Metadata

Frauke Kreuter – BLS – 2018
University of Maryland (JPSM), University of Mannheim & IAB

Improving Surveys with Paradata

Analytic Uses of Process Information

Edited by

Frauke Kreuter

WILEY

INNOVATIONS IN FEDERAL STATISTICS

Combining Data Sources While Protecting Privacy

BIG DATA AND SOCIAL SCIENCE

A Practical Guide to Methods and Tools

Edited by
**Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane**

CRC Press
Taylor & Francis Group
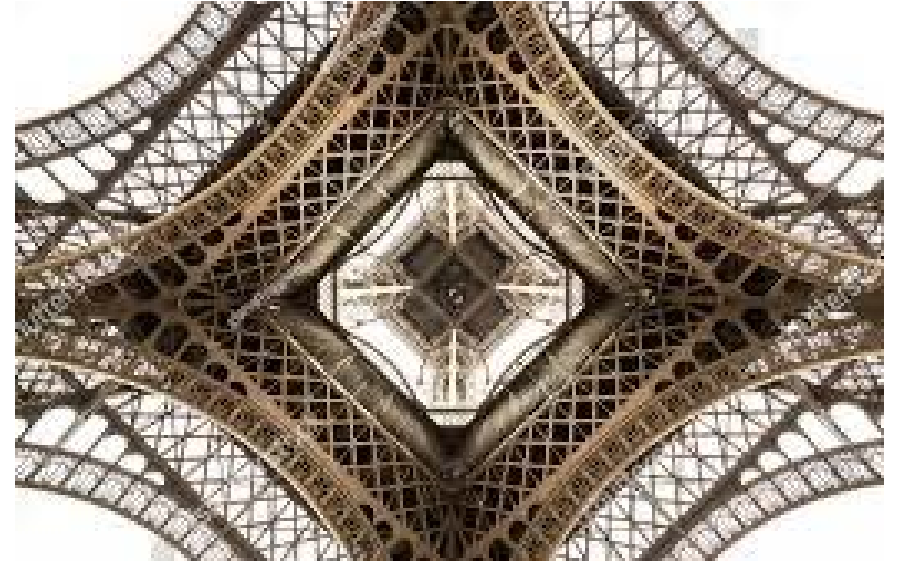
A CHAPMAN & HALL BOOK
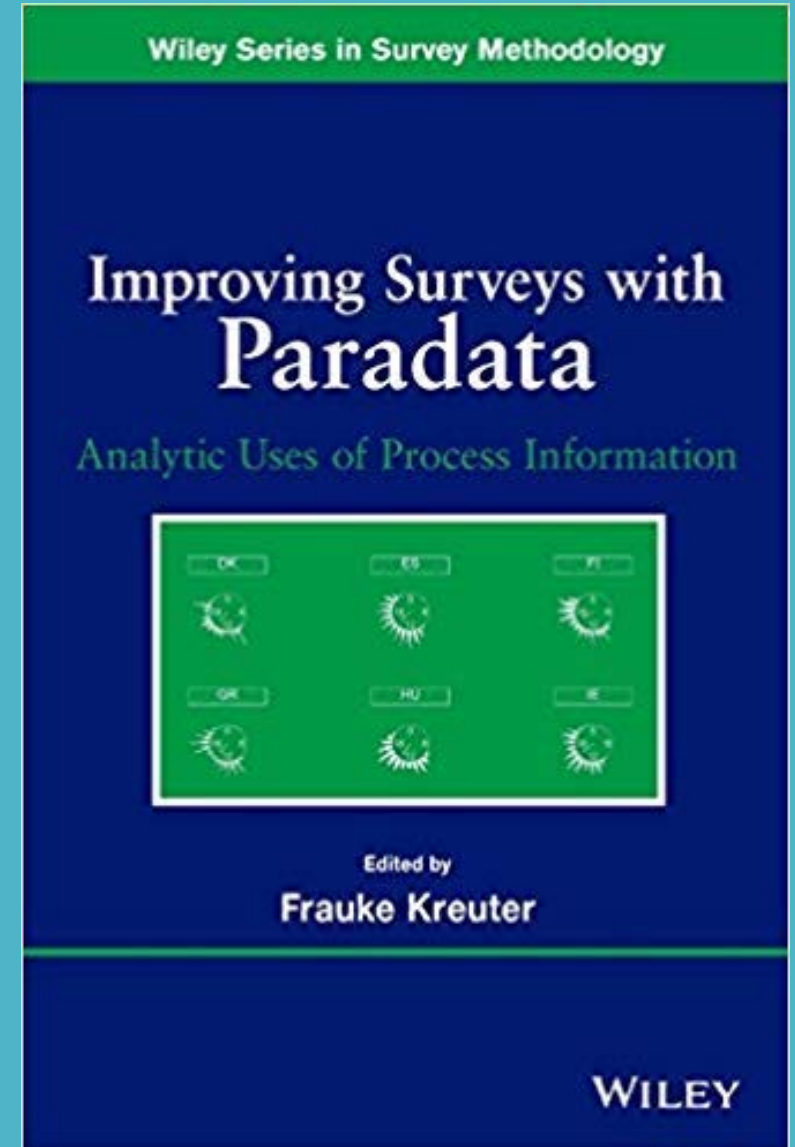
Metadata?

Process data

smile

whiskey

cheese

Spaghetti

File info

Close

File name
IMG_120

Metadata

Date taken
Saturday, March 10, 2018 6:19 PM

Size
3.8 MB

Dimensions
4032 x 3024

Shot
1/10 sec. f/1.8 4mm

ISO
100

Device
iPhone X

Folder path
C:\Users\fkreuter\Pictures
\Wohnzimmerkonzert

Source
This PC

Paradata and metadata

Paradata capture information about the data collection process on a more micro level.

Some of this information forms metadata if aggregated, for example the response rate for a survey (a piece of metadata) is an aggregated value across the case-level final result codes.

Metadata allow users to understand the structure of a data set and can inform analysis decisions.

Example: sampling frame, sampling methods, variable labels, value labels, percentage of missing data

Recommendations
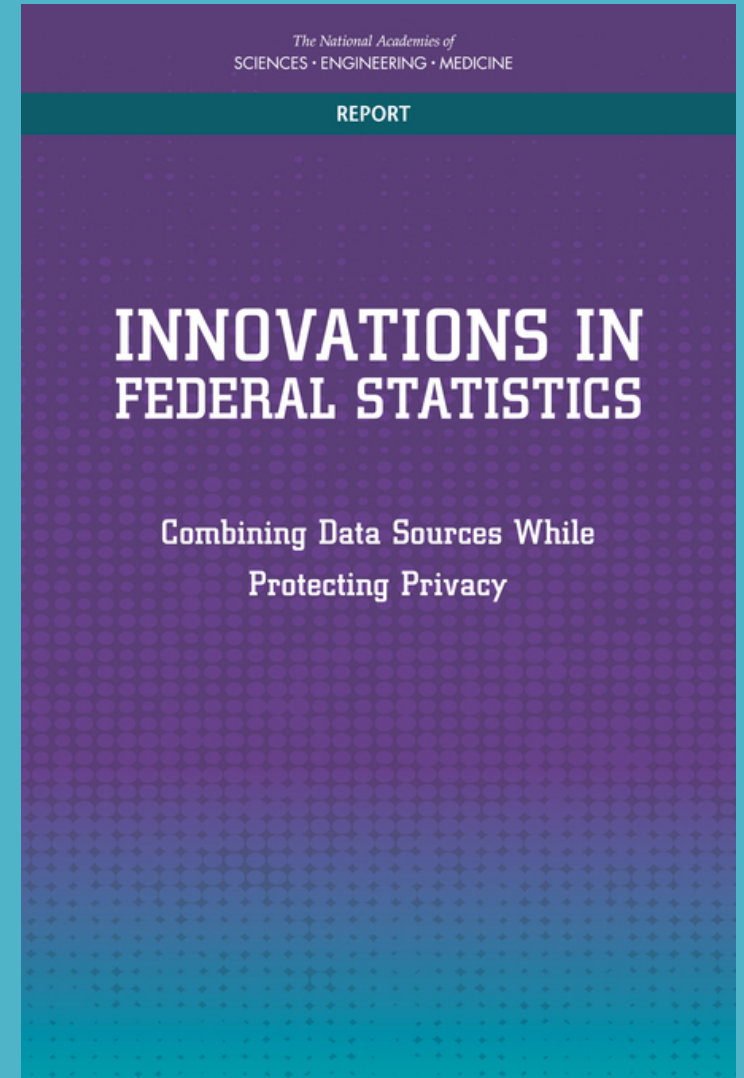
Observation

Administrative and private-sector data have their own challenges and errors. These errors arise for multiple reasons, such as mistakes in understanding or interpreting metadata, errors in entity linkage, and incomplete or missing information.

A precise understanding of those differences is critical for correct interpretation and difficult due to
varying metadata recording standards across data sources.

Conclusion 3-3

Creating statistics using multiple data sources often requires complex methodology to generate even relatively simple statistics.

With the advent of new and different sources and innovations in statistical products, federal statistical agencies need to figure out ways to provide transparency of their methods and to clearly communicate these methods to users.

Recommendation: Accessibility and Clarity

Statistics should be presented in a clear and understandable form, released in a suitable and convenient manner, and available and accessible on an impartial basis with supporting metadata and guidance.
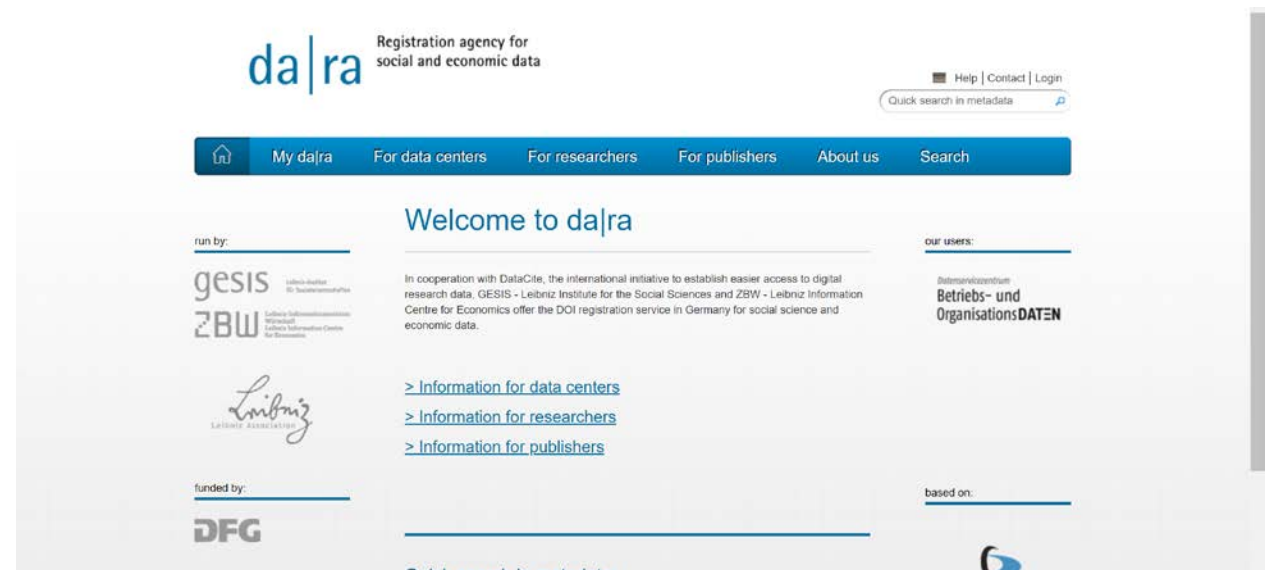
It is recommended that metadata are preserved and properly archived.

It is also recommended that metadata are standardized according to systems, dissemination services use proper communication and current technology, custom-designed analysis is provided when feasible, and public microdata files are available to researchers for specific purposes following protocols.

Implementations

Examples for Metadata Schema

- Statistical Data and Metadata Exchange (SDMX)
- Data Documentation Initiative (DDI)
- da|ra Metadata Schema

Digital Object Identifier (DOI)

- DOIs are **permanent** and **persistent** identifier which is **unique** and cannot be deleted.

- DOIs are a s**imple character string** which provides a **link** to a **resource**.

- In Germany DOIs are provided by the GESIS DOI registration service **da|ra** (GESIS is cooperating with **DataCite**).

  https://www.da-ra.de/en/home

Part 1: Identifier

1	Resource Type
2	Resource Identifier
3	Name of Dataset
4	Creator
5	DOI Proposal
6	URL
7	Language of Resource
8	Publication Date
9	Availability
10	Sampled Universe
11	Sampling
12	Temporal Coverage
13	Time Dimension
14	Collection Mode
15	Unit Descriptions
16	Descriptions
17	Geographical Coverage
18	Keywords
19	Alternative Identifiers
20	Relations
21	Publications

- *Creator* is a mandatory item in da|ra. May be used to provide more granular information on the data compiler

- *URL* refers to the webpage which displays information about the dataset

- *Availability (controlled)* describes the procedure under which the data can be accessed (eg download or on-site)

- *DOI Proposal* provides the suggested DOI name of the dataset. A Digital Object Identifier (DOI) is a permanent, persistent identifier used for citing and tracking datasets

Part 2: Methods

1	Resource Type
2	Resource Identifier
3	Name of Dataset
4	Creator
5	DOI Proposal
6	URL
7	Language of Resource
8	Publication Date
9	Availability
10	Sampled Universe
11	Sampling
12	Temporal Coverage
13	Time Dimension
14	Collection Mode
15	Unit Descriptions
16	Descriptions
17	Geographical Coverage
18	Keywords
19	Alternative Identifiers
20	Relations
21	Publications

- *Sampling* displays the type of sample design used to select the observations to present the population

- *Time Dimension* provides information on
  - frequency of observations.
  - whether dataset structure is panel, time-series or cross-sectional

- Structural breaks are defined as major events and revisions that have impacted the dataset

- Examples of structural breaks include:
  - changes to the time frequency with which data is collected
  - changes to the set of collected variables
  - changes in the population or sampling

Curtesy Stefan Bender - INEXDA's Metadata Database

Practice

"it's harder than you think"
"it's more important than you realize"

WIRED / Twitter

"Metadata is everywhere. Everything you tweet, every picture you take, and every status update you post on Facebook. It's used by police and security forces to identify people who try to hide their identities and locations, while associated metadata in selfies can inadvertently ensnare criminals unaware that the data can destroy their alibi." (Stokel-Walker July 9 2018)

You are your metadata

"This is information that describes the context on which the post was shared. Apart from the 140 character message, **each tweet contains about 144 fields of metadata**. Each of these fields provides additional information about: the account from which it was posted; the post **(e.g. time, number of views)**; other tweets contained within the message; various entities (e.g. hashtags, URLs, etc); and the information of any users directly mentioned in it." (Perez et al. 2018)
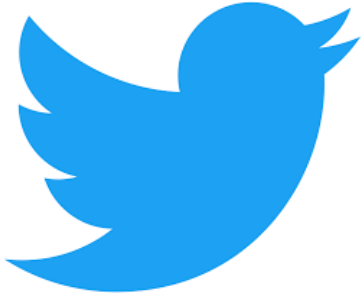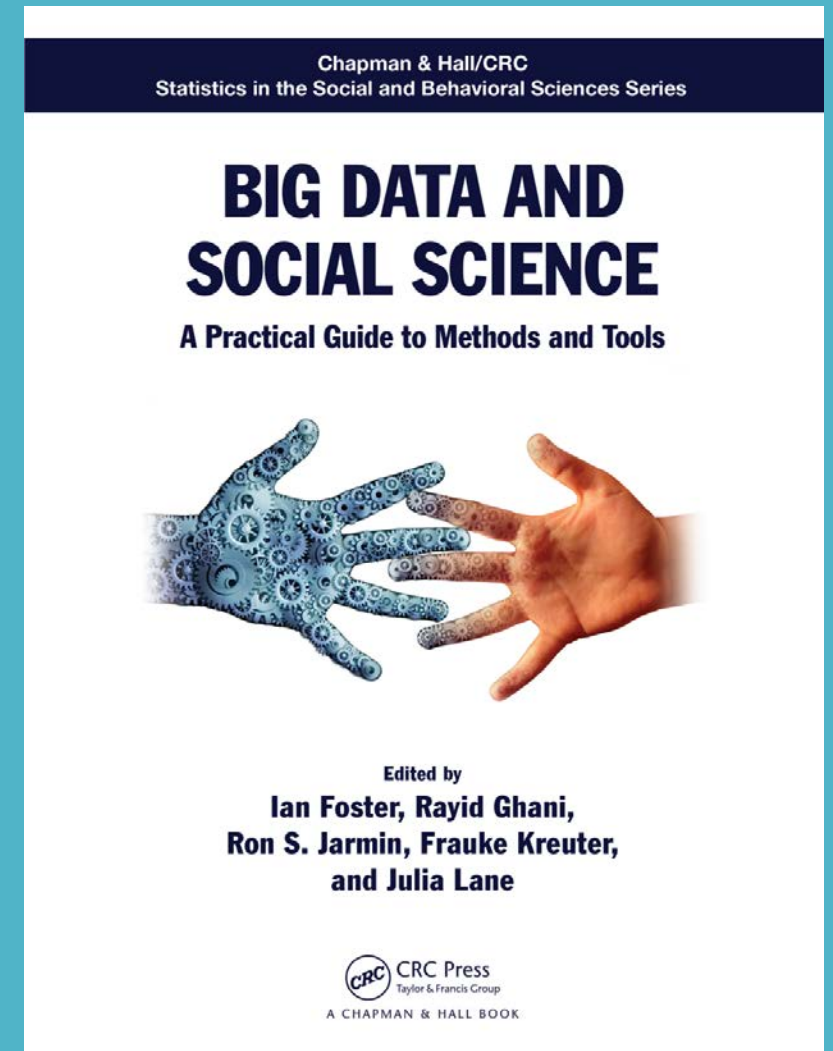
```
{
  "created_at" :   "Thu Apr 06 15:24:15 +0000 2017"  ,
  "id_str" :   "850006245121695744"  ,
  "text" :   "1\/ Today we\u2019re sharing our vision for the future of the Twitter API platform!\nhttps:
  "user" :   {
    "id" :   2244994945  ,
    "name" :   "Twitter Dev"  ,
    "screen_name" :   "TwitterDev"  ,
    "location" :   "Internet"  ,
    "url" :   "https:\/\/dev.twitter.com\/"  ,
    "description" :   "Your official source for Twitter Platform news, updates & events. Need technical h
  }  ,
  "place" :   {
  }  ,
  "entities" :   {
    "hashtags" :   [
    ]  ,
    "urls" :   [
      {
        "url" :   "https:\/\/t.co\/XweGngmxlP"  ,
        "unwound" :   {
          "url" :   "https:\/\/cards.twitter.com\/cards\/18ce53wgo4h\/3xo1c"  ,
          "title" :   "Building the Future of the Twitter API Platform"
        }
      }
    ]  ,
    "user_mentions" :   [
    ]
```

Tweet metadata, mutability, updates

While Tweet messages can be up to a fixed number of characters long, the JSON description of a Tweet consists of over 100 attributes. Attributes such as who posted, at what time, whether it's an original Tweet or a Retweet, and an array of first-class objects such as hashtags, mentions, and shared links. […]

Most account metadata is static, but some change slowly over time. People change jobs and move. Companies updates their information. When you are collecting historical Tweets, it is important to understand how some metadata is *as it was when Tweeted*, and other metadata is *as it is when the query is submitted*. The metadata that is potentially updated depends on the historical API.

Recommendations

Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

BIG DATA AND SOCIAL SCIENCE

A Practical Guide to Methods and Tools

Edited by
**Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane**

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

METADATA ABOUT VARIABLES

MATT SALGANIK

JULY 12, 2017

We are happy to announce that Challenge participant Connor Gilroy, a Ph.D. student in Sociology at the University of Washington, has created a new resource that should make working the Challenge data more efficient. More specifically, he created a csv file that identifies each variable in the Challenge data file as either categorical, continuous, or unknown. Connor has also open sourced the code that he used to create the csv file. We've had many requests for such a file, and Connor is happy to share his work with everyone! If you want to check and improve Connor's work, please consult the official Fragile Families and Child Wellbeing Study documentation.

ccgilroy / **ffc-data-processing**

⊙ Watch ▾   0   ★ Star   0   ⑂ Fork   0

<> Code   ⓘ Issues 0   ⑂ Pull requests 0   ▦ Projects 0   ☰ Wiki   �III Insights

Data processing for the background covariates Stata file from the Fragile Families challenge.   https://ccgilroy.github.io/ffc-data-p...

⊙ **26 commits**   ⑂ **1 branch**   ◇ **0 releases**   ⚼ **1 contributor**   ⚖ **MIT**

Branch: master ▾   New pull request   Create new file   Upload files   Find file   Clone or download ▾

ccgilroy update mi to generate imputations for constructed variables as well   Latest commit 408ed07 on Jul 31, 2017

■ R	break out examples into vignettes and remove from init file, which no…	a year ago
■ data	write character_to_factor_or_numeric, switch to Map() for speed, fix …	a year ago
■ output	two R scripts for producing data set ready for multiple imputation us…	a year ago
■ vignettes	update mi to generate imputations for constructed variables as well	a year ago
▤ .gitignore	two R scripts for producing data set ready for multiple imputation us…	a year ago
▤ LICENSE	Create LICENSE	a year ago

Metadata Schema (ADRF) ...... coleridgeinitiative.org

Metadata associated with intellectual entity of the dataset

metadata

| Field Name	Data Type	Description	In Explorer	Obligation
file_names	Array of strings	A list of file names in the dataset.	No	1-n
dataset_id	String	Dataset id	Yes	1
Title	String	Title of the dataset	Yes	1
description	String	Description of the dataset	Yes	1
temporal_coverage_start	Date in ISO 8601 format (yyyy-mm-dd)	Start date for years/months the dataset is valid for	Yes	0-1
temporal_coverage_end	Date in ISO 8601	End date for years/months the	Yes	0-1
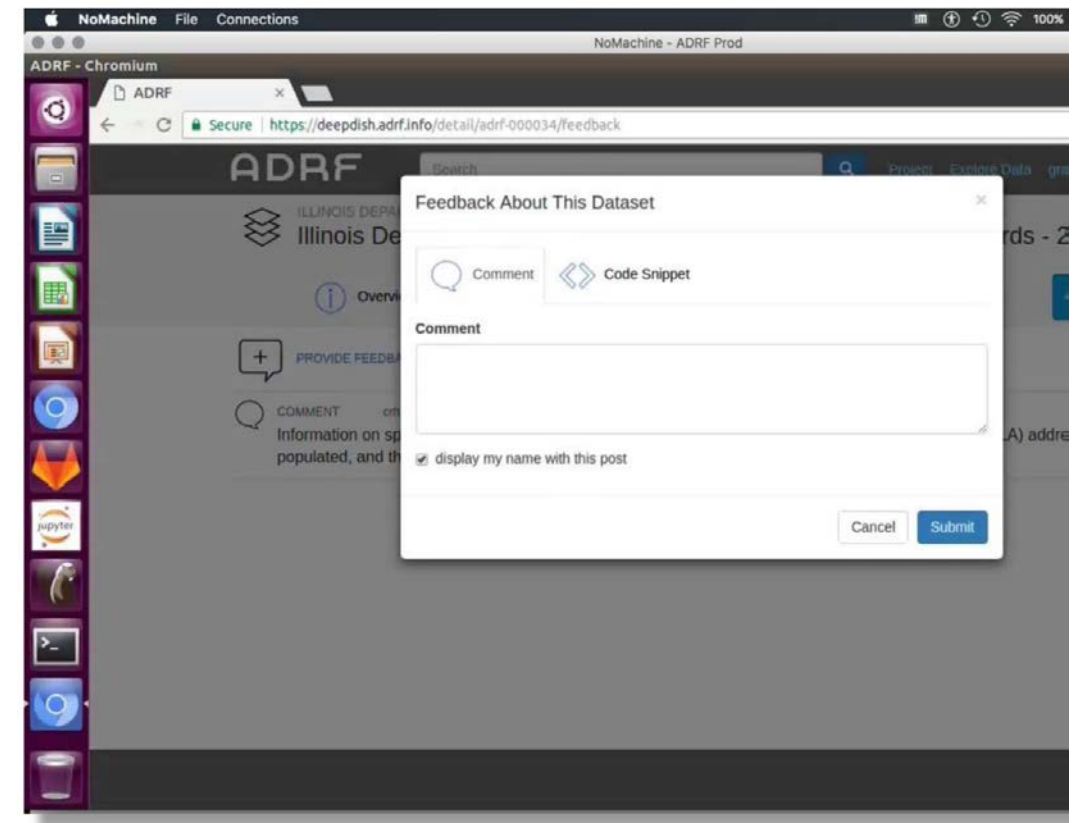
Figure 1: Feedback mechanism in ADRF Explore

Summary

o *Metadata are key to data documentation*
o *Metadata arise in context*
o *Metadata need to be understandable – especially when combining multiple data sources*

o *Documentation is hard, boring, often not rewarded*
o *Possible solutions: automatize, gamify, incentivize, personalize, crowdsource*

Thank You!

fkreuter@umd.edu