Metadata & Documentation

for

Data Curation

Lecture 1

INFX 551

Winter 2018

Course Outline

This course consists of three modules.
Each module consists of three topics.

Data	Data Systems	Policy, Privacy, & Ethics
Types & Roles	Repositories	Policy
Documentation	Preservation	Privacy
Standardization	Cost Models	Ethics

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

The goal of this lecture is disabuse you of the notion that —> “metadata is data about data” is an acceptable response the importance of metadata in doing data curation.

Definitions so far…

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

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

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

Knowledge Representation

(the five dollar term for ‘documentation and metadata’)

- **Ontology** vs **Epistemology**
  - Ontology - what is true or real according to existence of nature (or simply, **what exists** in the world)
  - Epistemology - what do we know and **how we know** it ? (e.g Justified, True, Beliefs)

Ontology is a way of talking about what exists in the world…

In Information Science, we also use ontology to **formally represent ‘objects’ that exist** in a domain, as well as how those things relate to other domains…

Ontology components

Instances - things

Classes - types of things

Sub-class & **Super-Class** - introduces hierarchy of things

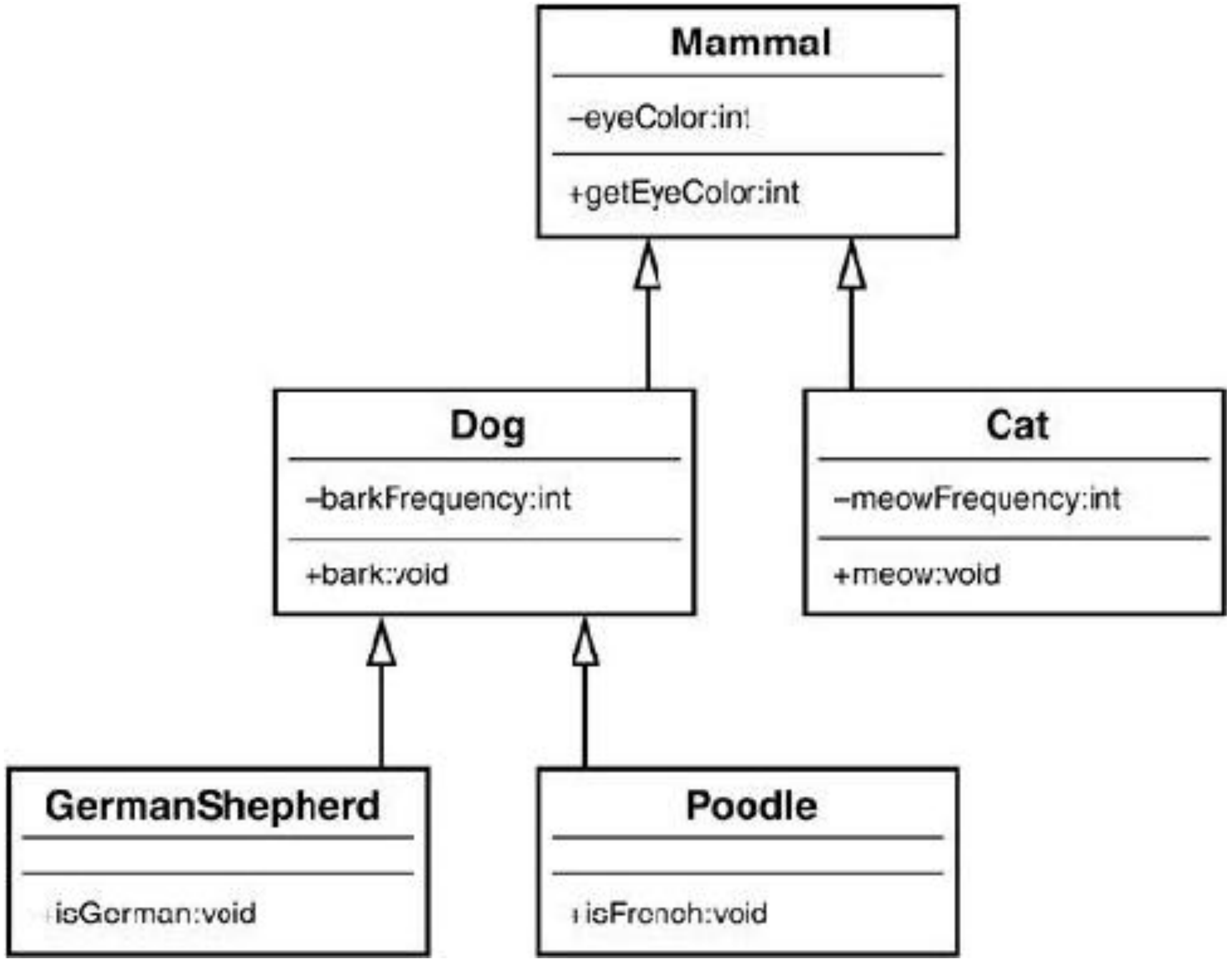
Attributes - properties, features, or characteristics of instances (and by inheritance, classes)

Relations - ways to link different instances or classes to one another

superClass

Class

subClass



Instance

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

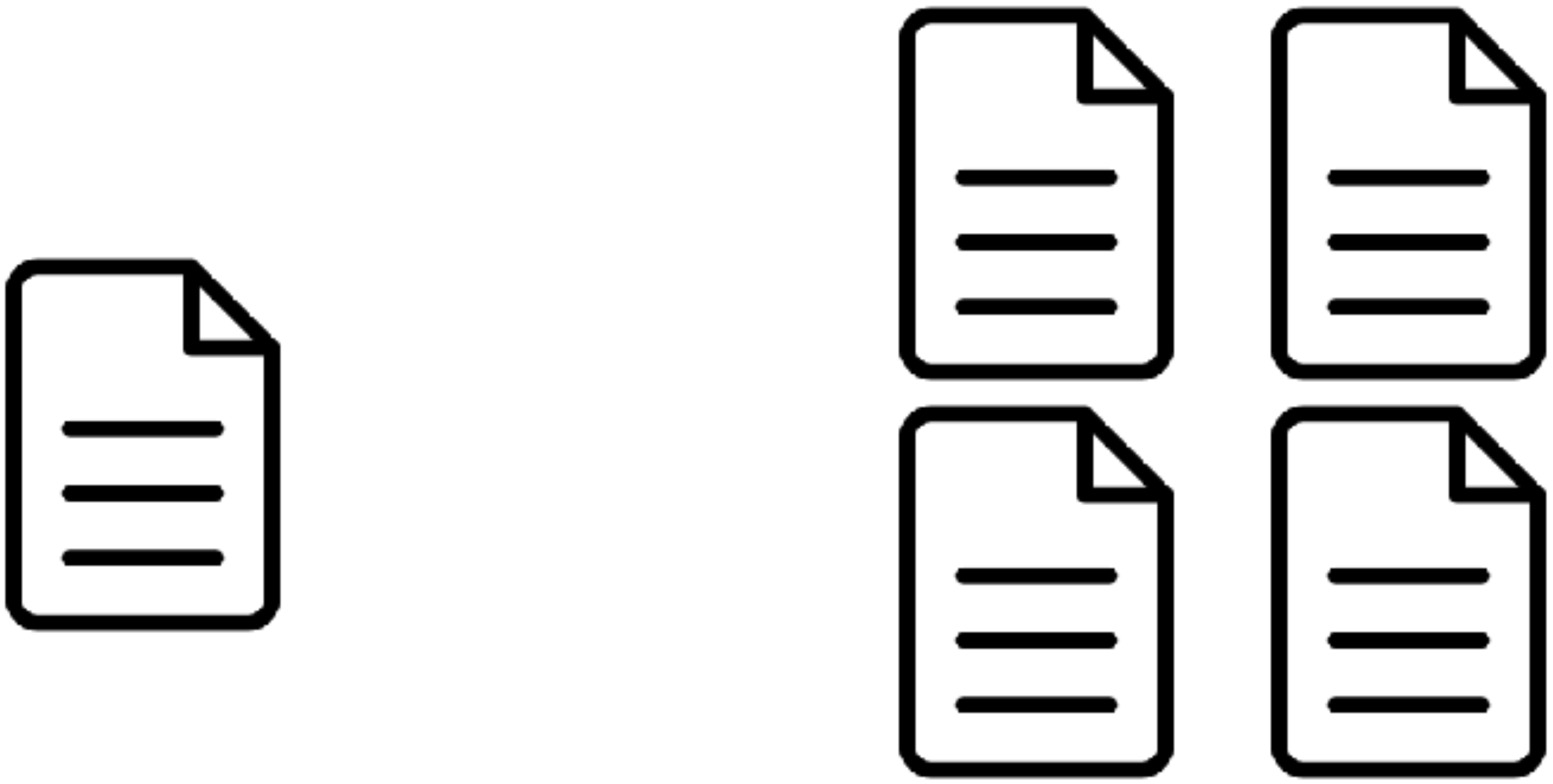
Attributes - properties, features, or characteristics of instances (and by inheritance, classes)

Attribute	Value
Name	Masha
Eye Color	Blue

We just described Attributes of an Instance….

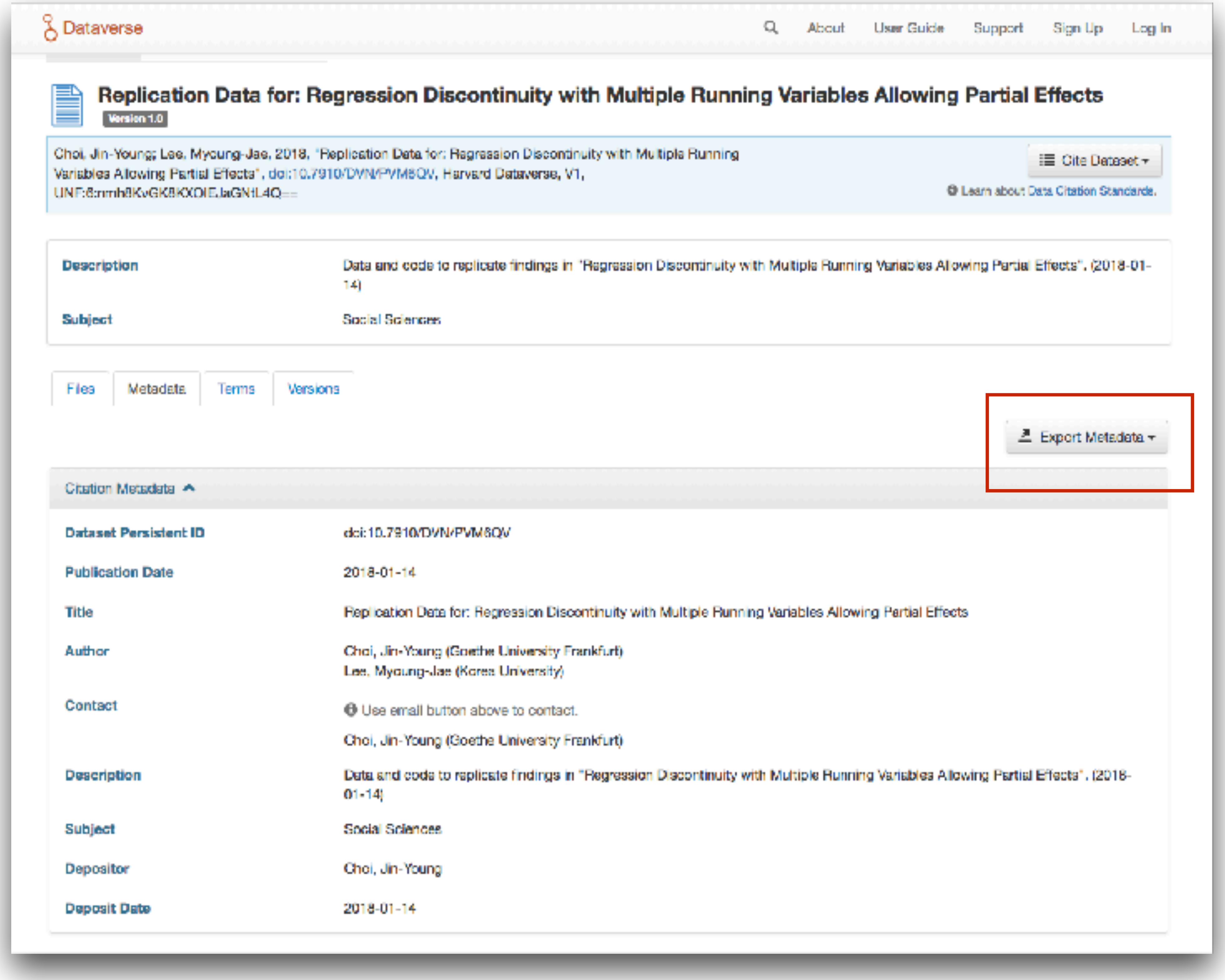
A Class can also have Attributes…

This introduces a distinction between **item-level**, and **collection-level** metadata.



Item

Collection

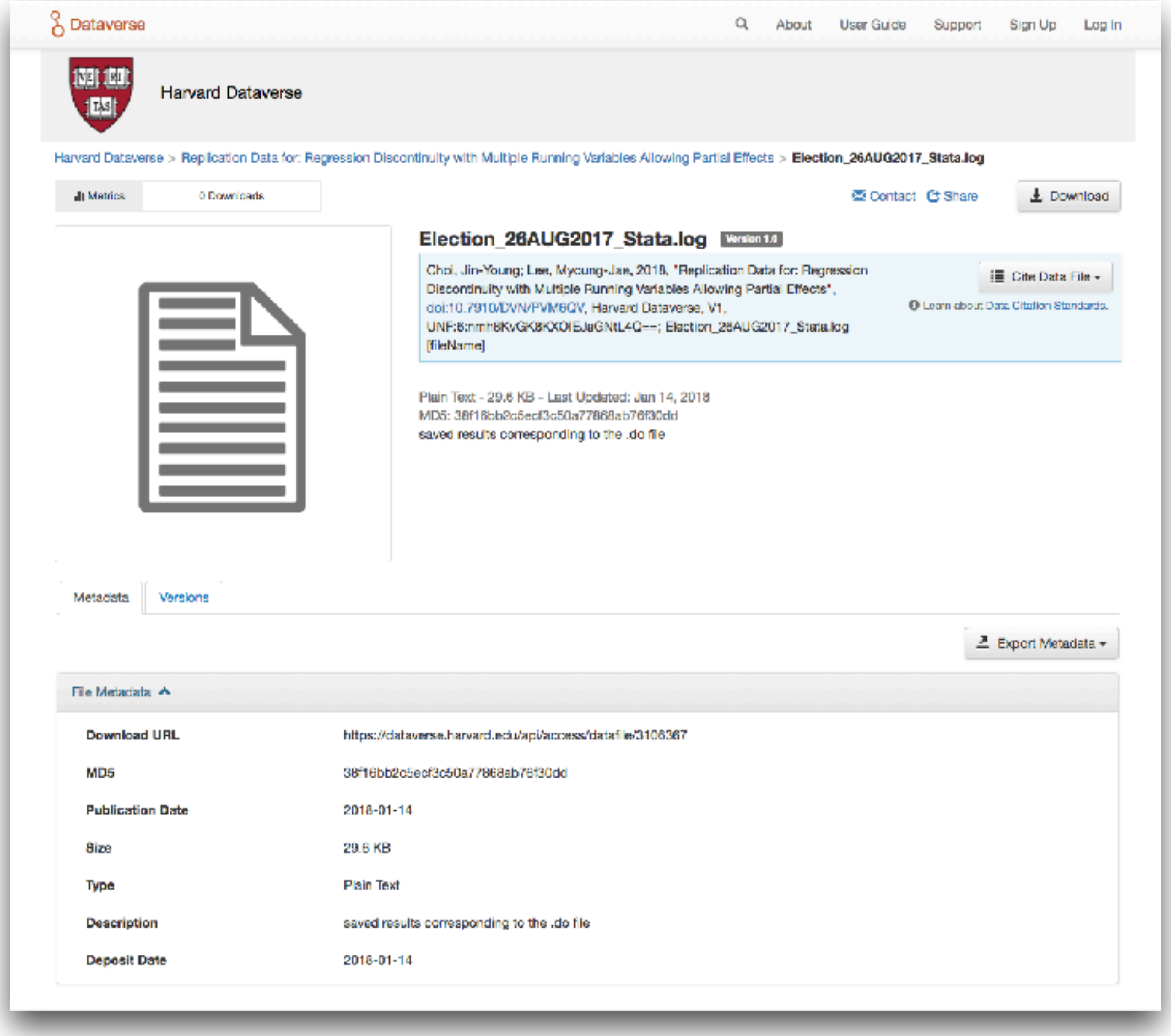


Replication Data for: Regression Discontinuity with Multiple Running Variables Allowing Partial Effects

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:mmh8KvGK8KXDIE.laGNtL4Q==

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



Harvard Dataverse > Replication Data for: Regression Discontinuity with Multiple Running Variables Allowing Partial Effects > Election_26AUG2017_Stata.log

Election_26AUG2017_Stata.log

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:nmh8KvGK8KXOIEJeGNtL4Q==; Election_26AUG2017_Stata.log [fileName]

Learn about Data Citation Standards.

Plain Text - 29.6 KB - Last Updated: Jan 14, 2018
MD5: 38f16bb2c5ecf3c50a77868ab76f30dd
saved results corresponding to the .do file

File Metadata

Download URL	https://dataverse.harvard.edu/api/access/datafile/3106367
MD5	38f16bb2c5ecf3c50a77868ab76f30dd
Publication Date	2018-01-14
Size	29.6 KB
Type	Plain Text
Description	saved results corresponding to the .do file
Deposit Date	2018-01-14

Class > Instance > Attribute

We can quickly go overboard with documenting data.
Especially complex, long-lived data collections.

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.