I 202: Information Organization & Retrieval
Fall 2025

Class 10: Structuring Data, Metadata, Databases

Today’s Outline

Structuring Information

Metadata

Intro to Databases

Schemas

This Week’s Assignment

Structuring Information

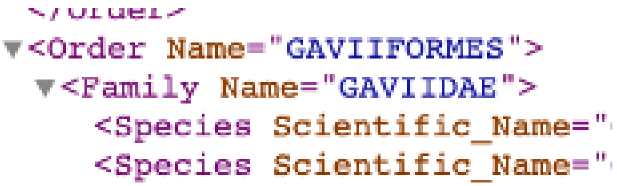
Start with something in the world that needs to be organized

Decide what will be retained and what will be ignored

Develop descriptions for what is being retained

Iterate on those descriptions, comparing them to evaluation criteria

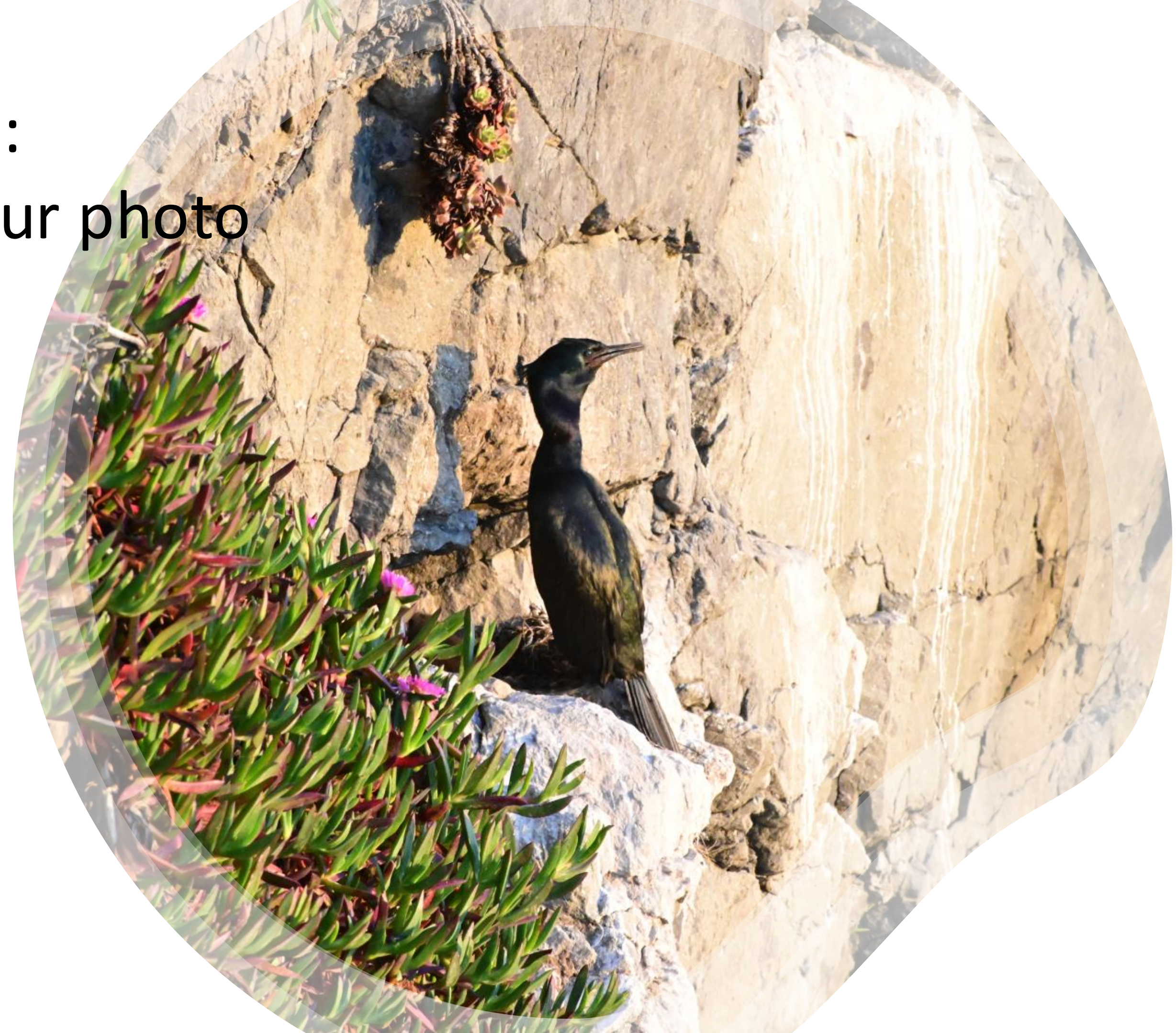
Assign descriptions to the things that are being organized



```
  </Order>
▼<Order Name="GAVIIFORMES">
  ▼<Family Name="GAVIIDAE">
      <Species Scientific_Name="
      <Species Scientific_Name="
```

Purpose:

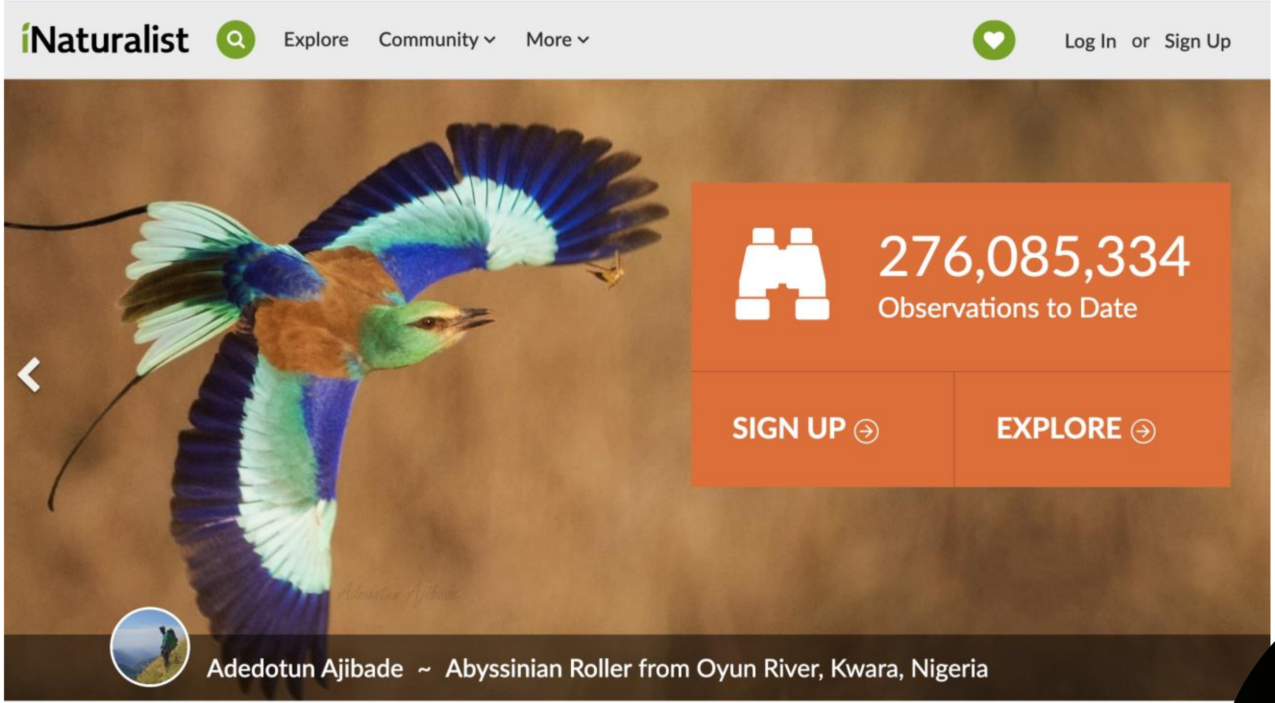
Label your photo



bird

plant

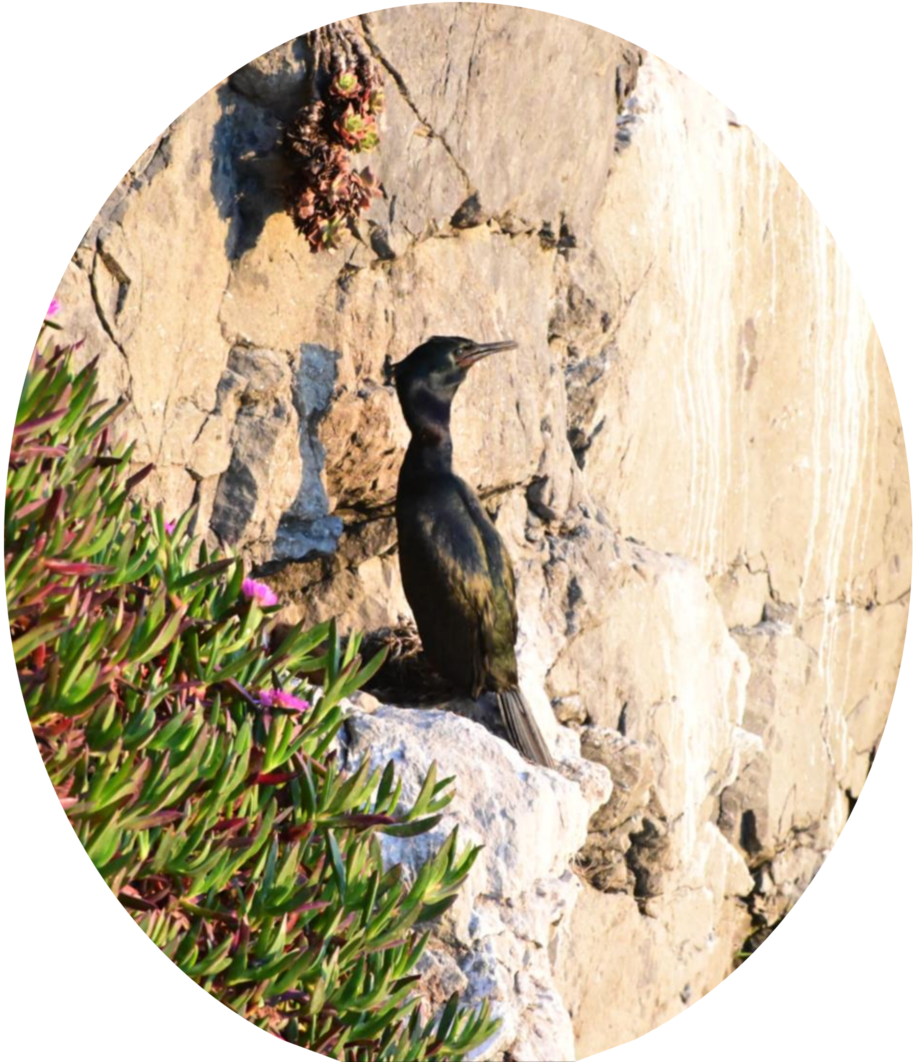
rock



276,085,334
Observations to Date

Adedotun Ajibade ~ Abyssinian Roller from Oyun River, Kwara, Nigeria

iNaturalist founded by Ken-ichi Ueda as a MIMS project
Here speaking at our commencement in 2023

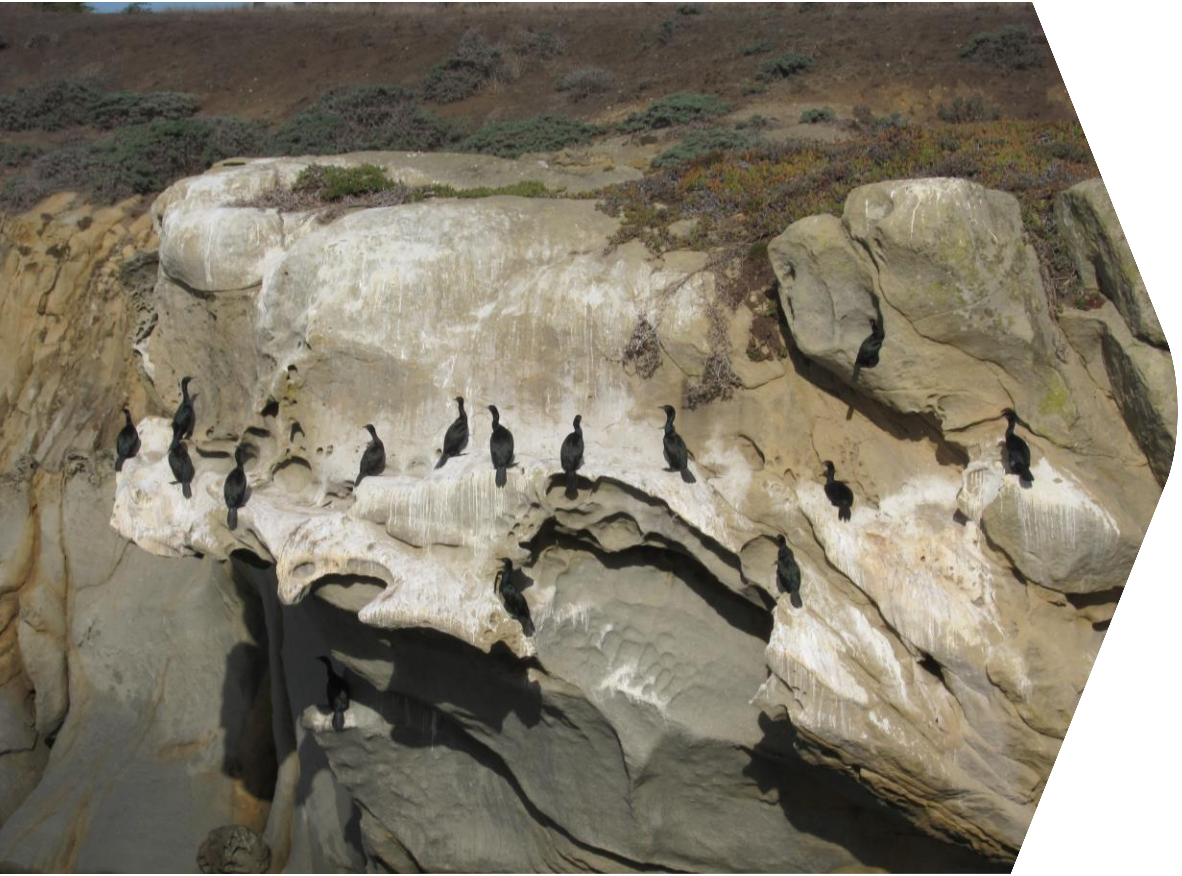


Purpose: Update your iNaturalist profile

Pelagic cormorant,
Breeding adult

Sea fig ice plant

Sandstone



Purpose: International Bird Count

Urile pelagicus
Count: 14

Date and Time:
4/10/2021
10:00 PT

Latitude / Longitude:
38.6910446187637,
-123.4369120201065

Data, Descriptions, Metadata, Metadata Description

Phenomena in the World	Data: What Items are Collected	Data Representation	MetaData	Metadata Standard

What is the Metadata?

Phenomena in the World	Data: What Items are Collected	Data Representation	MetaData	Metadata Standard
	Photos of birds, stored on hard drive	Pixels		

What Is Metadata?

- Data that describes other data / information
- Example:
  - *Information: blog post*
  - *Metadata: language, length, author, date, publisher, keywords, sentiment, …*
- Can be created manually or automatically

Why Do We Need Metadata?

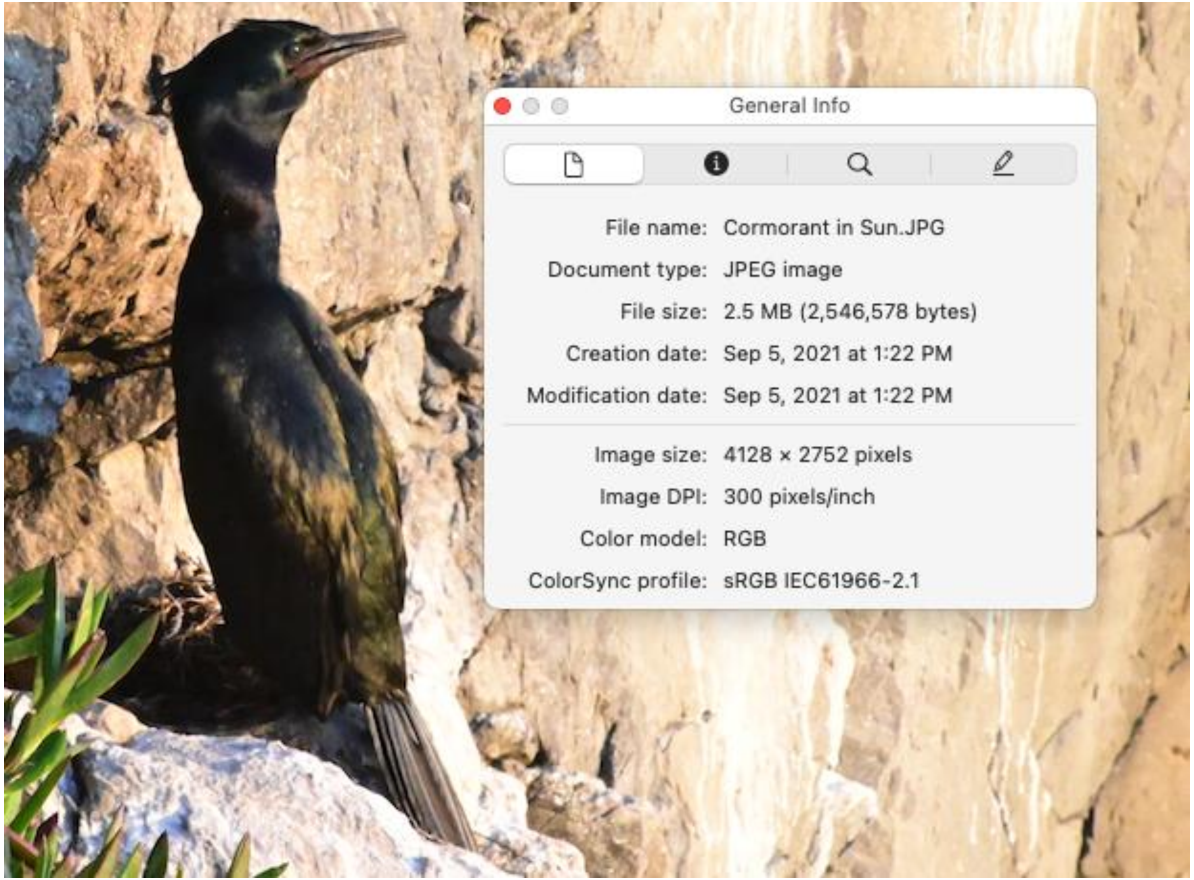
Main reason: **To Organize Collections**

Expanding on this:

- *To make representation of information* ***consistent***
- *To make retrieval / search* ***efficient and effective***
- *To enable* ***sharing and re-use*** *of information*
- *To support* ***auditing*** *(records of changes)*

Data, Descriptions, Metadata, Metadata Description

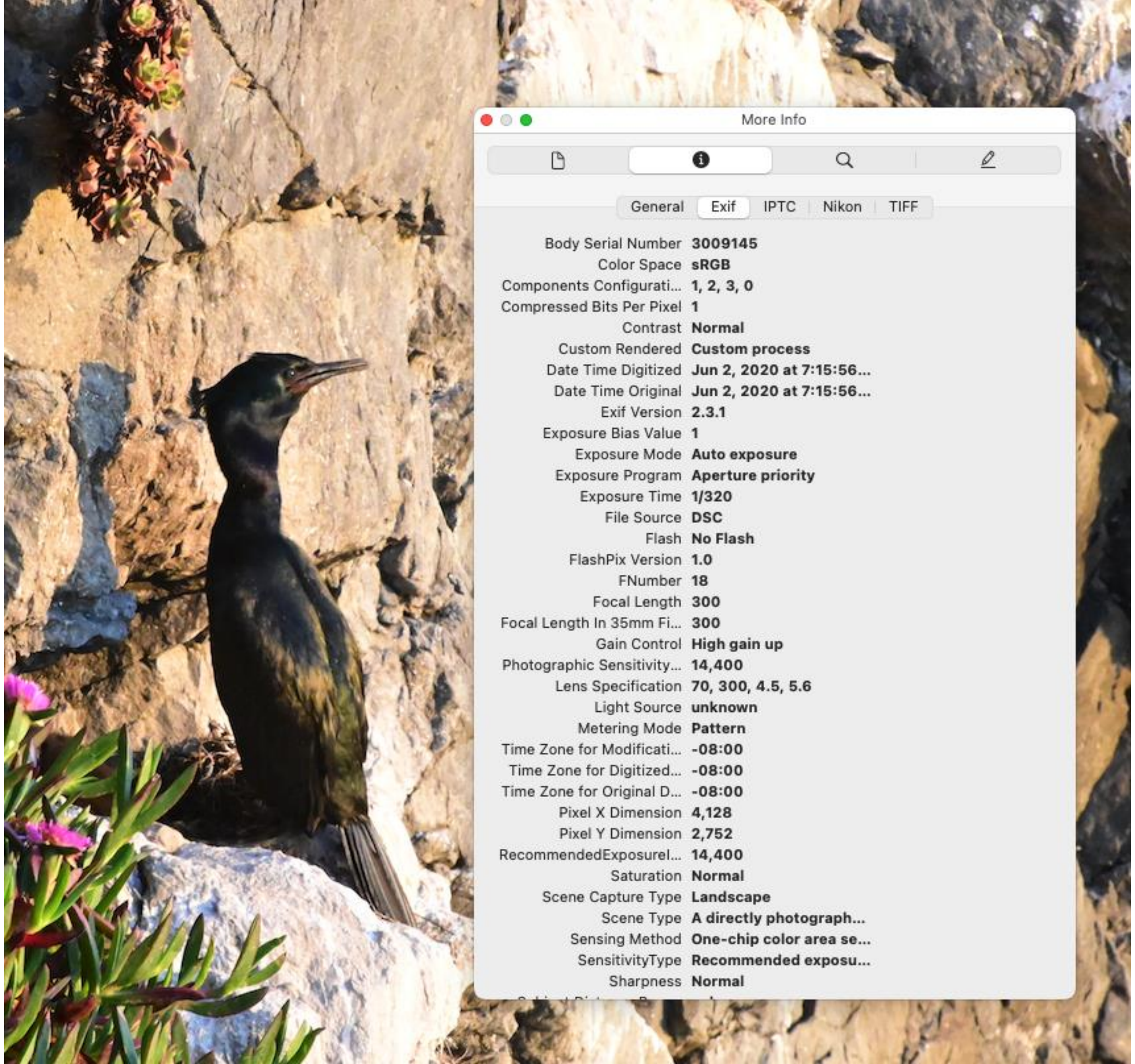
Phenomena in the World	Data: What Items are Collected	Data Representation	MetaData	Metadata Standard
	Photos of birds, stored on hard drive	Pixels	Image size, Image format, Dat/time taken, Lat/Long, Camera settings	EXIF (automatic) IPTC (manual); now XMP is a new standard



File name: Cormorant in Sun.JPG
Document type: JPEG image
File size: 2.5 MB (2,546,578 bytes)
Creation date: Sep 5, 2021 at 1:22 PM
Modification date: Sep 5, 2021 at 1:22 PM

Image size: 4128 × 2752 pixels
Image DPI: 300 pixels/inch
Color model: RGB
ColorSync profile: sRGB IEC61966-2.1

You can see the image metadata by right-clicking on the photo in some apps

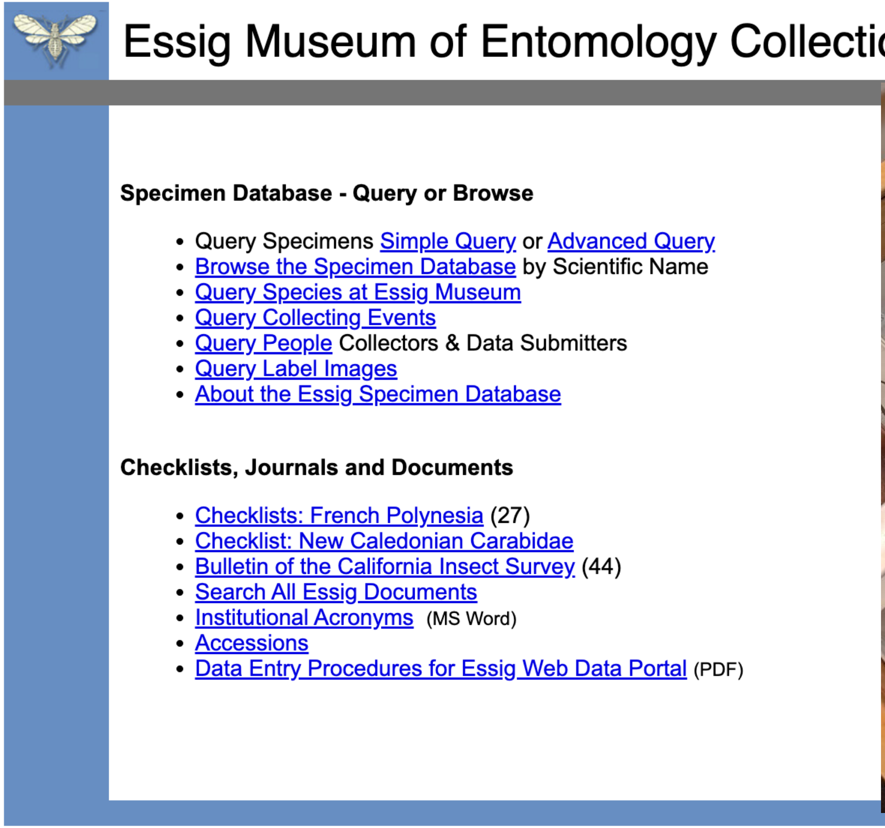


```
           Body Serial Number  3009145
                  Color Space  sRGB
     Components Configurati...  1, 2, 3, 0
   Compressed Bits Per Pixel  1
                     Contrast  Normal
              Custom Rendered  Custom process
          Date Time Digitized  Jun 2, 2020 at 7:15:56...
           Date Time Original  Jun 2, 2020 at 7:15:56...
                 Exif Version  2.3.1
          Exposure Bias Value  1
                Exposure Mode  Auto exposure
             Exposure Program  Aperture priority
                Exposure Time  1/320
                  File Source  DSC
                        Flash  No Flash
              FlashPix Version  1.0
                      FNumber  18
                 Focal Length  300
    Focal Length In 35mm Fi...  300
                 Gain Control  High gain up
   Photographic Sensitivity...  14,400
            Lens Specification  70, 300, 4.5, 5.6
                  Light Source  unknown
                Metering Mode  Pattern
    Time Zone for Modificati...  -08:00
     Time Zone for Digitized...  -08:00
    Time Zone for Original D...  -08:00
             Pixel X Dimension  4,128
             Pixel Y Dimension  2,752
    RecommendedExposureI...  14,400
                    Saturation  Normal
            Scene Capture Type  Landscape
                    Scene Type  A directly photograph...
                Sensing Method  One-chip color area se...
               SensitivityType  Recommended exposu...
                     Sharpness  Normal
```

Example: A Specimens Database

https://essig.berkeley.edu/

Example: A Specimens Database



Essig Museum of Entomology Collections

Specimen Database - Query or Browse

- Query Specimens Simple Query or Advanced Query
- Browse the Specimen Database by Scientific Name
- Query Species at Essig Museum
- Query Collecting Events
- Query People Collectors & Data Submitters
- Query Label Images
- About the Essig Specimen Database

Checklists, Journals and Documents

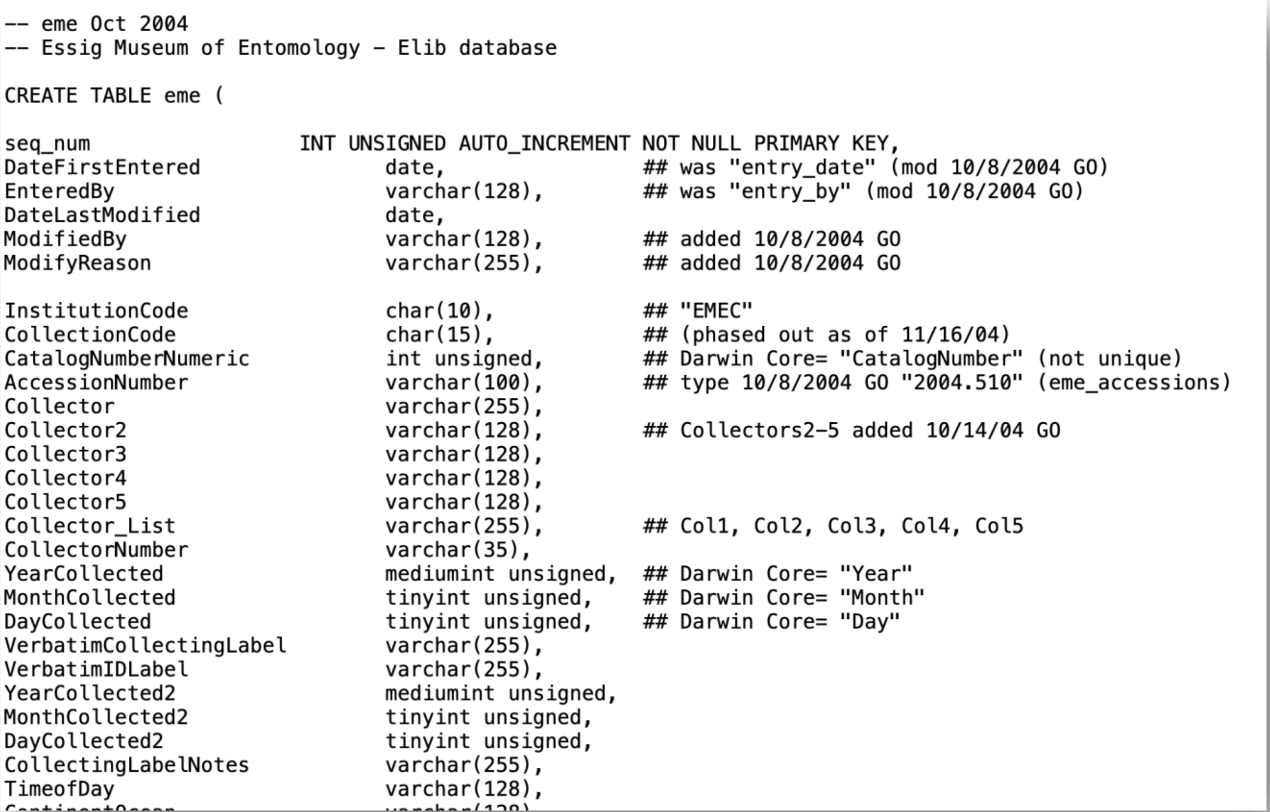
- Checklists: French Polynesia (27)
- Checklist: New Caledonian Carabidae
- Bulletin of the California Insect Survey (44)
- Search All Essig Documents
- Institutional Acronyms (MS Word)
- Accessions
- Data Entry Procedures for Essig Web Data Portal (PDF)

The Essig Database is a project of BNHM University of California, Berkeley Comments or questions? essig.museum@gmail.com

Data, Descriptions, Metadata, Metadata Description

Phenomena in the World	Data: What Items are Collected	Data Representation	MetaData	Metadata Description Language
	Photos of birds, stored on hard drive	Pixels	Image size, Image format, Dat/time taken, Lat/Long, Camera settings	EXIF (automatic) IPTC (manual); now XMP is a new standard
	Real bird specimen	Physical bird	Measurements, type, date, instrument, etc	Specimens database Schema (e.g. EME)

Part of a Specimens Database **Schema**



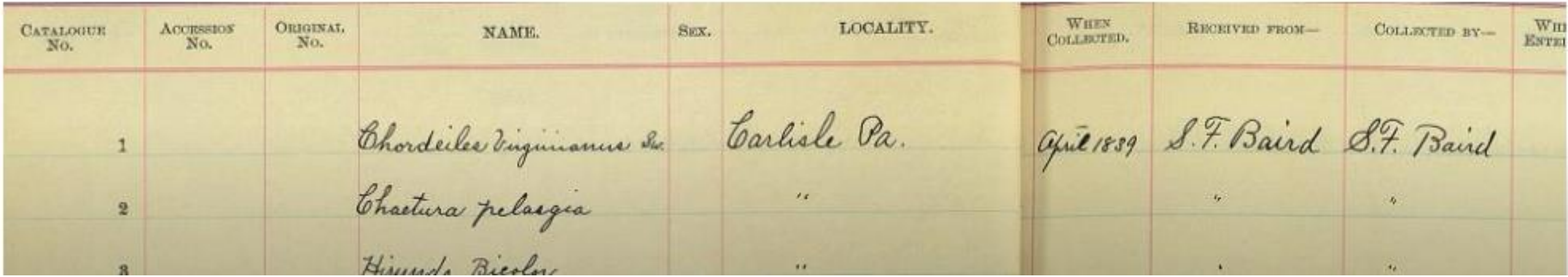
```
-- eme Oct 2004
-- Essig Museum of Entomology - Elib database

CREATE TABLE eme (

seq_num                 INT UNSIGNED AUTO_INCREMENT NOT NULL PRIMARY KEY,
DateFirstEntered               date,                ## was "entry_date" (mod 10/8/2004 GO)
EnteredBy                      varchar(128),        ## was "entry_by" (mod 10/8/2004 GO)
DateLastModified               date,
ModifiedBy                     varchar(128),        ## added 10/8/2004 GO
ModifyReason                   varchar(255),        ## added 10/8/2004 GO

InstitutionCode                char(10),            ## "EMEC"
CollectionCode                 char(15),            ## (phased out as of 11/16/04)
CatalogNumberNumeric           int unsigned,        ## Darwin Core= "CatalogNumber" (not unique)
AccessionNumber                varchar(100),        ## type 10/8/2004 GO "2004.510" (eme_accessions)
Collector                      varchar(255),
Collector2                     varchar(128),        ## Collectors2-5 added 10/14/04 GO
Collector3                     varchar(128),
Collector4                     varchar(128),
Collector5                     varchar(128),
Collector_List                 varchar(255),        ## Col1, Col2, Col3, Col4, Col5
CollectorNumber                varchar(35),
YearCollected                  mediumint unsigned,  ## Darwin Core= "Year"
MonthCollected                 tinyint unsigned,    ## Darwin Core= "Month"
DayCollected                   tinyint unsigned,    ## Darwin Core= "Day"
VerbatimCollectingLabel        varchar(255),
VerbatimIDLabel                varchar(255),
YearCollected2                 mediumint unsigned,
MonthCollected2                tinyint unsigned,
DayCollected2                  tinyint unsigned,
CollectingLabelNotes           varchar(255),
TimeofDay                      varchar(128),
ContinentOcean                 varchar(128),
```

https://essigdb.berkeley.edu/schema.txt

Another Bird Specimens Database

Smithsonian
National Museum of Natural History
Visit | Exhibits | Research | Education | Events | About | Join Us |
NMNH Home | NMNH Research & Collections | Vertebrate Zoology | Birds | Collections

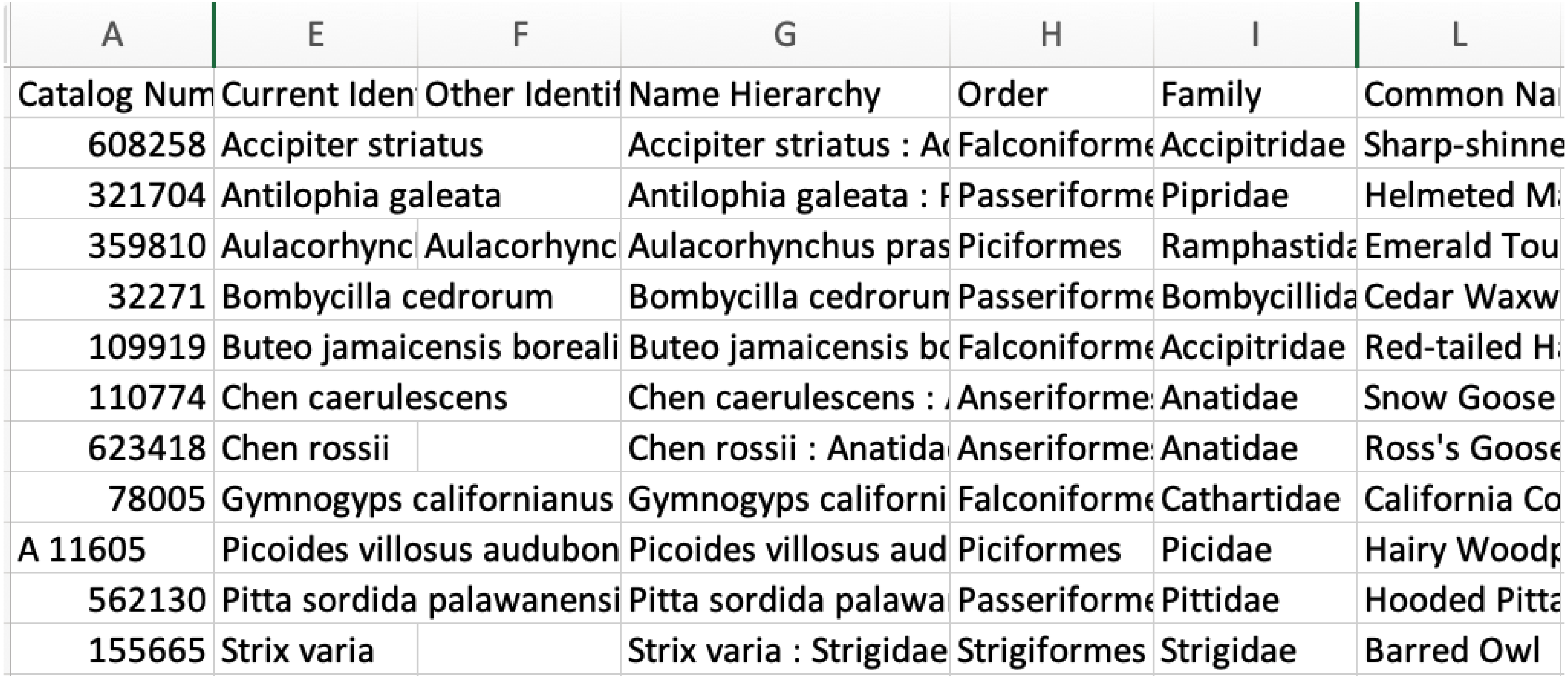
Search the Division of Birds Collections



Catalogue No.	Accession No.	Original No.	Name.	Sex.	Locality.	When Collected.	Received from—	Collected by—	Wh Ente
1			Chordeiles Virginianus Sw.		Carlisle Pa.	April 1839	S.F. Baird	S.F. Baird	
2			Chaetura pelasgia		"		"	"	
3			Hirundo Bicolor		"		"	"	

Bird Specimens Database

(selected attributes)

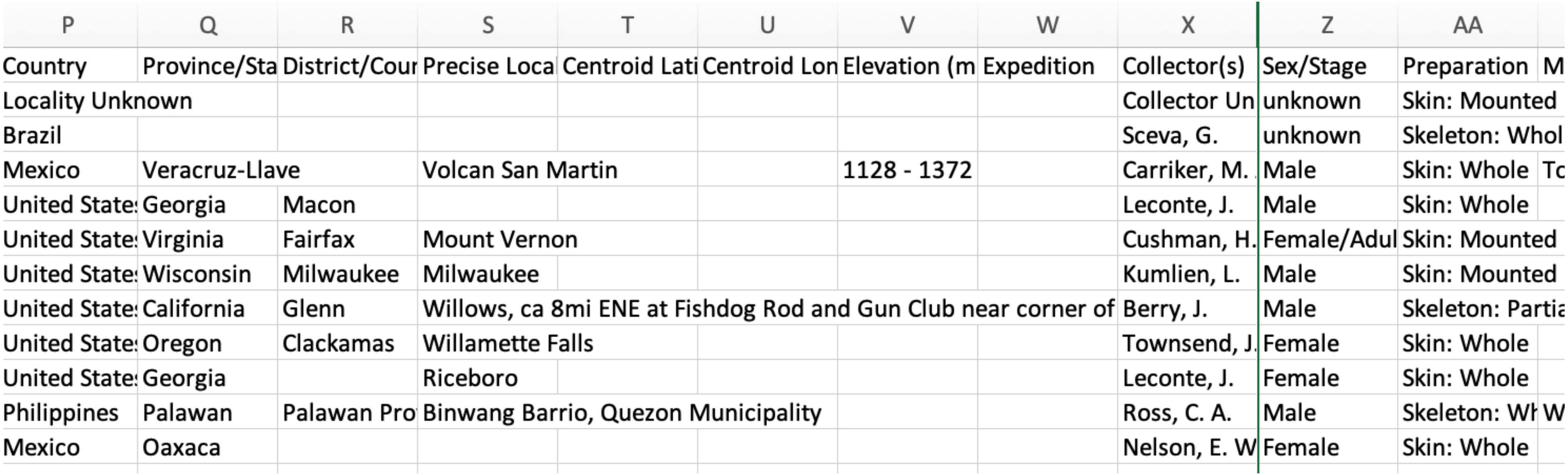


A	E	F	G	H	I	L
Catalog Num	Current Iden	Other Identif	Name Hierarchy	Order	Family	Common Na
608258	Accipiter striatus		Accipiter striatus : Ac	Falconiforme	Accipitridae	Sharp-shinne
321704	Antilophia galeata		Antilophia galeata : F	Passeriforme	Pipridae	Helmeted M
359810	Aulacorhync	Aulacorhync	Aulacorhynchus pras	Piciformes	Ramphastida	Emerald Tou
32271	Bombycilla cedrorum		Bombycilla cedrorun	Passeriforme	Bombycillida	Cedar Waxw
109919	Buteo jamaicensis boreali		Buteo jamaicensis bc	Falconiforme	Accipitridae	Red-tailed H
110774	Chen caerulescens		Chen caerulescens :	Anseriforme	Anatidae	Snow Goose
623418	Chen rossii		Chen rossii : Anatida	Anseriforme	Anatidae	Ross's Goose
78005	Gymnogyps californianus		Gymnogyps californi	Falconiforme	Cathartidae	California Co
A 11605	Picoides villosus audubon		Picoides villosus aud	Piciformes	Picidae	Hairy Woodp
562130	Pitta sordida palawanensi		Pitta sordida palawa	Passeriforme	Pittidae	Hooded Pitta
155665	Strix varia		Strix varia : Strigidae	Strigiformes	Strigidae	Barred Owl

https://collections.nmnh.si.edu/search/birds/?v=g0

Bird Specimens Database

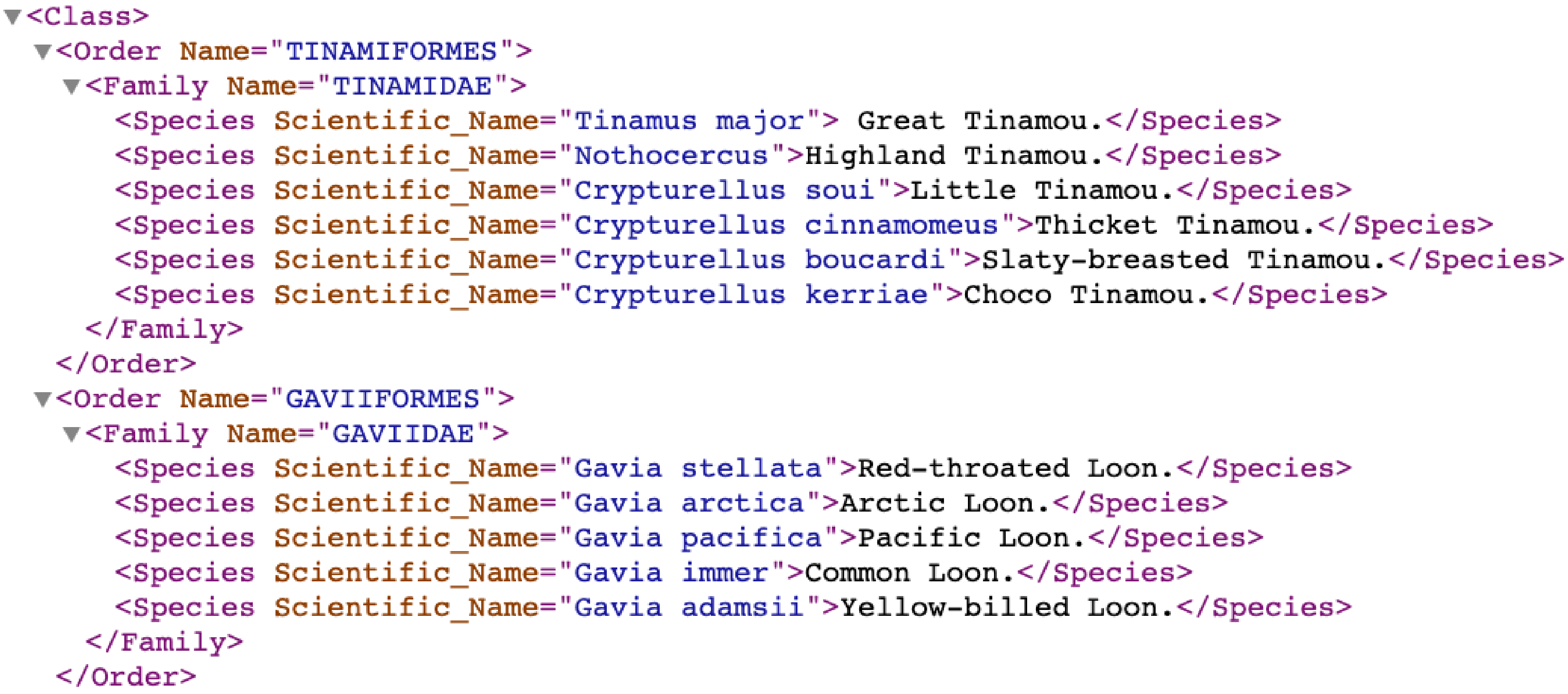
(selected attributes)



P	Q	R	S	T	U	V	W	X	Z	AA	
Country	Province/Sta	District/Cour	Precise Loca	Centroid Lati	Centroid Lon	Elevation (m	Expedition	Collector(s)	Sex/Stage	Preparation	M
Locality Unknown								Collector Un	unknown	Skin: Mounted	
Brazil								Sceva, G.	unknown	Skeleton: Whol	
Mexico	Veracruz-Llave		Volcan San Martin			1128 - 1372		Carriker, M. .	Male	Skin: Whole	To
United State	Georgia	Macon						Leconte, J.	Male	Skin: Whole	
United State	Virginia	Fairfax	Mount Vernon					Cushman, H.	Female/Adul	Skin: Mounted	
United State	Wisconsin	Milwaukee	Milwaukee					Kumlien, L.	Male	Skin: Mounted	
United State	California	Glenn	Willows, ca 8mi ENE at Fishdog Rod and Gun Club near corner of					Berry, J.	Male	Skeleton: Partia	
United State	Oregon	Clackamas	Willamette Falls					Townsend, J.	Female	Skin: Whole	
United State	Georgia		Riceboro					Leconte, J.	Female	Skin: Whole	
Philippines	Palawan	Palawan Pro	Binwang Barrio, Quezon Municipality					Ross, C. A.	Male	Skeleton: Wh	W
Mexico	Oaxaca							Nelson, E. W	Female	Skin: Whole	

https://collections.nmnh.si.edu/search/birds/?v=g0

Bird Species XML (more on this next lecture)



```
▼<Class>
  ▼<Order Name="TINAMIFORMES">
    ▼<Family Name="TINAMIDAE">
       <Species Scientific_Name="Tinamus major"> Great Tinamou.</Species>
       <Species Scientific_Name="Nothocercus">Highland Tinamou.</Species>
       <Species Scientific_Name="Crypturellus soui">Little Tinamou.</Species>
       <Species Scientific_Name="Crypturellus cinnamomeus">Thicket Tinamou.</Species>
       <Species Scientific_Name="Crypturellus boucardi">Slaty-breasted Tinamou.</Species>
       <Species Scientific_Name="Crypturellus kerriae">Choco Tinamou.</Species>
     </Family>
   </Order>
  ▼<Order Name="GAVIIFORMES">
    ▼<Family Name="GAVIIDAE">
       <Species Scientific_Name="Gavia stellata">Red-throated Loon.</Species>
       <Species Scientific_Name="Gavia arctica">Arctic Loon.</Species>
       <Species Scientific_Name="Gavia pacifica">Pacific Loon.</Species>
       <Species Scientific_Name="Gavia immer">Common Loon.</Species>
       <Species Scientific_Name="Gavia adamsii">Yellow-billed Loon.</Species>
     </Family>
   </Order>
```

http://webdam.inria.fr/Jorge/prog/birds.xml

Data, Descriptions, Metadata, Metadata Description

Phenomena in the World	Data: What Items are Collected	Data Representation	MetaData	Metadata Description Language
	Photos of birds, stored on hard drive	Pixels	Image size, Image format, Dat/time taken, Lat/Long, Camera settings	EXIF (automatic) IPTC (manual); now XMP is a new standard
	Real bird specimen	Physical bird	Measurements, type, date, instrument, etc	Specimens database Schema (e.g. EME)
Work done on a contract	PDF document of an invoice	Dollar amount embedded in PDF	Date, invoice number, payee, amount due, etc	Database Schema

Exercise: Identify the Metadata

Spotify®

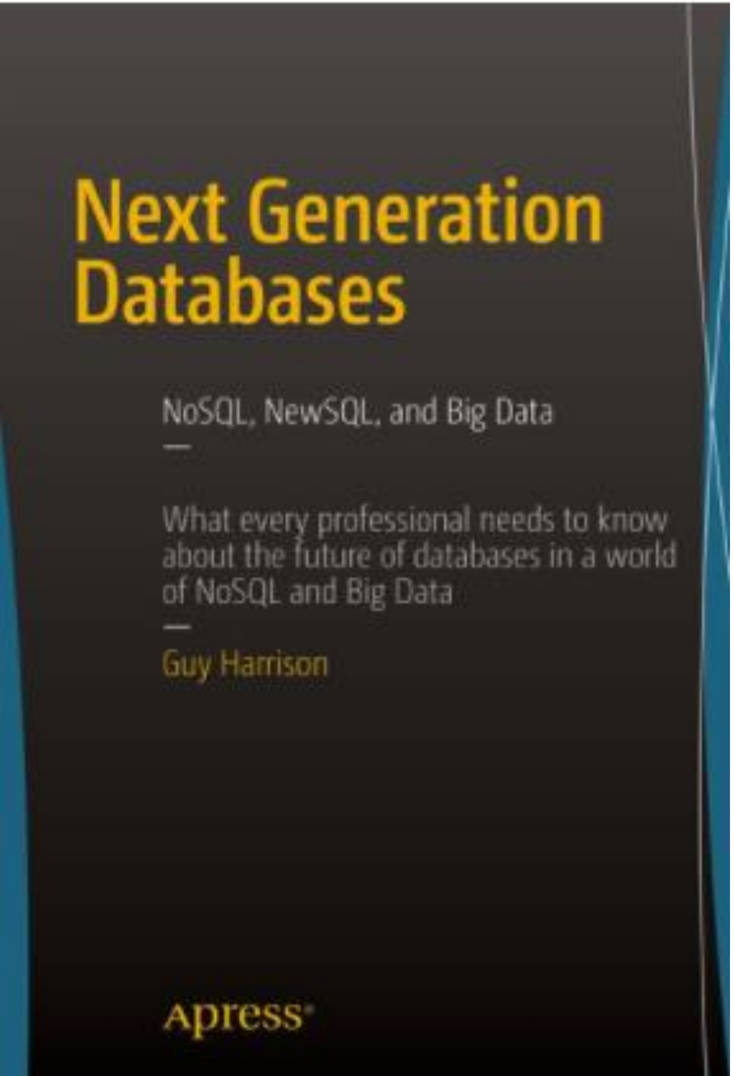
EXERCISE:

- Look several Kaggle.com competitions
- Examine the metadata
- What does it consist of?
- How are they similar and different across collections?

Databases

DATABASES

- Collections of records
- Highly structured
- Very much associated with computerization
- Mainly needed when there are requirements for:
  - *Scaling with large amounts of data*
  - *Accessing reliably over time by many parties*



Next Generation Databases

NoSQL, NewSQL, and Big Data

What every professional needs to know about the future of databases in a world of NoSQL and Big Data

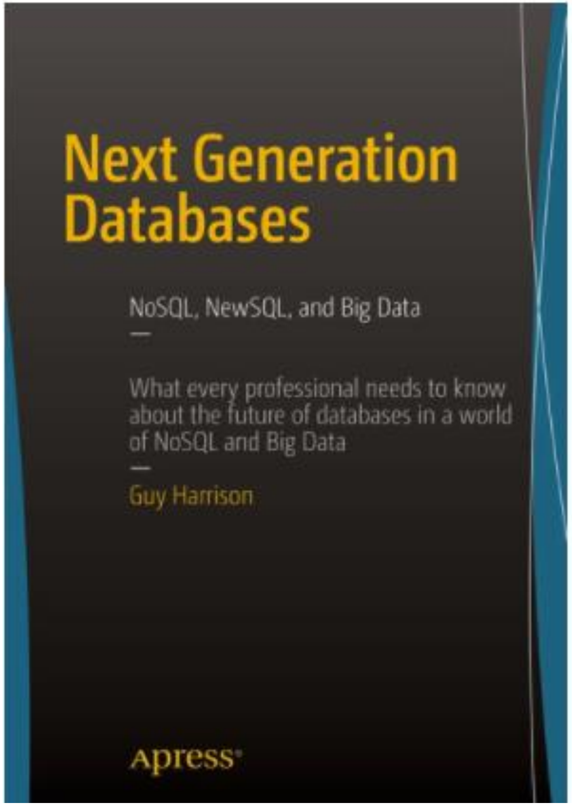
Guy Harrison

Apress

Three Database Revolutions: What To Draw from This Reading:

- An overall impression of the history
- The relationship between computer hardware and developments in data storage and retrieval
- The role of DBMSs in the past and in relation to data science today
- Don’t sweat the technical details; this should be a useful contextualizing reference when you hear/read unfamiliar terms.

Reading: Three Database Revolutions



Next Generation Databases

NoSQL, NewSQL, and Big Data

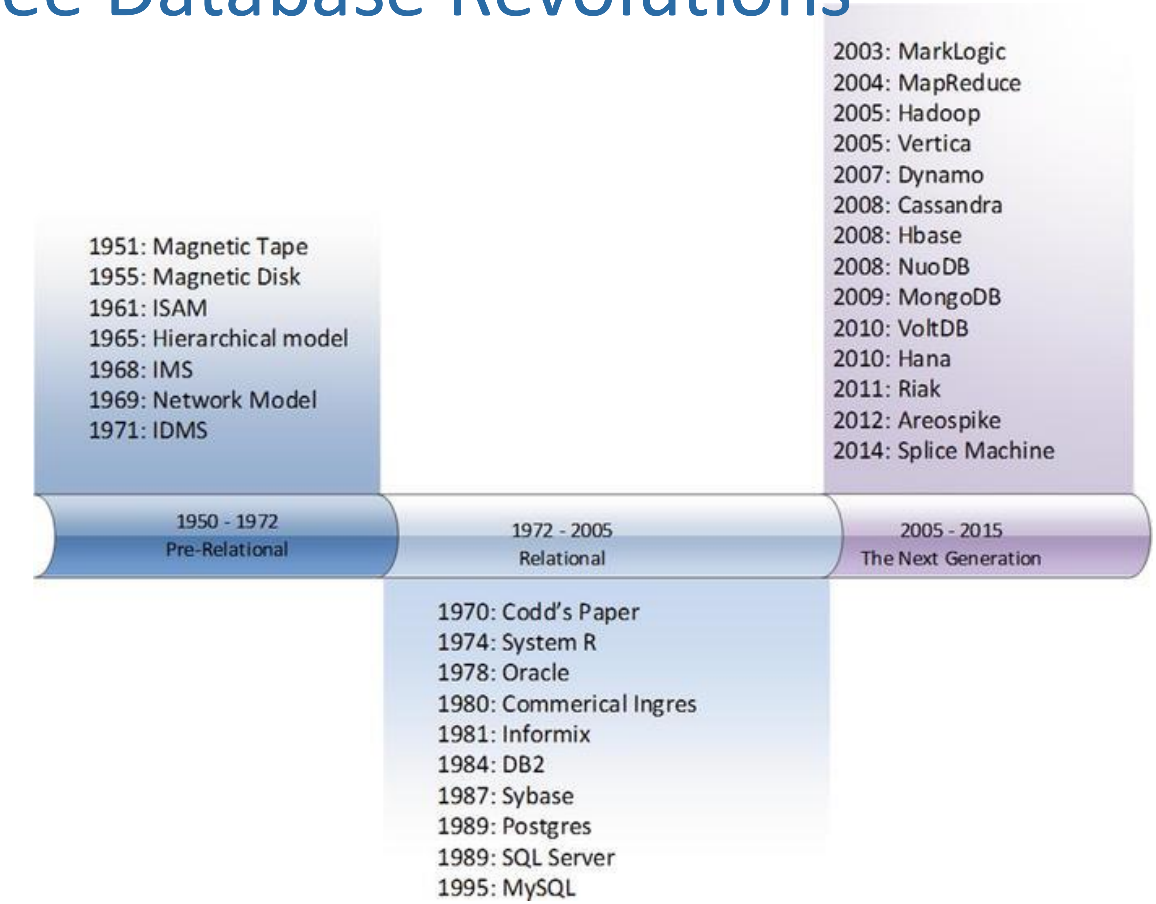
—

What every professional needs to know about the future of databases in a world of NoSQL and Big Data

—

Guy Harrison

Apress®



1951: Magnetic Tape
1955: Magnetic Disk
1961: ISAM
1965: Hierarchical model
1968: IMS
1969: Network Model
1971: IDMS

1950 - 1972
Pre-Relational

1972 - 2005
Relational

1970: Codd's Paper
1974: System R
1978: Oracle
1980: Commerical Ingres
1981: Informix
1984: DB2
1987: Sybase
1989: Postgres
1989: SQL Server
1995: MySQL

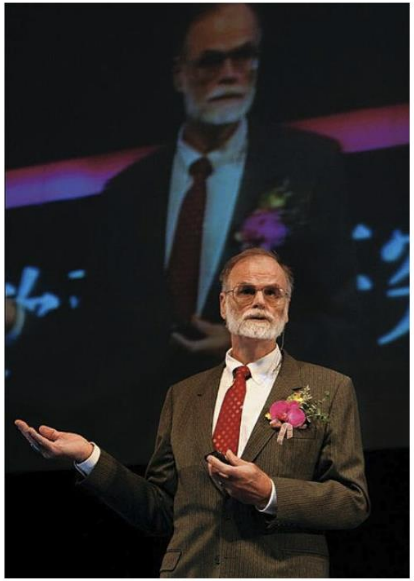
2005 - 2015
The Next Generation

2003: MarkLogic
2004: MapReduce
2005: Hadoop
2005: Vertica
2007: Dynamo
2008: Cassandra
2008: Hbase
2008: NuoDB
2009: MongoDB
2010: VoltDB
2010: Hana
2011: Riak
2012: Areospike
2014: Splice Machine

Database Systems History

- The database systems field began in the 60s with computerization
- Took off in the 1970s thanks to a new “data model” – a way of representing and thinking about data – Codd’s Relational Model
- Led to commercial and open source relational databases
- Relational databases are still at the core of the software technology stack of most companies today.

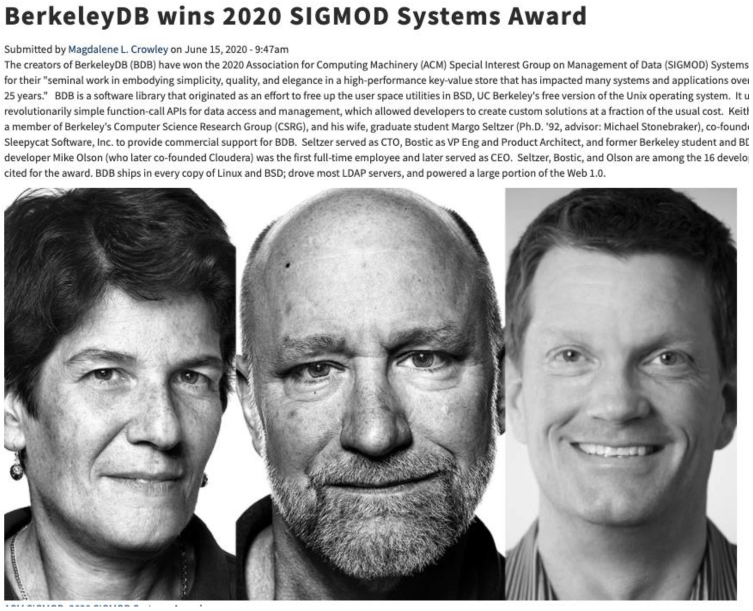
UC Berkeley and RDBMS History



Jim Gray:
1998 Turing Award recipient
First CS PhD student at UCB

Michael Stonebraker
2014 Turing award recipient
Former UCB professor

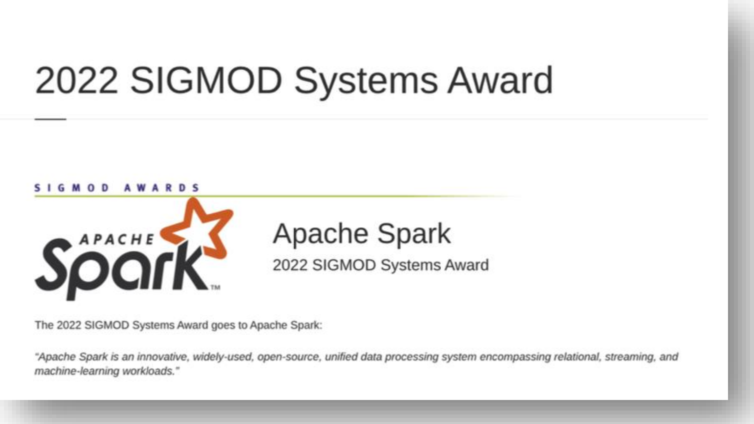
UC Berkeley and RDBMS History



BerkeleyDB wins 2020 SIGMOD Systems Award

Submitted by Magdalene L. Crowley on June 15, 2020 - 9:47am

The creators of BerkeleyDB (BDB) have won the 2020 Association for Computing Machinery (ACM) Special Interest Group on Management of Data (SIGMOD) Systems
for their "seminal work in embodying simplicity, quality, and elegance in a high-performance key-value store that has impacted many systems and applications ove
25 years." BDB is a software library that originated as an effort to free up the user space utilities in BSD, UC Berkeley's free version of the Unix operating system. It u
revolutionarily simple function-call APIs for data access and management, which allowed developers to create custom solutions at a fraction of the usual cost. Keit
a member of Berkeley's Computer Science Research Group (CSRG), and his wife, graduate student Margo Seltzer (Ph.D. '92, advisor: Michael Stonebraker), co-found
Sleepycat Software, Inc. to provide commercial support for BDB. Seltzer served as CTO, Bostic as VP Eng and Product Architect, and former Berkeley student and BD
developer Mike Olson (who later co-founded Cloudera) was the first full-time employee and later served as CEO. Seltzer, Bostic, and Olson are among the 16 develo
cited for the award. BDB ships in every copy of Linux and BSD; drove most LDAP servers, and powered a large portion of the Web 1.0.



2022 SIGMOD Systems Award

Apache Spark

2022 SIGMOD Systems Award

The 2022 SIGMOD Systems Award goes to Apache Spark:

"Apache Spark is an innovative, widely-used, open-source, unified data processing system encompassing relational, streaming, and machine-learning workloads."

Margo Seltzer, Keith Bostic, Mike Olson
Berkeley PhD and Researchers

Matei Zaharia, Ion Stoica,, et al.,
Berkeley Profs; Spark Project led to
Databricks

Relational Database Management Systems (RDBMS)

HUGELY successful – nearly 50 years!

- *System R and Ingres in 1973-4*
- *Oracle in 1977*

SQL a huge improvement over ad hoc programming

ENORMOUS amounts of research and development into making RDBMS very fast, very reliable

Nothing could do better for decades

The Relational Data Model

- A relation is a set of rows (also called tuples)
- Each row consists of a predefined set of attributes
- A database is a collection of relations
- These relations together define the data model

Relation / Table

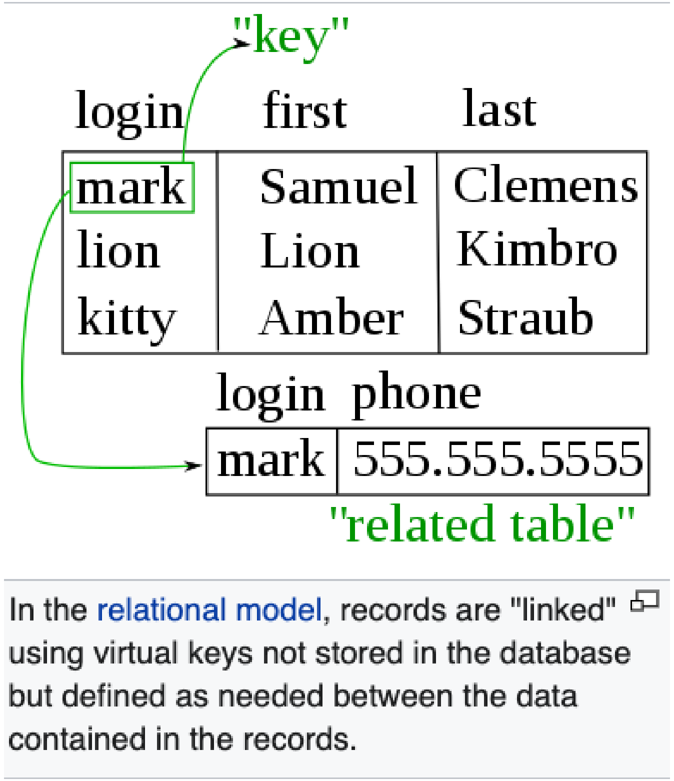
Attributes / Columns

Rows / Tuples / Records

Name	Price	Category	Brand
Climber	$120	Boot	REI
Lita	$98	Flats	West
Arigato	$55	Sneaker	Keds

Slide adapted from Aditya Parameswaran

The Relational Model



login	first	last
mark	Samuel	Clemens
lion	Lion	Kimbro
kitty	Amber	Straub

login	phone
mark	555.555.5555

In the relational model, records are "linked" using virtual keys not stored in the database but defined as needed between the data contained in the records.

wikipedia

Relational Database Management Systems (RDBMS)

- Efficient, reliable for transaction processing:
  - *Flight reservations*
  - *Financial transactions*

Transfer $1000 from A to B’s account

1. Debit A’s account
2. Credit B’s account
3. Update metadata

- What happens if the system shuts down in the middle?
  - **ACID Model** of Transaction Processing handles it

RELATIONAL DATABASES: ACID MODEL GOAL: CONSISTENCY, RELIABILITY

- **Atomicity**
  - Each transaction is treated as a single unit, which either succeeds completely, or fails completely:
- **Consistency** / **Correctness**
  - A transaction can only bring the database from one valid state to another
- **Isolation (Concurrency)**
  - Concurrent execution of transactions leaves the database in the same state that would have been obtained if the transactions were executed sequentially
- **Durability**
  - Once a transaction has been committed, it will remain committed even in the case of a system failure

Transactions vs Analysis

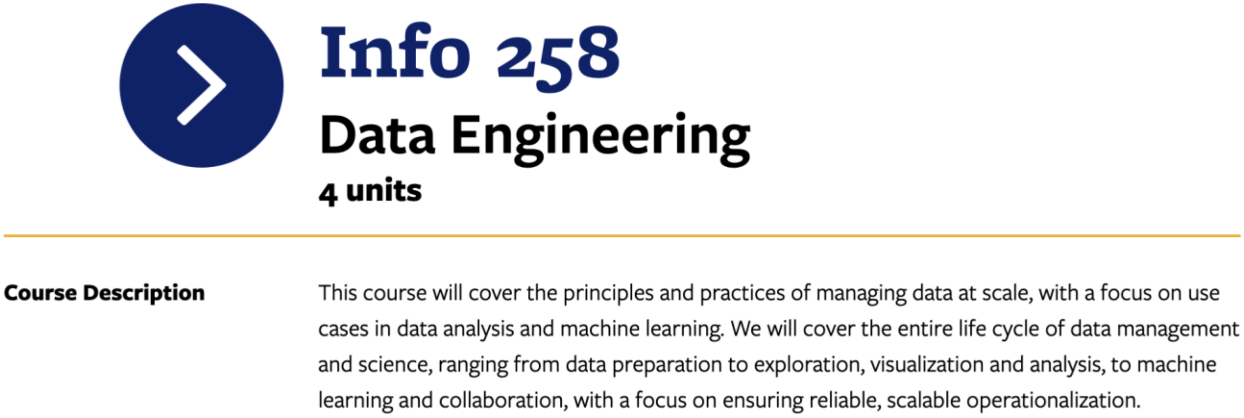
- Databases are often subdivided into:
  - *OLTP (Online Transaction Processing)*
  - *OLAP (Online Analytical Processing)*
- OLTP focuses on “transactions”, OLAP focuses on “large-scale analysis”

Third Generation Databases

- Is still an active area of development
- Different solutions tailored to different applications
  - *Semi-structured data*
  - *Streaming data*
- Sometimes “bolt-on” solutions to relational databases; sometimes entirely new solutions
  - *MongoDB is the most popular*

To Learn More…

To learn about SQL, database normalization, primary and foreign keys, ER diagrams, etc, etc….



Info 258

Data Engineering

4 units

Course Description

This course will cover the principles and practices of managing data at scale, with a focus on use cases in data analysis and machine learning. We will cover the entire life cycle of data management and science, ranging from data preparation to exploration, visualization and analysis, to machine learning and collaboration, with a focus on ensuring reliable, scalable operationalization.

Prof Aditya Parameswaran:

WHAT IS A SCHEMA?

- **Schema**: The overall structure of the metadata
- **Database schema**: relation names, attribute names, attribute types, keys that link relations, and rules that enforce structure
- **XML schema**: the possible types of content in a document and the rules that govern the structure and values of that content.

Relational Schema

- **Schema**: the structure, format or scaffolding
- Schema for a relation:
  - Relation names plus attribute names & types
  - *Product (Name String, Price Float, Category String, Manuf. String)*
- Schema for a Database:
  - Collection of schemas for many relations, and the keys that link them, and in some cases rules that enforce constraints among relations
  - *Product(…)*
  - *Brand (…)*
- Schema in RDB changes very rarely

Metadata: organized with the Schema
Data is: Instance of a database with "values filled in"

Slide adapted from Aditya Parameswaran

The Relational Model: What's Missing?

- Not good for semi-structured data
  - *Documents*
  - *Web pages*
- Not good for hierarchically structured data
- Not good for graph-structured datga

REPRESENTING SEMI-STRUCTURED DATA

- Semi-structured data is less “rigid” than structured data
- As semi-structured data became available online, this exposed a need for new representations (beyond relational)
- XML and JSON became the most popular
- They allow for a flexible format, multi-valued attributes, and nested attributes
- Because of their self-describing or markup nature, they are commonly used for interchange of data
- And for describing metadata about documents

IS SEMI-STRUCTURED DATA NEW?

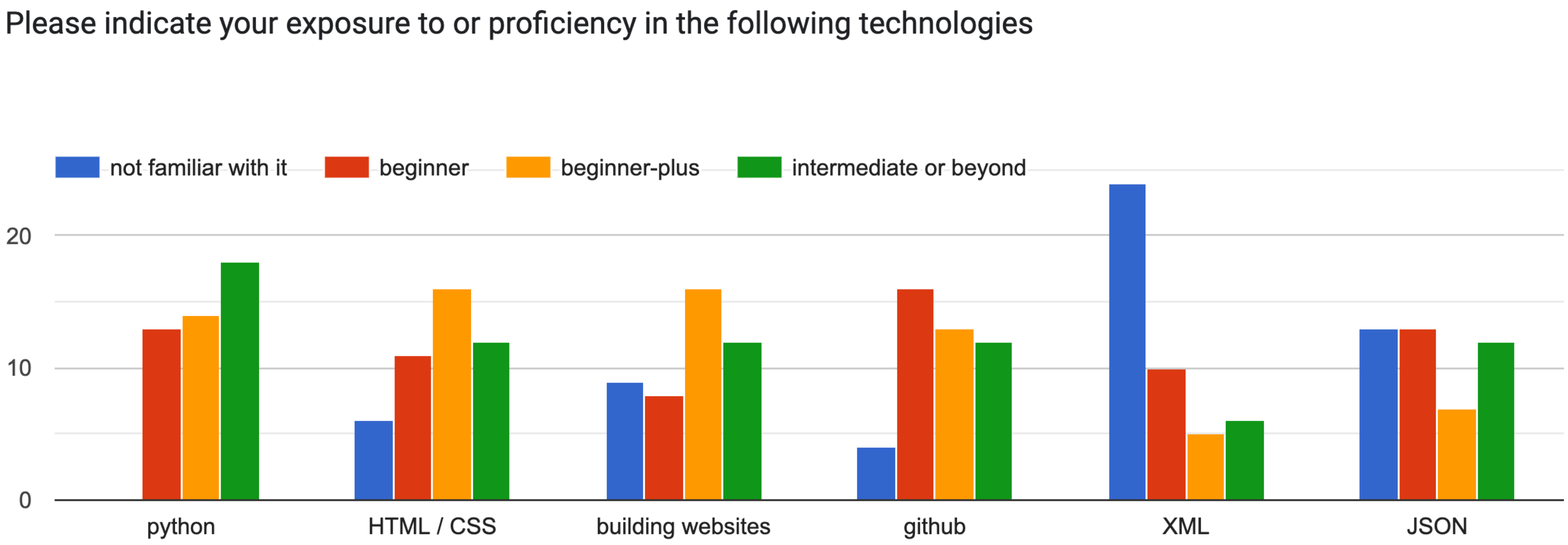
- In some ways, it is old; similar to the systems described in phase 1 of the 3 Database Revolutions paper.
  - *Some of these were nested / hierarchical in structure*
- Most were eventually abandoned
  - Representation redundancies, leads to errors when making changes
  - Difficult to efficiently retrieve across relations (link authors to papers)
- Today: there are specialized systems for specialized tasks
  - *There are XML and JSON-oriented database systems*
  - *MongoDB is very popular for semi-structured data*

Slide adapted from Aditya Parameswaran

This Week's Assignment

You have two weeks to complete it!

Class Familiarity w/XML, HTML, etc



¼ to ½ the class not familiar / beginner
If you are advanced, feel free to go beyond the assignment

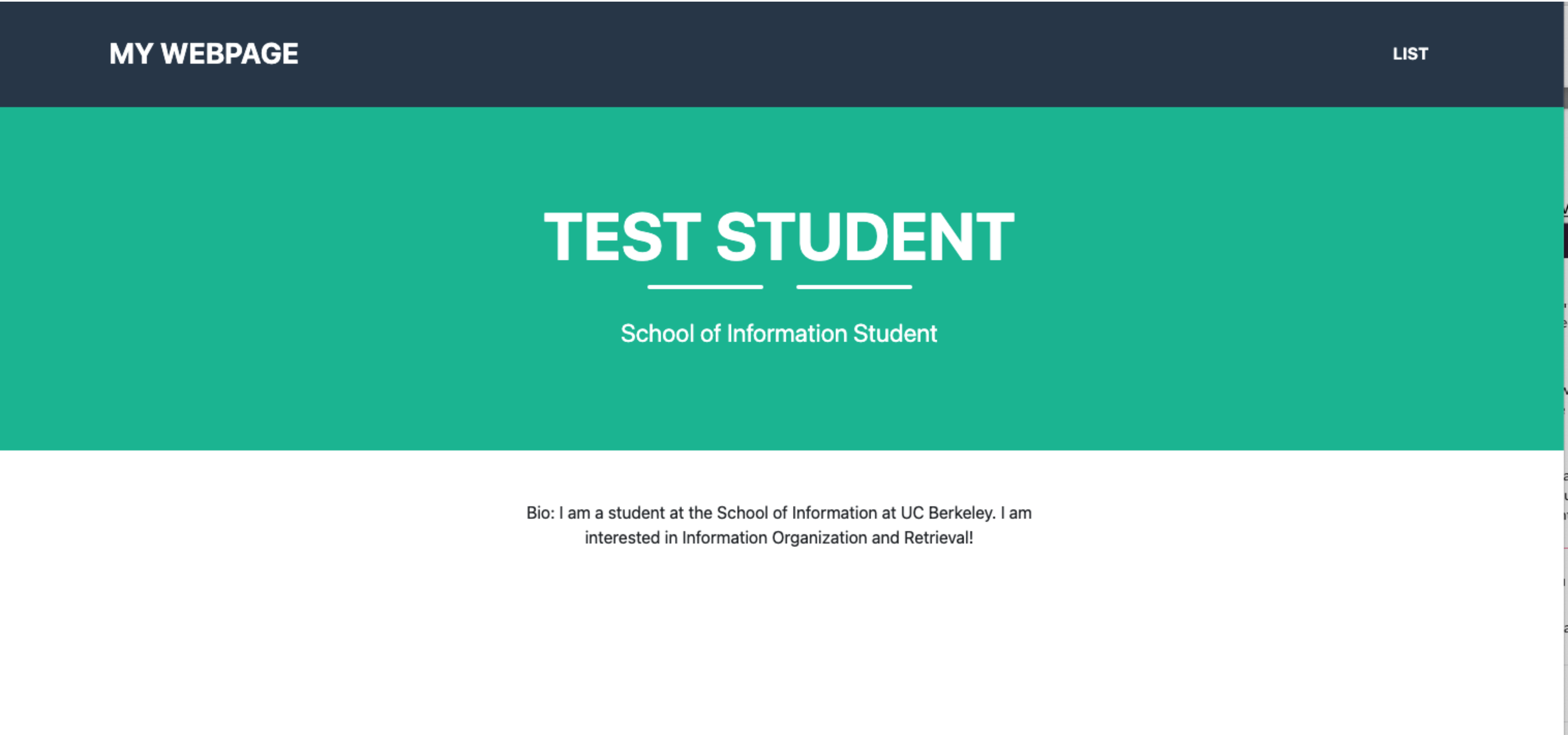
This Week's Assignment
Three main Goals

- **Goal 1: practice with metadata markup**
  - Mark up some content in JSON
  - Mark up some content in XML

This Week's Assignment

- Goal 2: **exposure to building a simple website**
  - A gentle introduction to editing HTML
  - Switching CSS files to see different effects
  - Learning how to post content to github pages
- Goal 3: **more JSON experience**
  - Create a dataset in JSON
  - Display it in an HTML table

The Website

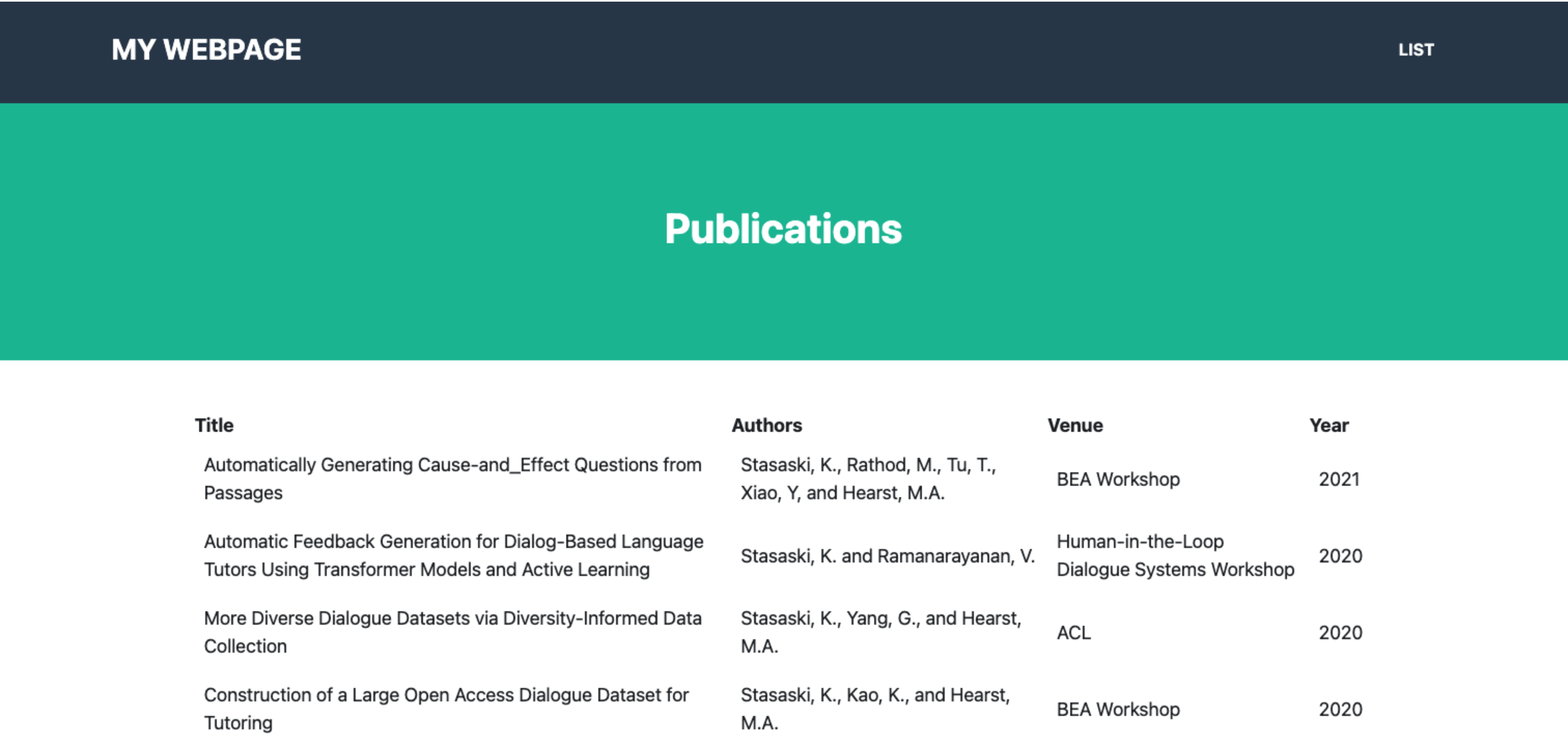


TEST STUDENT

School of Information Student

Bio: I am a student at the School of Information at UC Berkeley. I am interested in Information Organization and Retrieval!

The Website



Publications

Title	Authors	Venue	Year
Automatically Generating Cause-and_Effect Questions from Passages	Stasaski, K., Rathod, M., Tu, T., Xiao, Y, and Hearst, M.A.	BEA Workshop	2021
Automatic Feedback Generation for Dialog-Based Language Tutors Using Transformer Models and Active Learning	Stasaski, K. and Ramanarayanan, V.	Human-in-the-Loop Dialogue Systems Workshop	2020
More Diverse Dialogue Datasets via Diversity-Informed Data Collection	Stasaski, K., Yang, G., and Hearst, M.A.	ACL	2020
Construction of a Large Open Access Dialogue Dataset for Tutoring	Stasaski, K., Kao, K., and Hearst, M.A.	BEA Workshop	2020

NEXT TIME

- Semi-structured Data
- Markup Languages:
  - *HTML*
  - *XML*
- Data format
  - *JSON*