-people-in-the-west/373645/

- 1/2 - Conceptual Foundations of Data Curation
- 2/2 - How and why this might be useful for 530

For this lecture, let's use the following definition…

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 analyzed data."

ands
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)

Type vs Role Distinctions

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

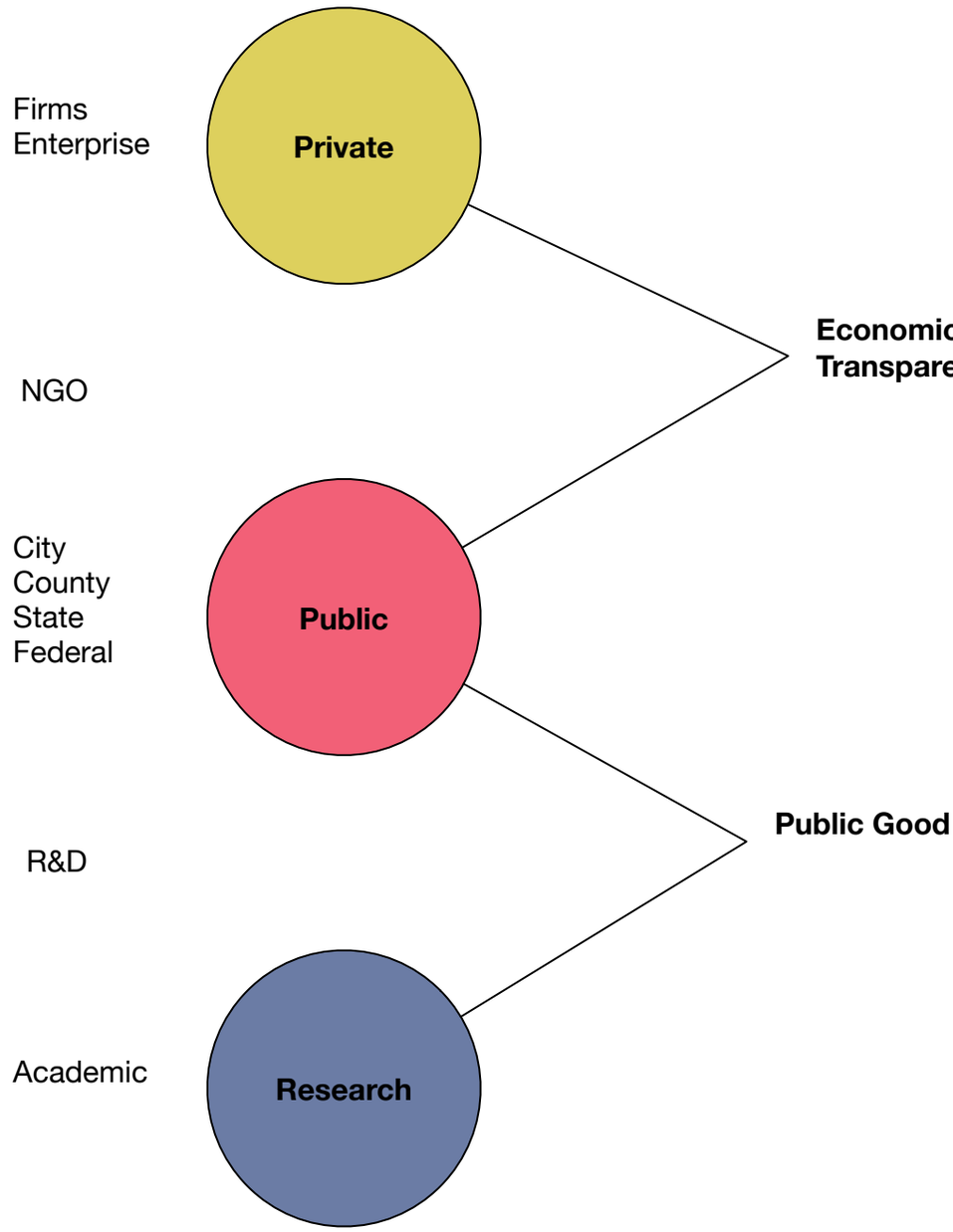
Excel

XBRL

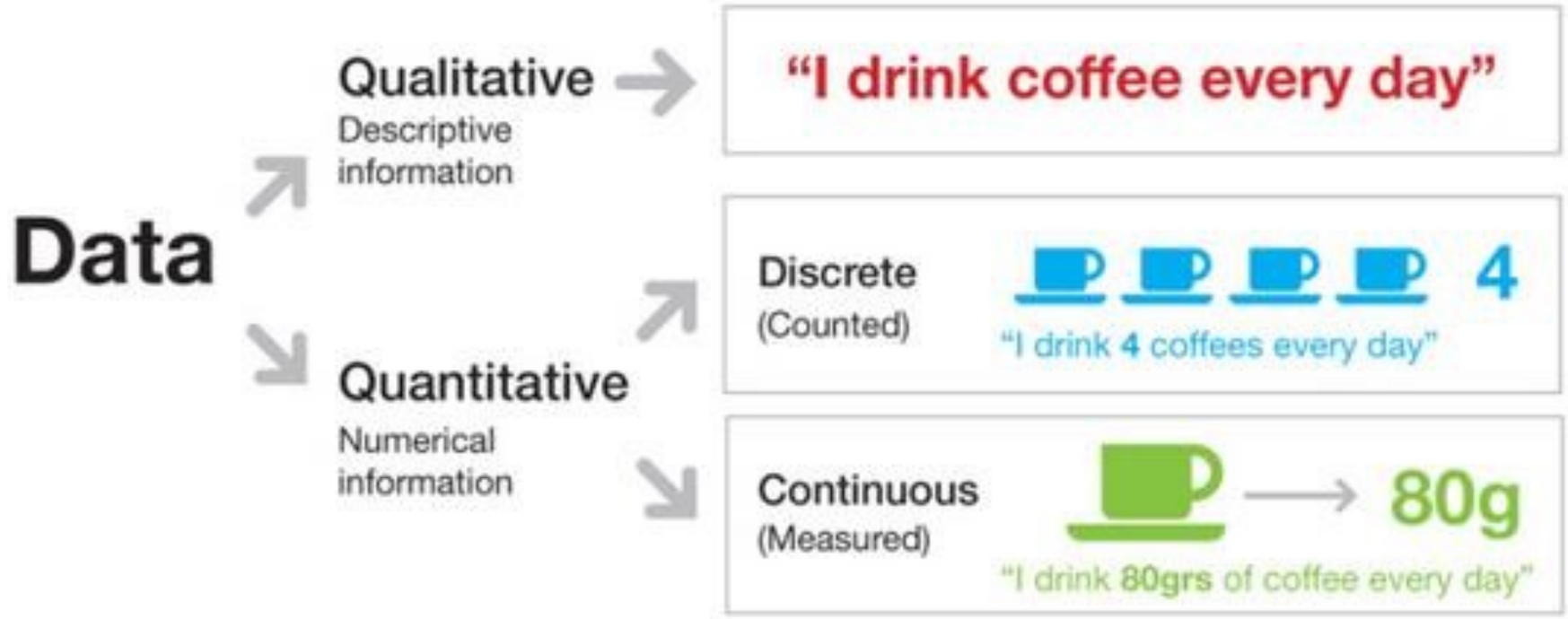
JSON

Web Services

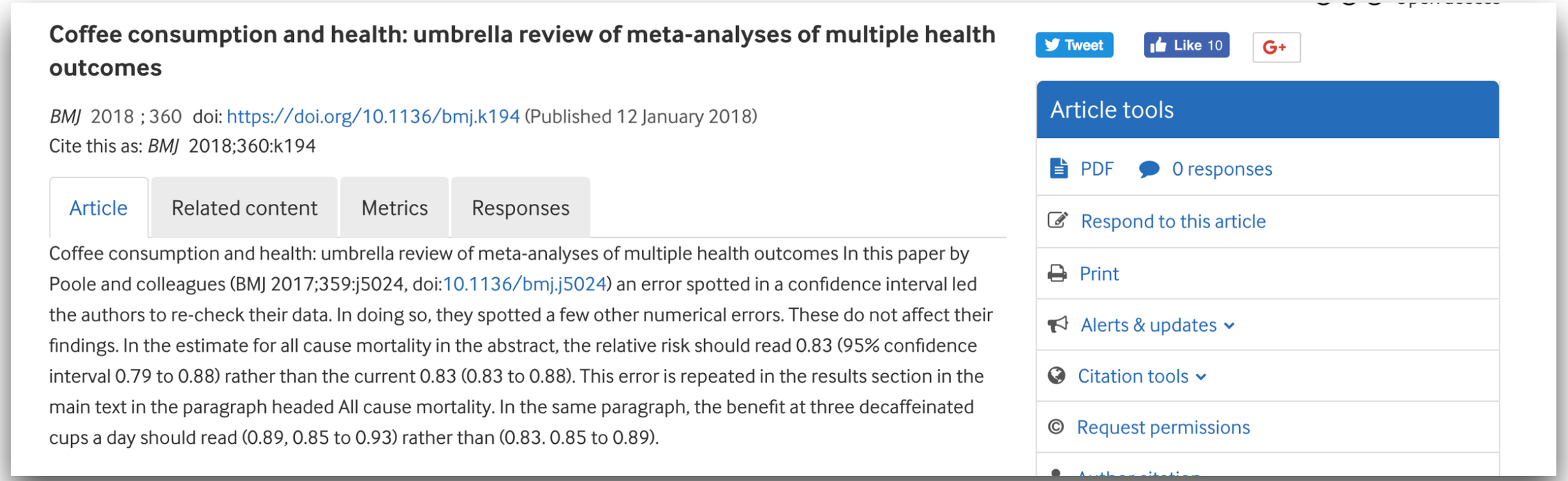
Types of Data
(by sector)



Data Types


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"

Data Roles



Coffee consumption and health: umbrella review of meta-analyses of multiple health outcomes

BMJ 2018 ; 360 doi: https://doi.org/10.1136/bmj.k194 (Published 12 January 2018)
Cite this as: *BMJ* 2018;360:k194

Article | Related content | Metrics | Responses

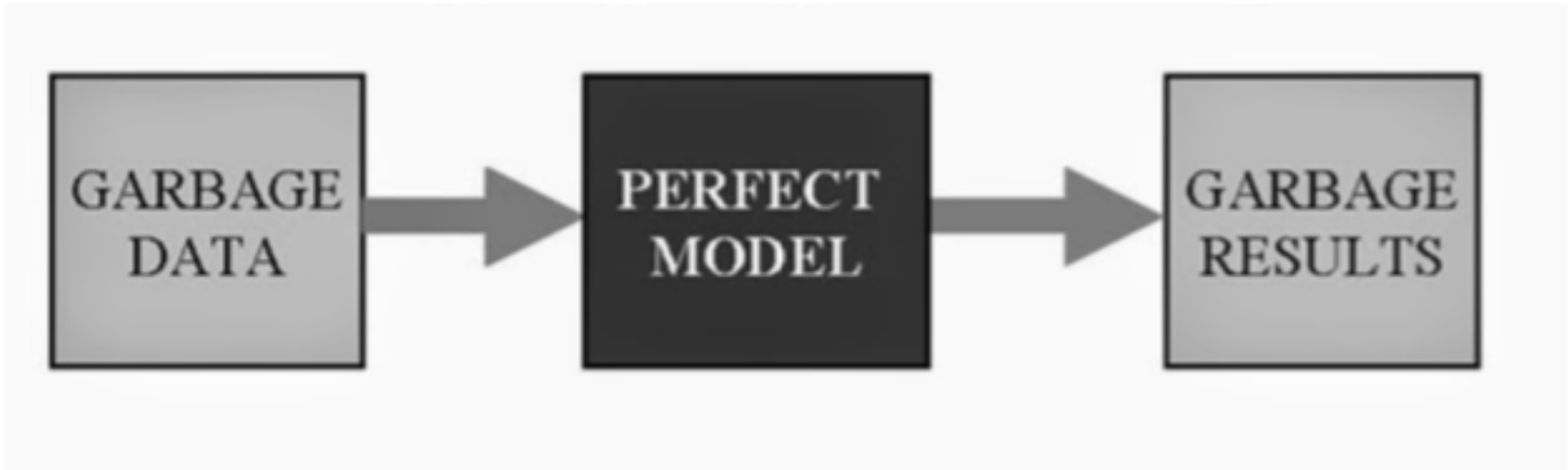
Coffee consumption and health: umbrella review of meta-analyses of multiple health outcomes In this paper by Poole and colleagues (BMJ 2017;359:j5024, doi:10.1136/bmj.j5024) an error spotted in a confidence interval led the authors to re-check their data. In doing so, they spotted a few other numerical errors. These do not affect their findings. In the estimate for all cause mortality in the abstract, the relative risk should read 0.83 (95% confidence interval 0.79 to 0.88) rather than the current 0.83 (0.83 to 0.88). This error is repeated in the results section in the main text in the paragraph headed All cause mortality. In the same paragraph, the benefit at three decaffeinated cups a day should read (0.89, 0.85 to 0.93) rather than (0.83. 0.85 to 0.89).

That was data… what about **Curation**

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

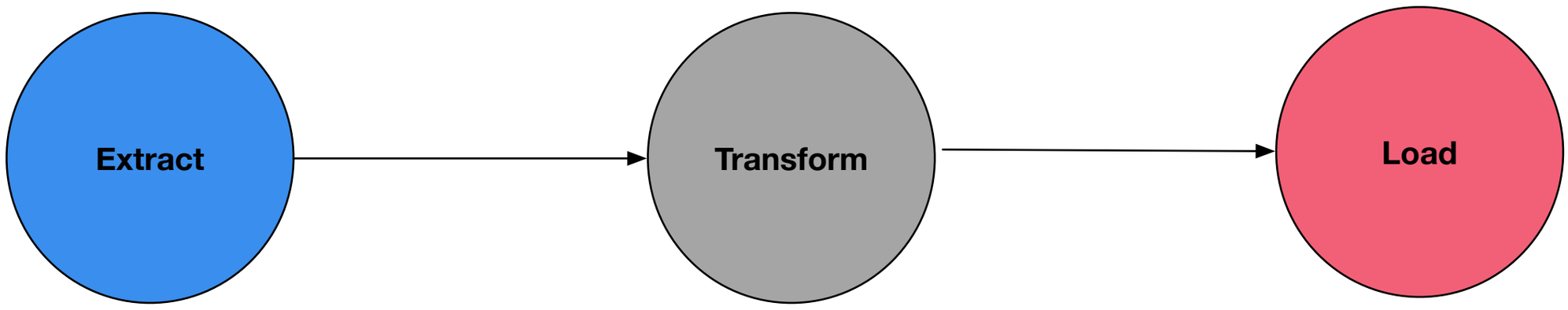
Data Curation

Old computer science saying ”Garbage in = Garbage out”

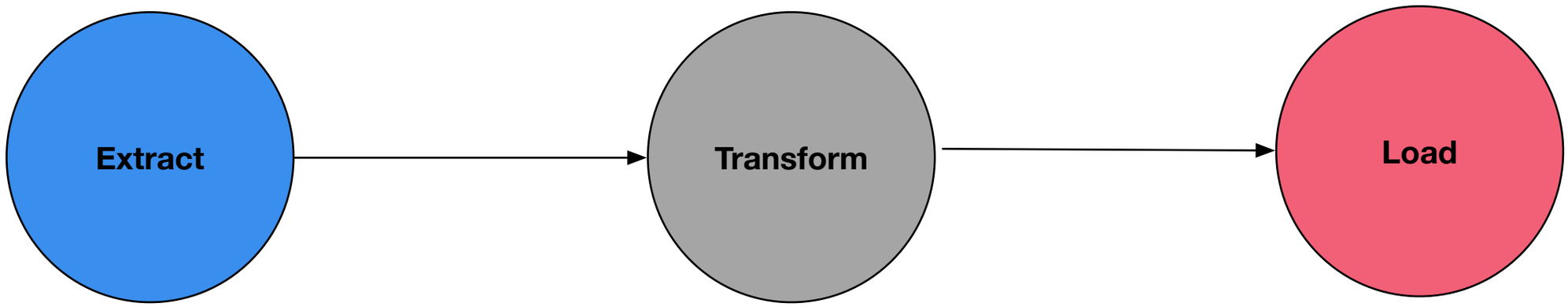


Data curation says “Quality in = Quality out”

ETL: Simplest form of Curation Model



ETL: Simplest form of Curation Workflow




```xml
<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 vanis'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>
```

Transformation

- Structured - Databases
- Unstructured - Scanned PDF that has been OCR’d
- Semi-Structured Data - Comma Separated Value
- Most curation work involves transforming data for reuse… and this can include migration to standard formats, applying version controlled in a repository, and improving quality…

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 (Structured Data): Practically making data conform to a *normal* schema

    - This includes transforming data structures (rows have 1 and only 1 record)

        - Organizing variables into columns and instances into rows

        - Editing values so that they are consistent, interpretable, and match best practices in a field. (e.g. applying a Controlled Vocabulary)

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

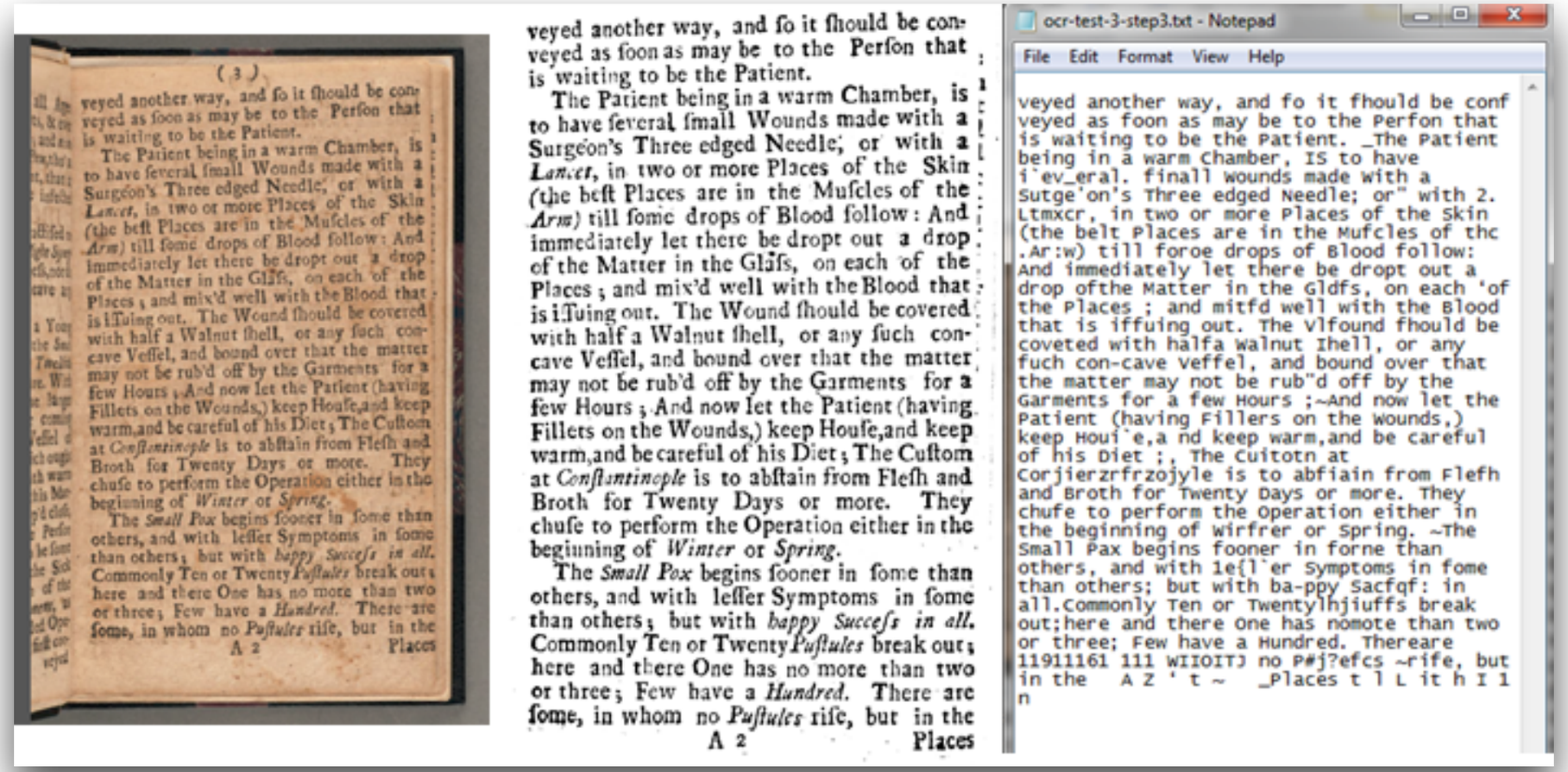
Parsing and standardization — aligning rows and columns, formatting of values into consistent layouts and convention of local standards (for example, postal authority standards for address data). Think of this as the Data Dictionary has to match the Data Value.

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.)


(3)

veyed another way, and ſo it ſhould be conveyed as ſoon as may be to the Perſon that is waiting to be the Patient.

The Patient being in a warm Chamber, is to have ſeveral ſmall Wounds made with a Surgeon's Three edged Needle, or with a *Lancet*, in two or more Places of the Skin (the beſt Places are in the Muſcles of the *Arm*) till ſome drops of Blood follow: And immediately let there be dropt out a drop of the Matter in the Glaſs, on each of the Places; and mix'd well with the Blood that is iſſuing out. The Wound ſhould be covered with half a Walnut ſhell, or any ſuch concave Veſſel, and bound over that the matter may not be rub'd off by the Garments for a few Hours; And now let the Patient (having Fillets on the Wounds,) keep Houſe, and keep warm, and be careful of his Diet; The Cuſtom at *Conſtantinople* is to abſtain from Fleſh and Broth for Twenty Days or more. They chuſe to perform the Operation either in the beginning of *Winter* or *Spring*.

The *Small Pox* begins ſooner in ſome than others, and with leſſer Symptoms in ſome than others; but with *happy Succeſs in all.* Commonly Ten or Twenty *Puſtules* break out; here and there One has no more than two or three; Few have a *Hundred.* There are ſome, in whom no *Puſtules* riſe, but in the

A 2 Places

veyed another way, and ſo it ſhould be conveyed as ſoon as may be to the Perſon that is waiting to be the Patient.

The Patient being in a warm Chamber, is to have ſeveral ſmall Wounds made with a Surgeon's Three edged Needle, or with a *Lancet*, in two or more Places of the Skin (the beſt Places are in the Muſcles of the *Arm*) till ſome drops of Blood follow: And immediately let there be dropt out a drop of the Matter in the Glaſs, on each of the Places; and mix'd well with the Blood that is iſſuing out. The Wound ſhould be covered with half a Walnut ſhell, or any ſuch concave Veſſel, and bound over that the matter may not be rub'd off by the Garments for a few Hours; And now let the Patient (having Fillets on the Wounds,) keep Houſe, and keep warm, and be careful of his Diet; The Cuſtom at *Conſtantinople* is to abſtain from Fleſh and Broth for Twenty Days or more. They chuſe to perform the Operation either in the beginning of *Winter* or *Spring*.

The *Small Pox* begins ſooner in ſome than others, and with leſſer Symptoms in ſome than others; but with *happy Succeſs in all.* Commonly Ten or Twenty *Puſtules* break out; here and there One has no more than two or three; Few have a *Hundred.* There are ſome, in whom no *Puſtules* riſe, but in the

A 2 Places

ocr-test-3-step3.txt - Notepad

File Edit Format View Help

```
veyed another way, and fo it fhould be conf
veyed as foon as may be to the Perfon that
is waiting to be the Patient. _The Patient
being in a warm Chamber, IS to have
i`ev_eral. finall Wounds made With a
Sutge'on's Three edged Needle; or" with 2.
Ltmxcr, in two or more Places of the Skin
(the belt Places are in the Mufcles of thc
.Ar:w) till foroe drops of Blood follow:
And immediately let there be dropt out a
drop ofthe Matter in the Gldfs, on each 'of
the Places ; and mitfd well with the Blood
that is iffuing out. The Vlfound fhould be
coveted with halfa Walnut Ihell, or any
fuch con-cave Veffel, and bound over that
the matter may not be rub"d off by the
Garments for a few Hours ;~And now let the
Patient (having Fillers on the Wounds,)
keep Houi`e,a nd keep warm,and be careful
of his Diet ;, The Cuitotn at
Corjierzrfrzojyle is to abfiain from Flefh
and Broth for Twenty Days or more. They
chufe to perform the Operation either in
the beginning of Wirfrer or Spring. ~The
Small Pax begins fooner in forne than
others, and with 1e{l`er Symptoms in fome
than others; but with ba-ppy Sacfqf: in
all.Commonly Ten or Twentylhjiuffs break
out;here and there One has nomote than two
or three; Few have a Hundred. Thereare
11911161 111 WIIOITJ no P#j?efcs ~rife, but
in the   A Z ' t ~   _Places t l L it h I 1
n
```

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

Text Normalization (unstructured data)

- Spelling (e.g. theatre or theater; organise vs organize)
- Generic vs Controlled Vocabulary (e.g Butterfly vs Lepidoptera)
- 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?)

Let’s look at some examples of quality curation through data on the web …

User story:
I am applying for a job as an assistant professor at the University of Washington. The school has a lot of statements about their commitment to diversity, equity and inclusion. I want to know if my potential employer practices this in their own faculty.

How would I find out?

- First, I would google.
    - Query: ”gender distribution of university of Seattle professors"
- My first result is http://www.washington.edu/faculty/senate/diversity/datareports/
- What do we get at the “data reports” page?

- Let’s say I’m a bit more sophisticated and I know Higher Ed databases.

- I would go to https://nces.ed.gov/

- Perform a search query on “University of Washington”

- What do you find?

Finding data on the web is like shopping for cereal in a grocery store that has no labels…. And some of the cereal boxes also contain shampoo.

–Cesar Hidalgo

DataUSA.IO

https://datausa.io/profile/university/university-of-washington-seattle-campus/#operations

Why is this so much better?

Back to some more lecture…

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

PSSST … This is where LIS 530 becomes REALLY important…

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

Metadata 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

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

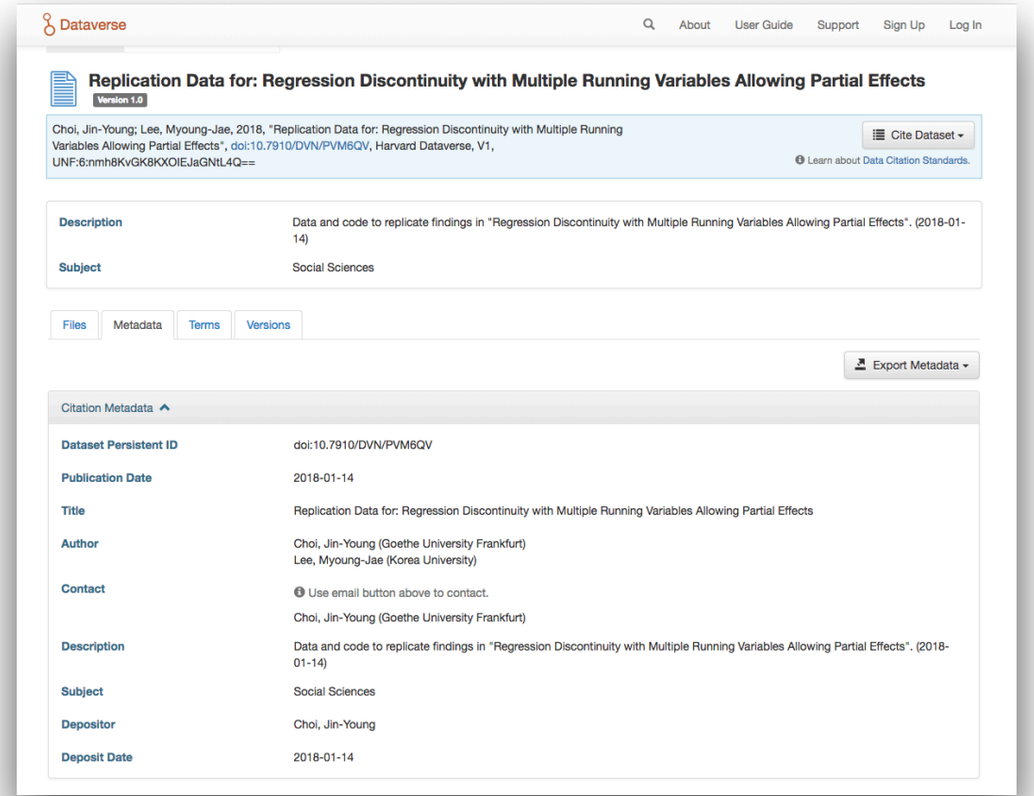
Machine Readable

Human Readable

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/PVM6QV</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

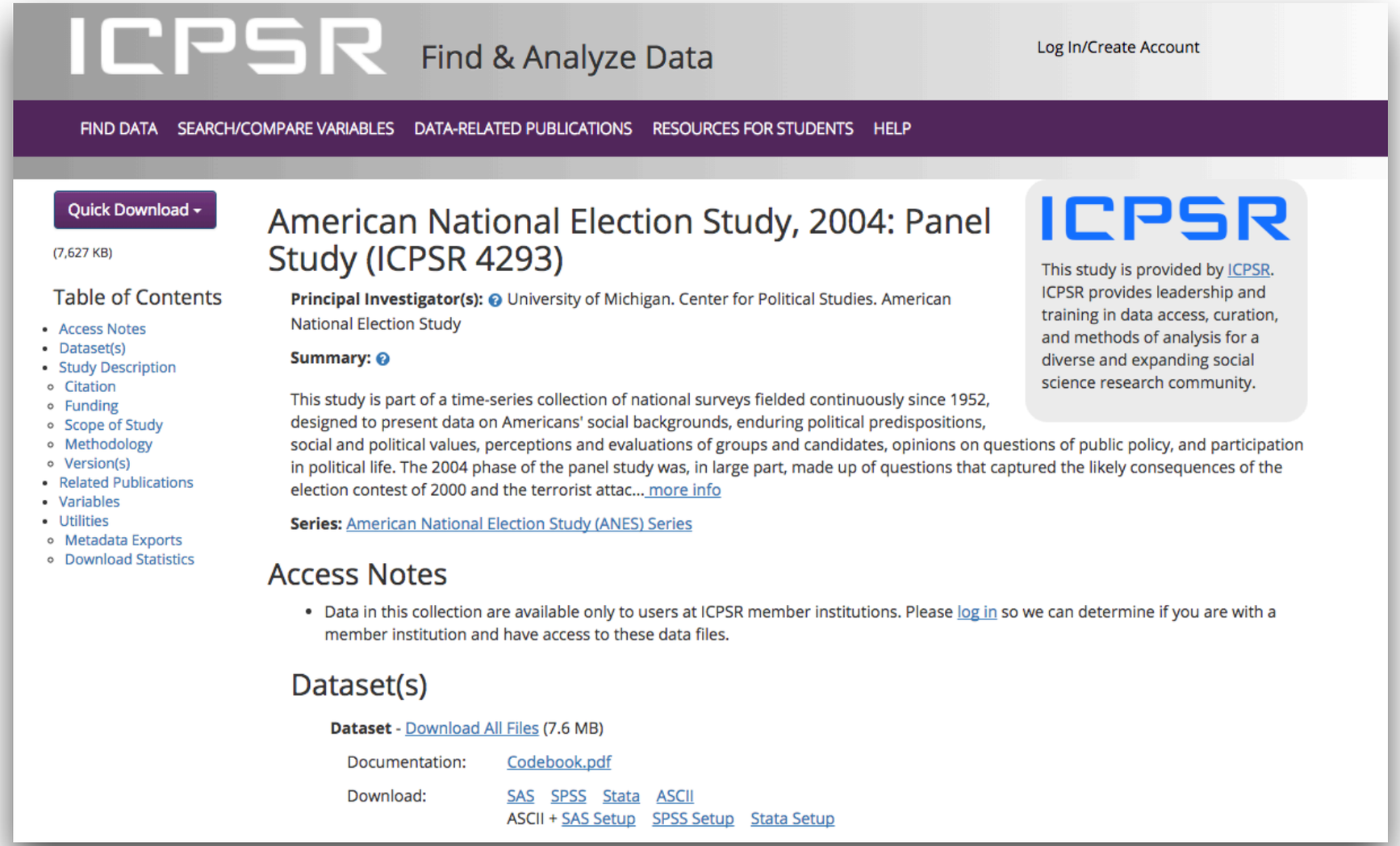
Three basic forms of structured metadata in data curation

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)

```
▼<codeBook version="2.1" ID="ICPSR04245">
  ▼<docDscr>
    ▼<citation>
      ▼<titlStmt>
          <titl>Metadata record for ANES 2004 Time Series Study</titl>
          <IDNo agency="ICPSR">4245</IDNo>
        </titlStmt>
      ▼<prodStmt>
        ▼<producer abbr="ICPSR">
            <ExtLink URI="http://www.icpsr.umich.edu/images/icpsr-logo.gif" title="ICPSR Logo" role="image"/>
            Inter-university Consortium for Political and Social Research
            <ExtLink URI="http://www.icpsr.umich.edu/ICPSR/" title="URL of ICPSR Web Site"/>
          </producer>
```

<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):

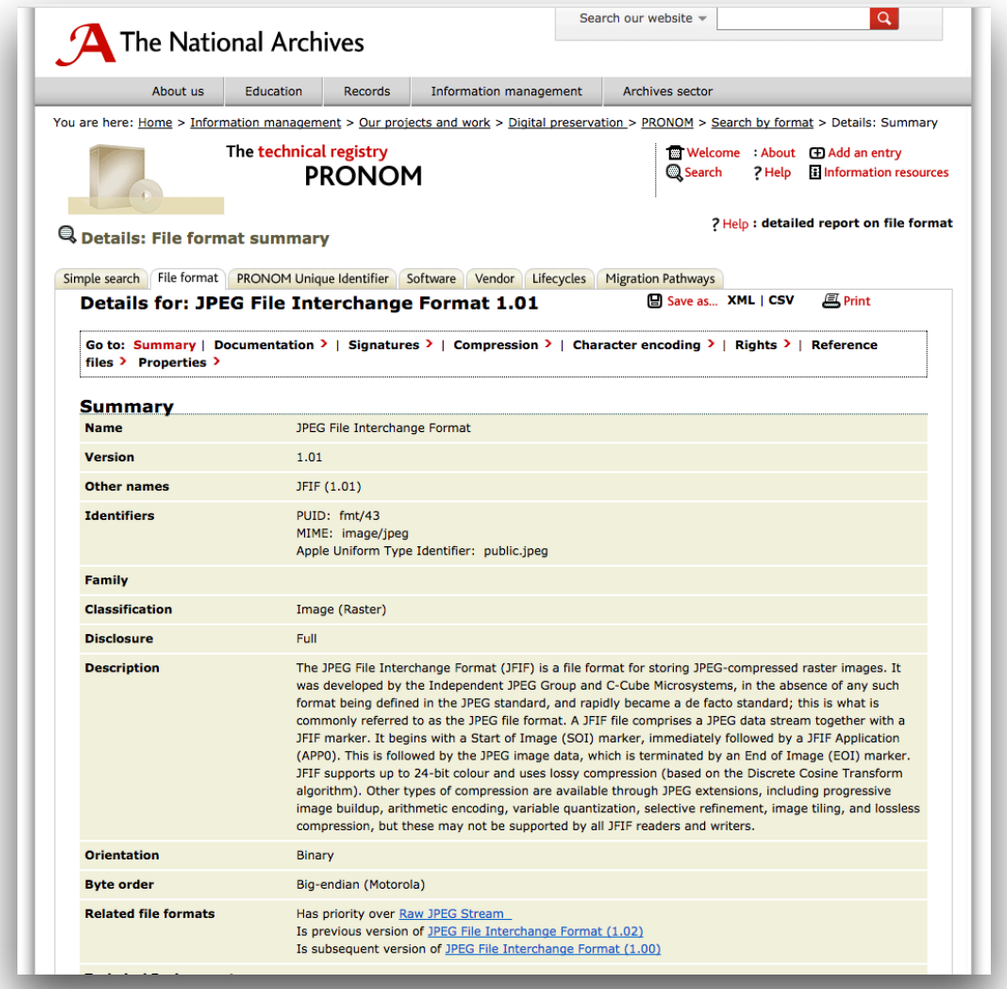
```
<titl>Domestic Violence Experience in Omaha, Nebraska, 1986-1987</titl>

<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)



The National Archives

Search our website

About us | Education | Records | Information management | Archives sector

You are here: Home > Information management > Our projects and work > Digital preservation > PRONOM > Search by format > Details: Summary

The technical registry
PRONOM

Welcome : About | Add an entry
Search | Help | Information resources

Help : detailed report on file format

Details: File format summary

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles | Migration Pathways

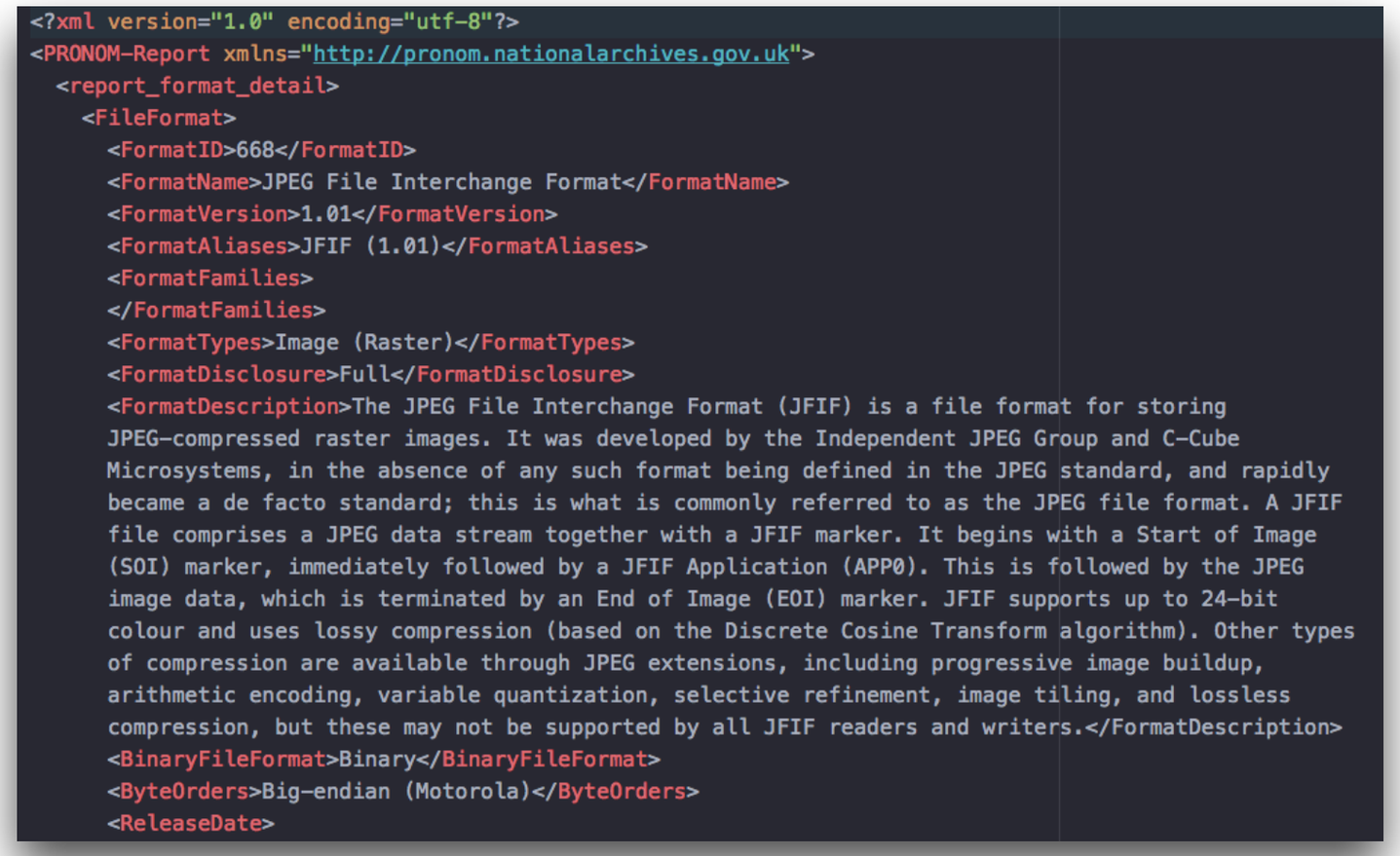
Details for: JPEG File Interchange Format 1.01

Save as... XML | CSV | Print

Go to: Summary | Documentation | Signatures | Compression | Character encoding | Rights | Reference files | Properties

Summary

Name	JPEG File Interchange Format
Version	1.01
Other names	JFIF (1.01)
Identifiers	PUID: fmt/43 MIME: image/jpeg Apple Uniform Type Identifier: public.jpeg
Family	
Classification	Image (Raster)
Disclosure	Full
Description	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.
Orientation	Binary
Byte order	Big-endian (Motorola)
Related file formats	Has priority over Raw JPEG Stream Is previous version of JPEG File Interchange Format (1.02) Is subsequent version of JPEG File Interchange Format (1.00)



```
<?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>
```

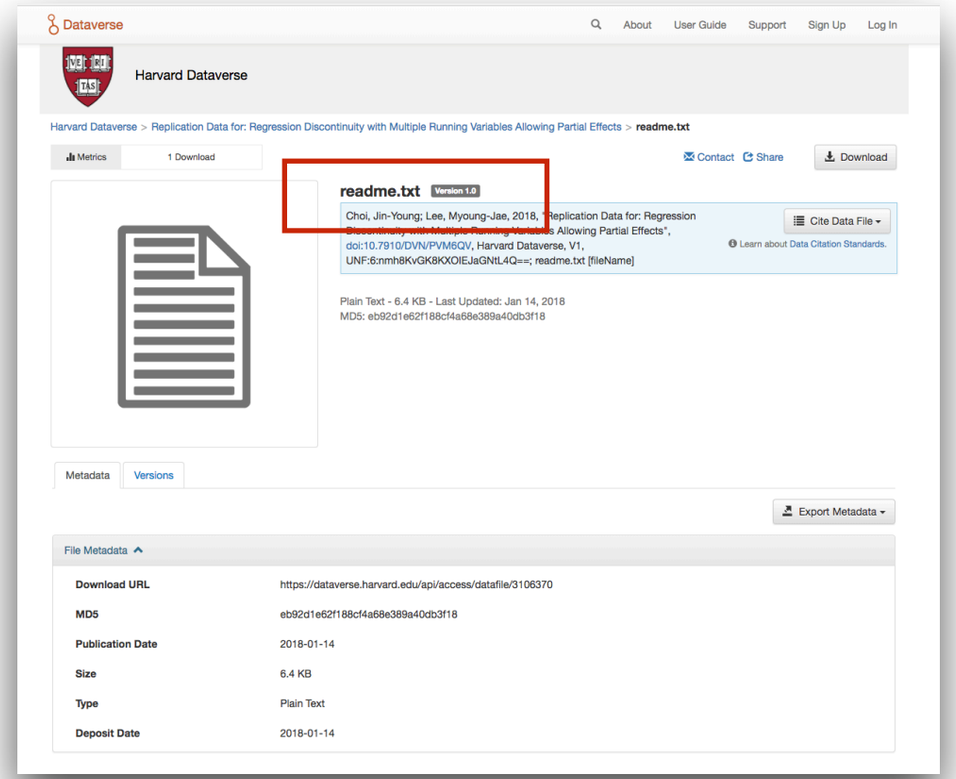
Unstructured Metadata or **Documentation**
(human readable)

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	numeric	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	Neighborhoc	text	neighborhood	The Department of Public Health and the Mayorâ??s Office of Housing and Community Development, with support from the Planning Dep	
Rent Arbitration Board	Eviction Notices	Location	Geometry	geometry: po	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			
V5	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 are various types of information objects playing the role of evidence.

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

Data Quality is "…the degree to which a set of characteristics of data fulfills stated requirements."

Slides, data and docs at:
https://github.com/nniiicc/LIS-530-2019

nmweber@uw.edu