I 202: Information Organization & Retrieval
Fall 2025

Collections

I’m so excited to introduce my tidying course.

Today’s Outline

- Collections
- Resources
- Organizing Systems
- Maintenance
- Ethics

Do you organize your spices? How?

Marion Botella, Unsplash

What is the **Right** Way to Organize?

Glushko, TDO

By cuisine

wikipedia

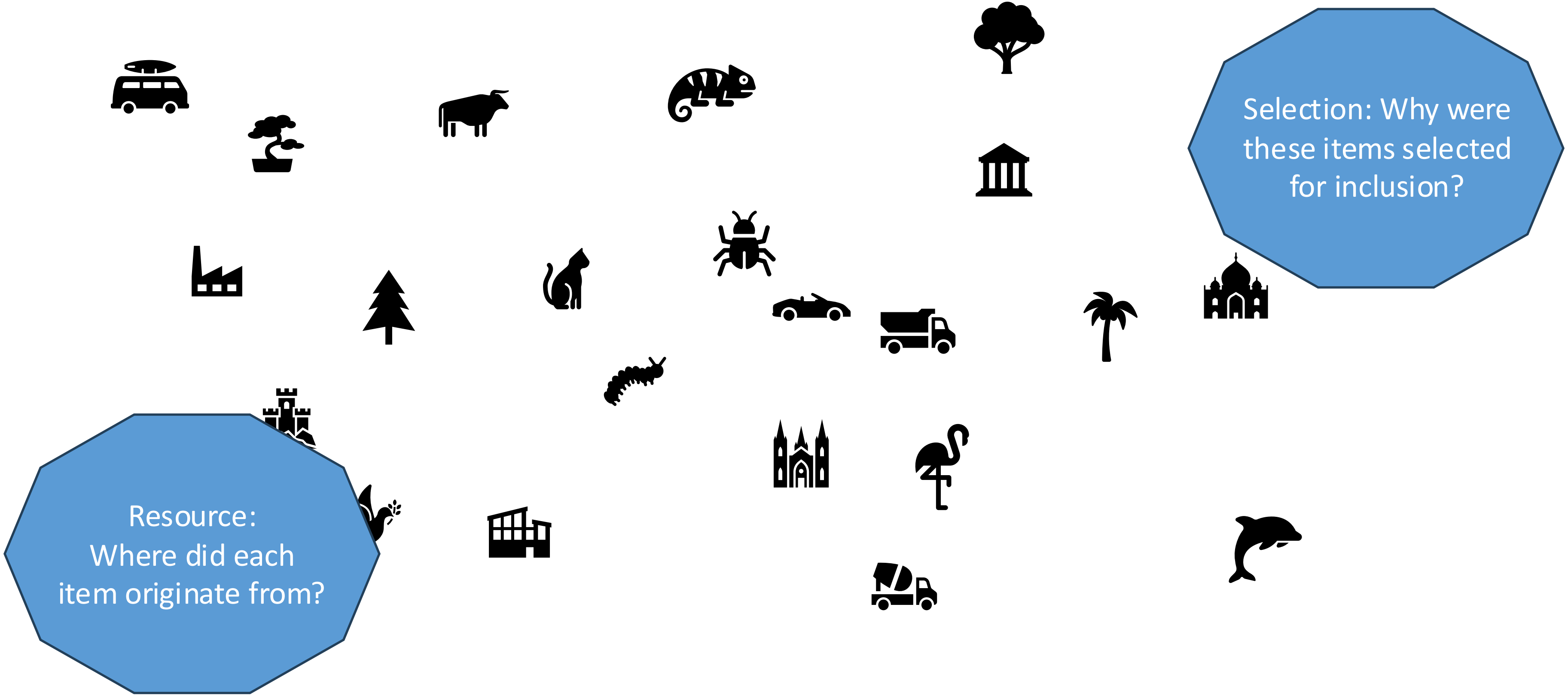
By price, eye appeal?

M Hearst

Alphabetically?

Collections and Organizing:

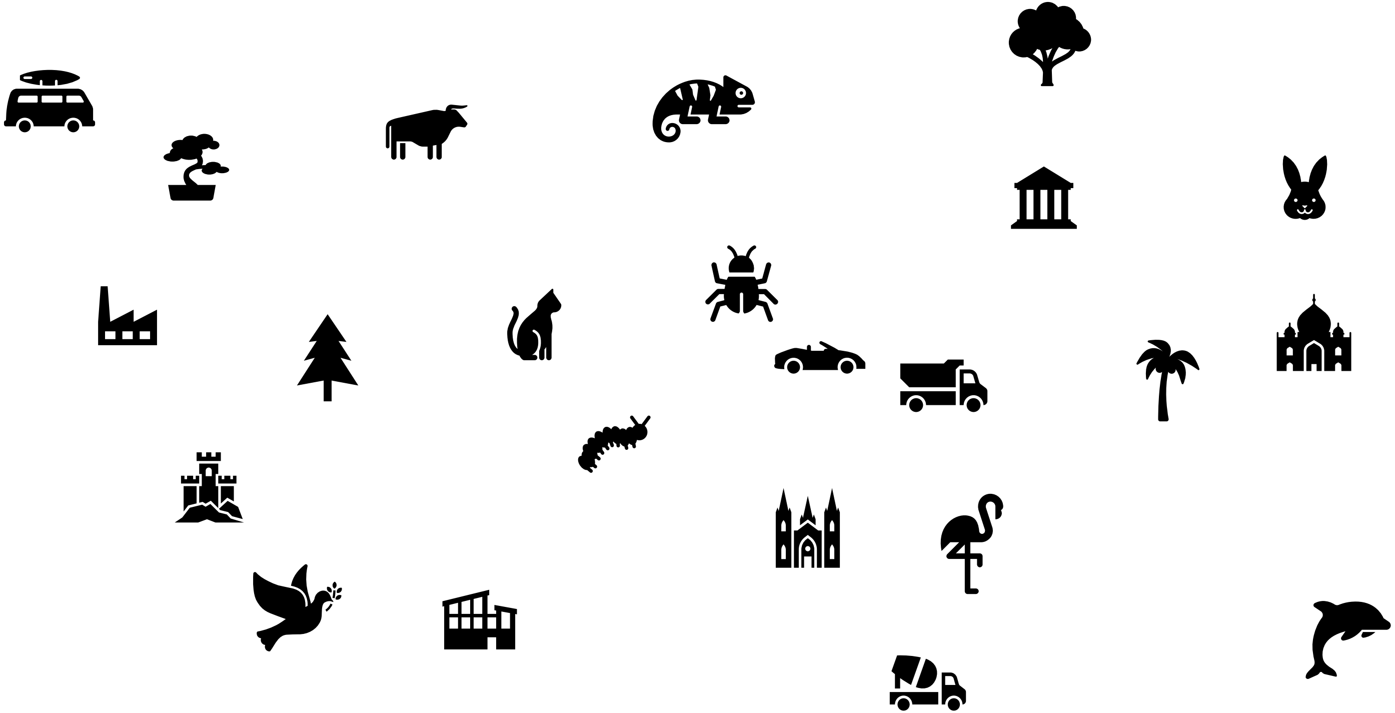
Collections and Organizing:
Start with a Universe of Data / Information / Resources



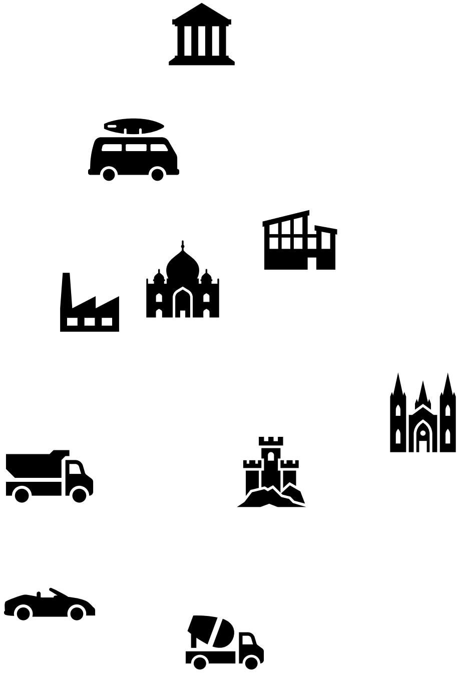
Selection: Why were these items selected for inclusion?

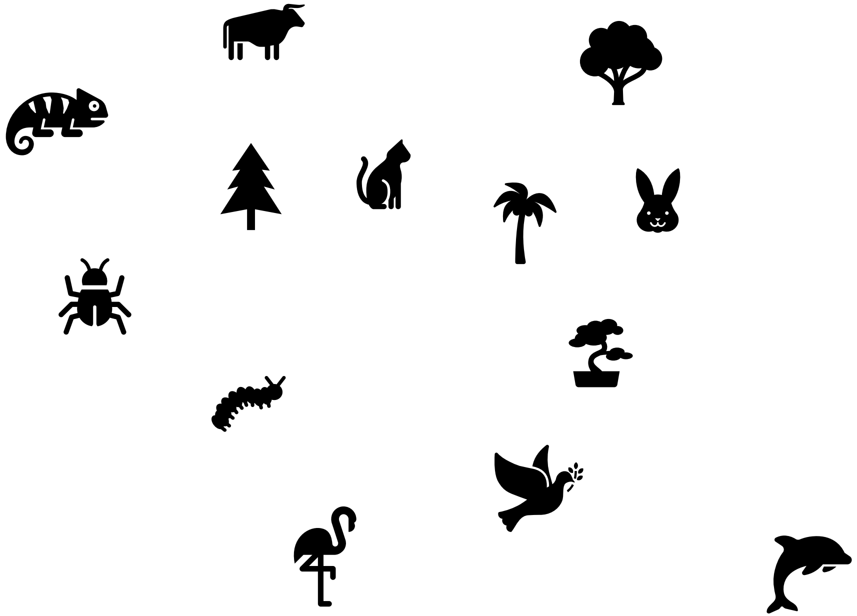
Resource:
Where did each item originate from?

Organization: How should these be grouped?

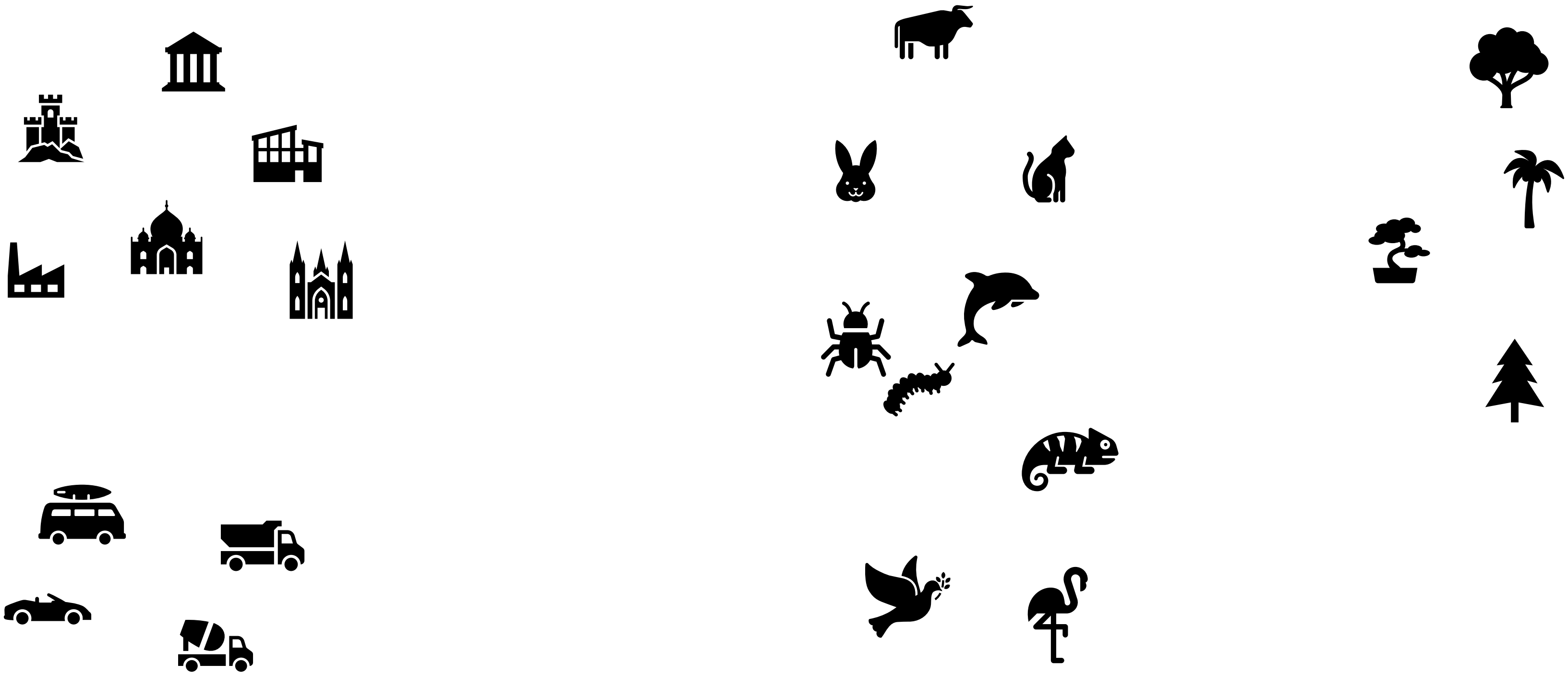


What is the Organizing Attribute?

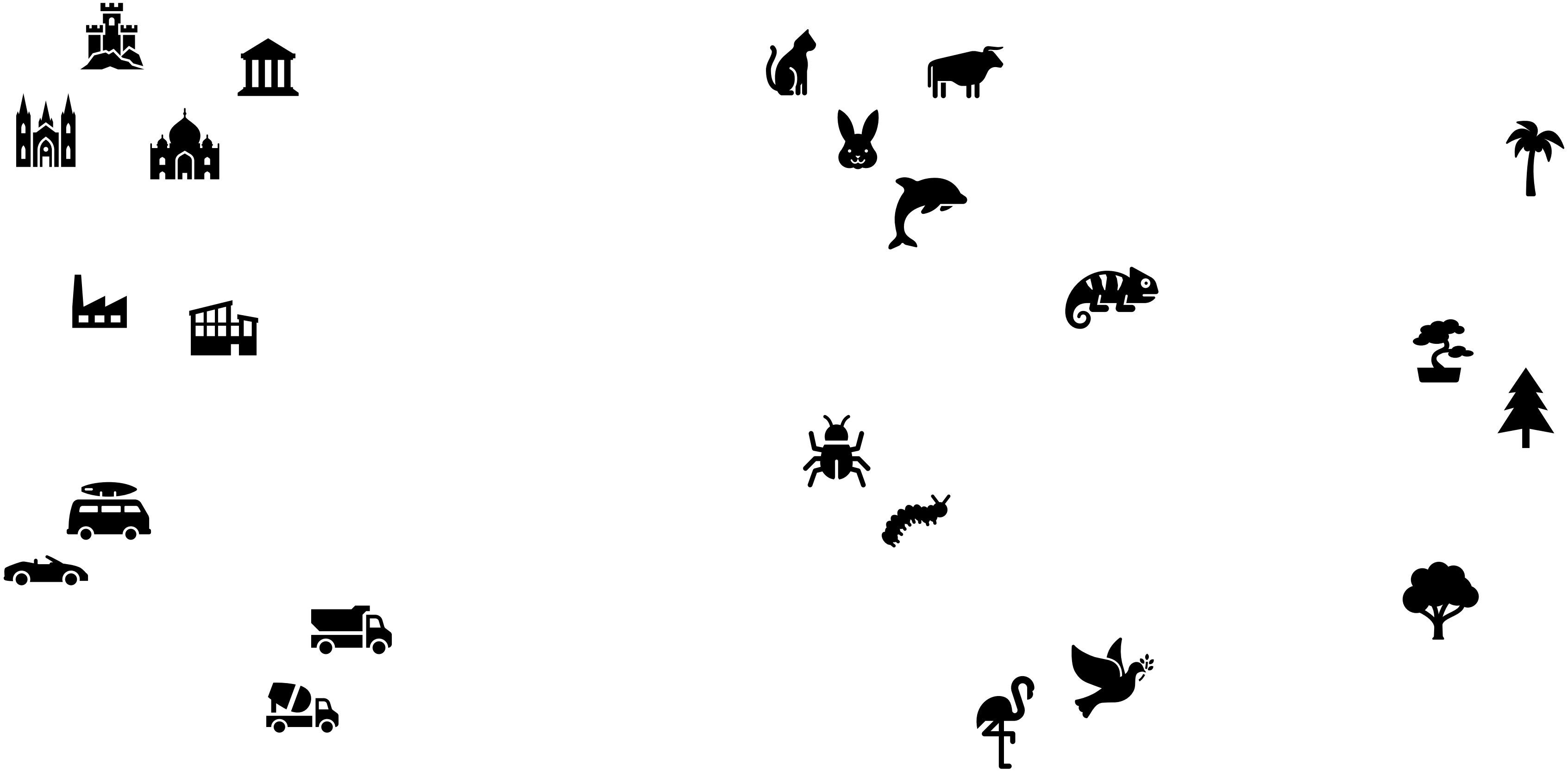




What are the Organizing Attributes?



What are the Organizing Attributes?



Collections & Organizing Systems

- **Collection**: A group of resources that have been selected for some purpose.
- **Organizing System**: An intentionally arranged collection of **resources** and the **interactions** they support.
- **Intentional Arrangement**: acts of organization by people (or computational proxies)

Definitions from TDO

What To Take From This Reading (Organizing System Roadmap)

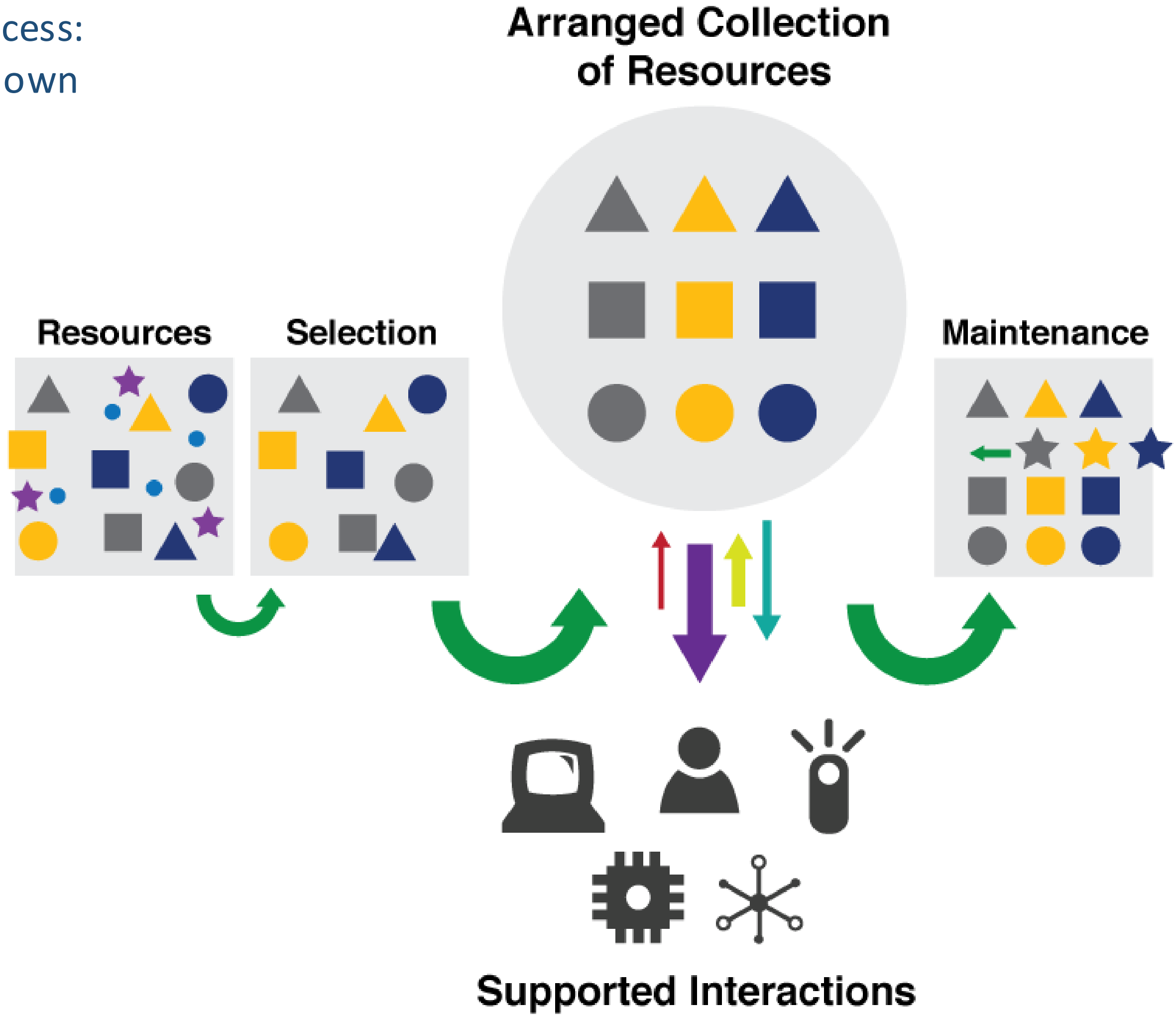
THE DISCIPLINE OF ORGANIZING

edited by ROBERT J. GLUSHKO

- The questions to ask when **designing a collection**
- The role of **resources** in collections
- Resource **selection, organization, interactions**, and **maintenance**.
- The role of **standards**

The whole process:

Let’s break it down



Identify the universe of possible resources / items / data / information

TDO Definitions of **Resource**:

”Anything of value that can support goal-directed activity.”

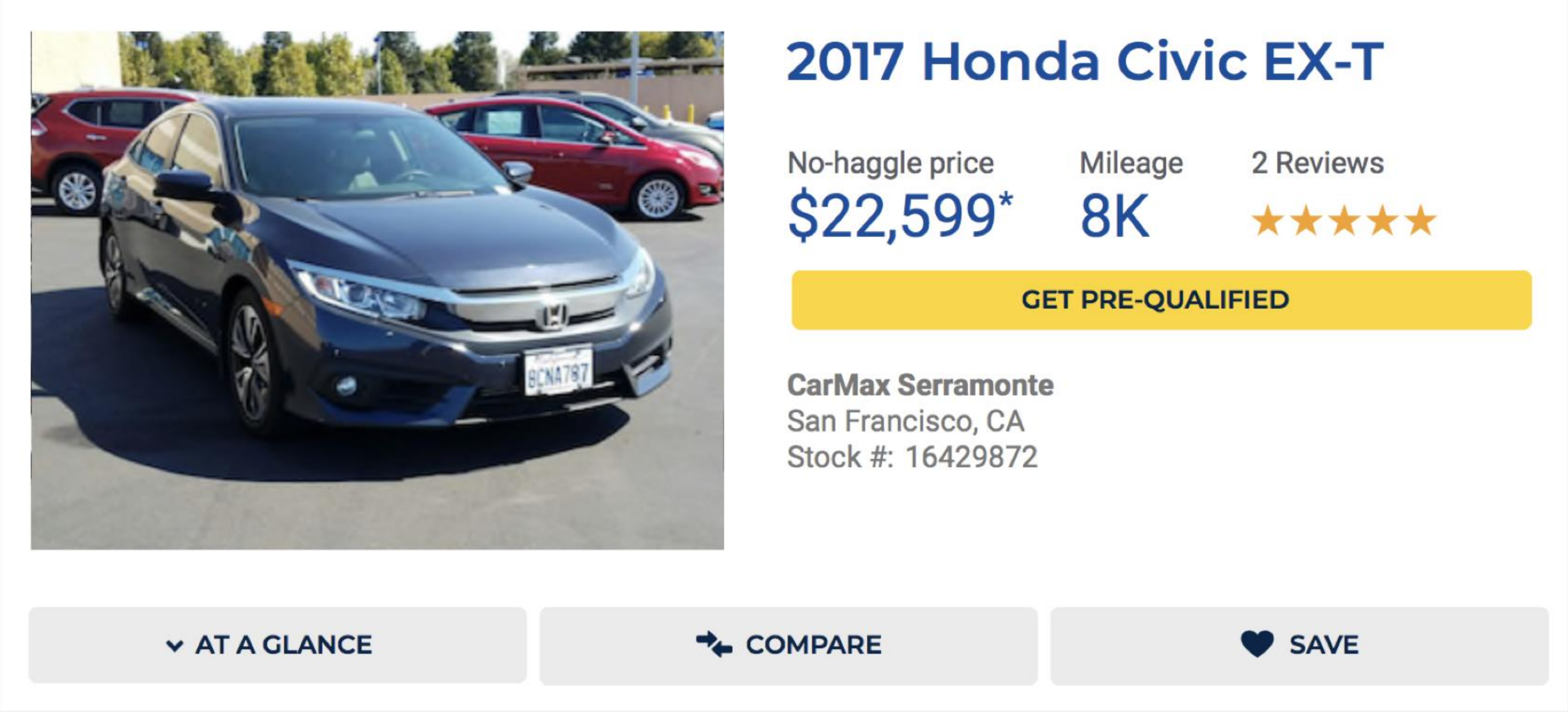
“Any entity that is the subject of organization”

Questions to Ask About Resources

- What are the individual resources?
- What is their granularity?
- How do we identify them?
- Which ones are identical?

Resource Granularity:

Example of Selling Cars



2017 Honda Civic EX-T

No-haggle price
$22,599*

Mileage
8K

2 Reviews
★★★★★

CarMax Serramonte
San Francisco, CA
Stock #: 16429872

Resource Granularity:

Example of Car Parts

Image provenance undetermined, circa 2008

QUESTIONS TO ASK ABOUT RESOURCES

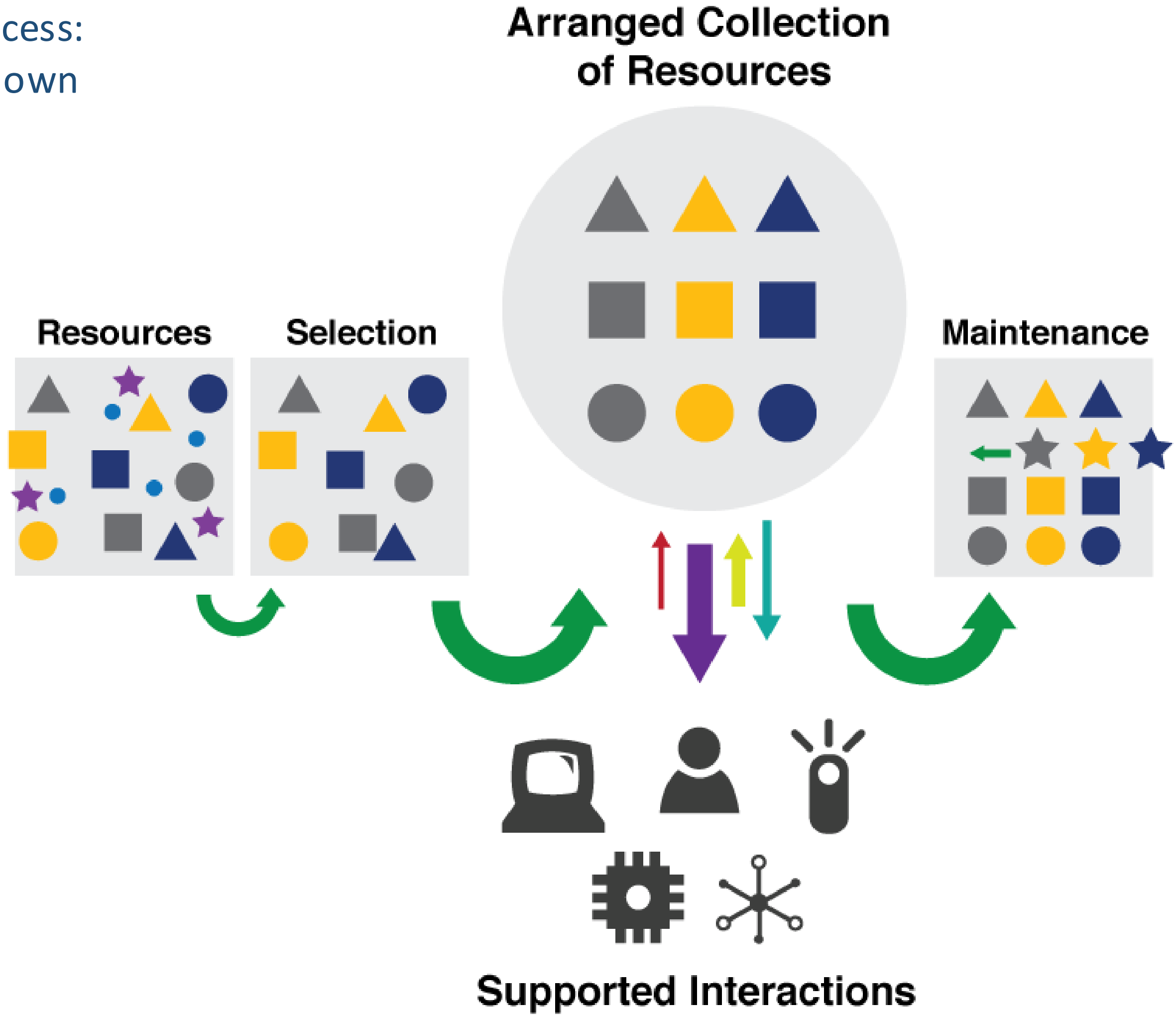
- What are the individual resources?
- What is their granularity?
- How do we identify them?
- Which ones are identical?

We'll come back to these

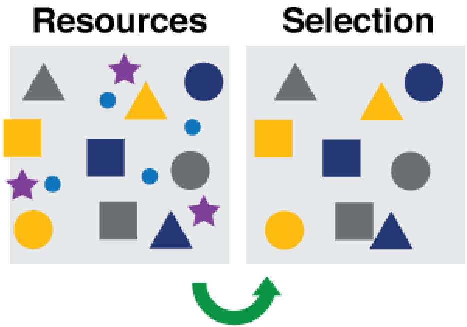
Resource Selection

The whole process:

Let’s break it down



Resource Selection



Netflix

Resource Selection Strategy

SELECTING RESOURCES

- **Selection** is the process by which resources are identified, evaluated, and added to a collection
- Selection is an *intentional* process
- Selection methods and criteria vary across domains

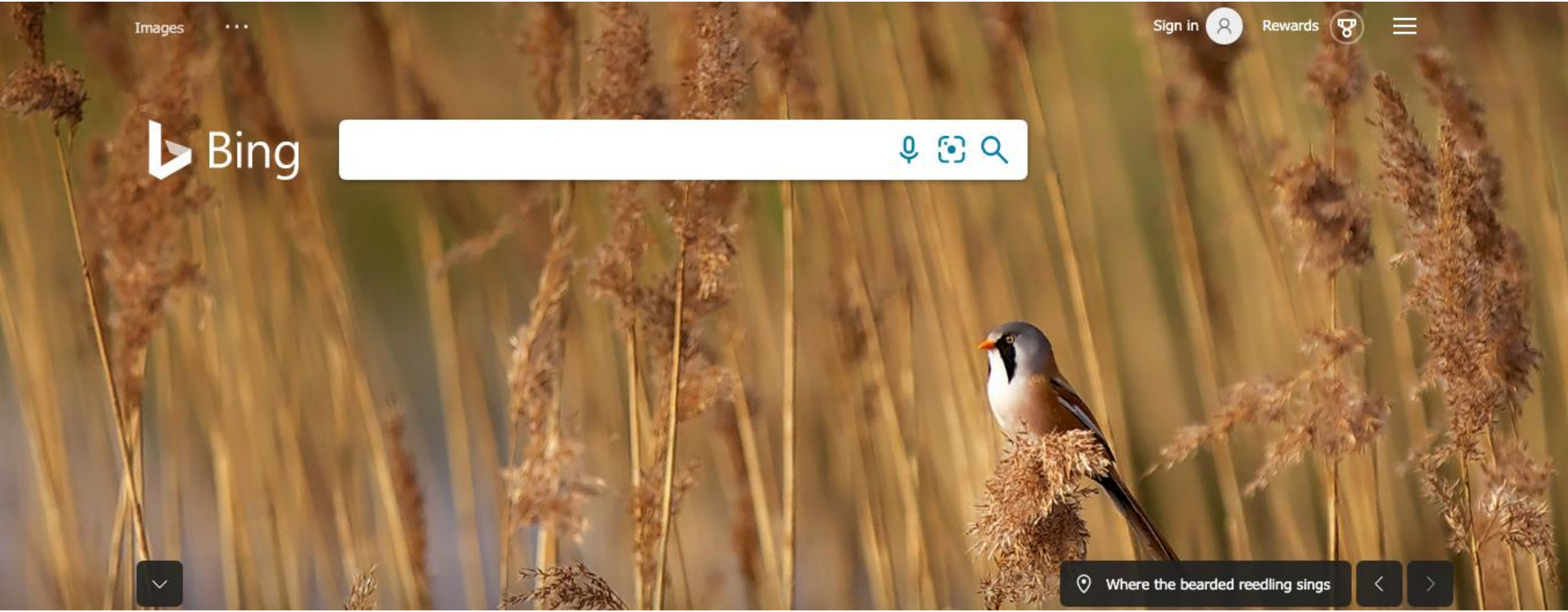
Some Selection Principles

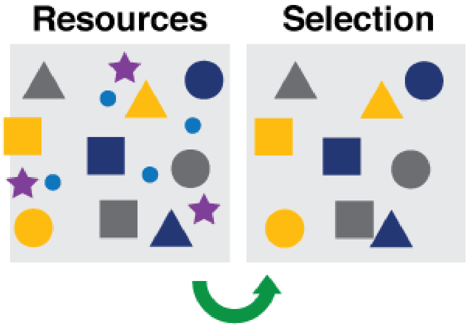
- Utility, usefulness, relevance
- Comprehensiveness
- Value
- Scarcity or uniqueness
- To support social goals (including avoiding bias)
- To establish a reputation or brand

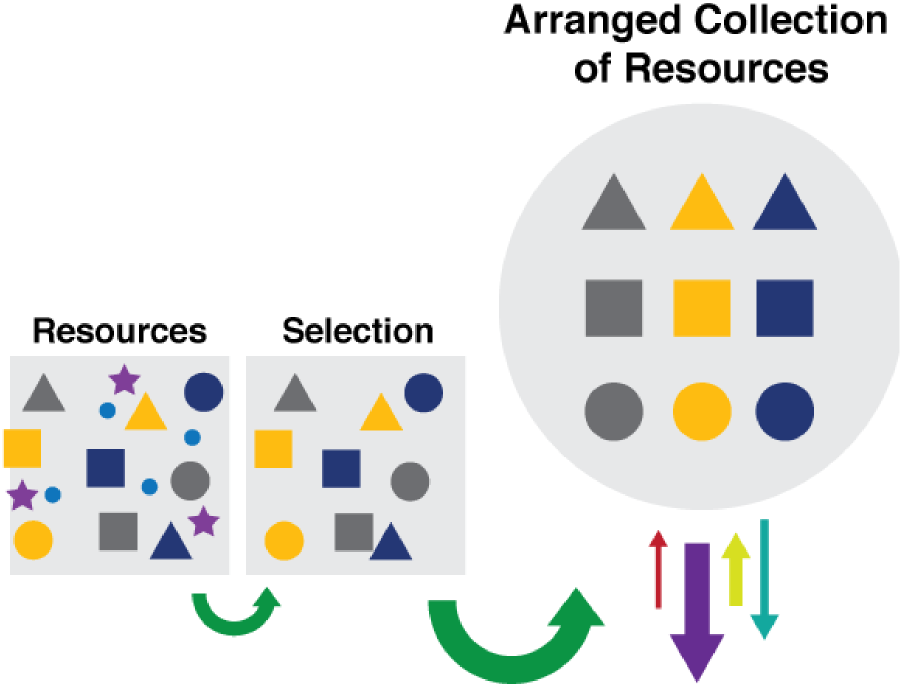
Example Organizing System:
What are the Resources? How Selected

https://unsplash.com/photos/zFlFvx-ygTo Random Sky

Example Organizing System:
What are the Resources? How Selected?




Resources
Selection

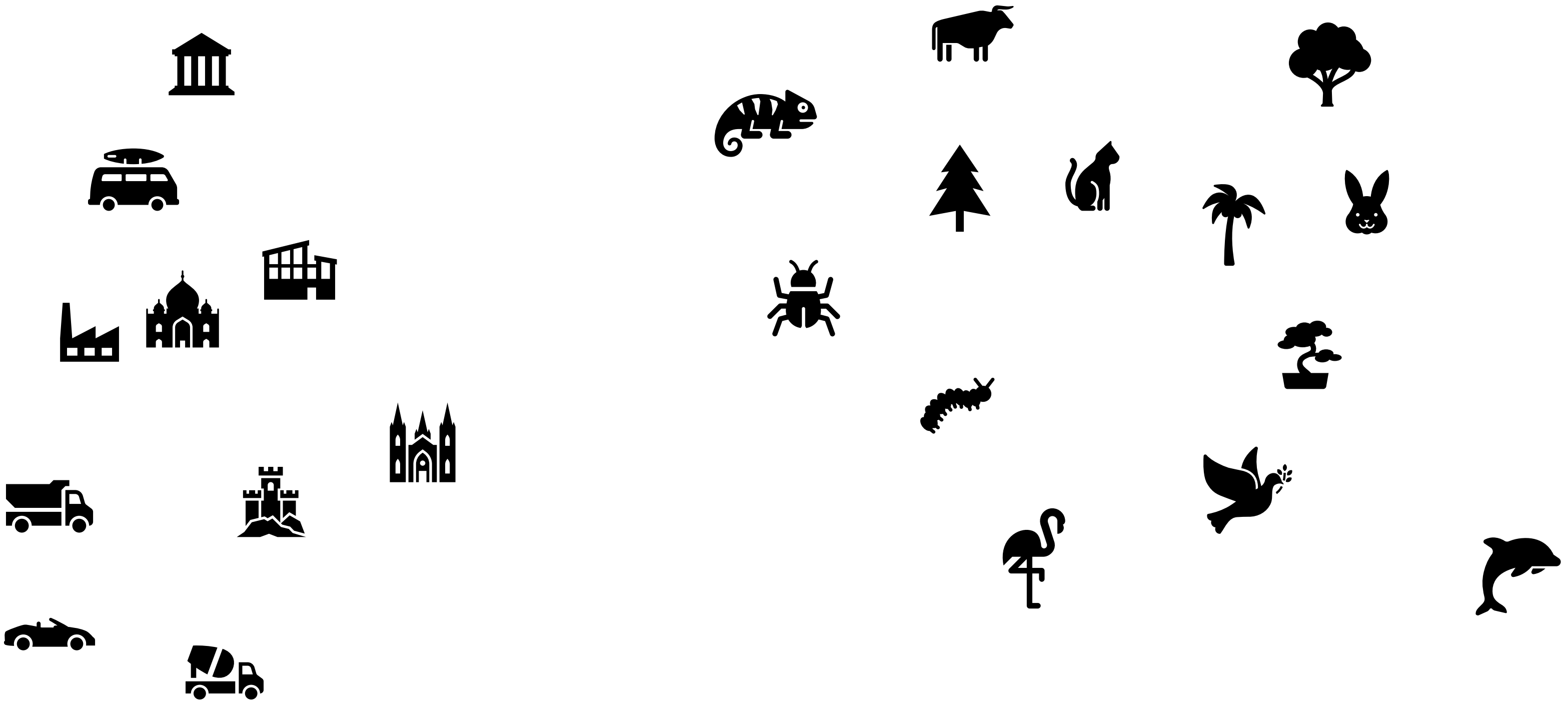


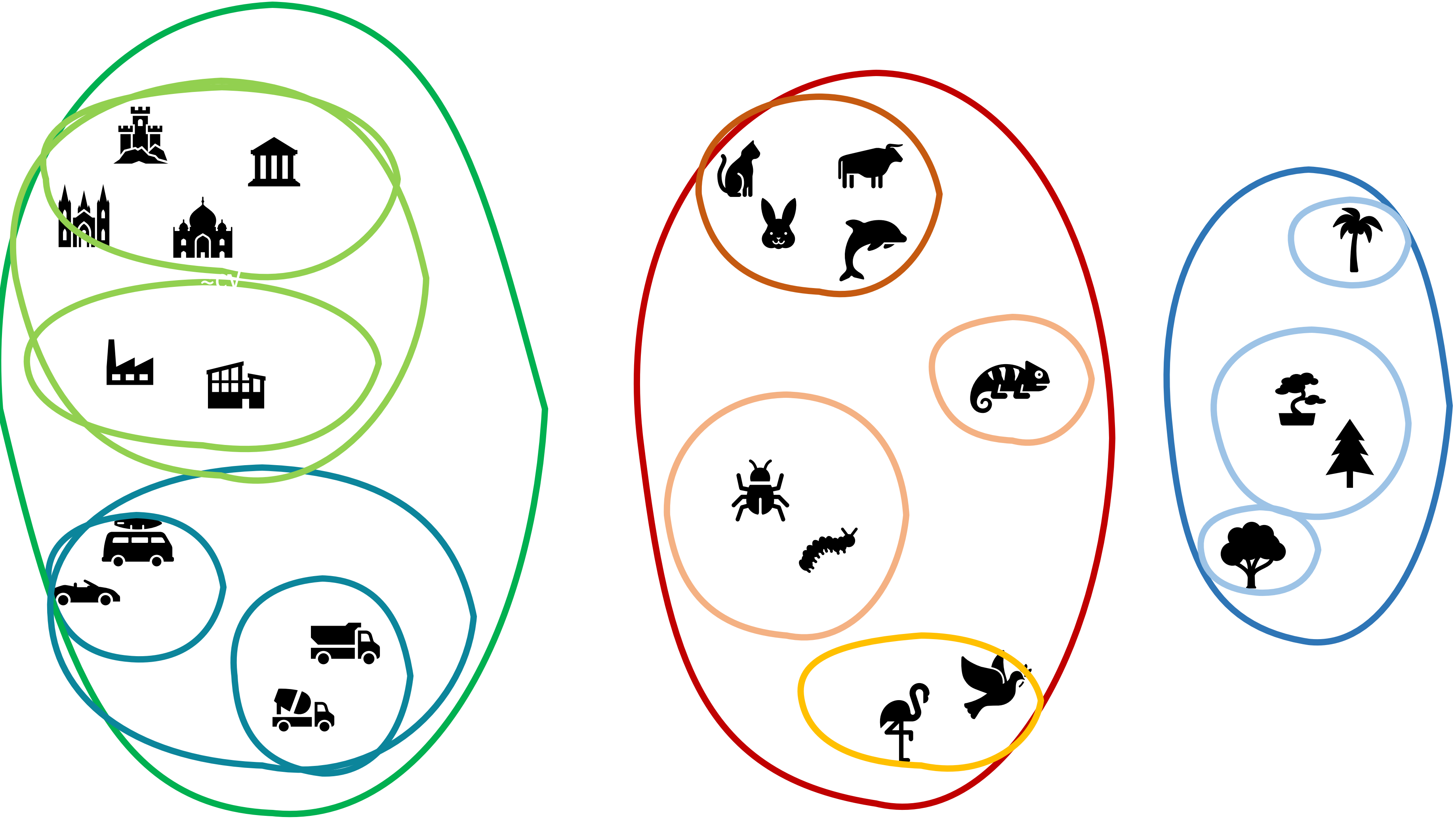
Arranged Collection of Resources

Resources

Selection

What are the Organizing Attributes?





NYTimes

3 RULES
ON HOW TO STORE PAPERS

RULE 1: Categorize every paper down to the last sheet.

RULE 2: Store your papers upright.

RULE 3: Make a pending box.

#joyatwork

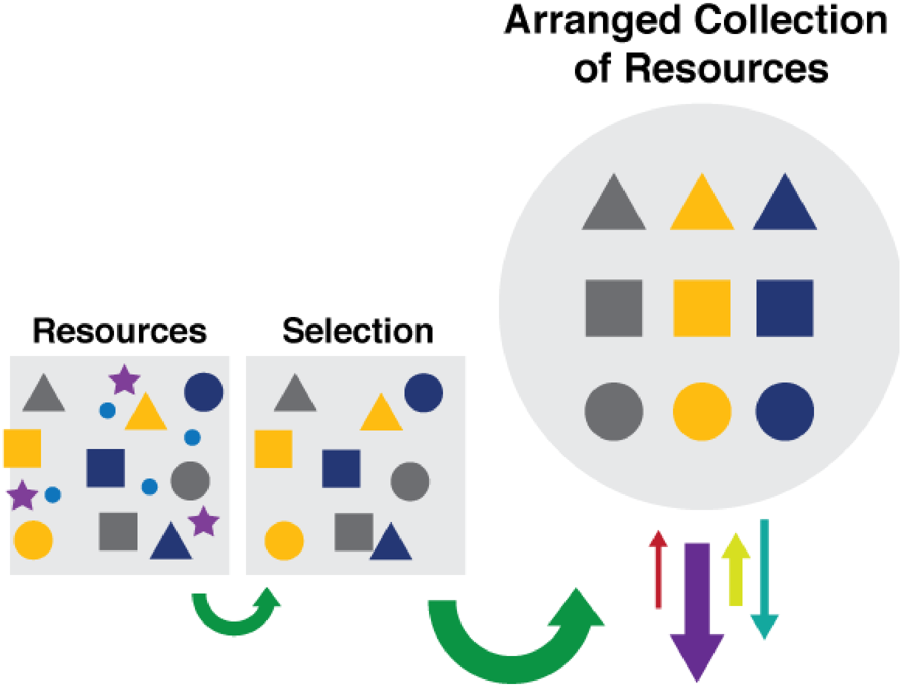
Little, Brown

Collection Organization Strategy

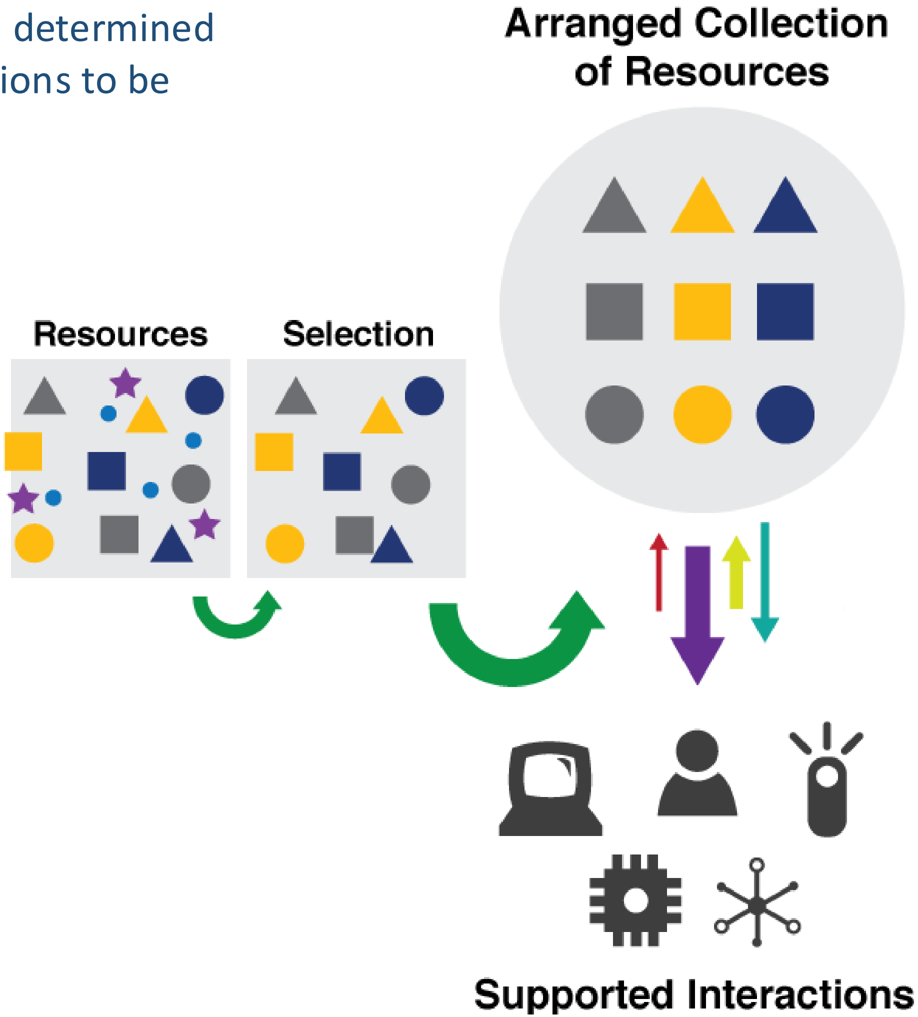
ORGANIZING SYSTEMS

An intentionally arranged collection of resources and the interactions they support

Much more about this in later lectures!


Arranged Collection of Resources
Resources
Selection

Organization is determined by the interactions to be supported



INTERACTIONS

An *interaction* is an action, function, service, or capability that makes use of the resources in a collection or the collection as a whole.

Types of Collection Interactions

Searching / Looking Up

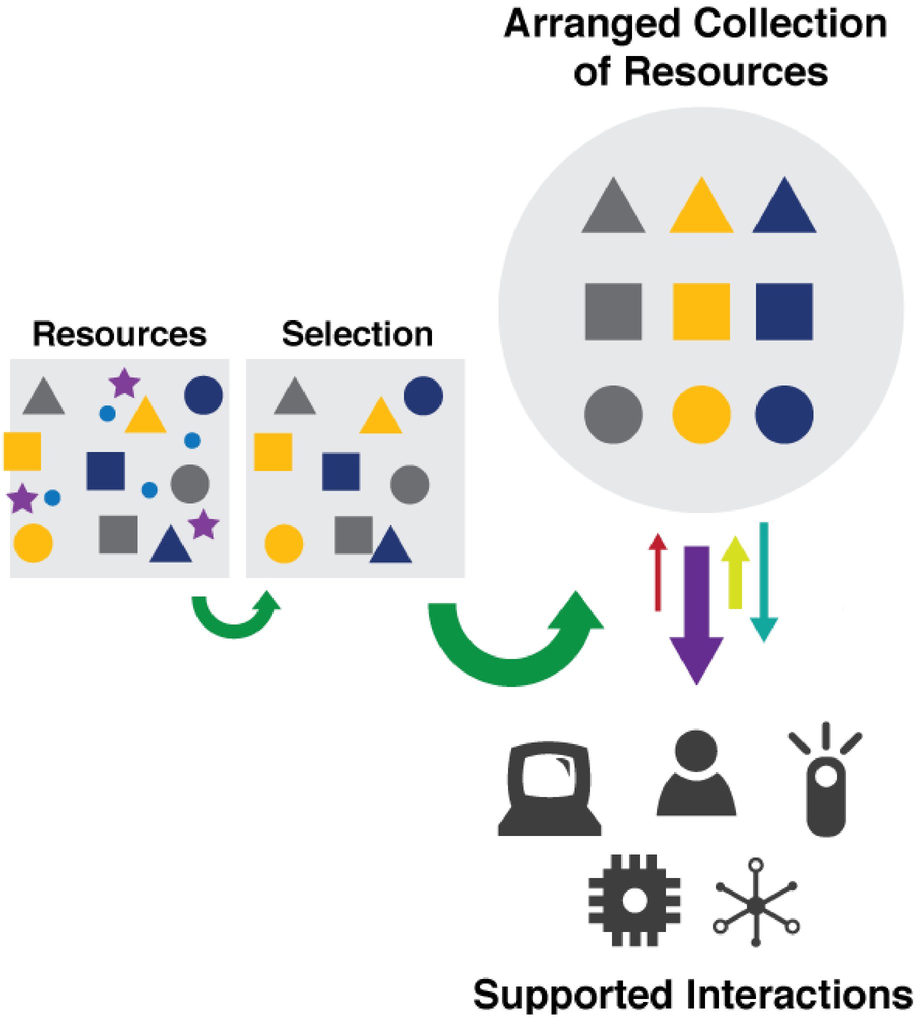
Browsing / Exploring

Retrieving (Accessing) / Returning

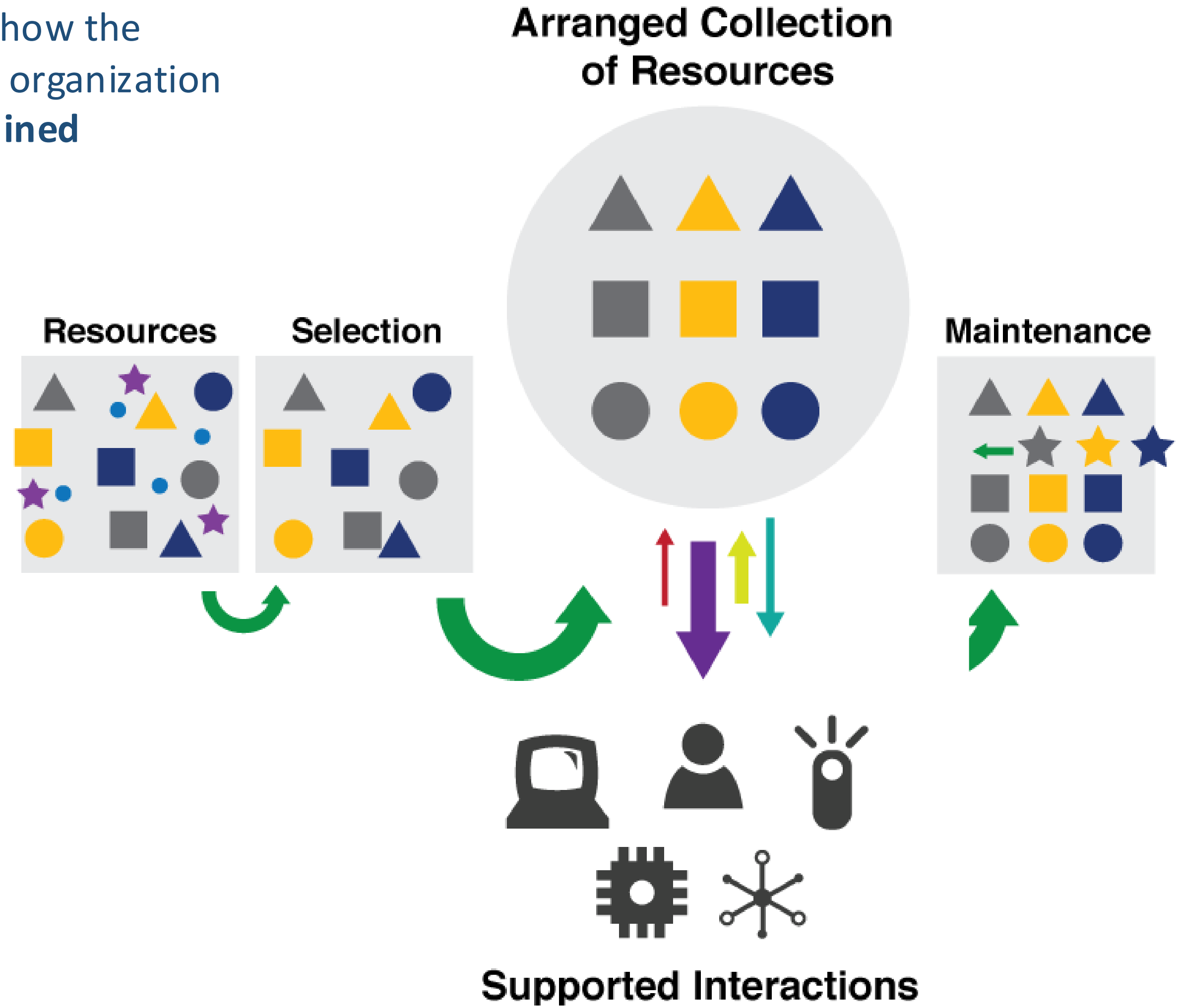
Using (reading, listening to, computing with, measuring…)

EXAMPLE: UC BERKELEY PSYCH LIBRARY

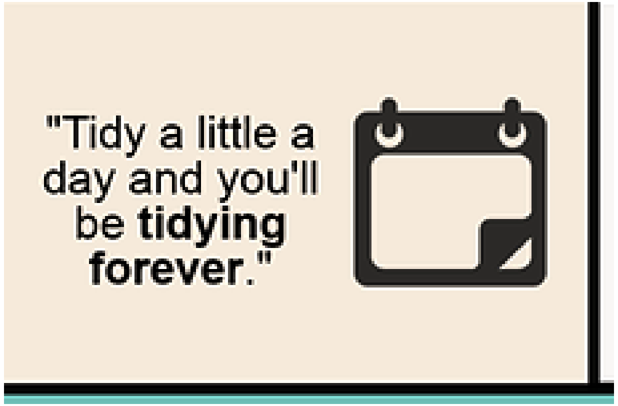
- **Resources**: books, journals, media, …
- **Selection Policy**:
  - *Fill gaps in collection; recently published items of academic merit, faculty requests*
- **Organization**:
  - *Physical arrangement in space, by call number*
  - *Online library catalog*
- **Interactions**
  - *Searching*
  - *Browsing / serendipitous discovery*
  - *Circulation (check out, return both virtual and physical)*
  - *Reading, Listening, …*


Arranged Collection of Resources
Resources
Selection
Supported Interactions

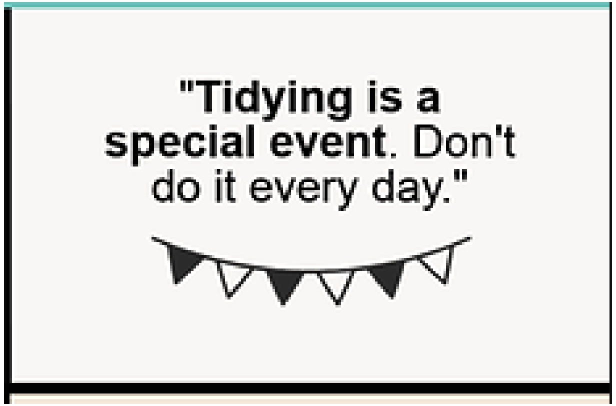
A strategy for how the collection and organization will be **maintained**



Organizing System Maintenance



"Tidy a little a day and you'll be **tidying forever**."



"**Tidying is a special event**. Don't do it every day."

https://www.jdwilliams.co.uk/lifestyle/story/50-marie-kondo-quotes-to-inspire-your-spring-clean

Organizing System Maintenance

- **Storage**: physical and technical aspects of maintaining the collection
- **Preservation**: maintaining resources to prevent damage or deterioration; also protect against obsolescence online
- **Curation / Governance**: the processes by which a resource in a collection is maintained over time, to improve access or to restore or transform its representation or presentation.

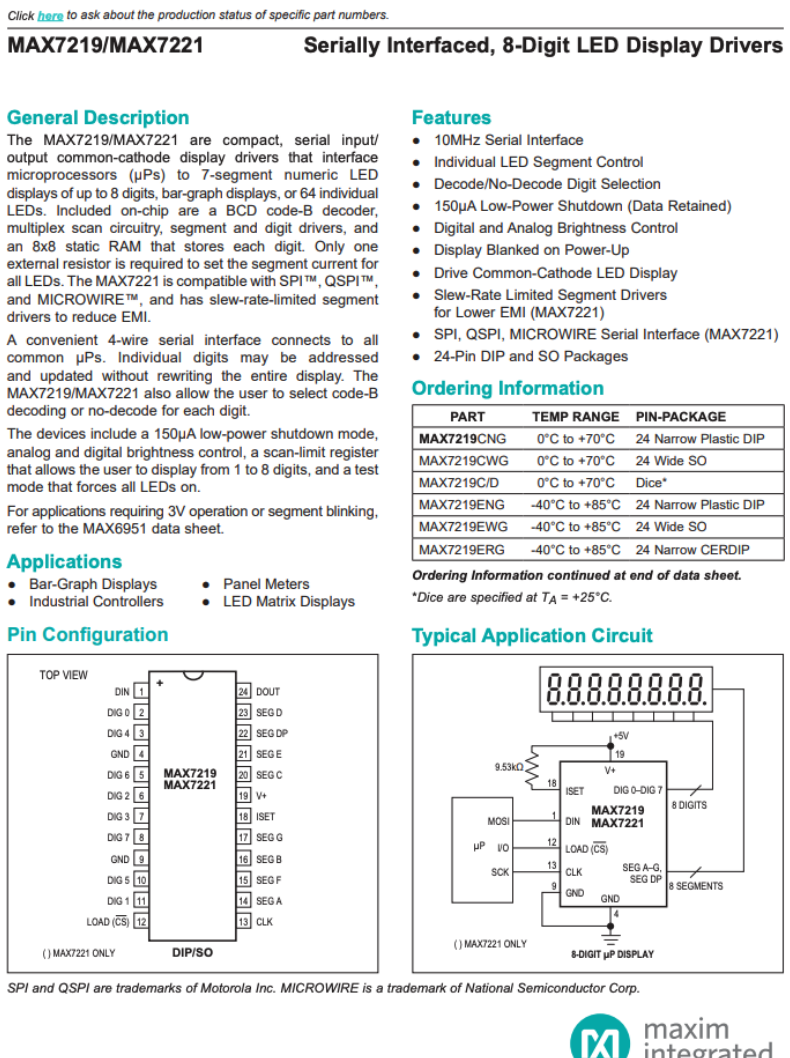
Provenance

The history of ownership of a resource

Ethics in Collection Creation

Datasheets for Datasets

- **Goal**: improve transparency and accountability in machine learning dataset collection
- **Method**: borrow an idea from engineering; describe metadata in terms of data sheets



Click here to ask about the production status of specific part numbers.

MAX7219/MAX7221 **Serially Interfaced, 8-Digit LED Display Drivers**

General Description

The MAX7219/MAX7221 are compact, serial input/output common-cathode display drivers that interface microprocessors (µPs) to 7-segment numeric LED displays of up to 8 digits, bar-graph displays, or 64 individual LEDs. Included on-chip are a BCD code-B decoder, multiplex scan circuitry, segment and digit drivers, and an 8x8 static RAM that stores each digit. Only one external resistor is required to set the segment current for all LEDs. The MAX7221 is compatible with SPI™, QSPI™, and MICROWIRE™, and has slew-rate-limited segment drivers to reduce EMI.

A convenient 4-wire serial interface connects to all common µPs. Individual digits may be addressed and updated without rewriting the entire display. The MAX7219/MAX7221 also allow the user to select code-B decoding or no-decode for each digit.

The devices include a 150µA low-power shutdown mode, analog and digital brightness control, a scan-limit register that allows the user to display from 1 to 8 digits, and a test mode that forces all LEDs on.

For applications requiring 3V operation or segment blinking, refer to the MAX6951 data sheet.

Applications

- Bar-Graph Displays
- Industrial Controllers
- Panel Meters
- LED Matrix Displays

Pin Configuration

Features

- 10MHz Serial Interface
- Individual LED Segment Control
- Decode/No-Decode Digit Selection
- 150µA Low-Power Shutdown (Data Retained)
- Digital and Analog Brightness Control
- Display Blanked on Power-Up
- Drive Common-Cathode LED Display
- Slew-Rate Limited Segment Drivers for Lower EMI (MAX7221)
- SPI, QSPI, MICROWIRE Serial Interface (MAX7221)
- 24-Pin DIP and SO Packages

Ordering Information

PART	TEMP RANGE	PIN-PACKAGE
MAX7219CNG	0°C to +70°C	24 Narrow Plastic DIP
MAX7219CWG	0°C to +70°C	24 Wide SO
MAX7219C/D	0°C to +70°C	Dice*
MAX7219ENG	-40°C to +85°C	24 Narrow Plastic DIP
MAX7219EWG	-40°C to +85°C	24 Wide SO
MAX7219ERG	-40°C to +85°C	24 Narrow CERDIP

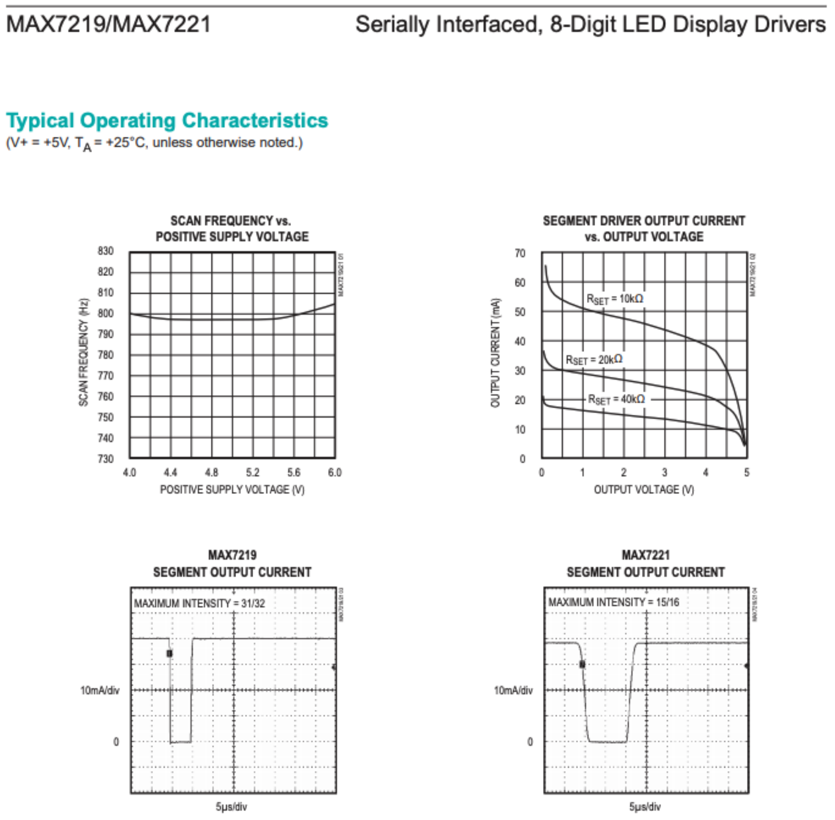
Ordering Information continued at end of data sheet.

**Dice are specified at T_A = +25°C.*

Typical Application Circuit

SPI and QSPI are trademarks of Motorola Inc. MICROWIRE is a trademark of National Semiconductor Corp.

maxim integrated



MAX7219/MAX7221 Serially Interfaced, 8-Digit LED Display Drivers

Typical Operating Characteristics

($V+$ = +5V, T_A = +25°C, unless otherwise noted.)

https://datasheets.maximintegrated.com/en/ds/MAX7219-MAX7221.pdf

DATASHEETS FOR DATASETS

Provides a series of questions to ask when creating or using a dataset. These include:

Motivations: includes funding, authors, what tasks is the dataset intended to be used for.

Composition: metadata, whether the dataset contains sensitive information.

Collection Process: sources, including human sources, any known errors?

Processing: including details on computational processing

Distribution: how, to whom, and any restrictions on distribution.

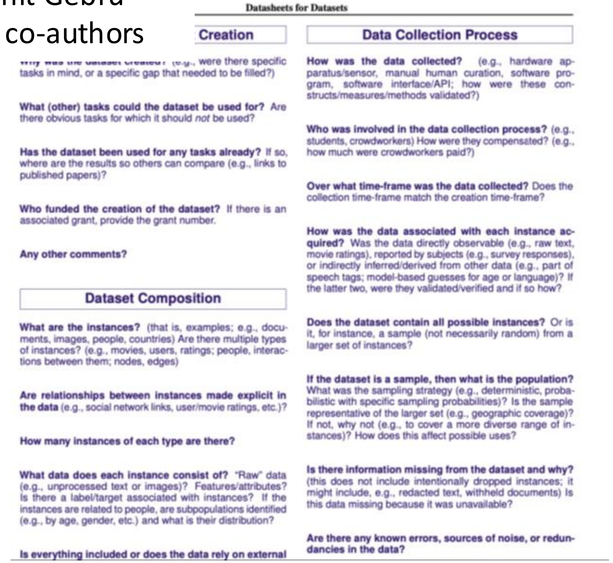
Maintenance: who and how, and if others will be able to build on it

Legal & Ethical Issues: human subjects’ questions: who was involved in the collection process, and, if people are involved, if consent was given for the data to be collected; privacy considerations

What To Take From This Reading (Datasheets for Datasets)

wikipedia

Timnit Gebru
& 6 co-authors



Datasheets for Datasets

Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

Any other comments?

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

How many instances of each type are there?

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Is everything included or does the data rely on external

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

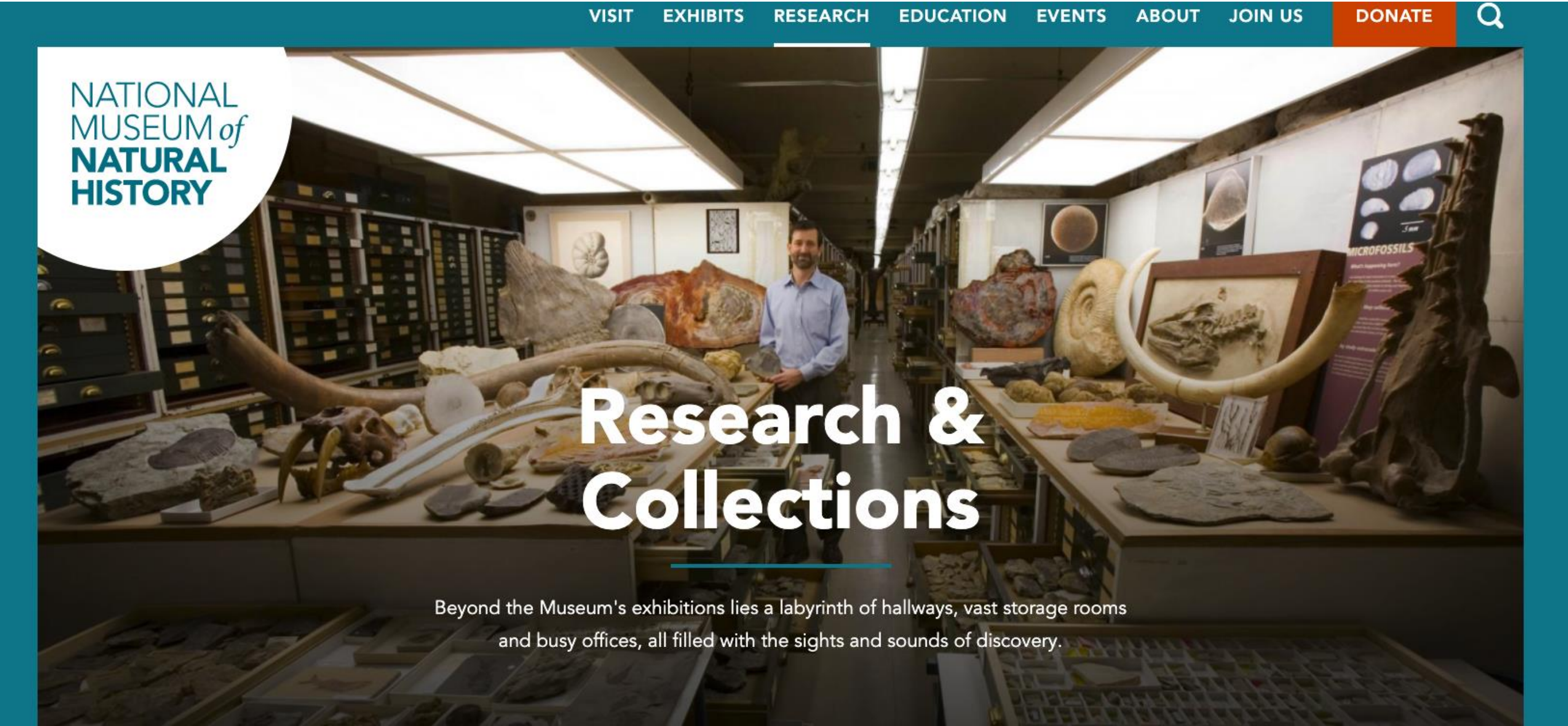
If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

Are there any known errors, sources of noise, or redundancies in the data?

- Datasets for machine learning are a kind of collection
- Reflect on the decision process behind creating, distributing, and maintaining a dataset
- Consider potential social harms that can result from non-reflective selection
- What are the questions that should be asked in creating a dataset? Do they differ from the Glushko reading?

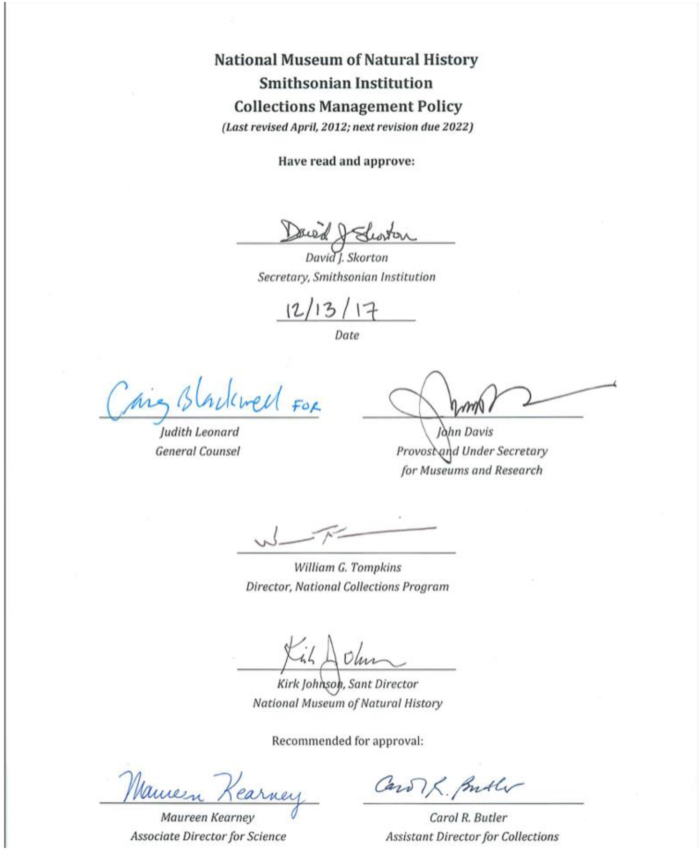
Example:
Smithsonian Natural History Collection



Research & Collections

Beyond the Museum's exhibitions lies a labyrinth of hallways, vast storage rooms and busy offices, all filled with the sights and sounds of discovery.

Smithsonian Collections Management Policy



National Museum of Natural History
Smithsonian Institution
Collections Management Policy
(Last revised April, 2012; next revision due 2022)

Have read and approve:

David J Skorton
David J. Skorton
Secretary, Smithsonian Institution

12/13/17
Date

Craig Blackwell FOR
Judith Leonard
General Counsel

John Davis
Provost and Under Secretary
for Museums and Research

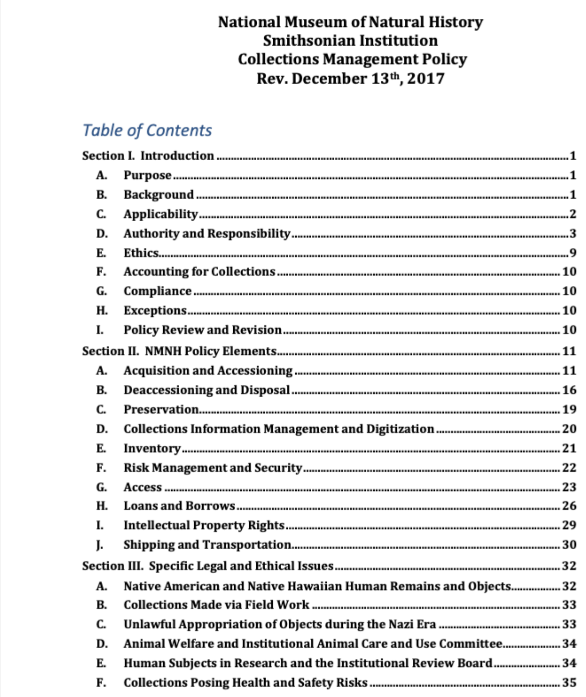
William G. Tompkins
Director, National Collections Program

Kirk Johnson
Kirk Johnson, Sant Director
National Museum of Natural History

Recommended for approval:

Maureen Kearney
Maureen Kearney
Associate Director for Science

Carol R. Butler
Carol R. Butler
Assistant Director for Collections



National Museum of Natural History
Smithsonian Institution
Collections Management Policy
Rev. December 13th, 2017

Table of Contents

https://naturalhistory.si.edu/research/nmnh-collections/museum-collections-policies

Smithsonian Collections Management Policy (Selected Excerpts)

“This document sets forth polices and guidance for the acquisition, management, use, and disposal of the collections of the National Museum of Natural History (NMNH)”

“The NMNH’s collections activities are conducted in compliance with The Smithsonian Institution Statement of Values and Code of Ethics; SD 103: Smithsonian Institution Standards of Conduct, the Advisory Board Ethics Statement; SD 600: Collections Management; and the SD 600 Implementation Manual.”

https://naturalhistory.si.edu/research/nmnh-collections/museum-collections-policies

Smithsonian Collections Management Policy (Selected Excerpts)

“Staff will consider and evaluate the concerns of indigenous source communities regarding collections items, recordings, information, collecting activities and use.”

“The Smithsonian repudiates the illicit traffic in objects and specimens that contribute to the despoliation of museums, monuments, environments, sites and species resulting in irreparable loss to science and humanity. Items that have been stolen, unscientifically gathered or excavated, or unethically acquired should not be made part of Smithsonian collections. “

https://naturalhistory.si.edu/research/nmnh-collections/museum-collections-policies

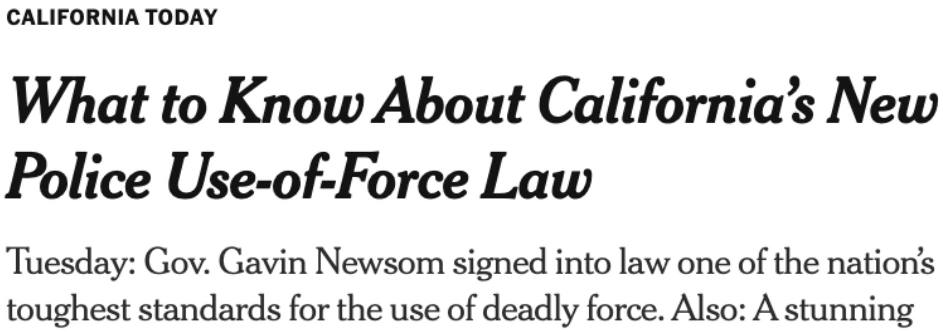
Smithsonian Collections Management Policy (Selected Excerpts)

“The concept of provenance refers to the history of ownership of a collection item. Collecting departments shall exercise due diligence in the acquisition of collections, including making reasonable inquiries into the provenance of collections items under consideration for acquisition consistent with Smithsonian policy.”

“Collections records must show decision-making processes of acquisitions evaluation…”

https://naturalhistory.si.edu/research/nmnh-collections/museum-collections-policies

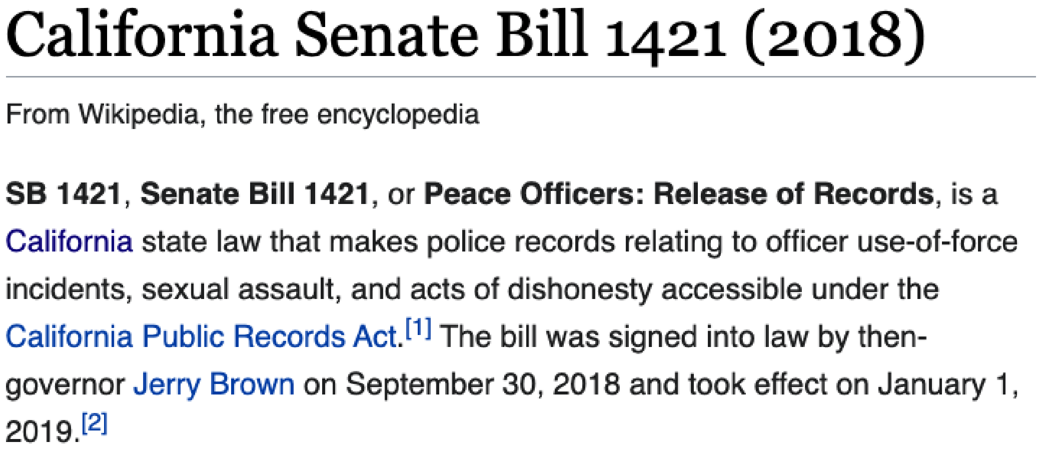
Example Collection Creation: SB1421 Datasets



CALIFORNIA TODAY

What to Know About California's New Police Use-of-Force Law

Tuesday: Gov. Gavin Newsom signed into law one of the nation's toughest standards for the use of deadly force. Also: A stunning



California Senate Bill 1421 (2018)

From Wikipedia, the free encyclopedia

SB 1421, **Senate Bill 1421**, or **Peace Officers: Release of Records**, is a California state law that makes police records relating to officer use-of-force incidents, sexual assault, and acts of dishonesty accessible under the California Public Records Act.[1] The bill was signed into law by then-governor Jerry Brown on September 30, 2018 and took effect on January 1, 2019.[2]

Example: SB1421 Dataset

https://sfpublicdefender.org/copmonitor/

- **Resources:**
- **Selection Policy:**
- **Organization:**
- **Interactions:**