Metadata & Documentation

for

Data Curation

Lecture 2

INFX 551

Winter 2018

Lecture 1

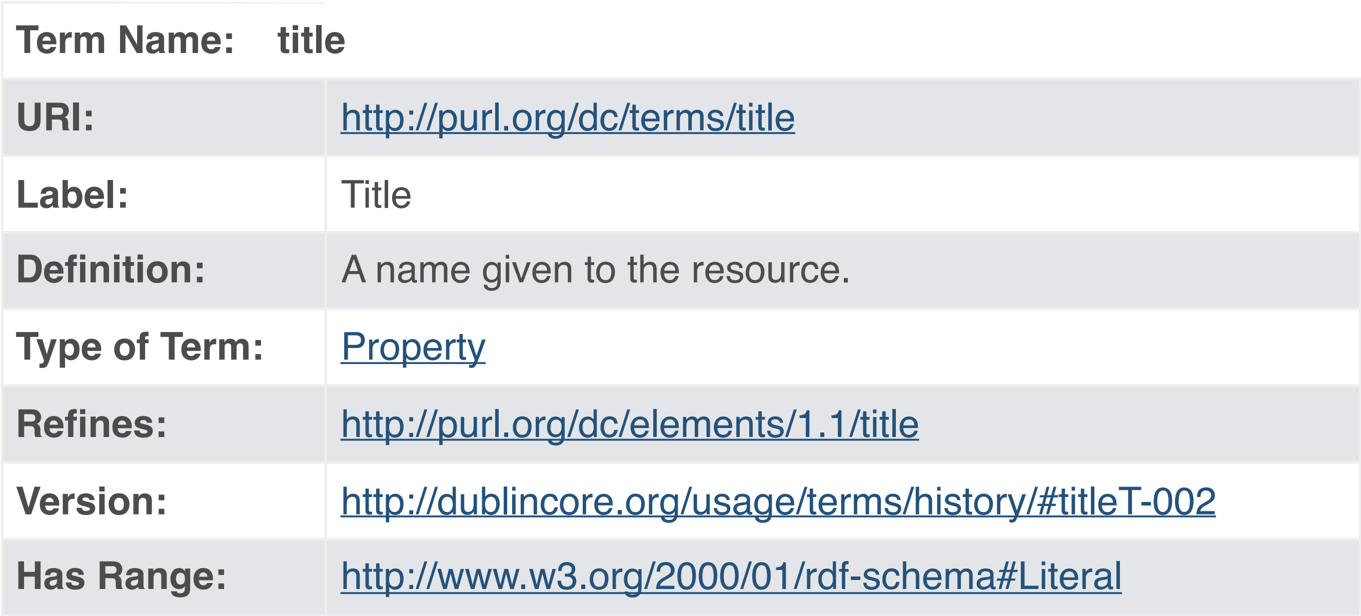
- Ontology & Epistemology
- Ontology (in Information Science)
- Knowledge Representation: Classes vs Instances
    - Attribute Value Pairs
- Expressivity vs Tractability

Lecture 2

- Metadata & Documentation
    - Structured vs Unstructured
- Encodings
- Standards Schema Development (application profiles)
- Forms of metadata
- Forms of documentation

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

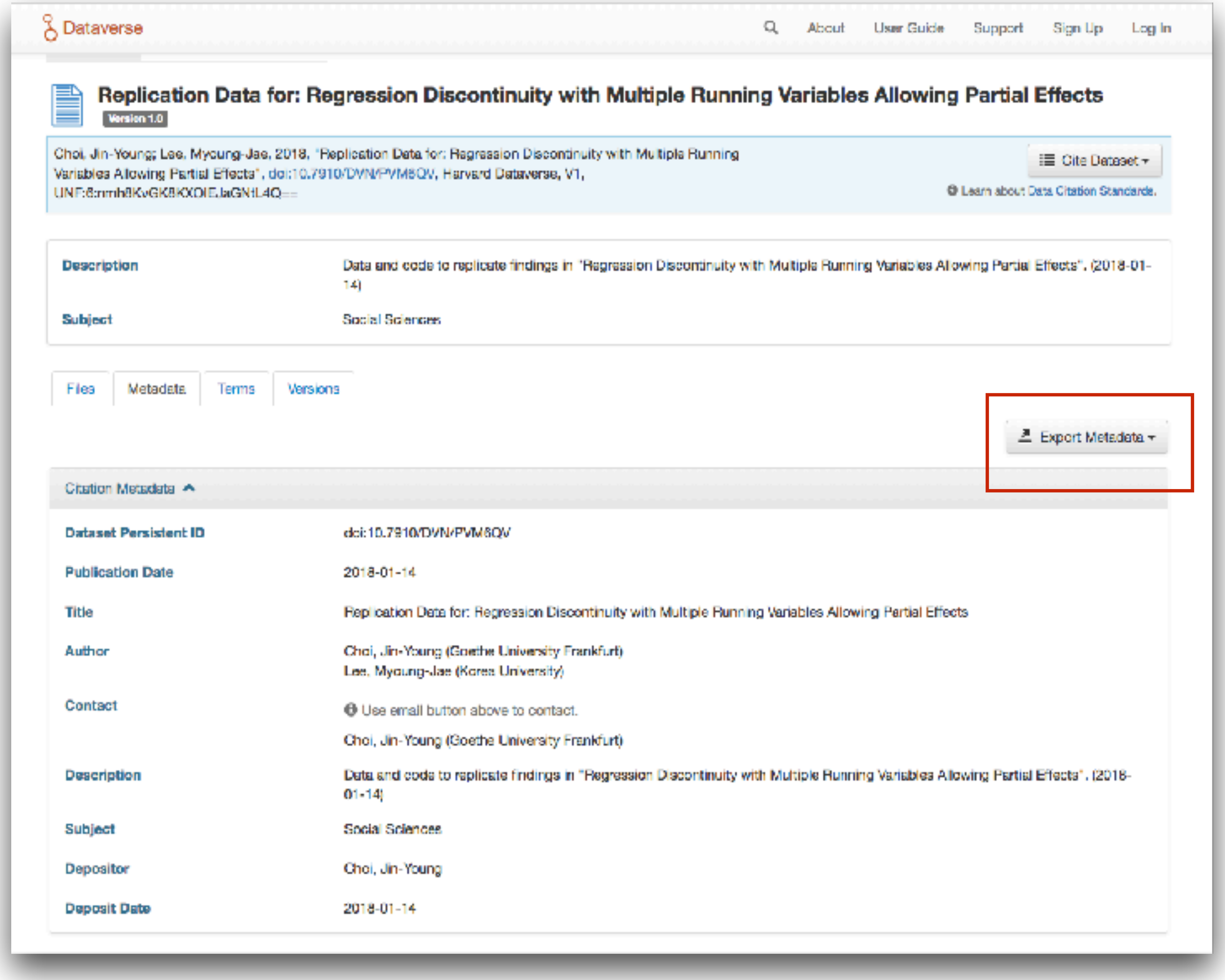
<dc: title>Hitchhikers Guide to the Galaxy**</dc:title>**



Term Name:	title
URI:	http://purl.org/dc/terms/title
Label:	Title
Definition:	A name given to the resource.
Type of Term:	Property
Refines:	http://purl.org/dc/elements/1.1/title
Version:	http://dublincore.org/usage/terms/history/#titleT-002
Has Range:	http://www.w3.org/2000/01/rdf-schema#Literal

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

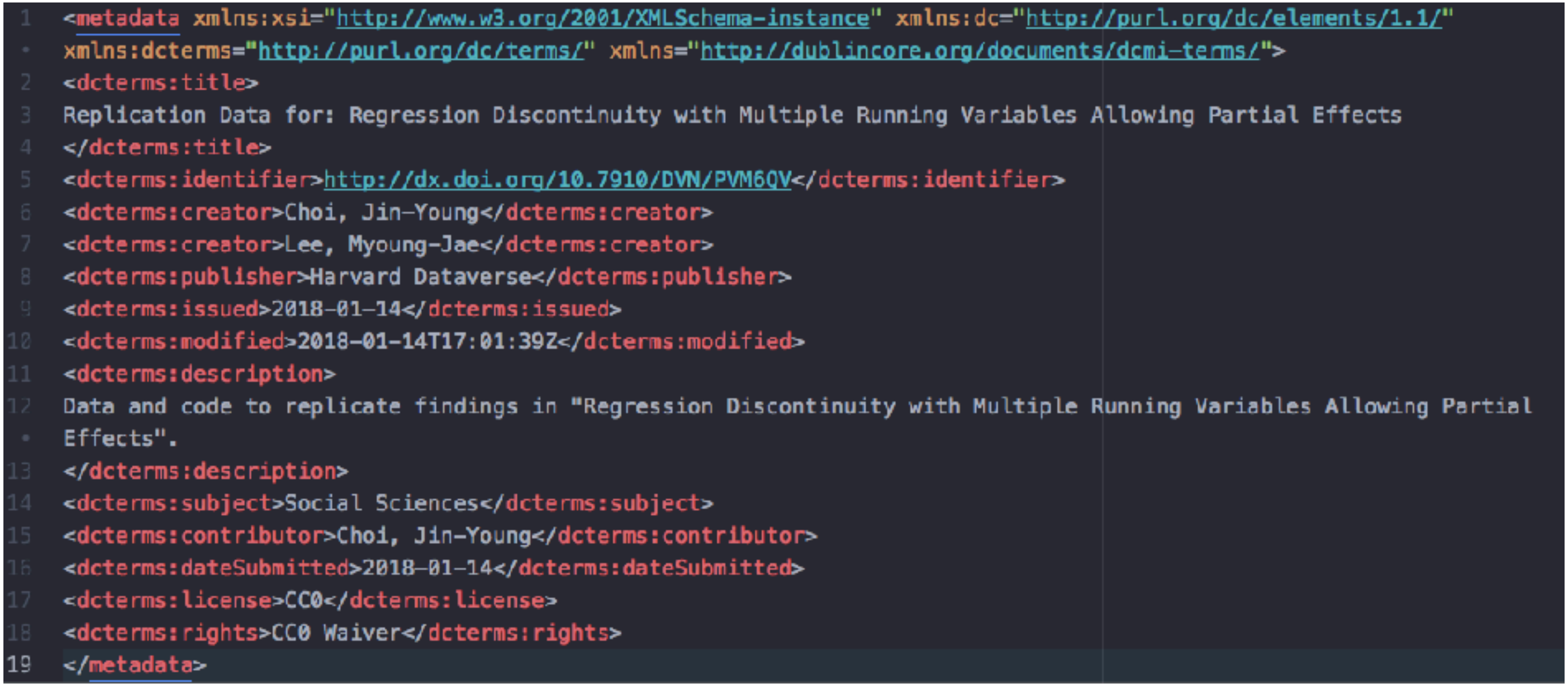


Replication Data for: Regression Discontinuity with Multiple Running Variables Allowing Partial Effects

Choi, Jin-Young; Lee, Myoung-Jae, 2018, "Replication Data for: Regression Discontinuity with Multiple Running Variables Allowing Partial Effects", doi:10.7910/DVN/PVM8QV, Harvard Dataverse, V1, UNF:6:nmh8KvGK8KXDIE.laGN1L4Q==

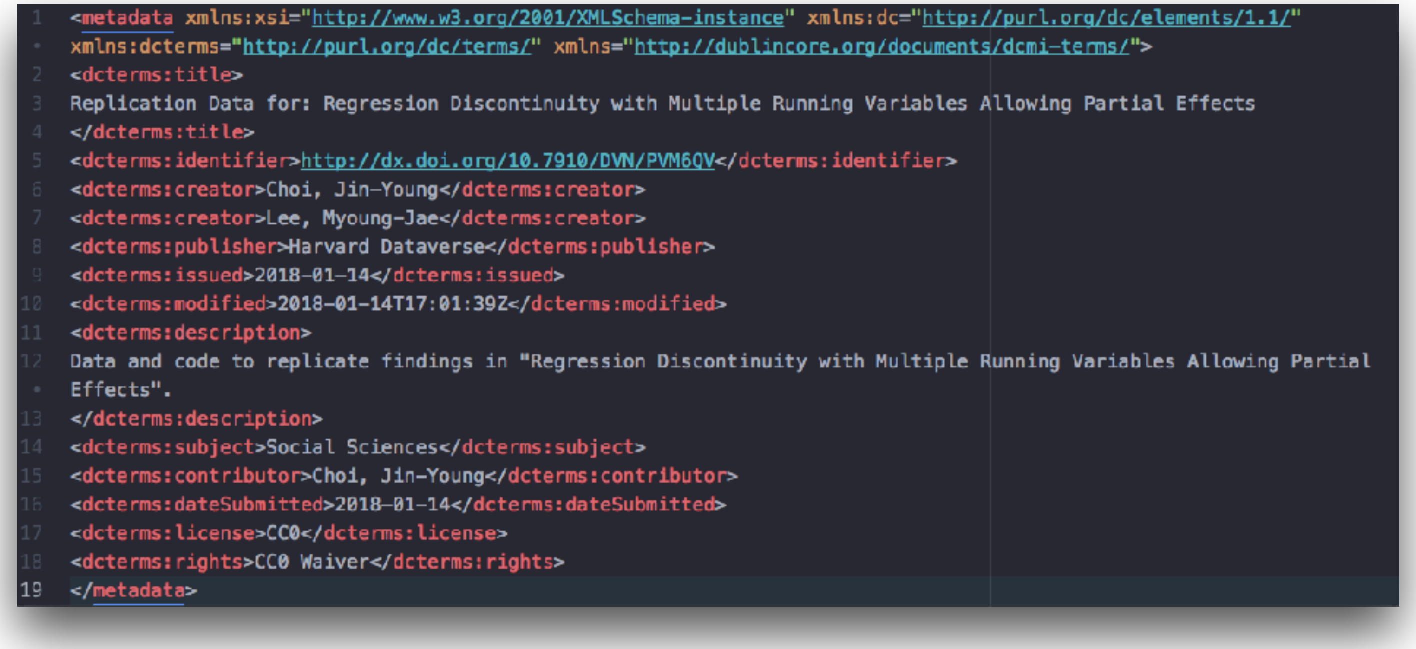
Description	Data and code to replicate findings in "Regression Discontinuity with Multiple Running Variables Allowing Partial Effects". (2018-01-14)
Subject	Social Sciences

Dataset Persistent ID	doi:10.7910/DVN/PVM8QV
Publication Date	2018-01-14
Title	Replication Data for: Regression Discontinuity with Multiple Running Variables Allowing Partial Effects
Author	Choi, Jin-Young (Goethe University Frankfurt) Lee, Myoung-Jae (Korea University)
Contact	Use email button above to contact. Choi, Jin-Young (Goethe University Frankfurt)
Description	Data and code to replicate findings in "Regression Discontinuity with Multiple Running Variables Allowing Partial Effects". (2018-01-14)
Subject	Social Sciences
Depositor	Choi, Jin-Young
Deposit Date	2018-01-14



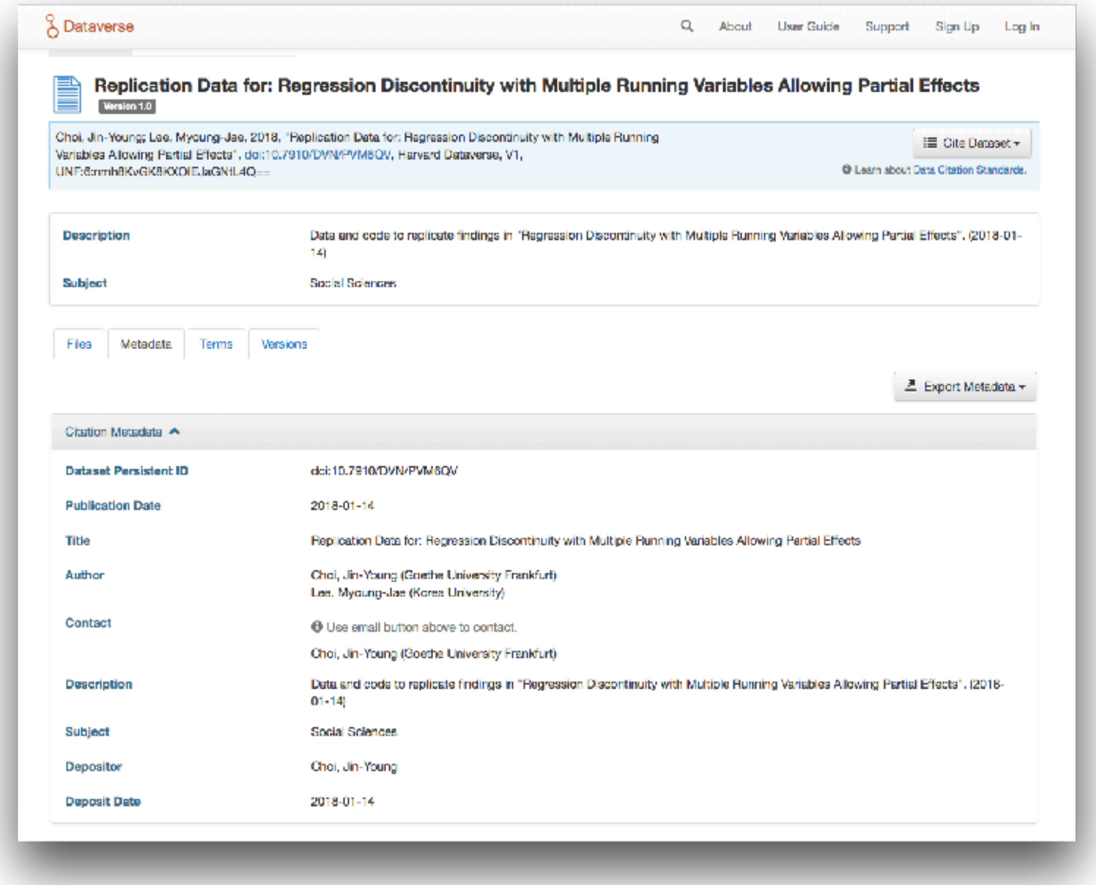
```xml
<metadata xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns="http://dublincore.org/documents/dcmi-terms/">
<dcterms:title>
Replication Data for: Regression Discontinuity with Multiple Running Variables 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>
```

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



Replication Data for: Regression Discontinuity with Multiple Running Variables Allowing Partial Effects

Choi, Jin-Young; Lee, Myoung-Jae, 2018, "Replication Data for: Regression Discontinuity with Multiple Running Variables Allowing Partial Effects", doi:10.7910/DVN/PVM6QV, Harvard Dataverse, V1, UNF:6:mnh8KvGK8KXOIE.laGN1L4Q==

Learn about Data Citation Standards.

Description	Data and code to replicate findings in "Regression Discontinuity with Multiple Running Variables Allowing Partial Effects". (2018-01-14)
Subject	Social Sciences

Dataset Persistent ID	doi:10.7910/DVN/PVM6QV
Publication Date	2018-01-14
Title	Replication Data for: Regression Discontinuity with Multiple Running Variables Allowing Partial Effects
Author	Choi, Jin-Young (Goethe University Frankfurt) Lee, Myoung-Jae (Korea University)
Contact	Use email button above to contact. Choi, Jin-Young (Goethe University Frankfurt)
Description	Data and code to replicate findings in "Regression Discontinuity with Multiple Running Variables Allowing Partial Effects". (2018-01-14)
Subject	Social Sciences
Depositor	Choi, Jin-Young
Deposit Date	2018-01-14

Human Readable

Structured Metadata

1. Is encoded in a machine readable format (xml, json)
2. Is compliant with (follows) a standard schema (dublin core, EML, DDI) **OR** accurately defines it's own schema/

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

Attribute	Value
Name	Masha
Eye Color	Blue

Encoding Attributes

XML

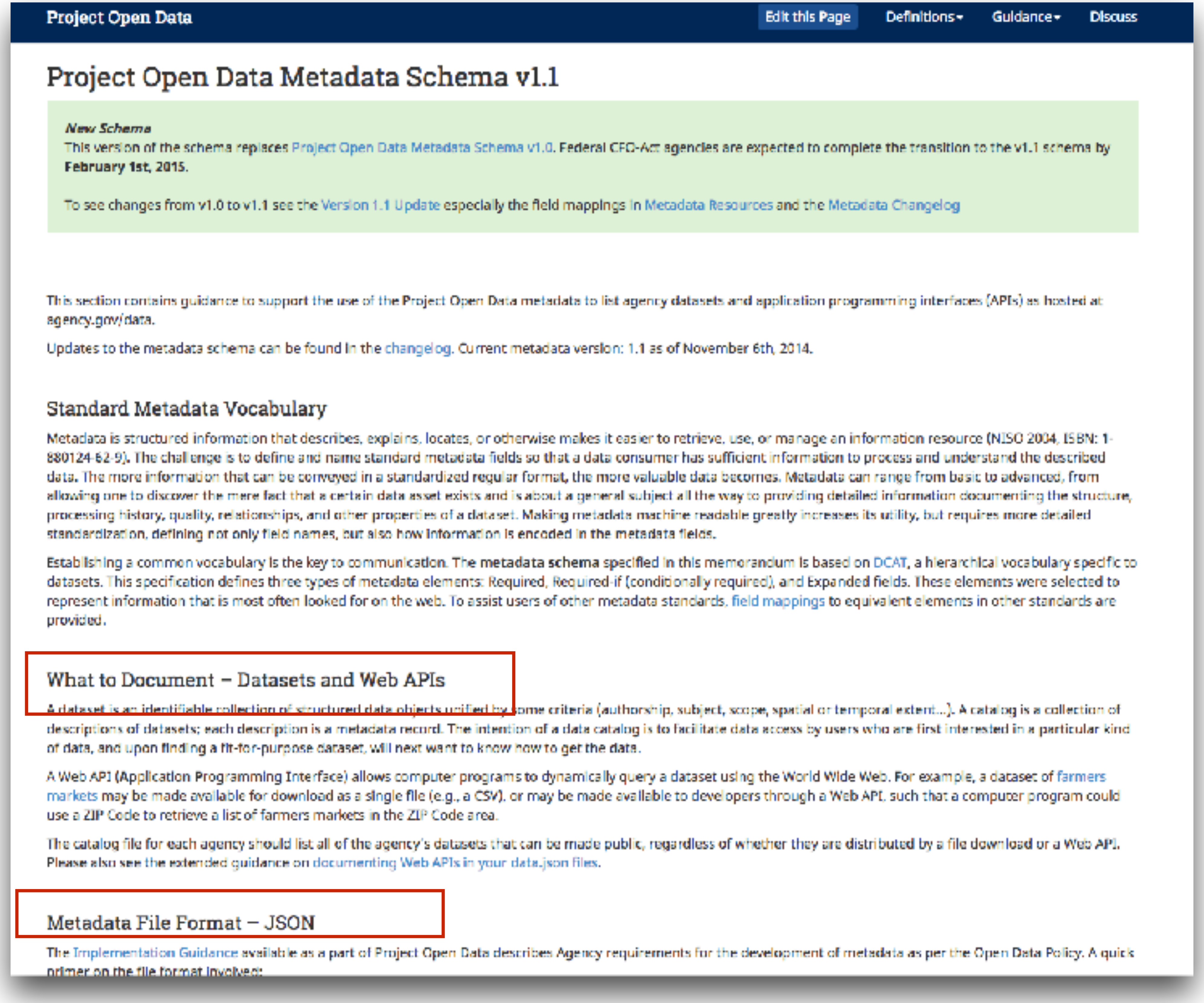
```
<eye_color>blue</eye_color>
```

JSON

```
{
  “eye_color": "blue",
}
```

A metadata schema standard will:

- Define attributes *(e.g. what do you mean by “creator” in ecology?)*
- Suggest controls of values (e.g. dates = MM-DD-YYYY)
- Define requirements for being “well-formed” (e.g. what fields are absolutely necessary for a valid metadata record?)
- Provide an implementation of the standard in an encoding *(e.g. XML)*
- Provide example use cases that are satisfied by the standard.



Project Open Data Metadata Schema v1.1

New Schema
This version of the schema replaces Project Open Data Metadata Schema v1.0. Federal CFO-Act agencies are expected to complete the transition to the v1.1 schema by **February 1st, 2015.**

To see changes from v1.0 to v1.1 see the Version 1.1 Update especially the field mappings in Metadata Resources and the Metadata Changelog

This section contains guidance to support the use of the Project Open Data metadata to list agency datasets and application programming interfaces (APIs) as hosted at agency.gov/data.

Updates to the metadata schema can be found in the changelog. Current metadata version: 1.1 as of November 6th, 2014.

Standard Metadata Vocabulary

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource (NISO 2004, ISBN: 1-880124-62-9). The challenge is to define and name standard metadata fields so that a data consumer has sufficient information to process and understand the described data. The more information that can be conveyed in a standardized regular format, the more valuable data becomes. Metadata can range from basic to advanced, from allowing one to discover the mere fact that a certain data asset exists and is about a general subject all the way to providing detailed information documenting the structure, processing history, quality, relationships, and other properties of a dataset. Making metadata machine readable greatly increases its utility, but requires more detailed standardization, defining not only field names, but also how information is encoded in the metadata fields.

Establishing a common vocabulary is the key to communication. The metadata schema specified in this memorandum is based on DCAT, a hierarchical vocabulary specific to datasets. This specification defines three types of metadata elements: Required, Required-if (conditionally required), and Expanded fields. These elements were selected to represent information that is most often looked for on the web. To assist users of other metadata standards, field mappings to equivalent elements in other standards are provided.

What to Document – Datasets and Web APIs

A dataset is an identifiable collection of structured data objects unified by some criteria (authorship, subject, scope, spatial or temporal extent...). A catalog is a collection of descriptions of datasets; each description is a metadata record. The intention of a data catalog is to facilitate data access by users who are first interested in a particular kind of data, and upon finding a fit-for-purpose dataset, will next want to know how to get the data.

A Web API (Application Programming Interface) allows computer programs to dynamically query a dataset using the World Wide Web. For example, a dataset of farmers markets may be made available for download as a single file (e.g., a CSV), or may be made available to developers through a Web API, such that a computer program could use a ZIP Code to retrieve a list of farmers markets in the ZIP Code area.

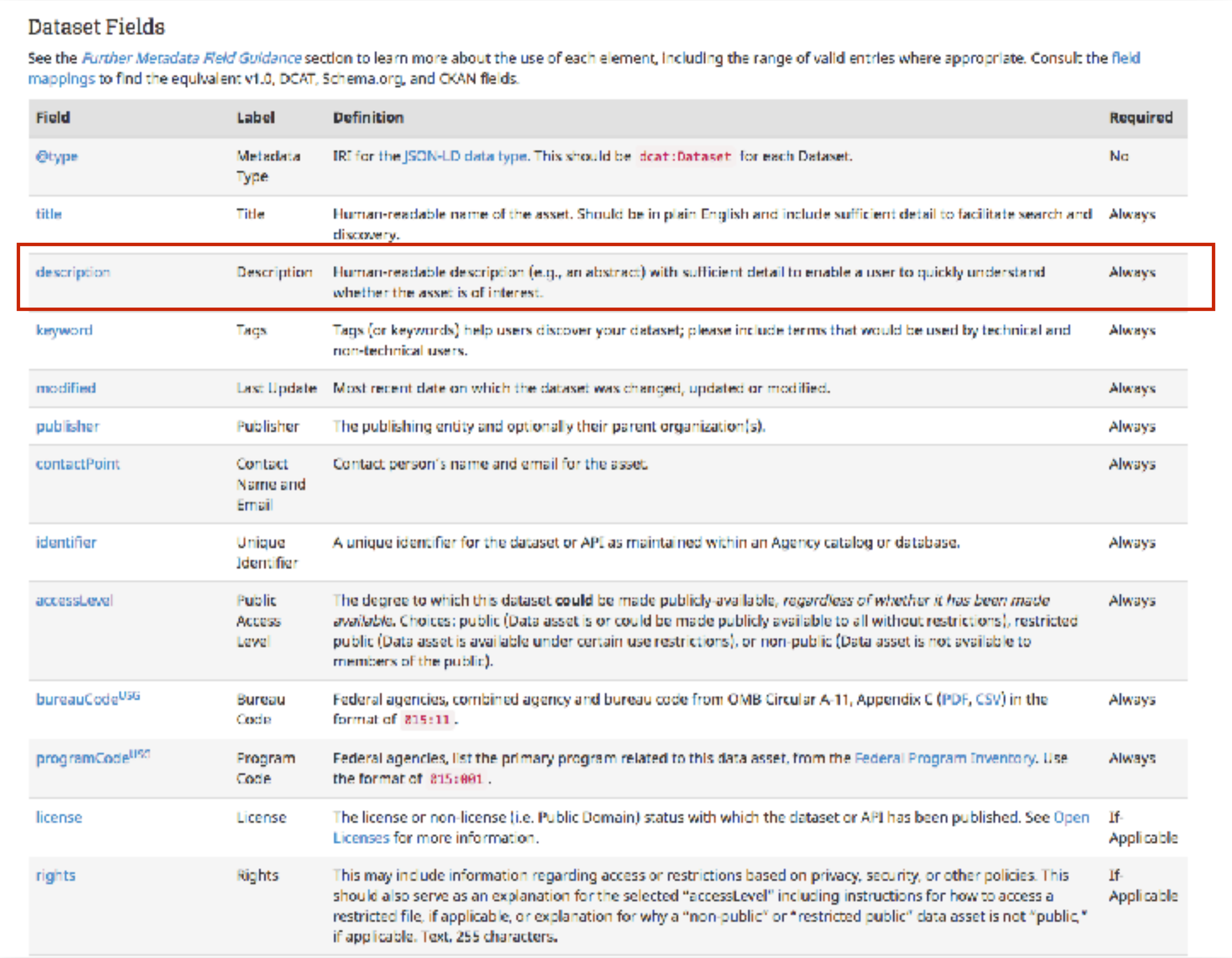
The catalog file for each agency should list all of the agency's datasets that can be made public, regardless of whether they are distributed by a file download or a Web API. Please also see the extended guidance on documenting Web APIs in your data.json files.

Metadata File Format – JSON

The Implementation Guidance available as a part of Project Open Data describes Agency requirements for the development of metadata as per the Open Data Policy. A quick primer on the file format involved:

https://project-open-data.cio.gov/v1.1/schema/

Defining the attributes of Open Data



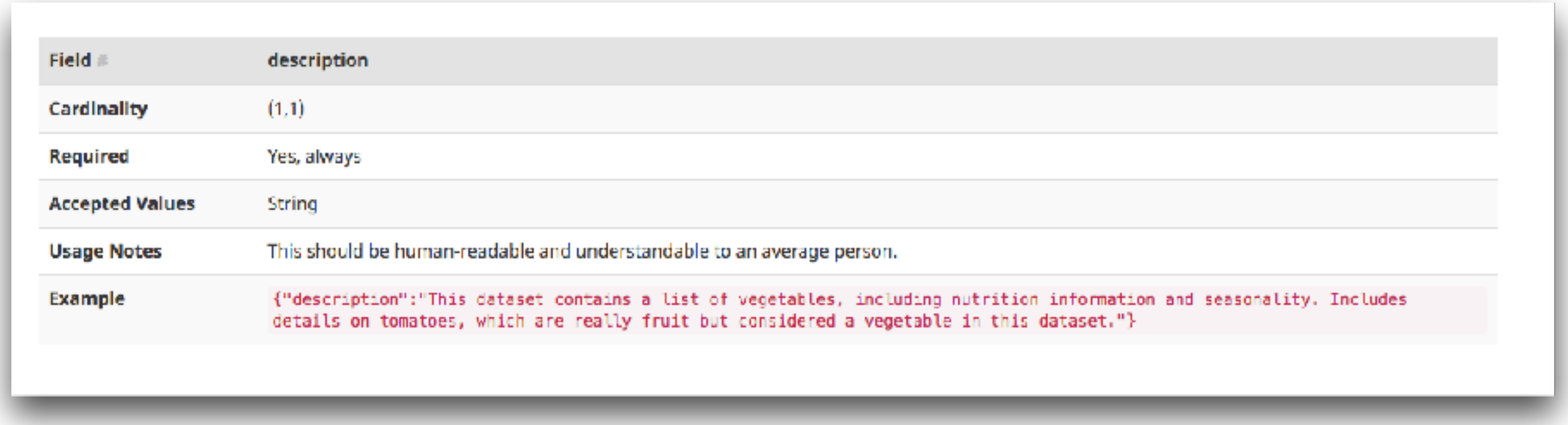
Dataset Fields

See the *Further Metadata Field Guidance* section to learn more about the use of each element, including the range of valid entries where appropriate. Consult the field mappings to find the equivalent v1.0, DCAT, Schema.org, and CKAN fields.

Field	Label	Definition	Required
@type	Metadata Type	IRI for the JSON-LD data type. This should be `dcat:Dataset` for each Dataset.	No
title	Title	Human-readable name of the asset. Should be in plain English and include sufficient detail to facilitate search and discovery.	Always
description	Description	Human-readable description (e.g., an abstract) with sufficient detail to enable a user to quickly understand whether the asset is of interest.	Always
keyword	Tags	Tags (or keywords) help users discover your dataset; please include terms that would be used by technical and non-technical users.	Always
modified	Last Update	Most recent date on which the dataset was changed, updated or modified.	Always
publisher	Publisher	The publishing entity and optionally their parent organization(s).	Always
contactPoint	Contact Name and Email	Contact person's name and email for the asset.	Always
identifier	Unique Identifier	A unique identifier for the dataset or API as maintained within an Agency catalog or database.	Always
accessLevel	Public Access Level	The degree to which this dataset **could** be made publicly-available, *regardless of whether it has been made available*. Choices: public (Data asset is or could be made publicly available to all without restrictions), restricted public (Data asset is available under certain use restrictions), or non-public (Data asset is not available to members of the public).	Always
bureauCode[USG]	Bureau Code	Federal agencies, combined agency and bureau code from OMB Circular A-11, Appendix C (PDF, CSV) in the format of `015:11`.	Always
programCode[USG]	Program Code	Federal agencies, list the primary program related to this data asset, from the Federal Program Inventory. Use the format of `015:001`.	Always
license	License	The license or non-license (i.e. Public Domain) status with which the dataset or API has been published. See Open Licenses for more information.	If-Applicable
rights	Rights	This may include information regarding access or restrictions based on privacy, security, or other policies. This should also serve as an explanation for the selected "accessLevel" including instructions for how to access a restricted file, if applicable, or explanation for why a "non-public" or "restricted public" data asset is not "public," if applicable. Text, 255 characters.	If-Applicable

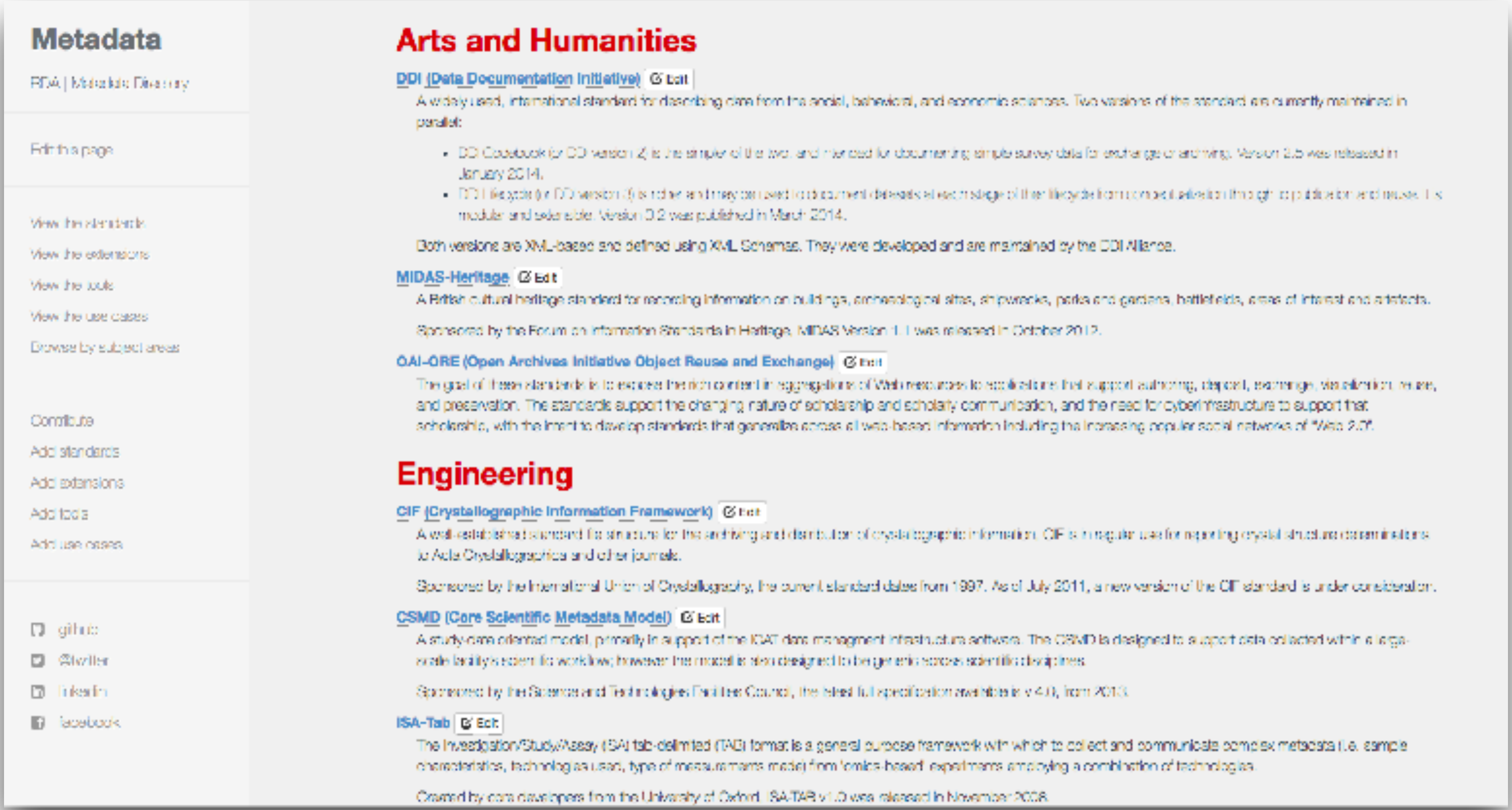
https://project-open-data.cio.gov/v1.1/schema/

Instructions on how to use…



Field #	description
Cardinality	(1,1)
Required	Yes, always
Accepted Values	String
Usage Notes	This should be human-readable and understandable to an average person.
Example	`{"description":"This dataset contains a list of vegetables, including nutrition information and seasonality. Includes details on tomatoes, which are really fruit but considered a vegetable in this dataset."}`

https://project-open-data.cio.gov/v1.1/schema/



Arts and Humanities

DDI (Data Documentation Initiative)

A widely used, international standard for describing data from the social, behavioral, and economic sciences. Two versions of the standard are currently maintained in parallel:

- DDI Codebook (or DDI version 2) is the simpler of the two, and intended for documenting simple survey data for exchange or archiving. Version 2.5 was released in January 2014.
- DDI Lifecycle (or DDI version 3) is richer and may be used to document datasets at each stage of their lifecycle from conceptualisation through to publication and reuse. It is modular and extensible. Version 3.2 was published in March 2014.

Both versions are XML-based and defined using XML Schemas. They were developed and are maintained by the DDI Alliance.

MIDAS-Heritage

A British cultural heritage standard for recording information on buildings, archaeological sites, shipwrecks, parks and gardens, battlefields, areas of interest and artefacts.

Sponsored by the Forum on Information Standards in Heritage. MIDAS Version 1.1 was released in October 2012.

OAI-ORE (Open Archives Initiative Object Reuse and Exchange)

The goal of these standards is to expose the rich content in aggregations of Web resources to applications that support authoring, deposit, exchange, visualization, reuse, and preservation. The standards support the changing nature of scholarship and scholarly communication, and the need for cyberinfrastructure to support that scholarship, with the intent to develop standards that generalize across all web-based information including the increasing popular social networks of "Web 2.0".

Engineering

CIF (Crystallographic Information Framework)

A well-established standard file structure for the archiving and distribution of crystallographic information. CIF is in regular use for reporting crystal structure determinations to Acta Crystallographica and other journals.

Sponsored by the International Union of Crystallography, the current standard dates from 1997. As of July 2011, a new version of the CIF standard is under consideration.

CSMD (Core Scientific Metadata Model)

A study-data oriented model, primarily in support of the ICAT data management infrastructure software. The CSMD is designed to support data collected within a large-scale facility's scientific workflow; however the model is also designed to be generic across scientific disciplines.

Sponsored by the Science and Technologies Facilities Council, the latest full specification available is v 4.0, from 2013.

ISA-Tab

The Investigation/Study/Assay (ISA) tab-delimited (TAB) format is a general purpose framework with which to collect and communicate complex metadata (i.e. sample characteristics, technologies used, type of measurements made) from 'omics-based' experiments employing a combination of technologies.

Created by core developers from the University of Oxford. ISA-TAB v1.0 was released in November 2008.

http://rd-alliance.github.io/metadata-directory/standards/

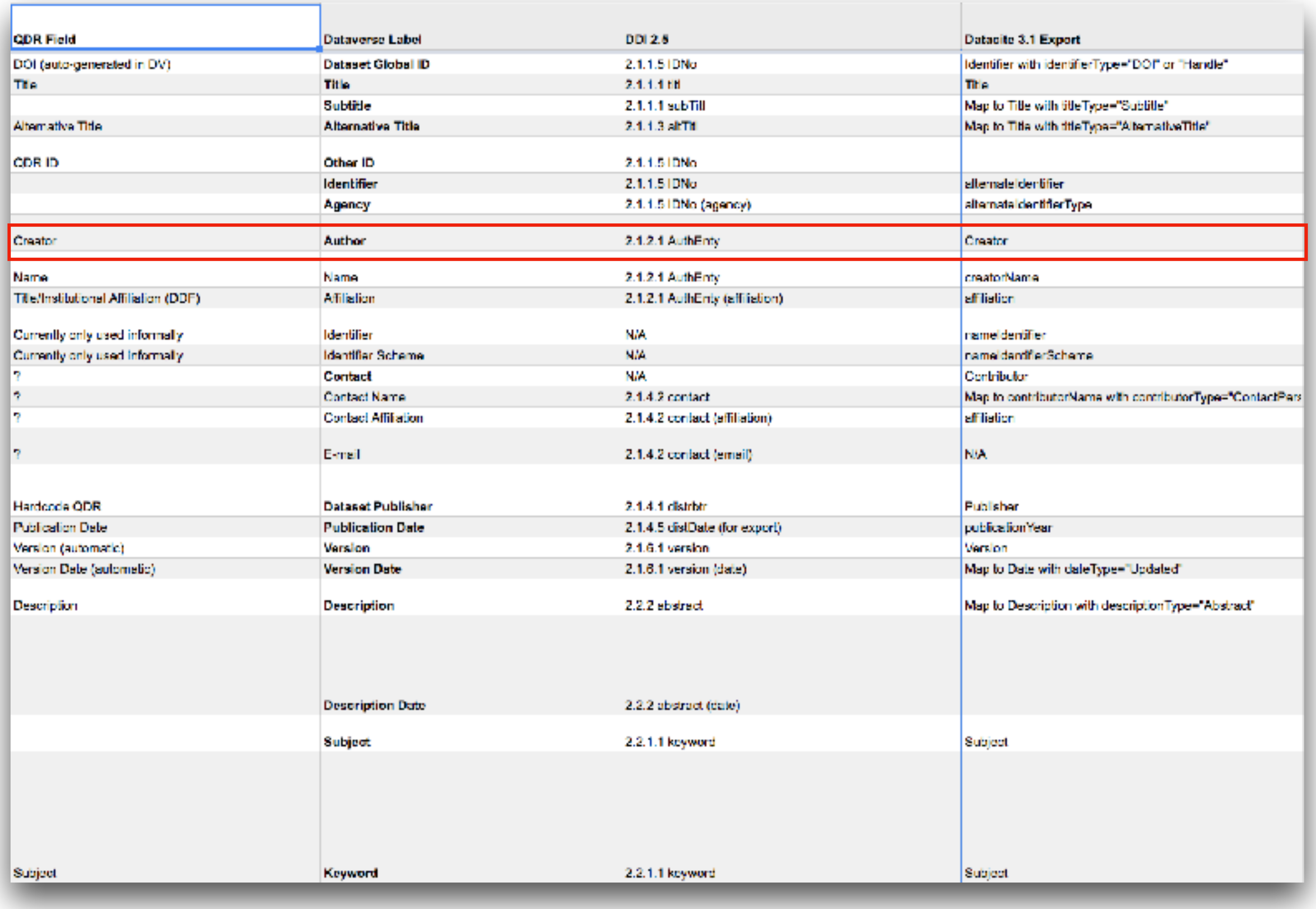
Structured Metadata

1. Is encoded in a machine readable format (xml, json)
2. Is compliant with (follows) a standard schema (dublin core, EML, DDI) **OR** accurately defines it’s own schema

Think of a custom schema like a playlist.

Individual tracks made by other people, but arranged by you to meet a particular purpose.

In data curation, we call a playlist an **application profile**.



QDR Field	Dataverse Label	DDI 2.5	Datacite 3.1 Export
DOI (auto-generated in DV)	**Dataset Global ID**	2.1.1.5 IDNo	Identifier with identifierType="DOI" or "Handle"
Title	**Title**	2.1.1.1 titl	Title
	Subtitle	2.1.1.1 subTitl	Map to Title with titleType="Subtitle"
Alternative Title	**Alternative Title**	2.1.1.3 altTitl	Map to Title with titleType="AlternativeTitle"
QDR ID	**Other ID**	2.1.1.5 IDNo	
	Identifier	2.1.1.5 IDNo	alternateIdentifier
	Agency	2.1.1.5 IDNo (agency)	alternateIdentifierType
Creator	**Author**	2.1.2.1 AuthEnty	Creator
Name	Name	2.1.2.1 AuthEnty	creatorName
Title/Institutional Affiliation (DDF)	Affiliation	2.1.2.1 AuthEnty (affiliation)	affiliation
Currently only used informally	Identifier	N/A	nameIdentifier
Currently only used Informally	Identifier Scheme	N/A	nameIdentifierScheme
?	**Contact**	N/A	Contributor
?	Contact Name	2.1.4.2 contact	Map to contributorName with contributorType="ContactPers
?	Contact Affiliation	2.1.4.2 contact (affiliation)	affiliation
?	E-mail	2.1.4.2 contact (email)	N/A
Hardcode QDR	**Dataset Publisher**	2.1.4.1 distrbtr	Publisher
Publication Date	**Publication Date**	2.1.4.5 distDate (for export)	publicationYear
Version (automatic)	**Version**	2.1.6.1 version	Version
Version Date (automatic)	**Version Date**	2.1.6.1 version (date)	Map to Date with dateType="Updated"
Description	**Description**	2.2.2 abstract	Map to Description with descriptionType="Abstract"
	Description Date	2.2.2 abstract (date)	
	Subject	2.2.1.1 keyword	Subject
Subject	**Keyword**	2.2.1.1 keyword	Subject

https://docs.google.com/spreadsheets/d/1kI1Qtq5JneY0cH6yIwORnCHyj7JluGns8VgvcFde4cg/edit#gid=0

What makes a good application profile?

*Balances expressivity vs tractability (see last lecture)

*Is applicable to a broad domain

*Uses standards that are complimentary

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)

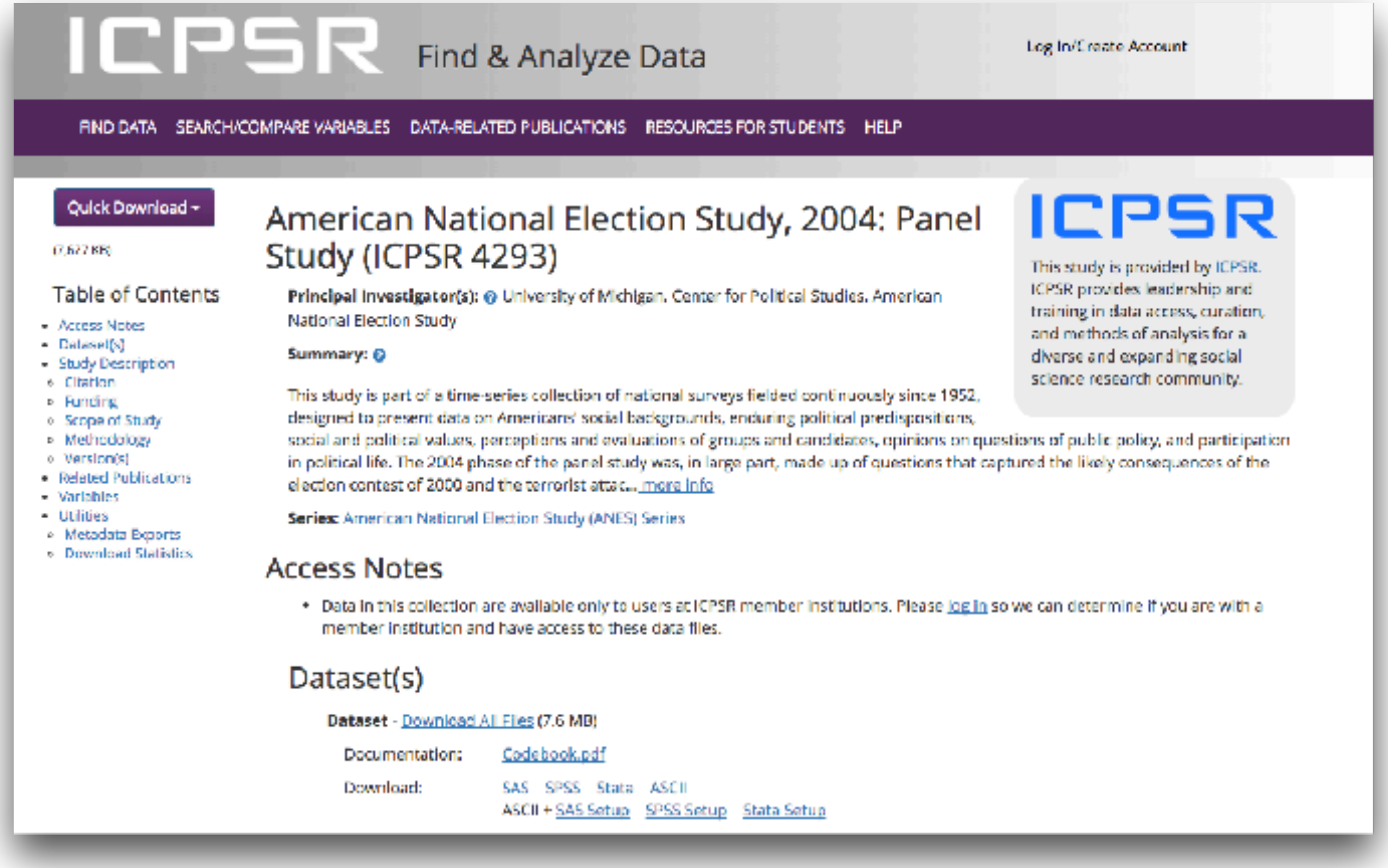


Table of Contents

- Access Notes
- Dataset(s)
- Study Description
  - Citation
  - Funding
  - Scope of Study
  - Methodology
  - Version(s)
- Related Publications
- Variables
- Utilities
  - Metadata Exports
  - Download Statistics

American National Election Study, 2004: Panel Study (ICPSR 4293)

Principal Investigator(s): University of Michigan. Center for Political Studies. American National Election Study

Summary:

This study is part of a time-series collection of national surveys fielded continuously since 1952, designed to present data on Americans' social backgrounds, enduring political predispositions, social and political values, perceptions and evaluations of groups and candidates, opinions on questions of public policy, and participation in political life. The 2004 phase of the panel study was, in large part, made up of questions that captured the likely consequences of the election contest of 2000 and the terrorist attac... more info

Series: American National Election Study (ANES) Series

ICPSR

This study is provided by ICPSR. ICPSR provides leadership and training in data access, curation, and methods of analysis for a diverse and expanding social science research community.

Access Notes

- Data in this collection are available only to users at ICPSR member institutions. Please log in so we can determine if you are with a member institution and have access to these data files.

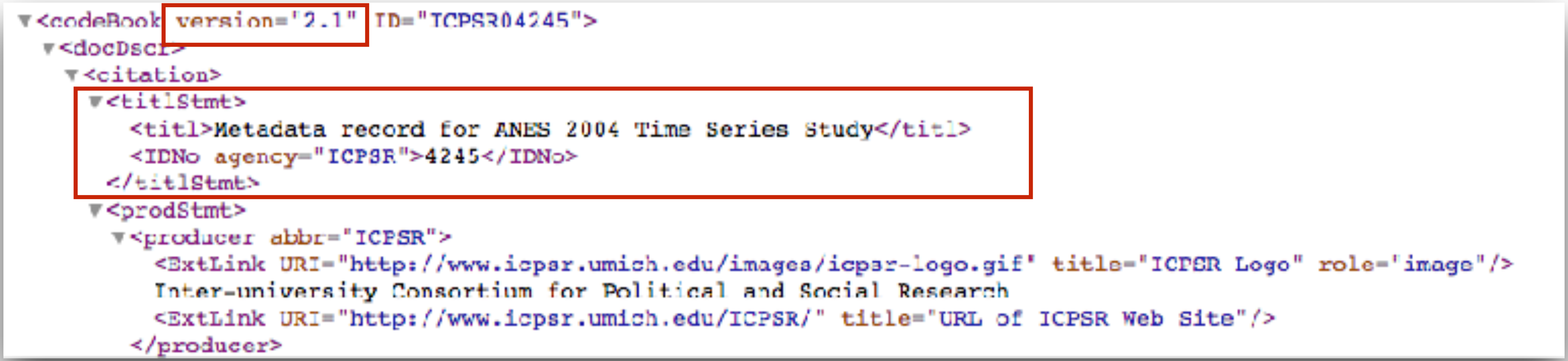
Dataset(s)

Dataset - Download All Files (7.6 MB)

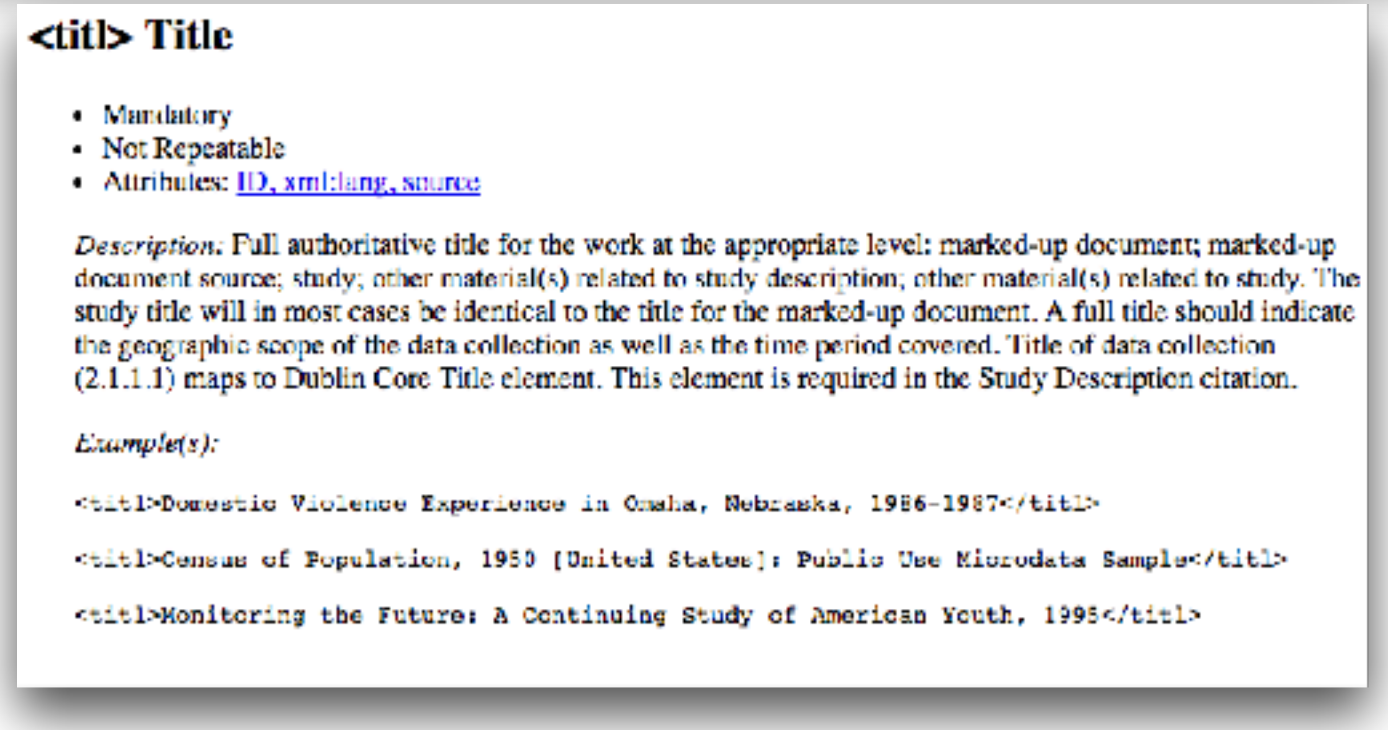
Documentation: Codebook.pdf

Download: SAS SPSS Stata ASCII
ASCII + SAS Setup SPSS Setup Stata Setup

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

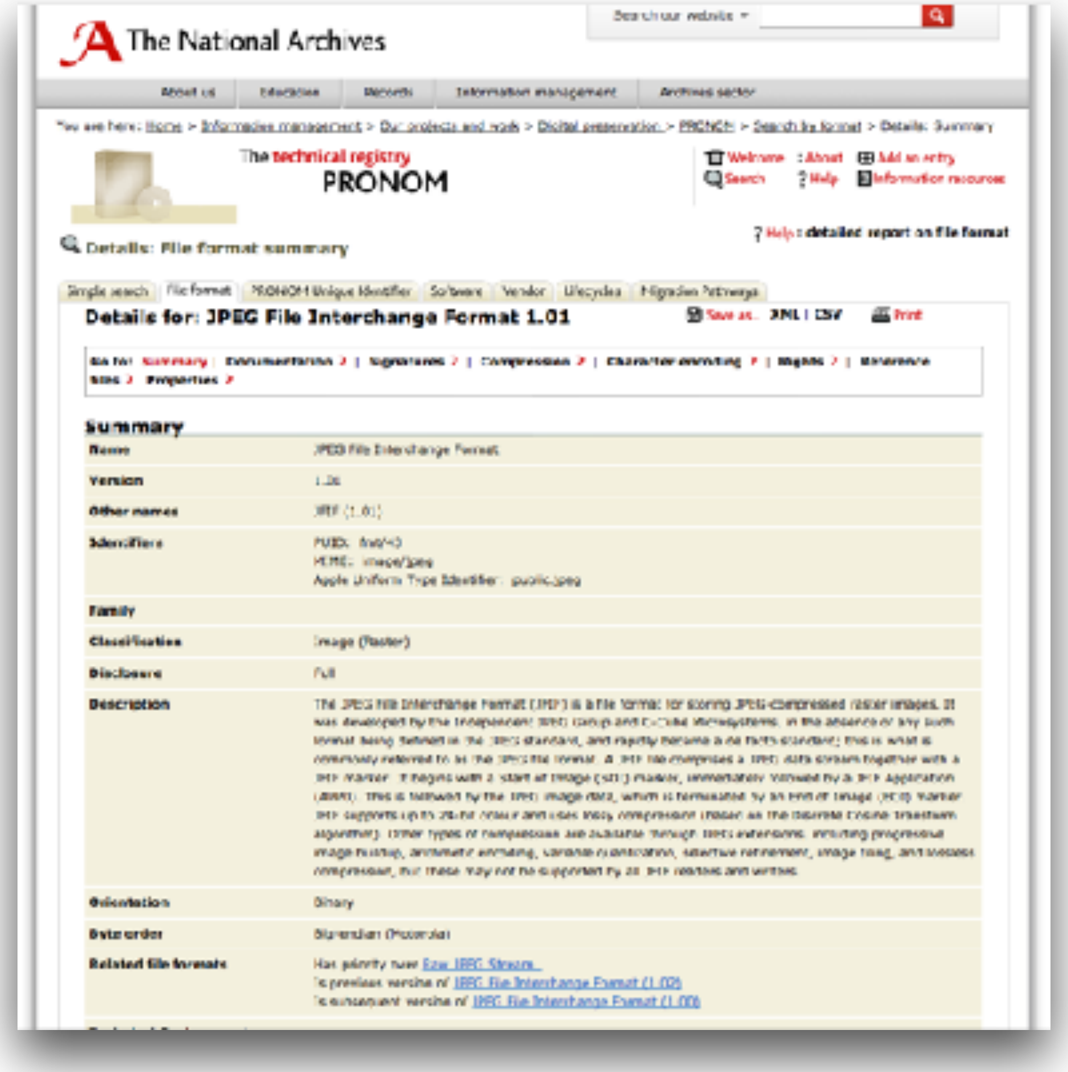
Description: Full authoritative title for the work at the appropriate level: marked-up document; marked-up document source; study; other material(s) related to study description; other material(s) related to study. The study title will in most cases be identical to the title for the marked-up document. 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)



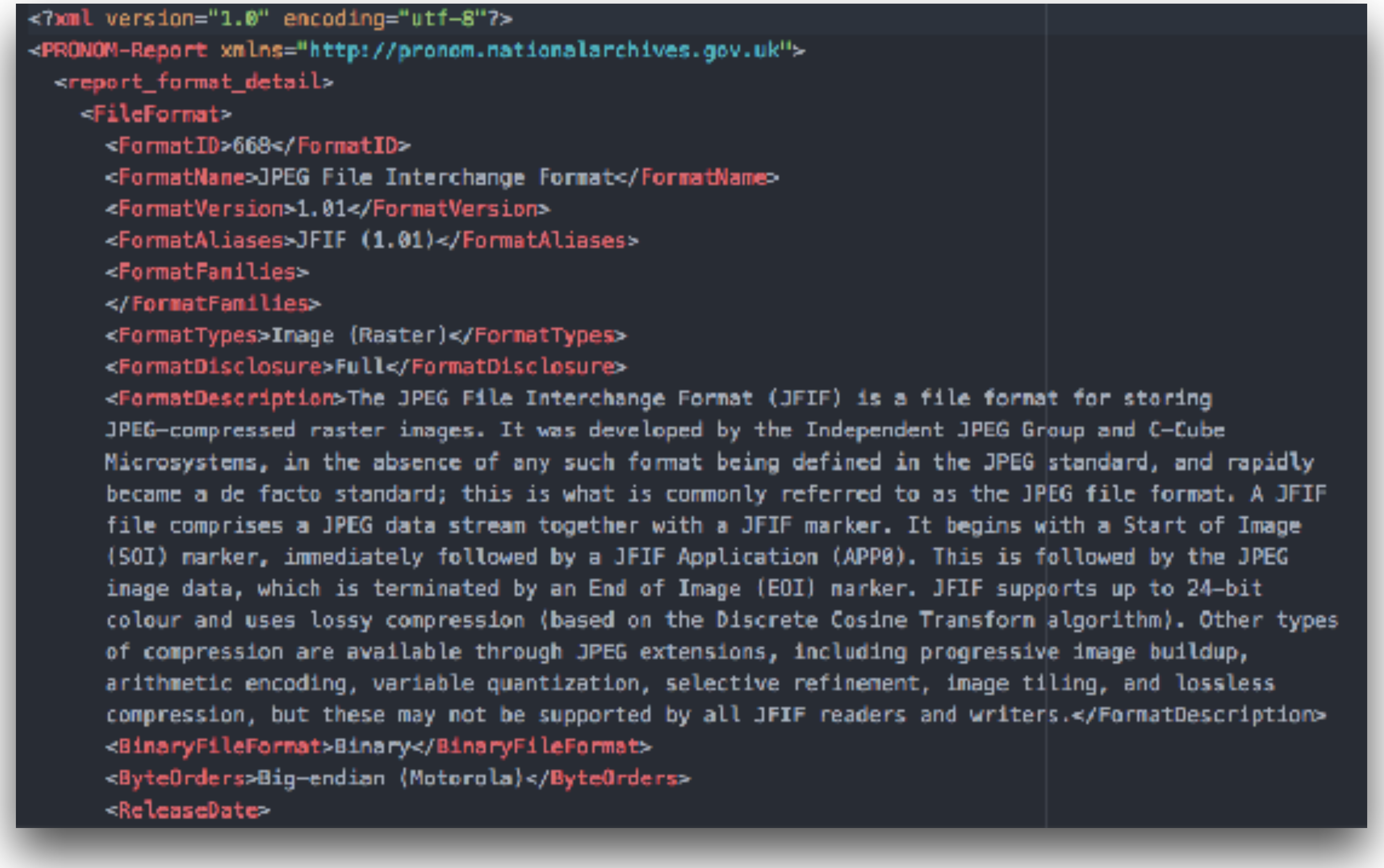
Details: File format summary

Details for: JPEG File Interchange Format 1.01

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 [illegible]
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
<?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#

Unstructured Metadata or **Documentation**

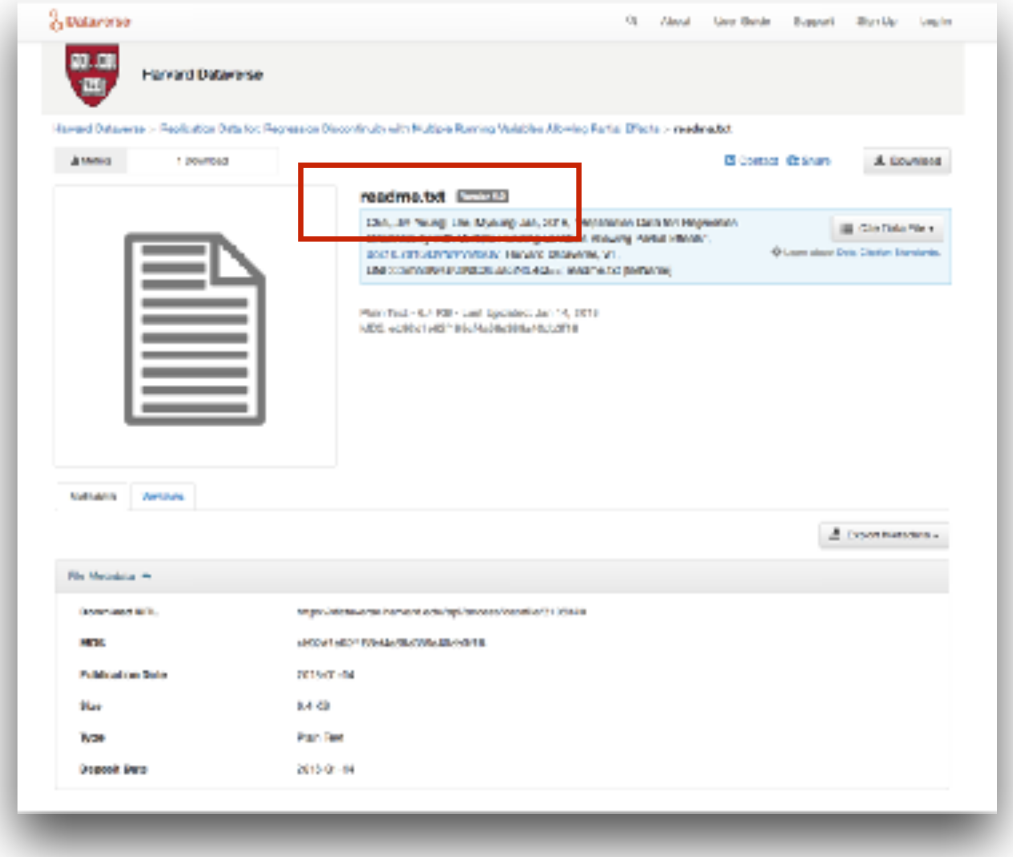
(human readable)

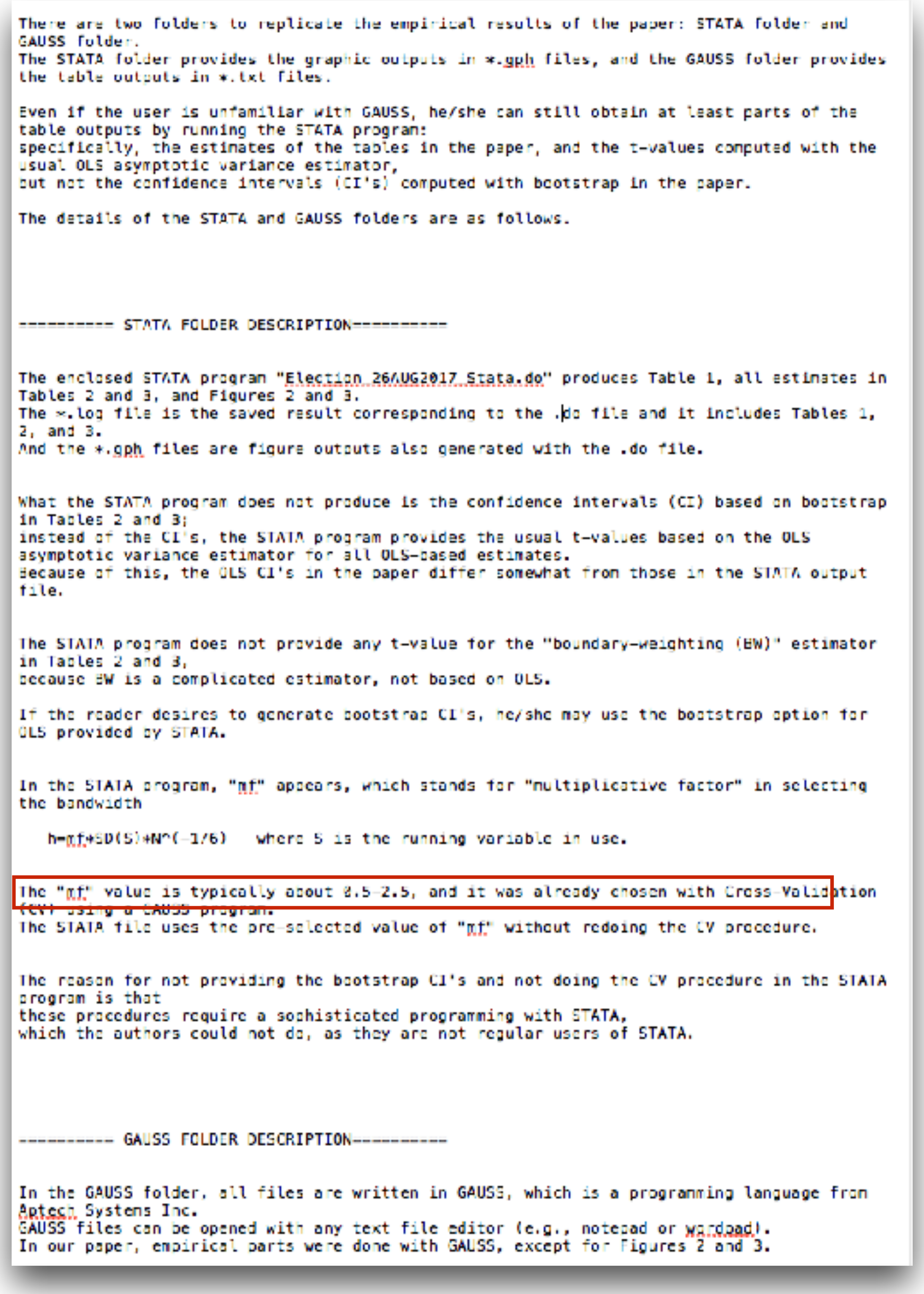
README.txt - provides narrative explanation of what a dataset contains, how it was produced, and how it can or should be used.

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

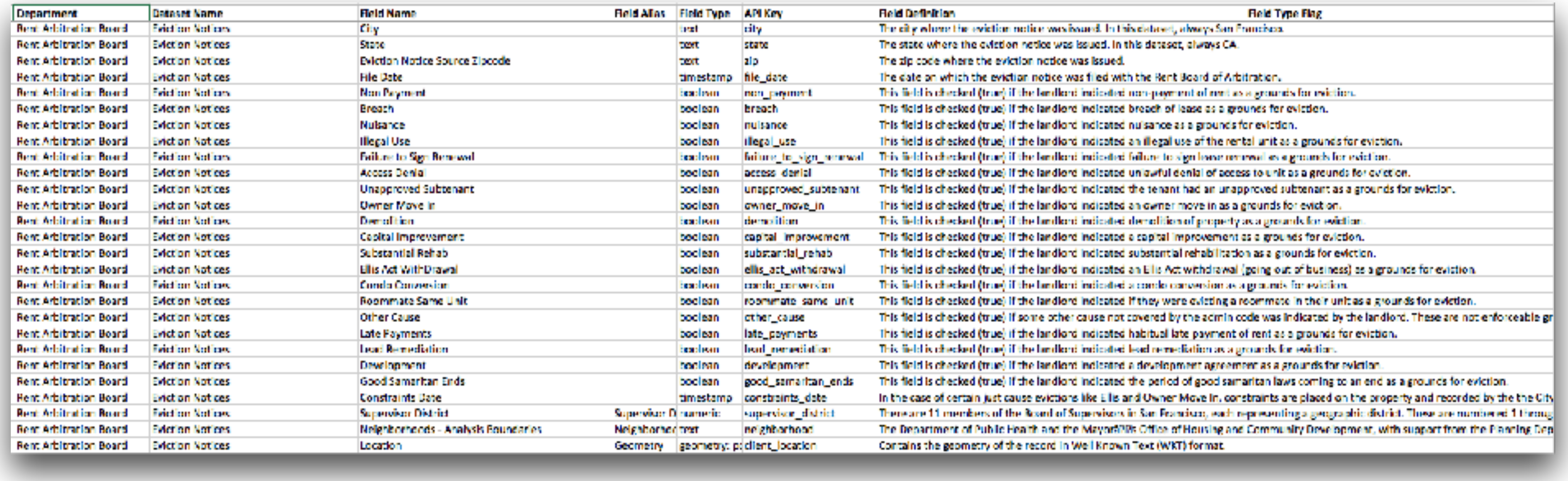
readMe.txt


Dataverse
About
User Guide
Support
Sign Up
Log In
Harvard Dataverse
Harvard Dataverse > Replication Data for: Regression Discontinuity with Multiple Running Variables Allowing Partial Effects > readme.txt
Contact
Share
Download
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.
```

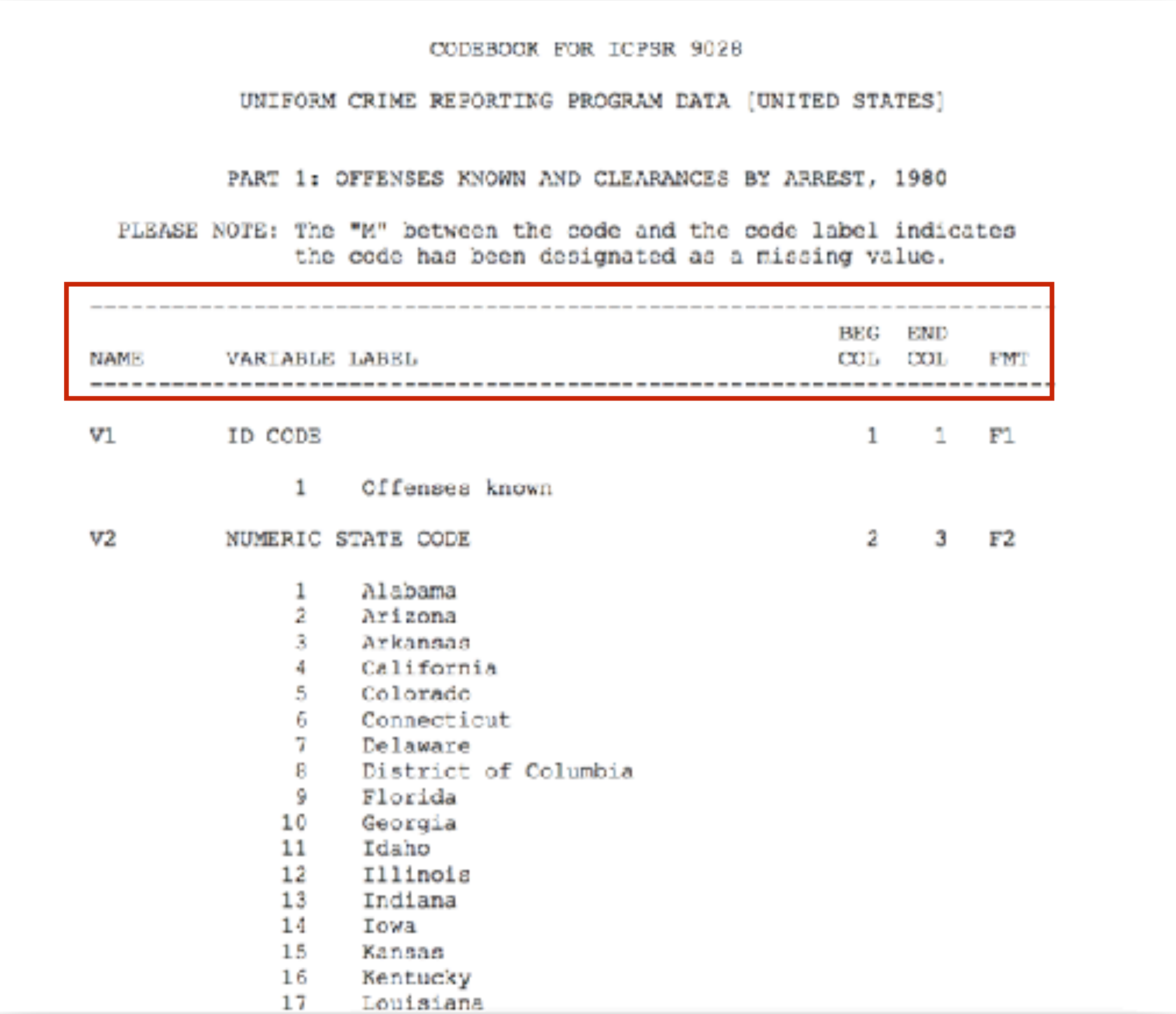
Data Dictionary



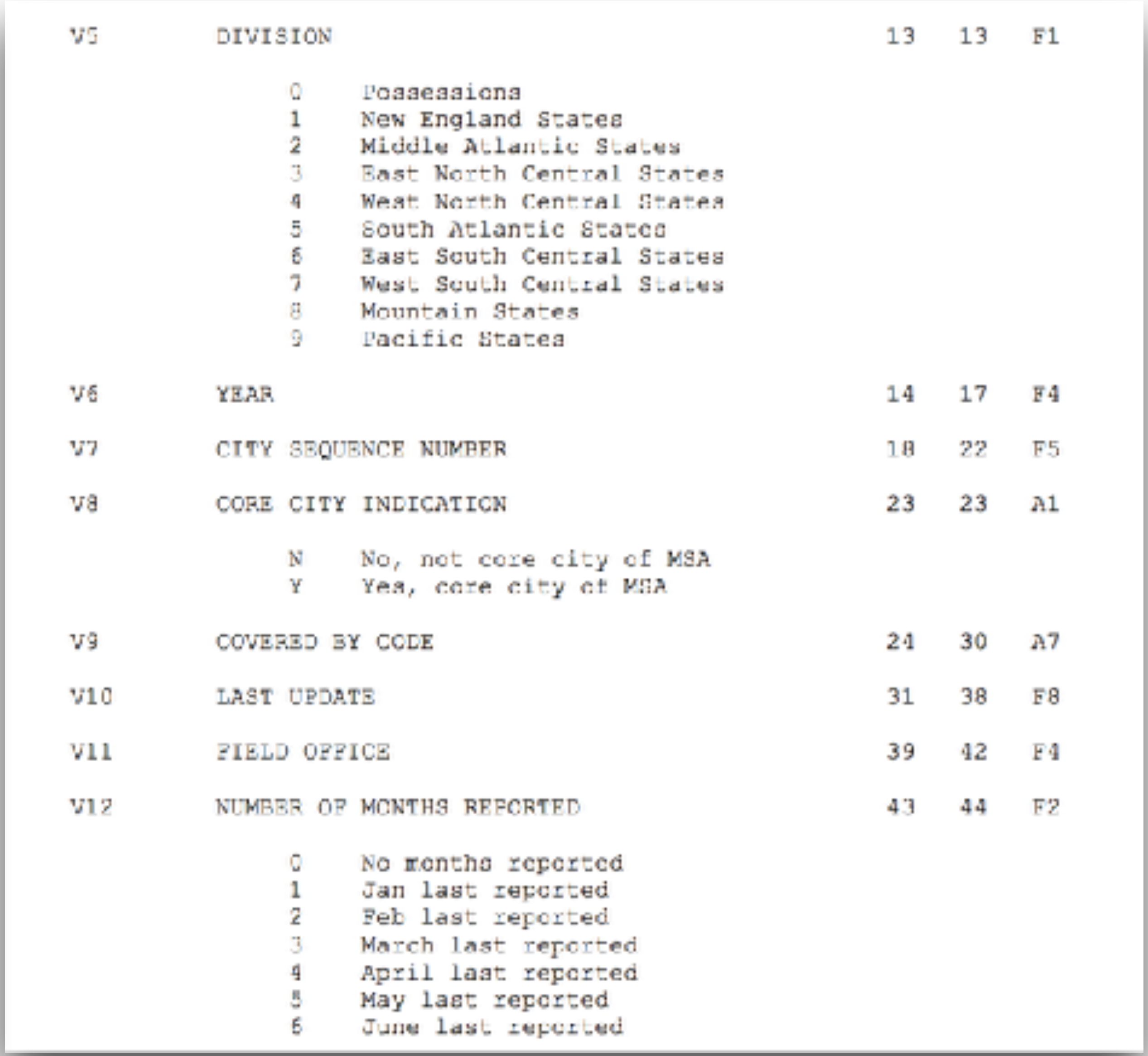
Department	Dataset Name	Field Name	Field Alias	Field Type	API Key	Field Definition	Field Type Flag
Rent Arbitration Board	Eviction Notices	City		text	city	The city where the eviction notice was issued. In this dataset, always San Francisco.	
Rent Arbitration Board	Eviction Notices	State		text	state	The 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	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.

-----------------------------------------------------------------------
                                                        BEG   END
NAME      VARIABLE LABEL                                COL   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**