COLERIDGE
INITIATIVE

ADRF DOCUMENTATION

coleridgeinitiative.org

Contents

1. Motivation 3
2. Context 3
3. Current ADRF approach 3
   - 3.1 Metadata Harmonization 3
   - 3.2 Urban Profiler 4
   - 3.3 ADRF Explorer 5
   - 3.4 Data Facility (DF) Admin 7
4. Features in Development 8
   - 4.1 Rich Context Enhancements 8
   - 4.2 ETL Improvements 9

Appendix 10
- GMeta METADATA SCHEMA V. 1.0 10
  - Dataset level metadata 10
  - File level metadata 13
  - Variable level metadata 15
  - Sample detailed GMETA file 16

1. Motivation

This white paper describes the data policies and procedures for making metadata and microdata more discoverable and usable across institutional boundaries. The ADRF is designed for that purpose. This document provides the following:

(i) a summary of the existing ADRF approach,
(ii) features in development, and
(iii) a discussion of how we will customize these tools.

2. Context

The goal of metadata tools is to make it easier for end users of confidential data to discover and use datasets appropriate for research and analysis. This is a major challenge, because most datasets are either not well documented or are not documented with the purpose of being conceptually or electronically connected to other datasets. Databases have historically been indexed and catalogued by input characteristics (name or provider), rather than by their potential use characteristics (topic or research area). Preparing data for use outside of its original purpose (e.g. creating descriptions and providing information for a nominal future user to understand the data) is difficult, time consuming, and error prone.

3. Current ADRF approach

The ADRF's Documentation module employs the following components to simplify and automate the accessibility of data from different organizations: metadata harmonization; a tool called Urban Profiler for generating more detailed column level metadata for datasets; the ADRF Explorer for helping users find datasets relevant to their work and enrich dataset descriptions; and DF Admin for administering dataset metadata and access within the ADRF.

3.1 Metadata Harmonization

The ADRF ingest process uses a flexible, generalized metadata schema called GMeta (see **Appendix**) for dataset description that is based on the established Data Catalog Vocabulary (DCAT)[1]. This provides a level of standardization for datasets coming from different industries,

[1] Data Catalog Vocabulary (DCAT): https://www.w3.org/TR/vocab-dcat/

organizations, and research purposes. DCAT Is already adopted by the US Federal government[2], and is an accepted dataset description standard of the European Commission in multiple European data portals[3].

Additionally, the GMeta schema provides standardization of controlled vocabulary categories while allowing for more open-ended tags, to make disparate datasets more discoverable in relation to each other. GMeta also defines ownership and the access restrictions and requirements of third-party datasets, providing users a better way to understand what datasets they might be able to access and how to access those datasets. GMeta also allows related publications to be associated with datasets, for better understanding how these datasets have been used by others in their analyses. Finally, GMeta provides for description of temporal and geographic coverage so that datasets about the same time and place might be found and linked together.

3.2 Urban Profiler

Urban Profiler is a tool written in Python which scans, computes, and outputs column-level metadata that facilitates comparison and evaluation of variables for a dataset at a glance. Urban Profiler output is folded into the GMeta metadata schema for each file that is part of a dataset (see **Appendix**). Urban Profiler generates such information as: the percentage of missing values; number of unique values; inferred datatypes (e.g., numeric, textual, date), statistical summaries for variables whose values are numeric, and frequency of top k values as well as data for generating histograms for distribution of values. This level of detail a llows analysts to better understand the contents and quality of a dataset quickly and provides an opportunity to examine variables across datasets for ability to link or compare with each other.

[2] US Project Open Data: https://project-open-data.cio.gov/v1.1/schema/

[3] DCAT Application Profile for data portals in Europe: https://joinup.ec.europa.eu/release/dcat-application-profile-data-portals-europe-final

3.3 ADRF Explorer

The ADRF Explorer is a discovery tool that provides an inventory of which datasets currently exist within an ADRF environment. The ADRF Explorer requires datasets to be described with the GMeta metadata schema and is connected to an installation of Elasticsearch[4], which supports the retrieval and indexing of GMeta information inside the ADRF. It also serves as a tool to allow users to add enhanced context for understanding a dataset and enforces restrictions on datasets that are restricted to only specific users of the ADRF. To elaborate on how this tool allows users to add enhanced context, the ADRF Explorer gives users the ability to provide feedback (in the form of comments and annotations as shown in **Figure 1**) on datasets at the dataset and dataset variable level. Comments might include information from an expert on how a variable was collected or it might also be a request for more information about the variable. In this way, it enables even richer metadata with which to understand a dataset's utility.

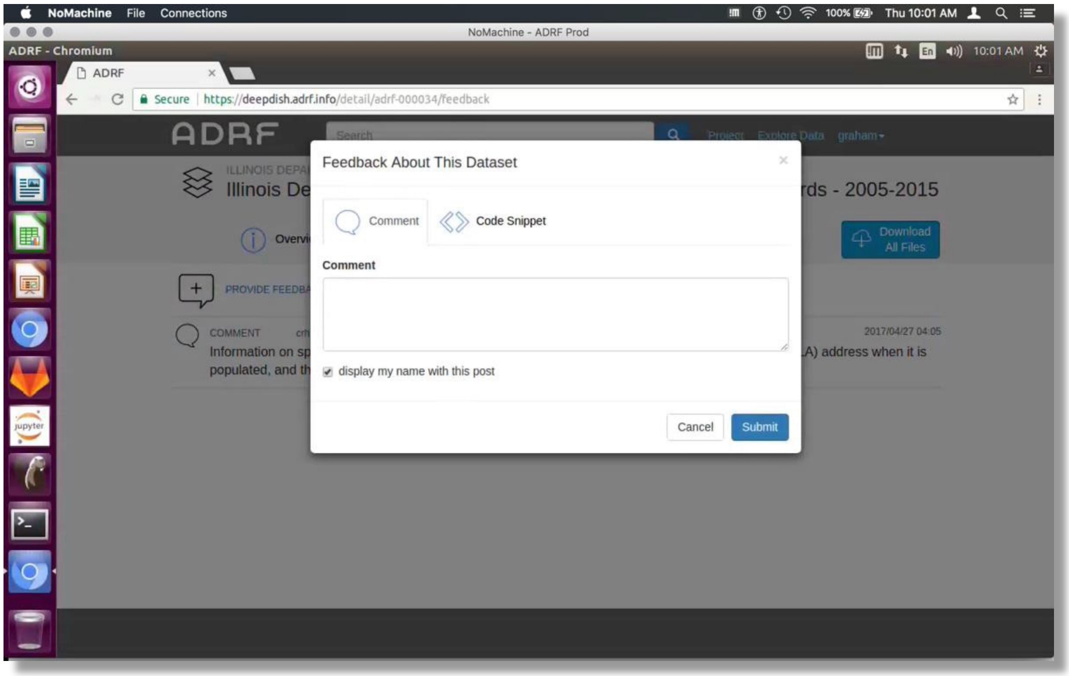


Figure 1: Feedback mechanism in ADRF Explorer

Entities that can be associated with datasets

Finally, the ADRF Explorer allows users to connect datasets to the context within which they may be used. The core approach is to relate datasets to key entities involved in their usage: people; research products; recipes (code); and tools.

[4] Elasticsearch: https://www.elastic.co/products/elasticsearch

People: ADRF Explorer initially designates two classes of “people”: users and authors. Users have profiles to which they may add personal details and where eventually they will also be able to display earned status levels or rewards (e.g., badges) via the Rich Context tool (described in more detail below). These may be earned by commenting on or annotating datasets, adding code snippets, and through peer-input such as having other users rank their comments or code snippets.

Research products: Research products—articles, publications, or reports—are ingested into the Explorer and each have individual landing pages just like datasets, providing ADRF users with access to the details of the research. Each research product may also be linked to other entities such as people, recipes, and tools, improving the search and discovery results. The system currently includes many selected papers and will expand as ADRF Explorer features are added, including a more streamlined research content upload interface as well as an automated research collection engine.

Recipes: Complete Jupyter notebooks[5] and other analytical scripts can be uploaded to the ADRF Explorer as “recipes” by ADRF users. These range from contained, reproducible examples of a specific analytical task to entire analytical workflows associated with a given research product. ADRF users can download recipes to their own workspace for further work in order to get a running start using a particular dataset or to adapt existing analyses for their own purpose.

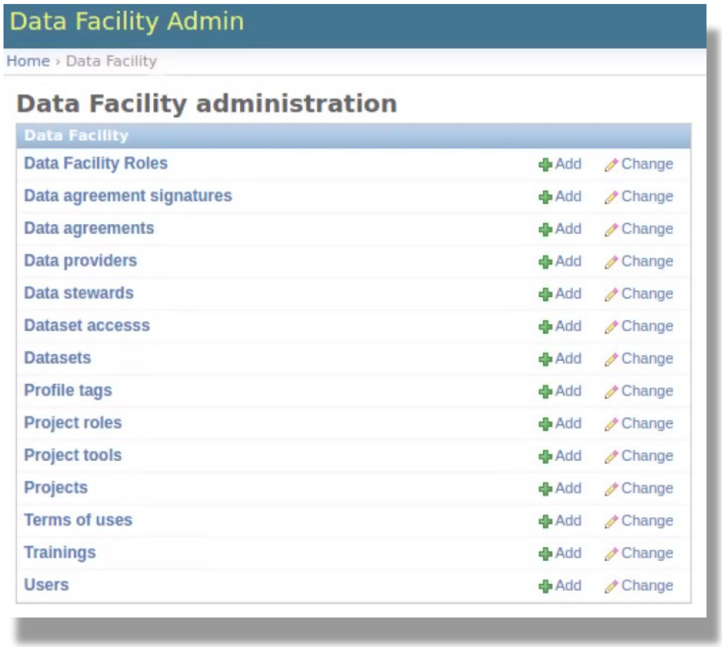
Tools: The ADRF provides a suite of analytical tools that can be adapted to include new or additional tools based on user needs. The current toolkit includes R and Python (with popular data science packages installed such as SciPy, NumPy, pandas, Matplotlib, and scikit-learn) programming environments, Jupyter notebooks, and PostgreSQL\PostGIS database environments. Each project also has a shared repository in a Gitlab[6] installation, which provides project-by-project version control to research teams (i.e., allowing users to track and avoid overwriting changes as well as providing a detailed history of development of code, documents, and notes in a project).

[5] Jupyter (http://jupyter.org/) Notebook “is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text”

[6] Gitlab (https://about.gitlab.com/) is a browser accessible tool which can be used by teams to collaborate on projects which involve code development. It is designed to add a social component to the popular version control system Git and is much like Github with the important difference that it can be installed locally for use inside an environment that is not connect to the Internet.

3.4 Data Facility (DF) Admin

DF Admin is a web accessible administrative portal built inside of the ADRF to manage users, projects, datasets, and user access to projects and datasets in the ADRF. Its technical underpinnings are a Django[7] web application frontend connected to a PostgreSQL[8] database and it is open source. It provides an interface for system administrators to see all the datasets in the ADRF, who has access to them, and to modify user and project access to those datasets based on access policies for those datasets.



Data Facility Admin

Home › Data Facility

Data Facility administration

Data Facility		
Data Facility Roles	Add	Change
Data agreement signatures	Add	Change
Data agreements	Add	Change
Data providers	Add	Change
Data stewards	Add	Change
Dataset accesss	Add	Change
Datasets	Add	Change
Profile tags	Add	Change
Project roles	Add	Change
Project tools	Add	Change
Projects	Add	Change
Terms of uses	Add	Change
Trainings	Add	Change
Users	Add	Change

Figure 2: DF Admin interface of entities which may be managed currently in the ADRF

DF Admin was designed to be integrated with the GMeta metadata schema and the ADRF Explorer. Future development of DF Admin will include more granular dataset metadata editing that will update records as they appear in ADRF Explorer. Additional features to DF Admin will also include tools for examining and defining relationships between datasets based on similarity between their metadata and specifically the similarity between variables within datasets as well as export of dataset metadata for sharing with other systems.

[7] https://www.djangoproject.com/

[8] https://www.postgresql.org/

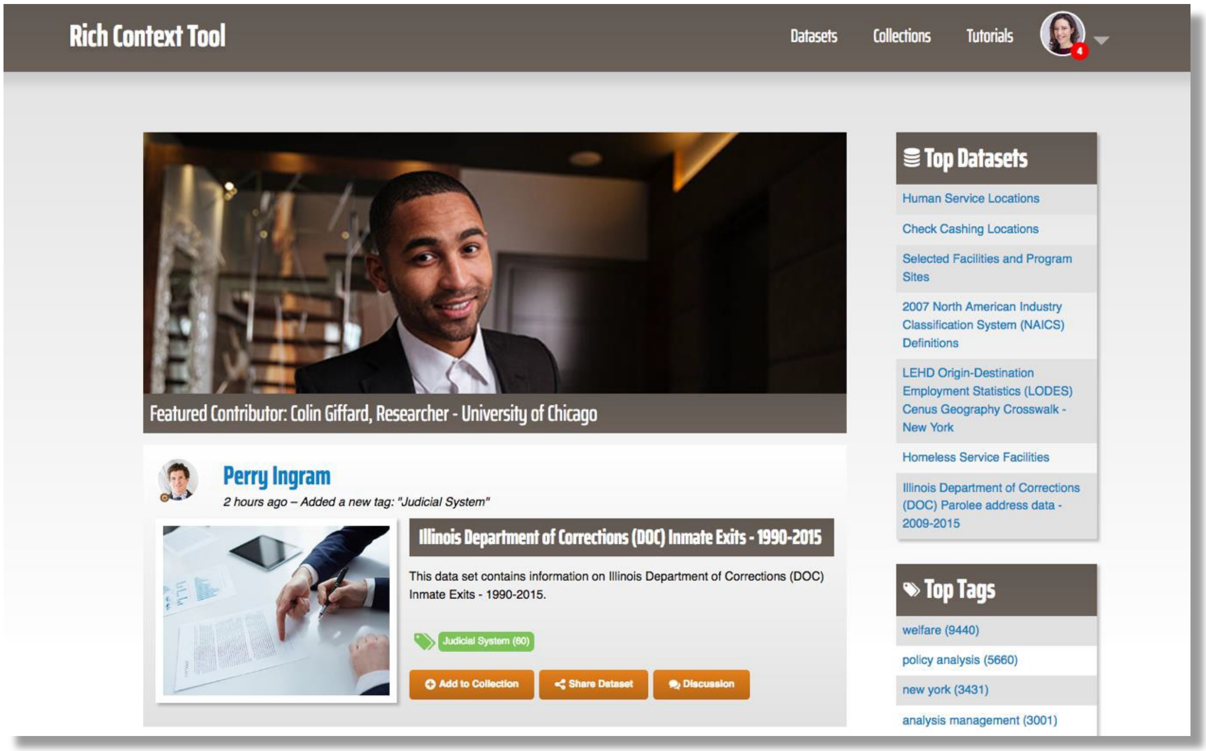
4. Features in Development

A key theme for the ongoing development of the ADRF Explorer is the incorporation of common “gamification” concepts borrowed from social media and online services. The goal of which is to incentivize experts and dataset users to help better describe and supplement dataset description with sample code, publications, and annotations on the dataset.

4.1 Rich Context Enhancements

Developing a richer context to understand a dataset requires input from users and experts who are using, have used, or have an in-depth understanding of how the data were generated. Inspired by TripAdvisor and Amazon, we believe discovery and consumption can be improved by documenting data by how it is used as well as by how it was produced. This module provides already enables researchers to both annotate datasets and upload code that can be used to read in and reuse data. The module would then be extended to make it easier for users to add this “rich context” – information about other researchers, publications, and topics that have been associated with the data.

We have worked recently and are in continued partnership with the NYU Game Center on a web application that is built with policy analysts and users of administrative data in mind to help them to collaborate. This work to date has resulted in a working prototype that demonstrates the potential of connecting researchers in more practical ways to the data they use. The approach has been to design ways in which users can earn rewards and reputation points as they contribute additional context, as well as better metadata and variable descriptions to datasets they use within their work. The Game Center is developing a "Prestige" feature that applies set values to meaningful actions that users take with the tool. Prestige will allow users to quantify and visualize their contributions to the larger knowledgebase, and encourage further contribution by highly featuring the work of users with high engagement. Applying point values to participation, comments, metadata contributions, and meaningful interaction will encourage expert users to share their knowledge in exchange for heightened visibility and recognition.



Featured Contributor: Colin Giffard, Researcher - University of Chicago

Perry Ingram

2 hours ago – Added a new tag: "Judicial System"

Illinois Department of Corrections (DOC) Inmate Exits - 1990-2015

This data set contains information on Illinois Department of Corrections (DOC) Inmate Exits - 1990-2015.

Judicial System (80)

Top Datasets

- Human Service Locations
- Check Cashing Locations
- Selected Facilities and Program Sites
- 2007 North American Industry Classification System (NAICS) Definitions
- LEHD Origin-Destination Employment Statistics (LODES) Cenus Geography Crosswalk - New York
- Homeless Service Facilities
- Illinois Department of Corrections (DOC) Parolee address data - 2009-2015

Top Tags

- welfare (9440)
- policy analysis (5660)
- new york (3431)
- analysis management (3001)

FIGURE 3: SCREENSHOT OF RICH CONTEXT TOOL PROTOTYPE HOMEPAGE

4.2 ETL Improvements

Through our current Applied Data Analytics training courses, we are in the process of developing better approaches by which the data we use for the course are transformed and loaded based on a pre-defined set of data-user needs (e.g., the creation of standardized data types around commonly used variables such as dates, geospatial data, and categorical data).

Appendix

GMeta METADATA SCHEMA V. 1.0

Dataset level metadata

Metadata associated with intellectual entity of the dataset

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 format (yyyy-mm-dd)	End date for years/months the dataset is valid for	Yes	0-1
geographic_coverage	Array of Strings	Human Readable geographic location the dataset applies to	Yes	0-n
geographic_unit	Array of Strings	Geographic resolution at which dataset is rendered, if applicable (e.g., State, City, Census Block, Census Tract).	Yes	0-n

keywords	Array of Strings	List of topics that characterize the dataset	Yes	0-n
category	String	Single thematic category the dataset falls under (e.g., Public Safety, Public Health, Transportation).	Yes	1
dataset_citation	String	How to cite this dataset in publications	No	1
source_archive	String	Name of original source where dataset can be found.	No	0-n
source_url	URL	External source for dataset if applicable (e.g., dataset found publically)	No	0-n
reference_url	URL	URL that points to more information about the dataset if applicable	No	0-n
data_provider	String	Provider of the dataset	Yes	1-n
data_steward	String	ADRF user who is responsible for reviewing project requests for the dataset and data exports for projects using the dataset.	Yes	1
data_steward_organization	String	Organization of the ADRF user who is the designated Data Steward for the dataset		0-1
data_usage_policy	String	How the dataset may be used. Including criteria for review, what can be exported, whether	Yes	0-n

		results of any analyses can be published or not		
access_actions_required	String	Proof of training/certification, signing of NDA, MOU, or other agreement which must occur before user can gain access to dataset.		0-n
access_requirements	String	Explicit requirements users must meet in order to be able to request access to a dataset. For example: US Citizen, US Government Employee, Researcher at Institution of higher education, sworn status.		0-n
data_classification	String	Like CUSP Data Classification, a level which dictates the type of users who have access or whether it requires the user to request access.		1
dataset_documentation	Array of Strings (Filepath s)	List of documents that document the data in this dataset.	No	0-n
files_total	Integer	Total number of files in Dataset (not including attachments such as data dictionaries)	No	1
related_articles	Array of Strings (article ids)	List of articles related to the dataset	Yes	0-n

dataset_version	Integer	Version of this dataset inside ADRF	No	1
dataset_version_date	Unix Timestamp	Date this version was published	No	1

File level metadata

Metadata associated with data file on disk

Field Name	Data Type	Description
File Size	String	File size of file on disk. E.g., 100 MB, 2 GB
Values Missing	Integer	Number of missing values in the file
Rows	Integer	Number of Rows on the file
Values	Integer	Number of values in the file
Columns Null	Integer	Number of NULL columns in the file
GPS-Lat-Max	Float	The maximum value in the file
Columns	Integer	Number of columns in the file
Values Missing Percent	Decimal	Percentage of Values missing in column
GPS-Long-Min	Float	The minimum GPS longitude value in the file

ETL-Profiler Status	Boolean	It has the OK value when Urban Profiler processed it successfully or the error message. Probably will not be included

GPS-Long-Max	Float	The maximum GPS longitude value in the dataset
ETL-Profiler Input File Size (KB)	Integer	FIle size of data file input to Urban Profiler
ETL-Profiler Input File	String	Filename of file input to Urban Profiler
GPS-Lat-Min	Float	The minimum GPS latitude value in the dataset
columns_metadata	Array of Strings	List of column_metadata objects. Each entry in the list has its name as key and the column_metada object as value.

Variable level metadata

Metadata associated with data file variables (same as columns)

Field Name	Data Type	Description
Adult_child_ind		This a sample of a column_metadata object.
column-name	String	The name of the column
description	String	Description of the column, it will be provided by the steward
values		Number of values
provided-type	String	Data type of column from the data provider
profiler-most-detected-%	Decimal	Percentage of the values that have the most detected type
profiler-most-detected		The most detected type
unique_values		Number of unique values.
missing	Integer	Number of missing values
profiler-type	String	Type of the column. Sometimes it is different than the most detected

		type.
max	Variable	Max value For numeric, temporal or geo (gps only)
min	Variable	Min value For numeric, temporal or geo (gps only)
Histogram Data JSON	Object	Data to create a histogram
top-k	Object	List of top-k values. K is a parameter of urban profiler. This list can have size<=k.
std	Float	Standard Deviation For numeric columns.
mean	Float	Mean For numeric for numeric columns
top-value	Variable	The value that occurred most.
freq-top-value	Variable	The frequency of the top value.

Sample detailed GMETA file

This is a sample of a GMeta JSON file with metadata for ADRF datasets and files.

```
{
        "gmeta": [
        {
```

```
"adrf-000077": {

"mimetype": "application/json",
"content": {
       "dataset_id": "adrf-000077",
                        "temporal_coverage_end": "2010",

                        "files_total": 1,

                        "data_classification": "Public",

                        "access_actions_required": "No further actions required to access this
dataset",

                        "geographical_coverage": [

                        "Illinois"

                        ],

                        "keywords": [

                        "census",
                        "geospatial",
                        "cross-walk",
                        "geography",
                        "state"

                        ],

                        "category": "Census Geography",
                        "dataset_version": 1,

                        "title": "2010 Census Block Assignment Files - Incorporated places /
census designated places - Illinois",

                        "data_usage_policy": "This dataset is intended for public access and
use.",

                        "data_steward_organization": "Center for Urban Science and Progress
(CUSP), NYU",
"data_steward": "Drew Gordon",
"files": [
{
    "file_name":"adrf-000077-census_baf_il_incplace_cdp.csv",
    "columns_metadata": {
    "BLOCKID": {
        "profiler-type": "Textual",
        "profiler-most-detected": "Textual-PHONE",
        "missing": 0.0,
        "values": 451554.0,
        "top-k": {
```

```
"171635043031106": 1,

"171635043031107": 1,

"171635043031104": 1,

"171635043031105": 1,

"171635043031102": 1,

"171635043031103": 1,

"171635043031100": 1,

"171635043031101": 1,

"171635043031108": 1,

"171635043031109": 1,

"170314914001002": 1,

"170314914001003": 1,

"170314914001000": 1,

"170318077001027": 1,

"171978803041011": 1,

"172030302002038": 1,

"170810506004000": 1,

"172030302002036": 1,

"172030304001061": 1,
"171430002003001": 1,

"170630005001056": 1,

"170630005001055": 1,

"171950009002078": 1,

"171950009002079": 1,

"171950009002072": 1,

"171950009002073": 1,

"171950009002071": 1,

"171950009002076": 1,

"171950009002077": 1,

"171950009002074": 1,

"171950009002075": 1,

"170898501064017": 1,
"170898501064016": 1,

"170898501064015": 1,

"170898501064014": 1,

"170898501064013": 1,

"170898501064012": 1,

"171830112001017": 1,

"170898501064019": 1,

"170898501064018": 1,

"170318212001029": 1,

"170318212001028": 1,

"170318212001021": 1,

"170318212001020": 1,

"170318212001023": 1,

"170318212001022": 1,

"170318212001025": 1,

"170318212001024": 1,

"170318212001027": 1,

"170318212001026": 1

},
                                     "top-value": "170810506004000",

                                     "freq-top-value": 1,

                                     "description": "15-character code that is the
concatenation of fields consisting of the 2-character state FIPS code, the 3-character
county FIPS code, the 6-character census tract code, and the 4-character tabulation block
code."

                             },

                             "PLACEFP": {

                                     "profiler-type": "Numeric",

                                     "profiler-most-detected": "Numeric-Integer",
                                     "missing": 162967.0,

                                     "values": 288587.0,

                                     "min": 113.0,
                                     "max": 84220.0,
                                     "std": 24325.791326155173,
"mean": 37294.64430830217,
"Histogram Data JSON": {
"113.0": 25758,
"8523.7": 69622,
"16934.4": 27528,
"25345.1": 20584,
"33755.8": 19974,
"42166.5": 25742,
"50577.2": 29697,
"58987.9": 26385,
"67398.6": 22743,
"75809.3": 20554
},
"top-k": {
"57225.0": 873,
"1114.0": 1078,
"38934.0": 826,
"53234.0": 829,
"30926.0": 1137,
"3012.0": 2362,
"14000.0": 46324,
"12385.0": 1580,
"14351.0": 1273,
"19642.0": 1008,
"81048.0": 881,
"72546.0": 971,
"49867.0": 949,
"6613.0": 1744,
"51622.0": 1806,
"38570.0": 2694,
"2154.0": 1252,
"33383.0": 1107,
"20591.0": 921,
"62367.0": 1266,
"54885.0": 1138,
"54820.0": 1010,
"22164.0": 905,
"10487.0": 1088,
"72000.0": 3901,
"65078.0": 917,
"35411.0": 864,
"24582.0": 1504,
"68003.0": 1109,
"79293.0": 1200,
"57628.0": 959,
"46916.0": 872,
"18823.0": 2202,
"12164.0": 880,
"22255.0": 1289,
"43536.0": 907,
"23074.0": 1899,
"56640.0": 957,
"28326.0": 884,
"7133.0": 888,
"18563.0": 1882,
"14026.0": 904,
"77005.0": 904,
"5573.0": 1139,
"75484.0": 825,
"70122.0": 1779,
"58447.0": 1207,
"65000.0": 4496,
"4845.0": 1303,
```

```
"59000.0": 2482
```

```
                                        },

                                        "top-value": 14000.0,

                                        "freq-top-value": 46324,

                                        "description": "5-character place FIPS code"

                                }

                                },

                                "file_type": "data",
                                "file_size": 9570922,
                                "mimetype": "text/csv"

                        },

                        {

                                "file_name":"adrf-000077-census_baf_il_incplace_cdp_dd.csv",
                                "file_type": "documentation",

                                "file_size": 259,

                                "mimetype": "text/csv"

                        }

                        ],

                        "access_requirements": "No access restrictions on this dataset",

                        "description": "Cross walk of 2010 Census Blocks to Census Places for
the state of Illinois. Geographic data here has been pulled from the 2010 Census Block
Assignment Files dataset.\r\n\r\n\r\nBlock Assignment Files (BAFs) have been created for
each of the 50 states, the District of Columbia, and Puerto Rico. Each file contains 2010
Census tabulation block codes and geographic area codes for a specific geographic entity
type. Each BAF contains every block within the given state, even if the block is not within
one of the geographic areas represented in the file. For those blocks where no geographic
area is present, the block is listed, followed by one or more commas (depending on the file
layout for the specific geographic area type).",
"source_url":"https://www.census.gov/geo/maps-data/data/baf.html",
"geographical_unit": [

"State",


"Census Block",
"Census Place"

],


"related_articles": [],
"data_provider": "US Census Bureau",
"dataset_documentation": [
                        ""

                        ],

                        "dataset_version_date": 1496163737,

                        "source_archive": "US Census Bureau",

                        "dataset_citation": "US Census Bureau, 2010, \"2010 Census Block
Assignment Files - Incorporated places / census designated places - Illinois\",
https://www.census.gov/geo/maps-data/data/baf.html, United States Census Bureau
[Distributor], 1 [Version]",

                        "reference_url": "https://www.census.gov/geo/maps-
data/data/baf_description.html",

                        "temporal_coverage_start": "2010",
                        "file_names": [

                        "adrf-000077-census_baf_il_incplace_cdp.csv"

                        ]
                },

                "visible_to": [

                        "public"

                ]

                }

        }

        ]

}
```