2023 ASIS&T WEBINAR SERIES MARCH 30, 2023

"CRITICAL DATA MODELING (DCMI)"

webinars@asist.org

asis&t

Association for Information Science and Technology

CRITICAL DATA MODELING

Karen Wickett
School of Information Sciences
University of Illinois Urbana-Champaign

ASIST/DCMI Webinar
March 30, 2023

CRITICAL DATA MODELING

- Using data modeling and systems analysis techniques to closely examine the creation and implementation of information systems
  - to expose unjust or biased assumptions in data models or algorithms
  - to highlight the technical roles of those assumptions within a system
- Exposing and examining these assumptions in the language of the technical systems in which they are embedded will validate and expand existing critiques and enable the development of more equitable information systems.
- Wickett, K. M. (2023). Critical data modeling and the basic representation model. *Journal of the Association for Information Science and Technology*, 1– 11. https://doi.org/10.1002/asi.24745

WHY WE NEED CRITICAL DATA MODELING

- Digital platforms are pervasive and the information representation decisions that structure those systems play out in the lives of our communities.
- In *Race After Technology* (2019), Ruha Benjamin shows how information systems have created an insidious system of oppression obfuscated by the technical nature of the systems involved.
- The complex relationships and layers of encodings involved in the realization of any digital object are an additional obfuscating factor for the analysis of modern information objects and systems.

DATA MODELING

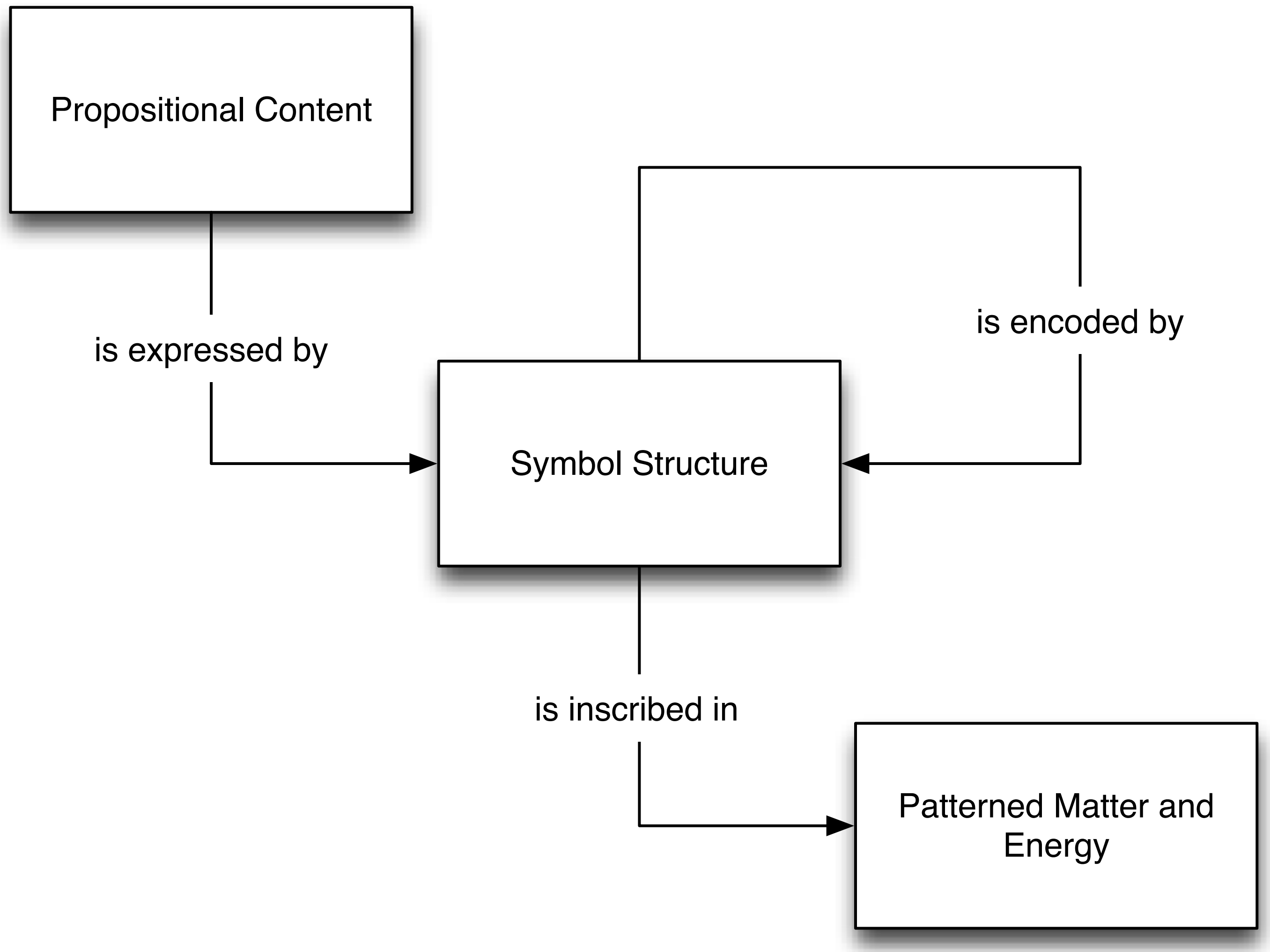
- A *data model* is a set of labels for categories, and a set of assumptions about how those categories will be handled in a computational system.
- Every information system uses data models to structure data and to specify processing.
- Data modeling necessarily reduces real-world entities into the set of attributes defined by a data model.
- Data modeling therefore takes a position on what matters about those entities by elevating some aspects and leaving others out.

THE BASIC REPRESENTATION MODEL

A general model for information representation and encoding in digital objects.

An organizing framework for critical readings of datasets and systems

- Naming the entities and relationships involved at any level of representation or encoding
- Highlighting the interconnections between information representation choices at the various levels.



THE BASIC REPRESENTATION MODEL

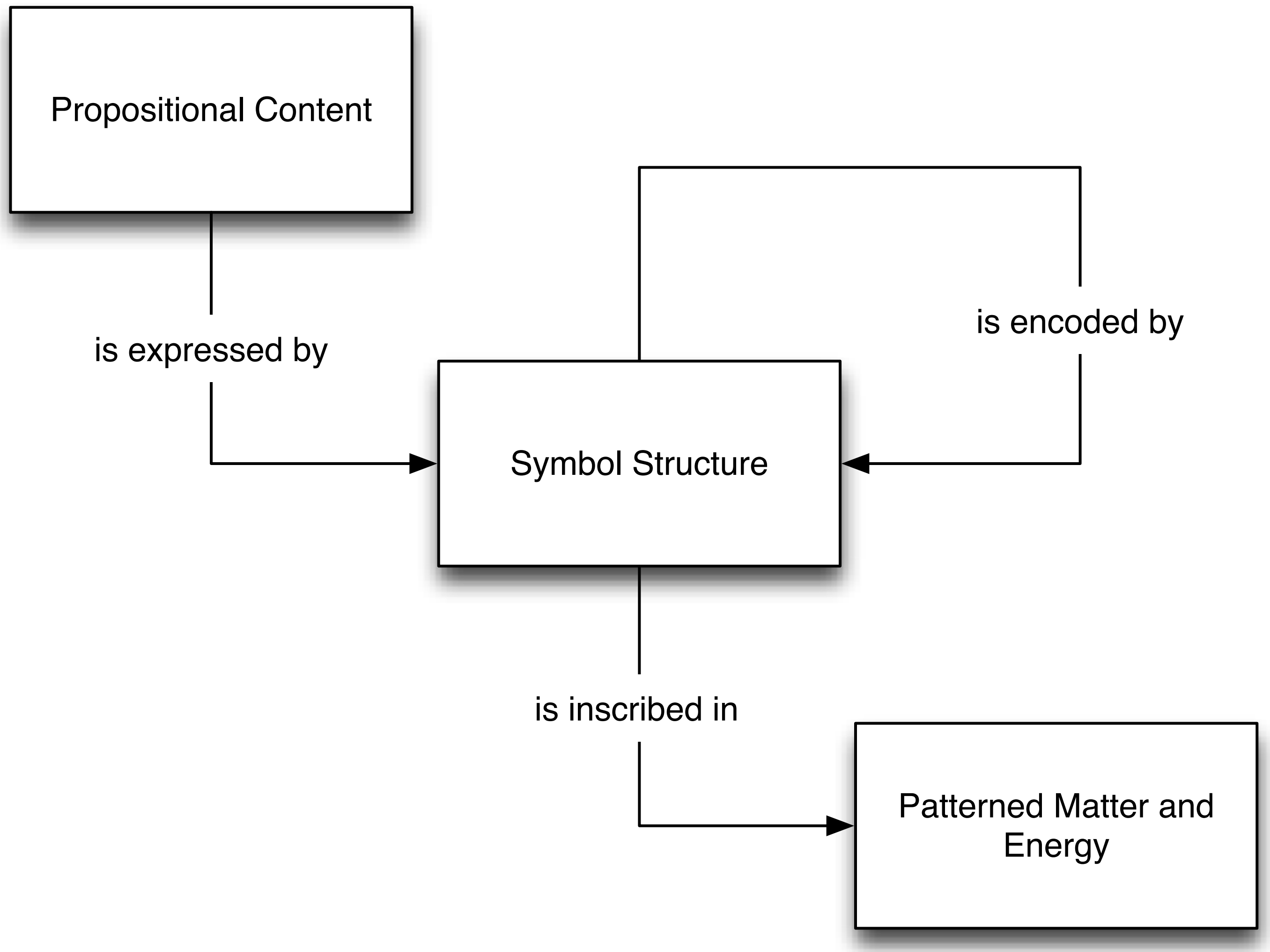
Expressions and encodings in various forms convey propositional content to some audience.

- The content of a class roster could be expressed as a table of data, an XML document, or a series of sentences in natural language

Symbol structures express semantic content in some context or encode other symbol structures within a computational system.

- layers of encoding and representation in a digital system are by a series of *is Encoded By* relationships between *Symbol Structures*.

We are material beings in a material world, we encounter information via patterned matter or energy that carries the inscription of symbol structures

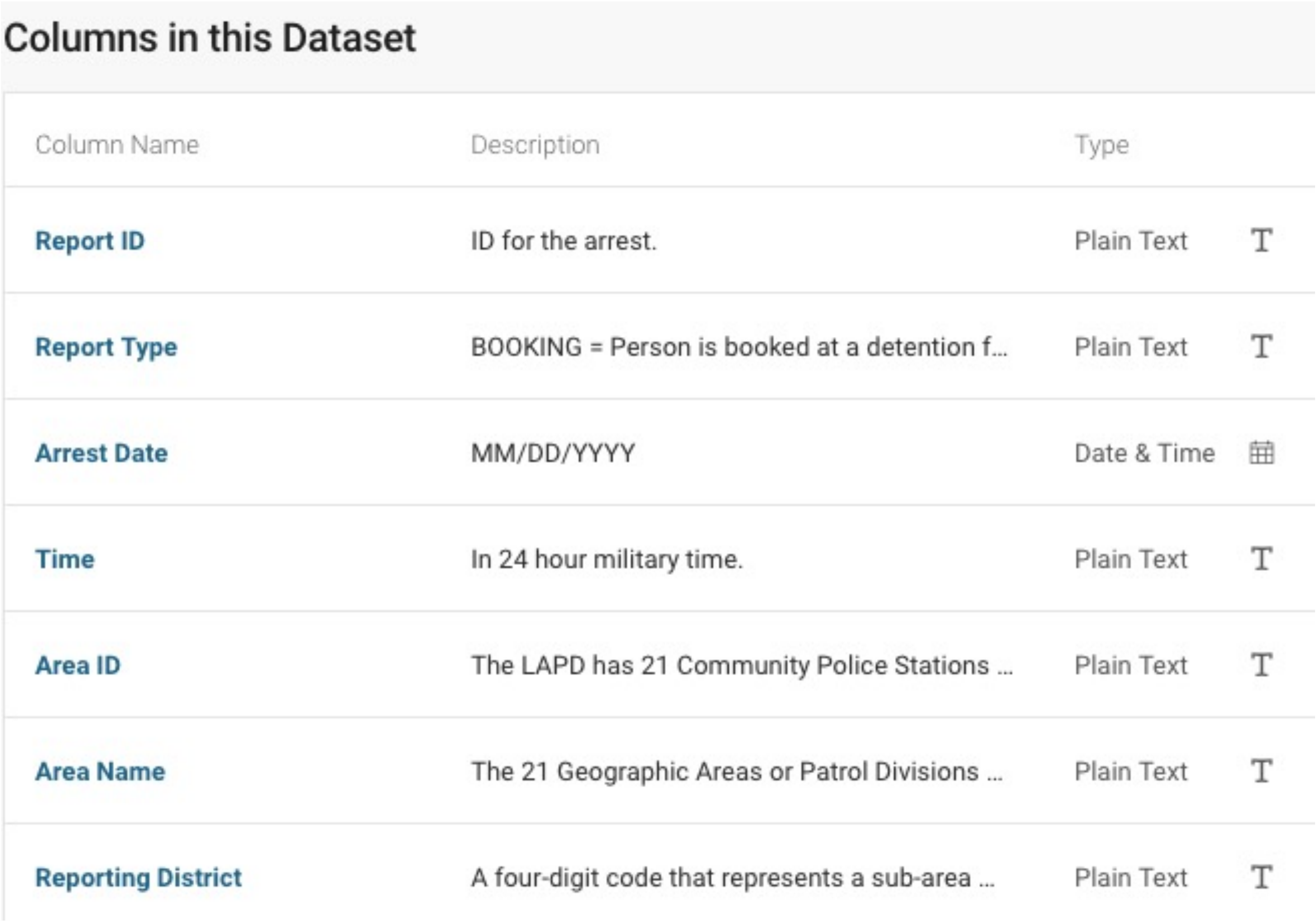


MODELING A POLICE ARREST RECORD DATASET

Examples from an ongoing critical reading of the “Arrest Data from 2020 to Present” dataset published through the Los Angeles Open Data platform.

Close readings of the dataset versions and data model documentation.

- Metadata for the entire dataset and a table that describes each of the 25 columns of data provided in the dataset.
- Structured queries designed around reading the metadata.



Columns in this Dataset

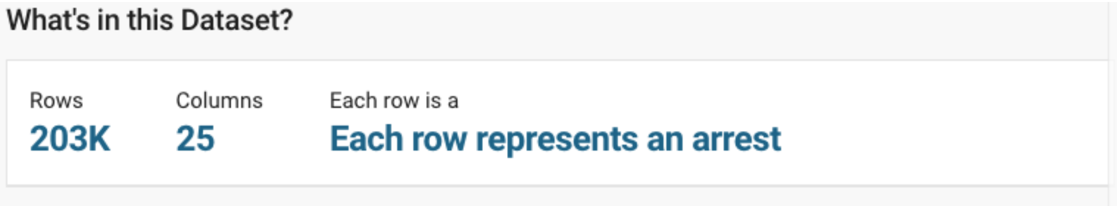
Column Name	Description	Type	
Report ID	ID for the arrest.	Plain Text	T
Report Type	BOOKING = Person is booked at a detention f...	Plain Text	T
Arrest Date	MM/DD/YYYY	Date & Time	▦
Time	In 24 hour military time.	Plain Text	T
Area ID	The LAPD has 21 Community Police Stations ...	Plain Text	T
Area Name	The 21 Geographic Areas or Patrol Divisions ...	Plain Text	T
Reporting District	A four-digit code that represents a sub-area ...	Plain Text	T

COUNTING ROWS AS CRIMES

Arrests are the primary entity in the dataset.

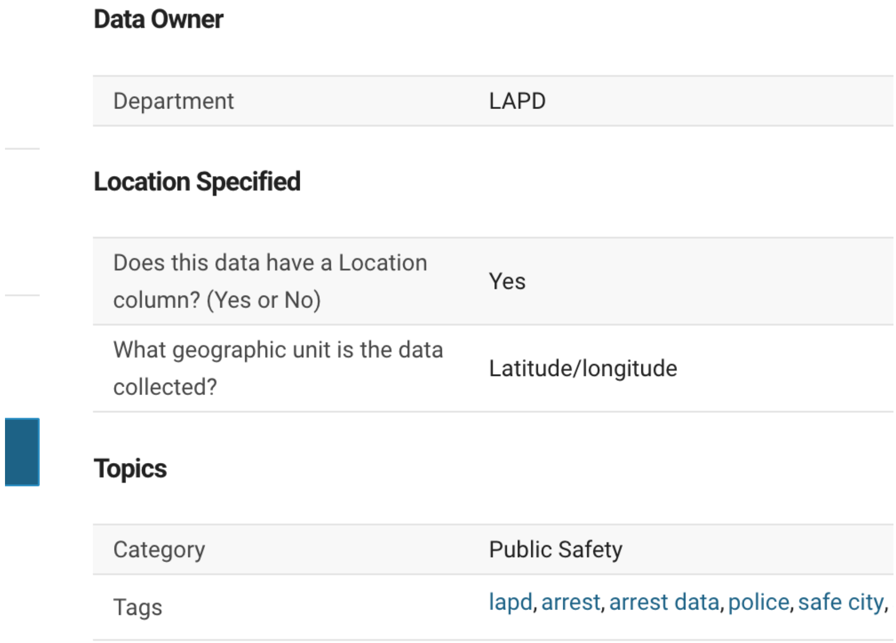
There is an assumed correspondence between a row in the dataset and a criminal incident.

Location information in the dataset reflects location of a **criminal incident** associated with an arrest, *not* the location where the arrest occurred.



What's in this Dataset?

Rows	Columns	Each row is a
203K	25	Each row represents an arrest



Data Owner

Department	LAPD

Location Specified

Does this data have a Location column? (Yes or No)	Yes
What geographic unit is the data collected?	Latitude/longitude

Topics

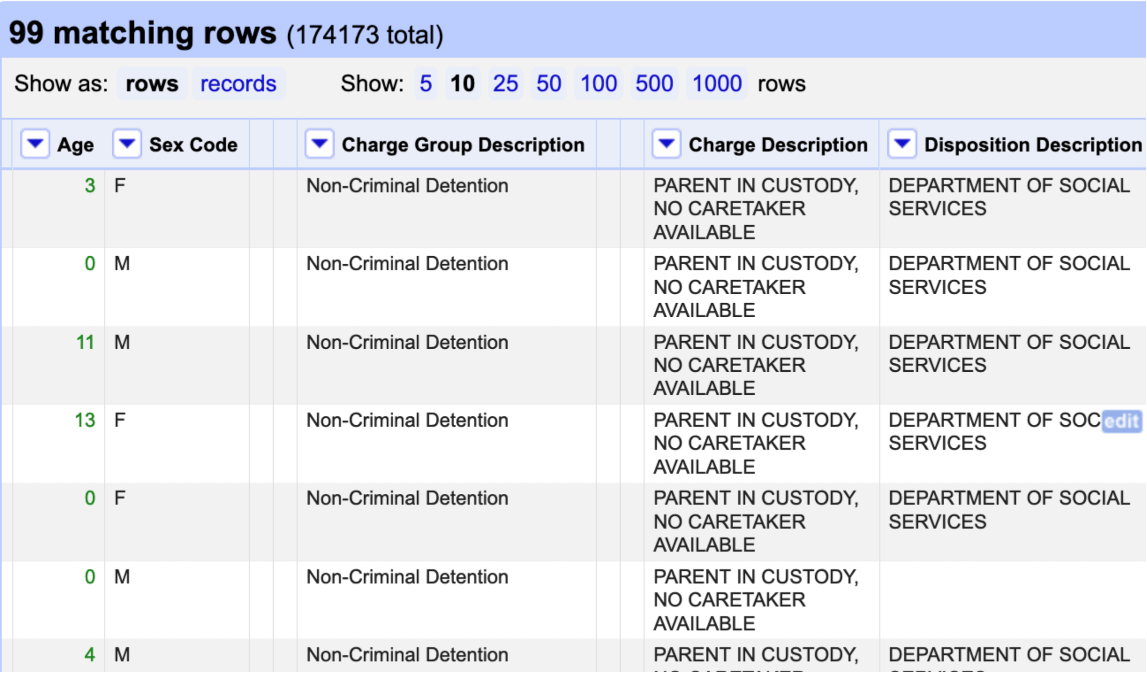
Category	Public Safety
Tags	lapd, arrest, arrest data, police, safe city,

COUNTING ROWS AS CRIMES

Some rows do not correspond to suspected criminal activity of the person described by that row.

Filtering the dataset for the value "PARENT IN CUSTODY, NO CARETAKER AVAILABLE" in Charge Description gives 99 rows, all with Age values between 0 and 17.

- These rows represent children taken into custody by the LAPD when one or both of their parents were arrested.



99 matching rows (174173 total)

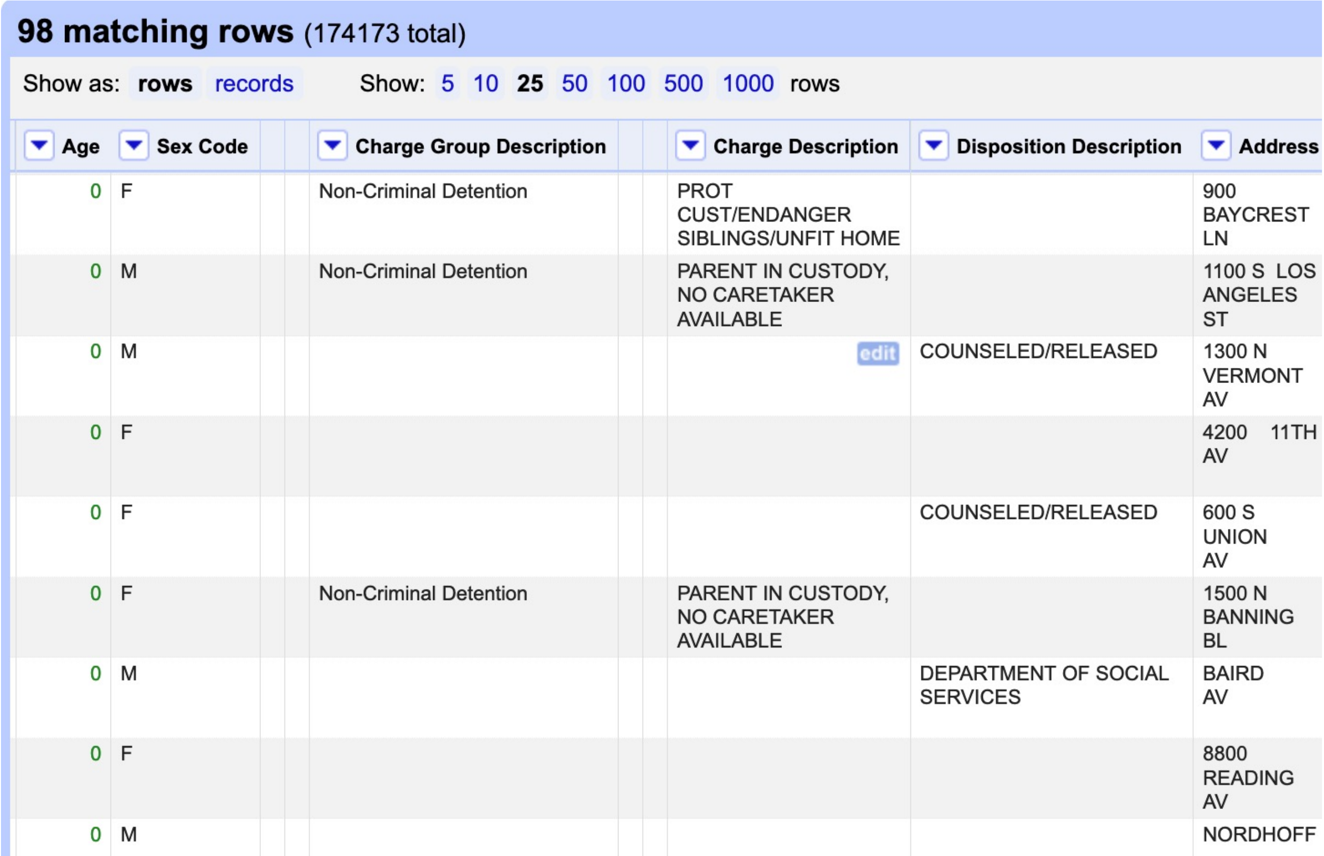
Show as: **rows** records   Show: 5 **10** 25 50 100 500 1000 rows

Age	Sex Code	Charge Group Description	Charge Description	Disposition Description
3	F	Non-Criminal Detention	PARENT IN CUSTODY, NO CARETAKER AVAILABLE	DEPARTMENT OF SOCIAL SERVICES
0	M	Non-Criminal Detention	PARENT IN CUSTODY, NO CARETAKER AVAILABLE	DEPARTMENT OF SOCIAL SERVICES
11	M	Non-Criminal Detention	PARENT IN CUSTODY, NO CARETAKER AVAILABLE	DEPARTMENT OF SOCIAL SERVICES
13	F	Non-Criminal Detention	PARENT IN CUSTODY, NO CARETAKER AVAILABLE	DEPARTMENT OF SOC edit SERVICES
0	F	Non-Criminal Detention	PARENT IN CUSTODY, NO CARETAKER AVAILABLE	DEPARTMENT OF SOCIAL SERVICES
0	M	Non-Criminal Detention	PARENT IN CUSTODY, NO CARETAKER AVAILABLE	
4	M	Non-Criminal Detention	PARENT IN CUSTODY,	DEPARTMENT OF SOCIAL

COUNTING ROWS AS CRIMES

Rows for children in the Arrest Dataset have a tendency for sparseness but still have **location information.**

- 66 of the 98 rows with Age of 0 have no Charge Description listed
- Only 2 of the 98 rows with Age of 0 have location as (0,0), which indicates missing data
- 96 rows list a latitude and longitude for location



98 matching rows (174173 total)

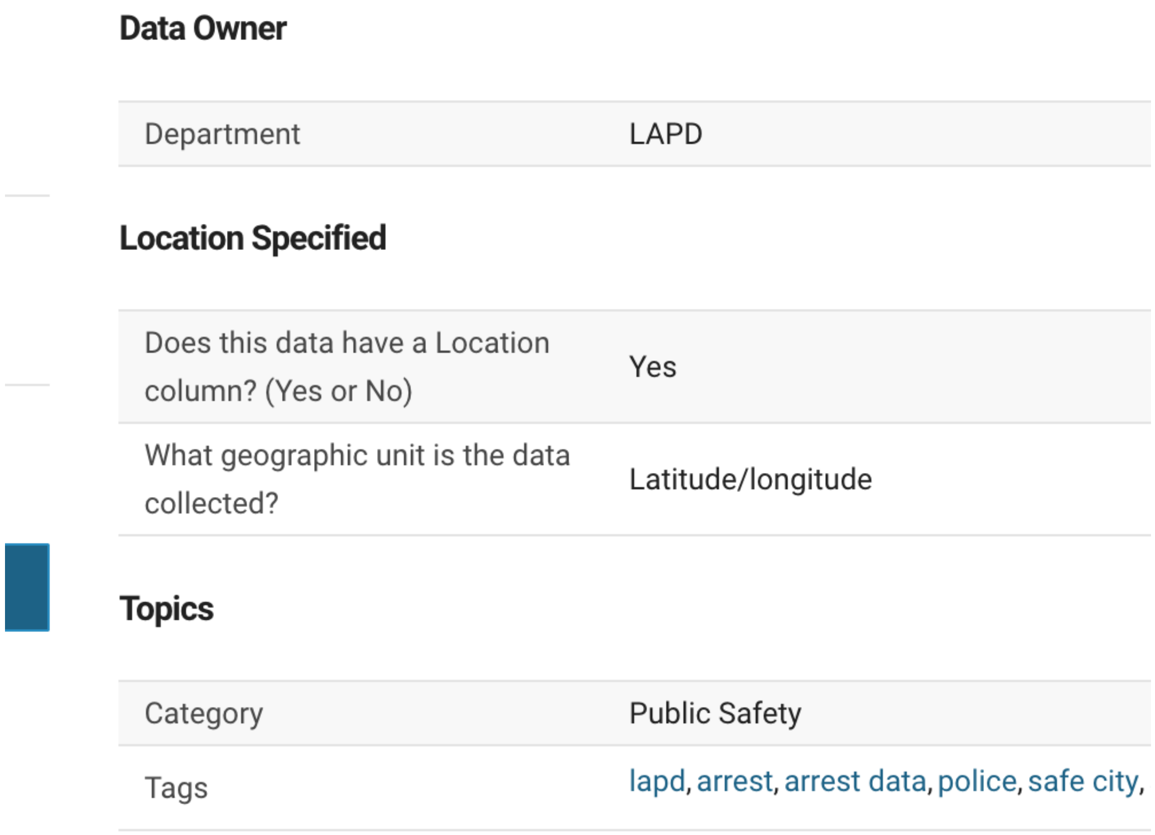
Show as: **rows** records   Show: 5 10 **25** 50 100 500 1000 rows

Age	Sex Code	Charge Group Description	Charge Description	Disposition Description	Address
0	F	Non-Criminal Detention	PROT CUST/ENDANGER SIBLINGS/UNFIT HOME		900 BAYCREST LN
0	M	Non-Criminal Detention	PARENT IN CUSTODY, NO CARETAKER AVAILABLE		1100 S LOS ANGELES ST
0	M		edit	COUNSELED/RELEASED	1300 N VERMONT AV
0	F				4200 11TH AV
0	F			COUNSELED/RELEASED	600 S UNION AV
0	F	Non-Criminal Detention	PARENT IN CUSTODY, NO CARETAKER AVAILABLE		1500 N BANNING BL
0	M			DEPARTMENT OF SOCIAL SERVICES	BAIRD AV
0	F				8800 READING AV
0	M				NORDHOFF

ATTRIBUTES FOREGROUND LOCATION

10 of the 25 available attributes offer information about geographic location, at varying levels of granularity

- Area ID, Area Name, Reporting District, Address, Cross Street, LAT, LON, Location, Booking Location, Booking Location Code.



Data Owner

Department	LAPD

Location Specified

Does this data have a Location column? (Yes or No)	Yes
What geographic unit is the data collected?	Latitude/longitude

Topics

Category	Public Safety
Tags	lapd, arrest, arrest data, police, safe city,

EXPORT FORMATS FOREGROUND LOCATION

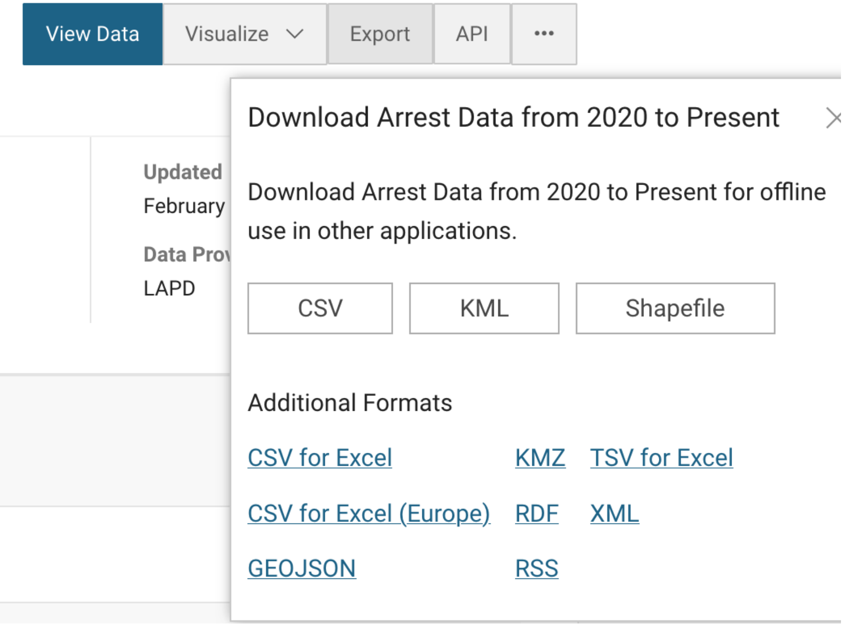
There are three main formats for downloading the Arrest Dataset:

- CSV, KML, and Shapefile.
- KML and Shapefile are both geographic data formats

The dataset is positioned as geographic information.

Criminal justice systems transform information into geographic information.

This has a significant impact on how we understand places and people in our communities (Jefferson, 2020).



Download Arrest Data from 2020 to Present

Download Arrest Data from 2020 to Present for offline use in other applications.

Additional Formats

DATATYPE OVER ACCURACY - TIME

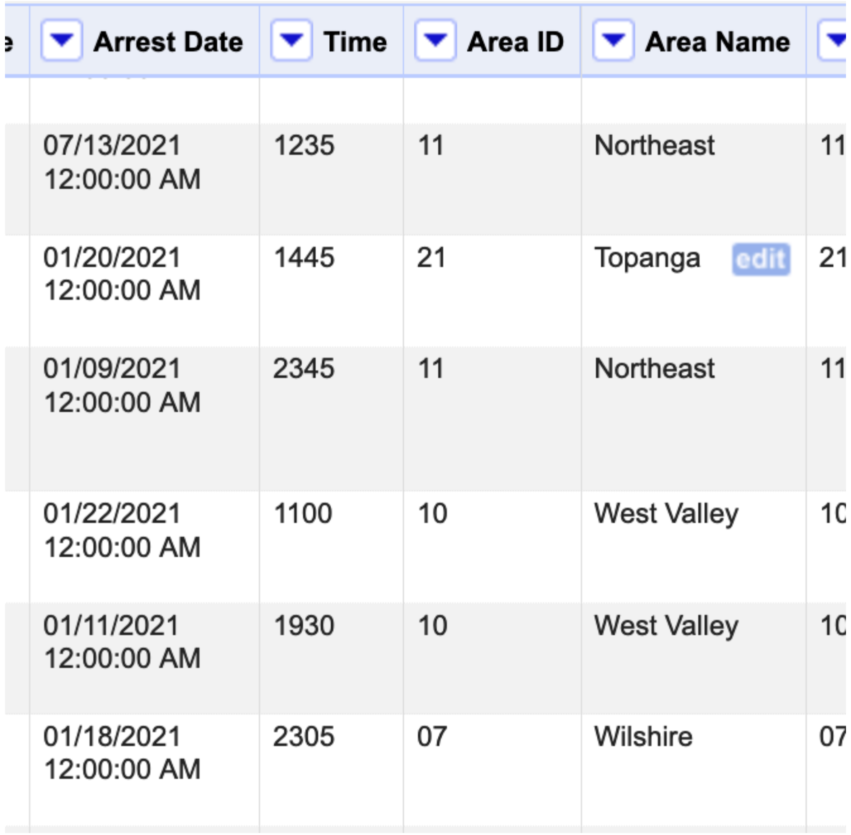
Arrest Date and Booking Date use the Floating Timestamp datatype.

To conform with the datatype, values for this attribute must include a time, not just a date.

Every time value listed for Arrest Date and Booking Date is “12:00:00 AM”

Other attributes give times with a distribution of values that suggest that they are accurate transcriptions from the original arrest record

- Time corresponds with Arrest Date
- Booking Time corresponds with Booking Date



Arrest Date	Time	Area ID	Area Name	
07/13/2021 12:00:00 AM	1235	11	Northeast	11
01/20/2021 12:00:00 AM	1445	21	Topanga	21
01/09/2021 12:00:00 AM	2345	11	Northeast	11
01/22/2021 12:00:00 AM	1100	10	West Valley	10
01/11/2021 12:00:00 AM	1930	10	West Valley	10
01/18/2021 12:00:00 AM	2305	07	Wilshire	07

DATATYPE OVER ACCURACY - LOCATION

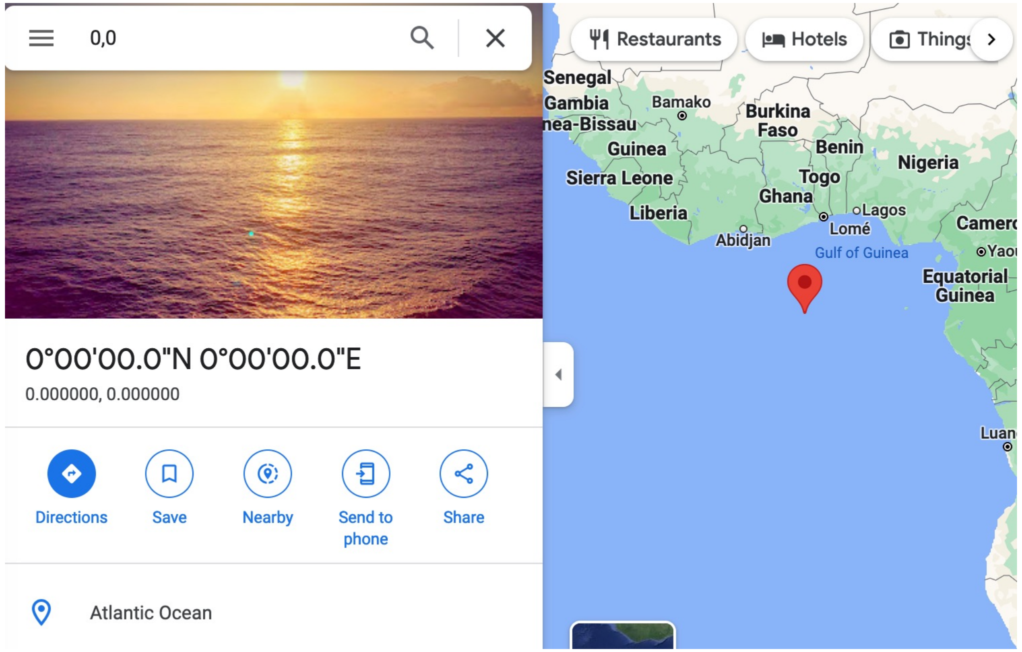
The dataset summary states, “Some location fields with missing data are noted as (0.0000°, 0.0000°).” But this is an actual location.

This point appears in 4640 number of rows in the dataset.

Every row in the dataset has a value for Location that matches the Point datatype, even in rows that were missing location data in the original arrest record.

Maintaining conformance with the datatype has taken precedence over accuracy of the data.

- Awareness of levels of representation and interactions between lets us analyze these cases



0,0

0°00'00.0"N 0°00'00.0"E

0.000000, 0.000000

Atlantic Ocean

CONCLUSIONS

- Information science is called as a field to examine the ways in which social injustices are built into our information systems through data models.
- Critical data modeling is a method for analyzing data models and existing data objects in order to examine the sociotechnical commitments, underlying assumptions, and social and cultural consequences of information systems.
- The Basic Representation Model provides a logical framework for a critical interrogation of an information object.

THANK YOU FOR ATTENDING

Submit a webinar proposal at https://www.asist.org/meetings-events/webinars/

webinars@asist.org

A copy of the recording and a follow-up survey will be emailed within 24 hours.

asis&t

Association for Information Science and Technology