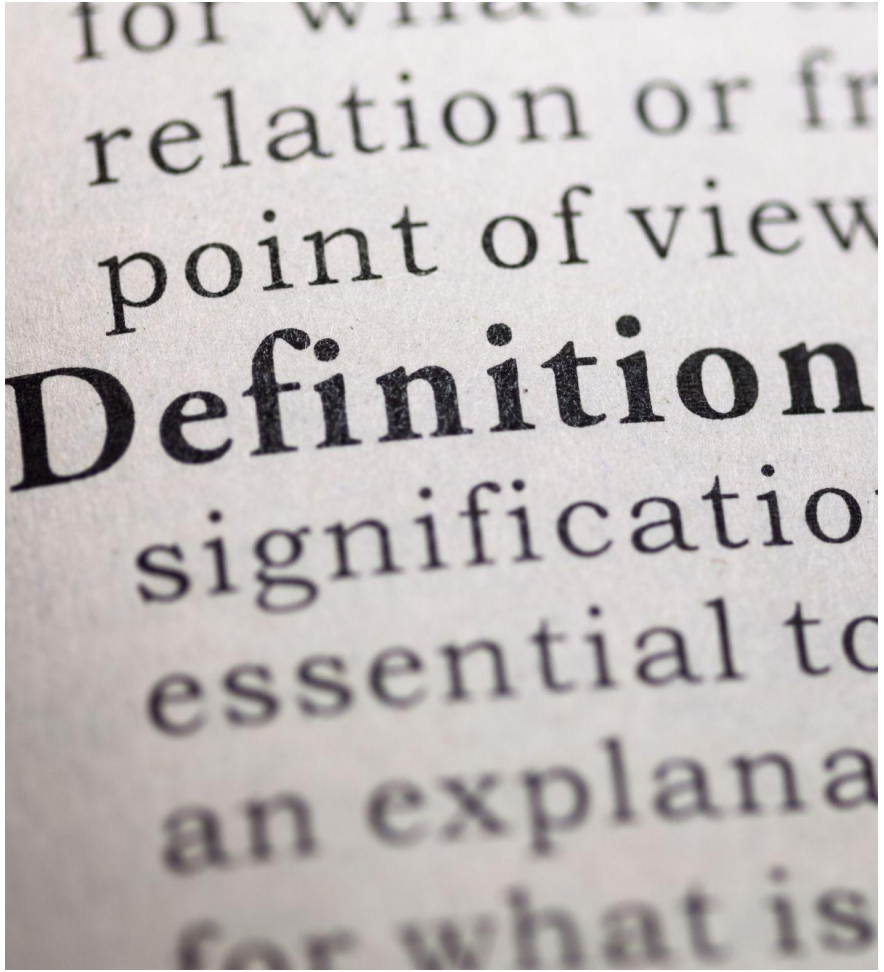
Fundamentals: Terminology

There’s a word for that…



Agenda

Cleanliness is King	01
Let’s Agree to Agree	02
A Rose by Any Other Name	03
Planes, Trains and Automobiles…	04
Here be Dragons.	05

Cleanliness is King

Cleanliness is King

Fitness for Use	01
Quality	02
Correctness	03
Consistency	04

Fitness for Use

"… data quality is related to use and cannot be assessed independently of the user. In a database, the data have no actual quality or value (Dalcin 2004); they only have potential value that is realized when someone uses the data to do something useful. Information quality relates to its ability to satisfy its customers and to meet customers' needs (English 1999)."

- Do you understand your data and can you explain its purpose to someone else?
  - **Accessible**: How easily can someone get to your data? You can't use it if you can't find it.
  - **Accurate**: Can you trust the data? For example are identifications recent and by known experts?
  - **Timely**: Will the data be made available soon enough for it to be of use to you? How often is it updated? How out of date is it?
  - **Complete/Comprehensive**: Which parts of the dataset are fully fleshed out? It may be taxonomically comprehensive for some groups and not others.
  - **Consistent with other sources**: Is the data in each field always of the same type? For example are all your dates in the format `Day Month Year`.
  - **Relevant**: How similar is this dataset to others that have been used successfully for the same purpose?
  - **Well documented (outside of your head)**: How much resolution is there in your data? For example at what scale can it be used to map species distributions?
  - **Be easy to read and easy to interpret**: Is the dataset documented in a clear and concise way? For handwritten documents are they legible?

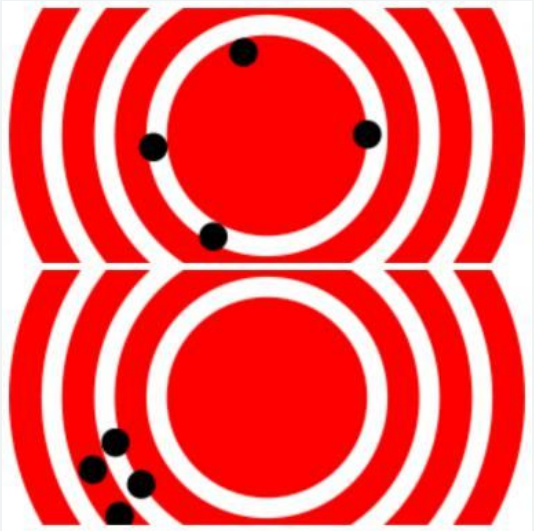
Reference: Chapman, A. D. 2005. Principles of Data Quality, version 1.0. Report for GBIF, Copenhagen. ISBN 87-92020-03-8.

FIELD MUSEUM

Quality

“All data include error — there is no escaping it! It is knowing what the error is that is important, and knowing if the error is within acceptable limits for the purpose to which the data are to be put. (Chapman 2005)”

- Correctness (Accuracy)
  - How close was the recorded value to the actual value?
- Consistency (Precision)
  - How often did you get it right?

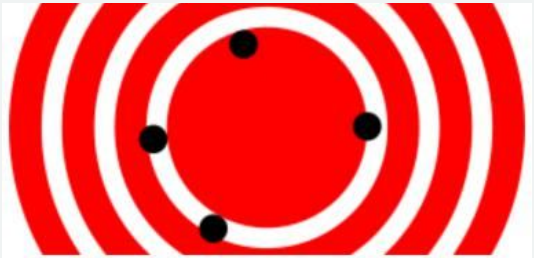


Reference: http://www.businessdictionary.com/definition/data-cleaning.html

Correctness — Examples

Data cleaning is the process of **correcting** or removing dirty data caused by contradictions, disparities, keying mistakes, missing bits, etc. It also includes validation of the changes made, and may require normalization.

- Correctness (Accuracy)
  - How close was the recorded value to the actual value?
    - Is *Thismia* a fossil bird?
    - Is 60305 the right zip code for Kalamazoo?
    - Did Richard Spruce travel to the Peru in 1863?
    - Can an Adelie penguin be 180 m tall?
    - The speed of sound is 10.29 ms^{-1}.

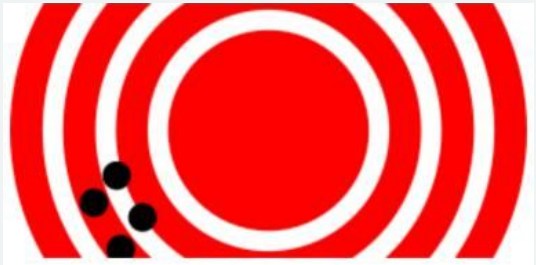


Reference: http://www.businessdictionary.com/definition/data-cleaning.html

Consistency – Examples

Data cleaning is the process of correcting or removing dirty data caused by **contradictions**, **disparities**, **keying mistakes**, **missing bits**, etc. It also includes validation of the changes made, and may require normalization.

- Consistency (Precision)
  - How often did you get it right?
    - Full Name = Joseph Dalton Hooker
    - Full Name = Hooker, J.
    - Full Name = W. J. Hooker
    - Full Name = Hook.f.
    - Full Name = Hook.



Reference: http://www.businessdictionary.com/definition/data-cleaning.html

Let’s Agree to Agree

Let’s Agree to Agree

What is a Standard?

"An agreed way of doing something."

- Norm
- Convention
- Specification
- Requirement
- Restriction
- Rule

Everyday Standards

"The main purpose for standards is to create a framework to ease sharing. They should provide clarity and help communication."

Existing Standards

- Units of measurement (metric; imperial)
- Numeral systems (Hindu-Arabic; Roman numerals)
- Alphabets
- Languages

- Emojis
- ISO Country Codes, US Postal Addressing Standards
- PCI Data Security
- Morse Code

Reference: https://www.bsigroup.com/en-US/Standards/Information-about-standards/What-is-a-standard/

Natural History Standards

"Data standards are the rules by which data are described and recorded. In order to share, exchange, and understand data, we must standardise the format as well as the meaning." (USGS)

- Existing Standards
  - Ecological Metadata Language Standard (EML)
  - Audubon Media Description (a.k.a. Audubon Core)
  - Global Genome Biodiversity Network (GGBN)
  - Ocean Data Standards and Best Practices Project (ODSBP),
  - **Darwin Core**

The main purpose for a data standard is to create a framework to allow easy sharing of data. The result of using a standard is that you will increase data integrity, accuracy and consistency by clarifying ambiguous meaning, minimizing redundant data, and documenting "business" rules.

What is Darwin Core?

"List of fields and their definitions, as they relate to biodiversity data."

Reference: Governance, http://www.tdwg.org; Standard, http://rs.tdwg.org/dwc.

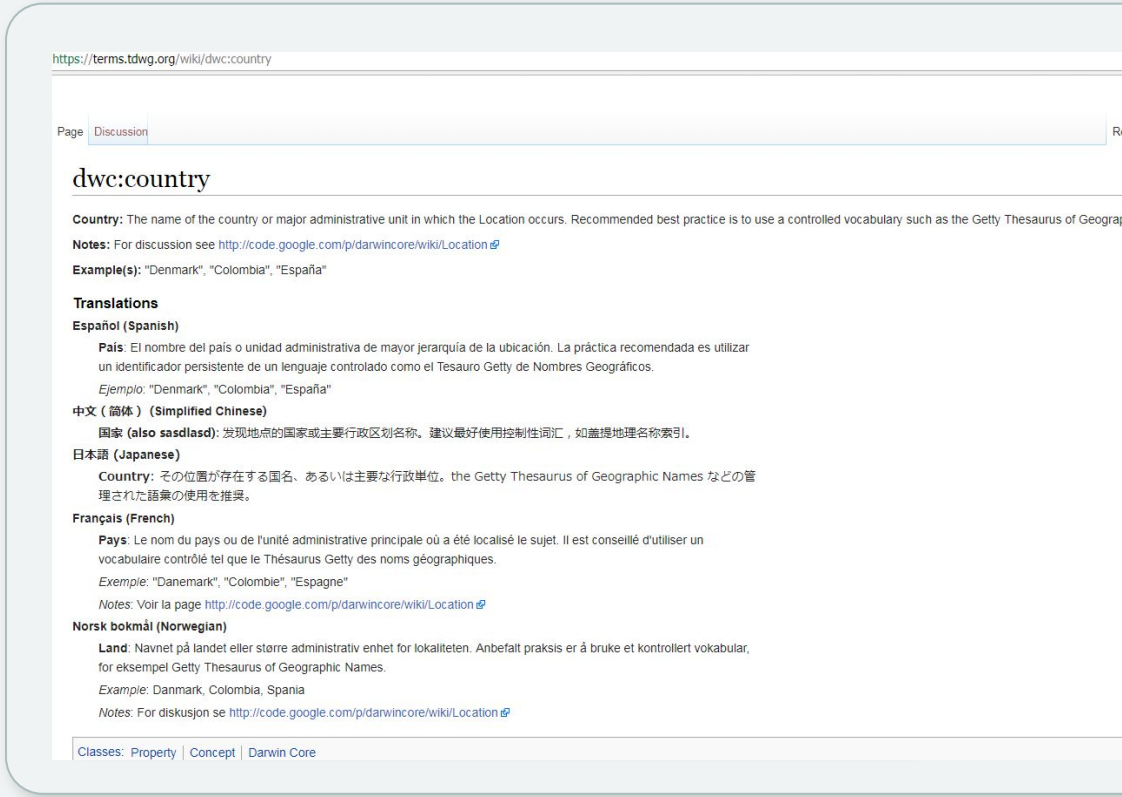
Term Name dwc:country	
Term IRI	http://rs.tdwg.org/dwc/terms/country
Modified	2021-07-15
Term version IRI	http://rs.tdwg.org/dwc/terms/version/country-2021-07-15
Label	Country
Definition	The name of the country or major administrative unit in which the Location occurs.
Notes	Recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names. Recommended best practice is to leave this field blank if the Location spans multiple entities at this administrative level or if the Location might be in one or another of multiple possible entities at this level. Multiplicity and uncertainty of the geographic entity can be captured either in the term higherGeography or in the term locality, or both.
Examples	`Denmark`, `Colombia`, `España`
ABCD equivalence	DataSets/DataSet/Units/Unit/Gathering/Country/Name
Type	Property

Metadata for the 2021-07-15 version of the term dwc:country

Term Name:	dwc:country
Label:	Country
Term version IRI:	http://rs.tdwg.org/dwc/terms/version/country-2021-07-15
Version of:	http://rs.tdwg.org/dwc/terms/country
Issued:	2021-07-15
Definition:	The name of the country or major administrative unit in which the Location occurs.
Type:	Property
Status:	recommended
Replaces:	http://rs.tdwg.org/dwc/terms/version/country-2017-10-06

Example term: `dwc:country`

What is Darwin Core?



Reference: https://terms.tdwg.org/wiki/dwc:country

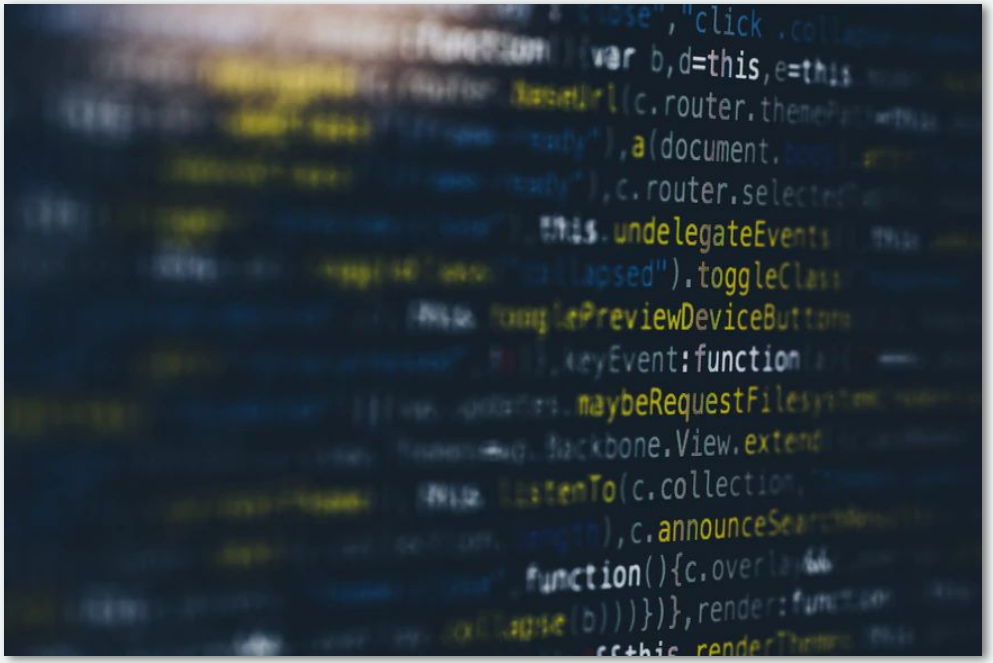
A Rose by Any Other Name

A Rose by Any Other Name

What's the Difference? Words with "data"

"All the World's a [database]."

1. database language
2. database program / software / platform
3. a data-cleaning tool
4. database



Reference: Spiske, M. Nov. 2018, Unsplash License, https://unsplash.com/photos/hvSr_CVecVI

What's the Difference? Database Language

"And all the [languages]..."

1. database language
2. database program / software / platform
3. a data-cleaning tool
4. database

a) **is a method to create all the logical objects** like tables, views, procedures and packages in the database and we need some interface between the user and the database, so that we can access the data stored in it.

b) **allows a user to select records from a database.** It may be in the form of typed commands such as the widely used SQL language, a predefined query menu or a query by example (QBE).

c) **is a generic term referring to a class of languages** used for defining and accessing databases. A particular database language will be associated with a particular database management system.

d) **is used for reading, updating and storing data in a database.** There are several such languages that can be used for this purpose; one of them is SQL (Structured Query Language).

method, create, tables, views, procedures, packages, interface, select, commands, query, defining, accessing, reading, updating, storing

http://www.linfo.org/database.html

FIELD MUSEUM

What's the Difference? Program

"And [programs]..."

1. database language
2. database program / software / platform
3. a data-cleaning tool
4. database

a) **is a utility used for creating, editing and maintaining** database files and records.

b) **is designed to create databases and to store, manage, change, search, and extract** the information contained within them.

c) **is a business information system** that provides file creation, data entry, update, query and reporting functions.

d) **is sometimes also known as database management software (DBMS).**

e) **is a user interface.**

utility, creating, editing, update, store, manage, change, user interface, search, extract, query, reporting

http://www.linfo.org/database.html

FIELD MUSEUM

What's the Difference? Software

"Merely [software]..."

1. database language
2. database program / software / platform
3. a data-cleaning tool
4. database

a) software that **assists** the process of **identifying** and **repairing** inaccurate, incomplete, redundant, or non-conforming data.

b) software that assists in **discovering** and **analyzing** the **quality** of your data. Helps to **find** the patterns, missing values, character sets and other characteristics of your data values.

c) is a **tool** for working with **messy** data: **cleaning** it; **transforming** it from one format into another; and extending it with web services and external data.

d) cleans your database of duplicate data, bad entries and incorrect information.

tool, assists, identifying, repairing, discovering, analyzing, quality, messy, cleaning, transforming

http://www.datasciencecentral.com/profiles/blogs/5-data-cleansing-tools

FIELD MUSEUM

What's the Difference? Database

"That ends this strange, eventful [database]."

1. database language
2. database program / software / platform
3. a data-cleaning tool
4. database

a) a **structured** set of **data** held in a **computer**, especially one that is accessible in various ways.

b) a **collection** of **information organized** to provide efficient retrieval.

c) a data **structure** that stores **organized information**.

d) a **collection** of pieces of **information** that is **organized** and used on a computer.

e) definition, a comprehensive **collection** of related data **organized** for convenient access, generally in a **computer**.

f) a set of **data** that has a regular **structure** and that is **organized** in such a way that a **computer** can easily find the desired **information**.

structured, organized, collection, data, information, computer

FIELD MUSEUM

What's the difference? Field vs. Label

"Things you see and things you don't."

- Field names/column names:
  - An assigned name for a field (e.g., NAME, ADDRESS, CITY, STATE, etc.) that will be the same in every record.
  - In computer science, a name identifying a field in a database record.
  - The underlying unique descriptor for a field.

unique, assigned, identifying

- Label names/output names
  - Descriptive human readable name for a column or field.
  - What you see in a form or output such as a report or user-interface

descriptive, readable, user-interface

Reference: Chapman, A. D. 2005. Principles of Data Quality, version 1.0. Report for GBIF, Copenhagen. ISBN 87-92020-03-8.

FIELD MUSEUM

What's the difference? Lists

"They make lists most usual and lists most unusual…"

- Lookup table:
  - An array that replaces runtime computation with a simpler array indexing operation.
- Lookup field:
  - a read-only field that displays values at runtime based on search criteria you specify.

- Drop-down list:
  - is a graphical control element, similar to a list-box, that allows the user to choose one value from a list. (abbreviated drop-down; also known as a drop-down menu. drop menu, pull-down list, picklist)
- Controlled vocabulary:
  - Controlled vocabulary is an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain.

What's the difference? Characters

"What a difference a byte makes."

- A character set (repertoire):
  - has nothing to do with computers.
  - "It is the the set of characters one might use for a particular purpose."
  - For example, the letters used in Western European languages.

- A coded character set (code page):
  - translates characters to numbers.
  - "Is a set of characters for which a unique number has been assigned".

- Character encoding:
  - translates numbers to binary.
  - "Reflects the way the coded character set is mapped to bytes for manipulation by a computer."
  - It's important to **be aware of character encoding** when migrating data from one environment to another.

Reference: https://www.w3.org/International/articles/definitions-characters/#charsets

Characters

"Representing
and transforming"

- Latin Alphabet = Sharon
- ASCII = [083] [104] [097] [114] [111] [110]
- HEX = [53] [68] [61] [72] [6F] [6E]
- BINARY = [1010011] [1101000] [1100001] [1110010] [1101111] [1101110]

Reference: http://www.asciitable.com/

A = 065 = 01000001

FIELD. MUSEUM

Planes, Trains and Automobiles…

Planes, Trains, and Automobiles…

Field Types / Data Types	01
Character Sets	02
Data Structures	03
Functions	04
Structure and Integrity	05

FIELD MUSEUM

Containers for Data

"There are different shaped containers for your data. Some are simple, some are crazy. Beware."

1. Field types / Data types
2. Character sets
3. Data structures
4. Functions
5. Structure and Integrity

Reference: Teysen, P. Oct. 2020, Unsplash License, https://unsplash.com/photos/bukjsECgmeU

FIELD MUSEUM

Containers for Data

"There are different shaped containers for your data. Some are simple, some are crazy. Beware."

1. Field types / Data types
2. Character sets
3. Data structures
4. Functions
5. Structure and Integrity

a) **Numbers** (numeric) — Two types:
   - integer, long integer
   - float, double

b) **Text** (alphanumeric) — character, string

c) **System Value** (date) — date/time, date, time

d) **Unstructured Text** (memo) — long text, longchar, blob

e) **Binary** (boolean) — 1/0, 0/1: Yes/No, Y/N, True/False

f) **Structured Text** (factor)

Field Types: Simple

"What happens when you put data into a field?

Integer	"56"	"56"	
Float	"0.98"	0.98	
Alphanumeric	"Mrs Pink"	"Mrs Pink"	
Boolean	"0" or "1"	"0" or "1"	
Date	"01 March 1971"	"25993"	"01 March 1971"

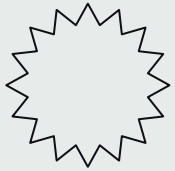
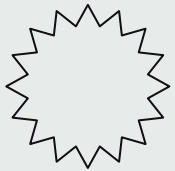
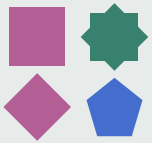
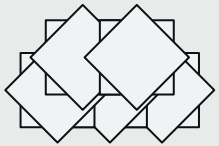
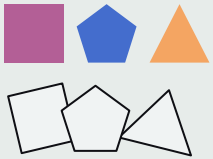
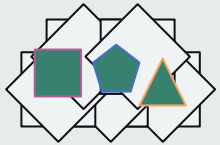
Field Types: Simple

“What happens when you put data into a field?

Alphanumeric	“Mrs Pink”	“Mrs Pink”
Alphanumeric	56	“56”
Alphanumeric	0.98	“0.98”

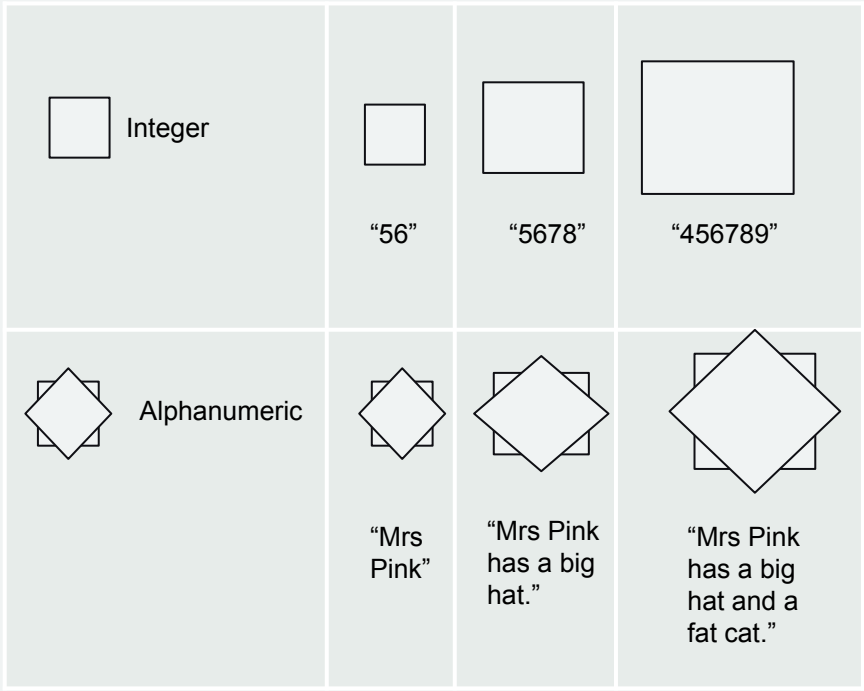
Field Types: Complex

“What happens when you data into a field?”

 Memo	 The quick brown fox...		
 Memo	 56, 0.5, skull, 25788		
 Factor	 10, 1 March 1971, 0		

Modifying Field Types

“What happens when you put data into a field?



The Wrong Field Types

“What happens when you put data into a field?

Integer	0.98		**Weird behavior**
Integer	1 March 1971		“25993”
Integer	“Mrs Pink”		**Weird behavior**
			Weird behavior

Containers for Data

"There are different shaped containers for your data. Some are simple, some are crazy. Beware."

1. Field types / Data types
2. Character sets
3. Data structures
4. Functions
5. Structure and Integrity

a) ASCII

b) Latin1

c) UTF / Unicode

Characters — ASCII

“Representing and transforming”

American **S**tandard **C**ode for **I**nformation **I**nterchange.

- It was the first **character encoding standard.**
- It defined **128** different alphanumeric characters that could be used on the internet.

- Each is represented with a 7-bit binary number (a string of seven 0s or 1s)
- These included:
  - Numbers (0-9),
  - English letters (Aa-Zz),
  - Some special characters like ! $ + - ( ) @ < >

Reference: https://www.w3schools.com/html/html_charset.asp

Characters — ISO Latin 1

“Representing and transforming”

ISO Latin 1 ISO-8859-1:

- Is a standard **character set** developed by the International Organization for Standardization.
- Released in 1998

- It is a superset of the ASCII character encoding standard
- It is the default in HTML 4.01
- There are several variants which encompass other alphabets

Reference: http://casa.colorado.edu/~ajsh/iso8859-1.html

Characters — Unicode

“Representing and transforming”

UNICODE: Unique, Universal, and Uniform character enCoding.

- The Unicode Consortium developed the Unicode Standard.
- Because the character set described in ISO-8859 was too limited.
- Unicode is an encoding standard.

- The latest version contains a repertoire of 136,755 characters
- Unicode can be implemented by different character sets.
- The 1st 128 characters correspond 1:1 with ASCII

Reference: https://www.w3schools.com/charsets/ref_html_utf8.asp

Characters — UTF-8

“Representing and transforming”

- **U**nicode **T**ransformation **F**ormat
- UTF is variable width character encoding
- The most common encodings are **UTF-8** and UTF-16
  - UTF-8 uses a minimum of 8 bits describe a character
  - UTF-16 uses 16 bits to describe a character

- The latest version contains a repertoire of **1,112,064** characters
- UTF-8 is the default encoding in HTML-5

Reference: http://casa.colorado.edu/~ajsh/iso8859-1.html

Containers for Data

“There are different shaped containers for your data. Some are simple, some are crazy. Beware.”

1. Field types / Data types
2. Character sets
3. Data structures
4. Functions
5. Structure and Integrity

a) Grid vs. Table?

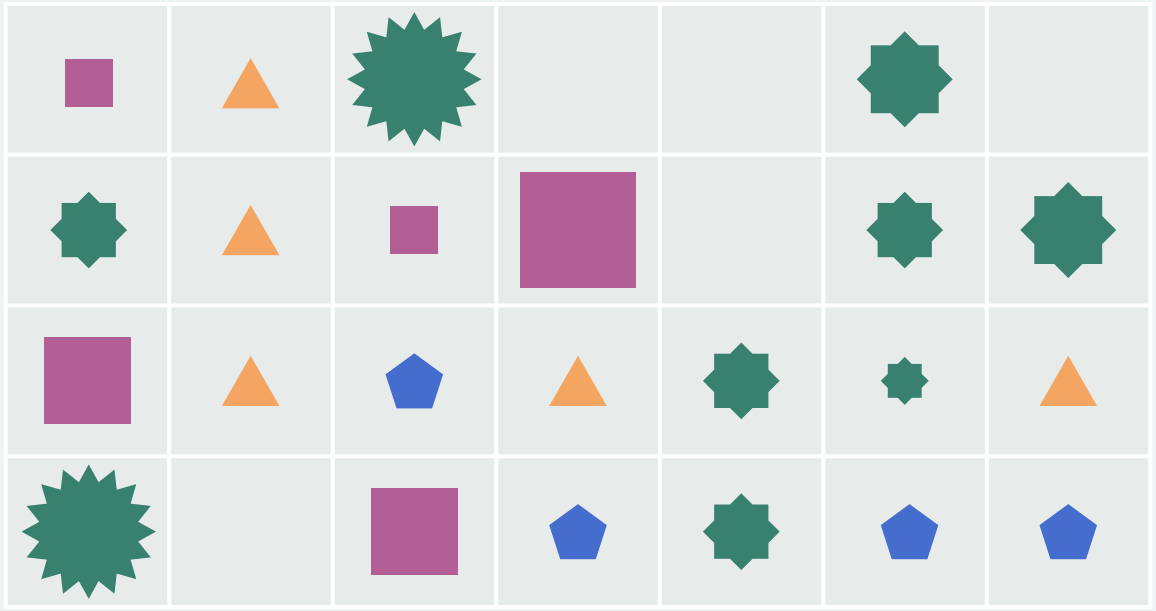
b) What is a record? (What does a row actually represent? What is a column?)

c) null — NA, NULL, 0?

FIELD MUSEUM

Grid

“What happens when you put data into the cells of a grid?”



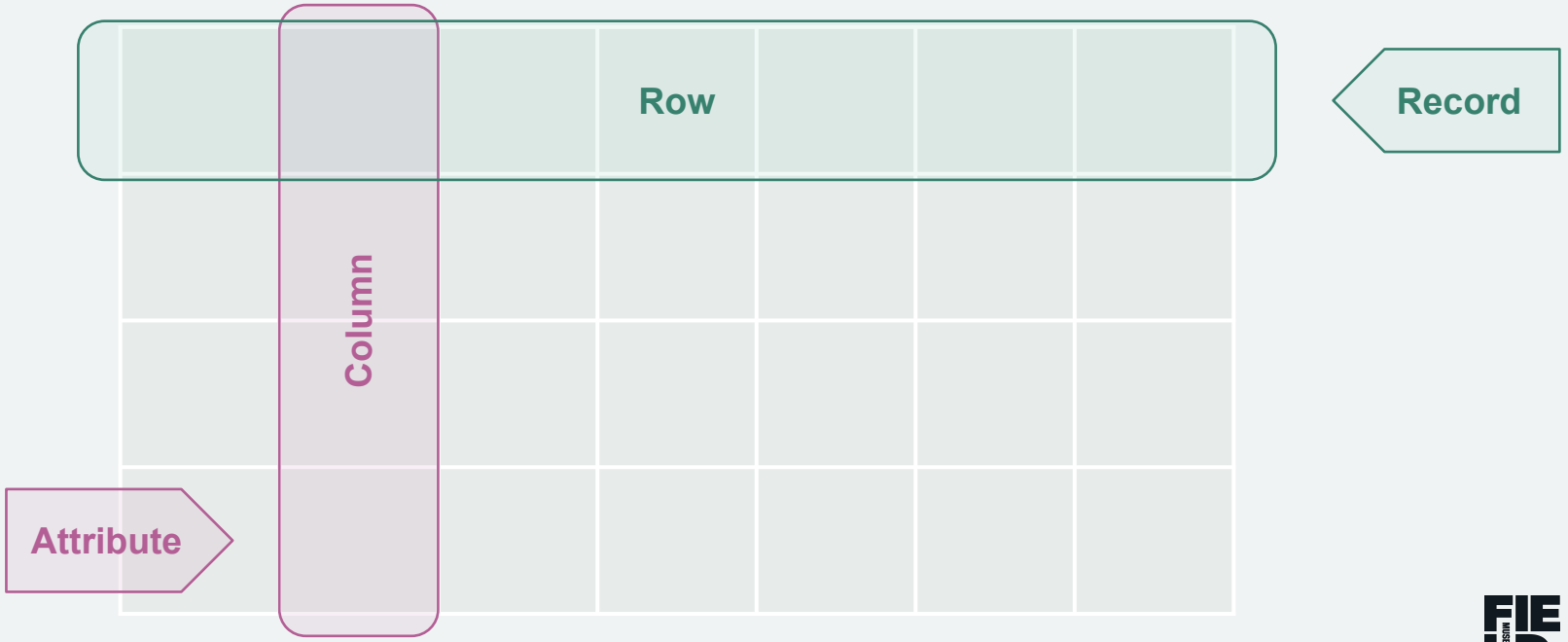
Table

“What happens when you add structure to a grid?”

ID	Expired?	Date	Zip	State	Name	Age

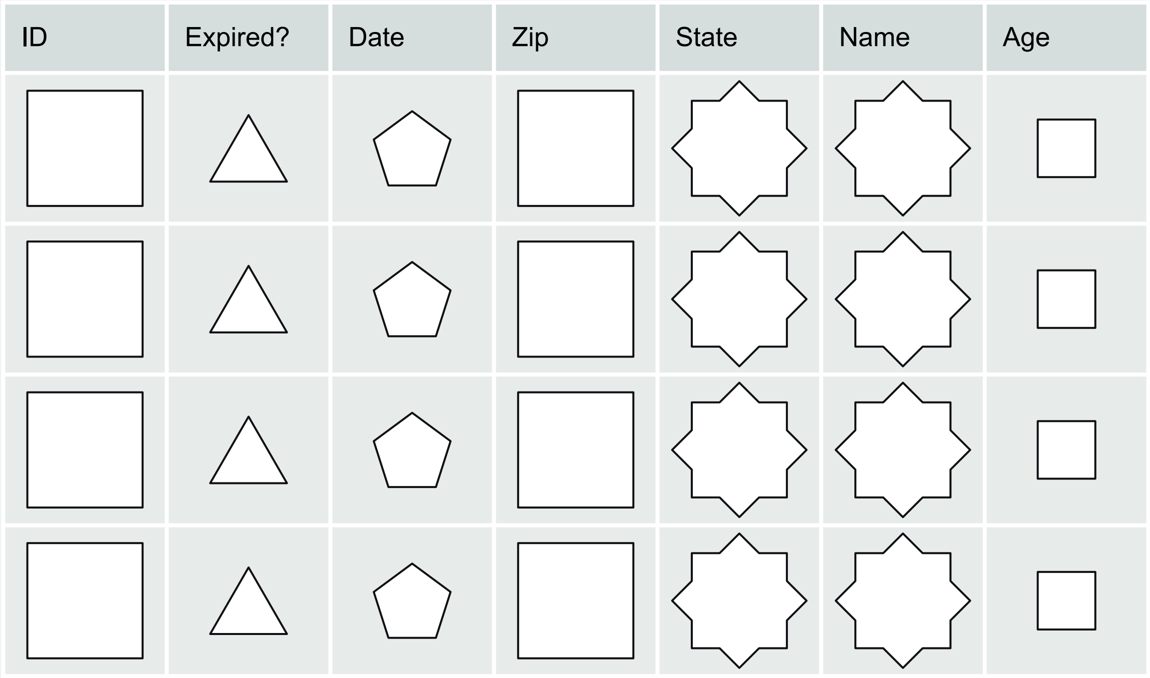
Table

“What happens when you add structure to a grid?”



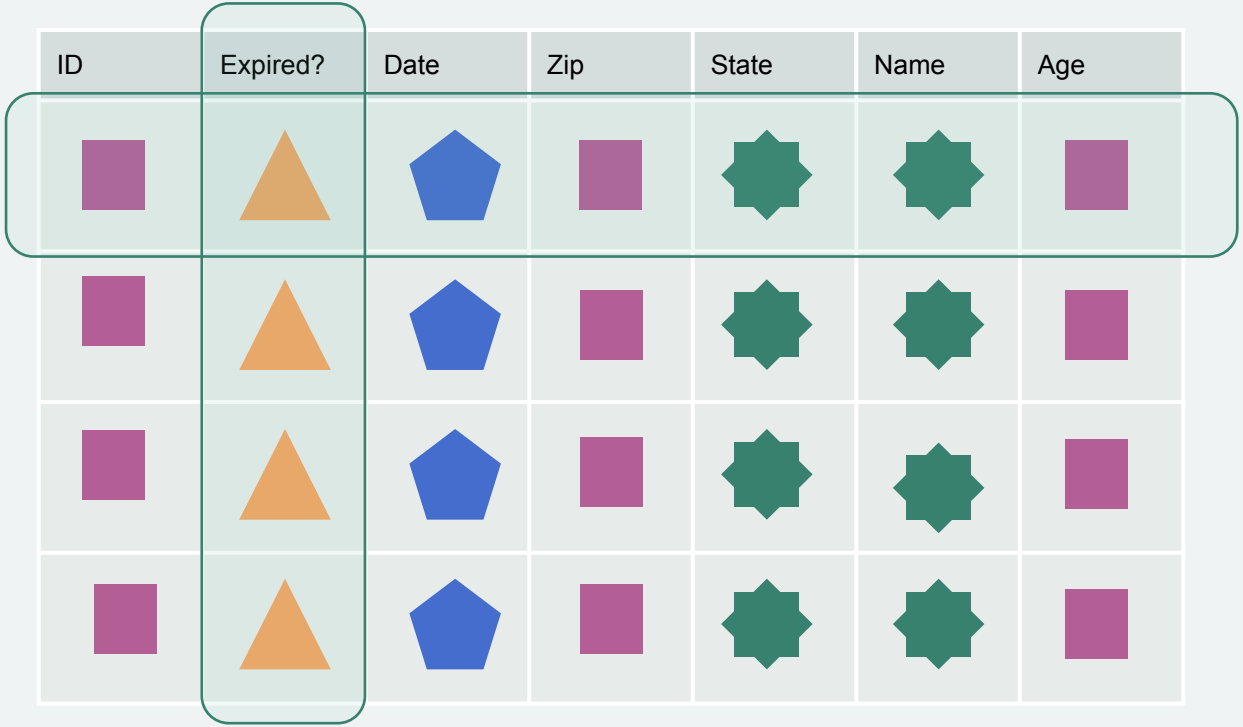
Table

“What happens when you add structure to a grid?”



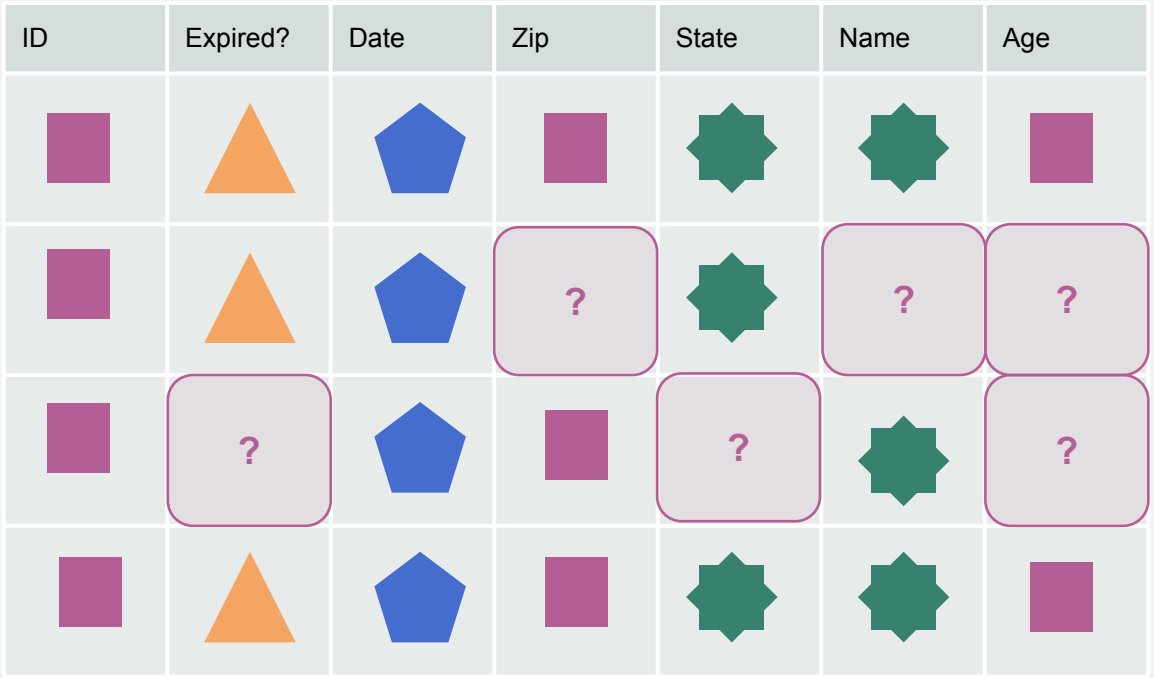
Table

“What happens when you add structure to a grid?”



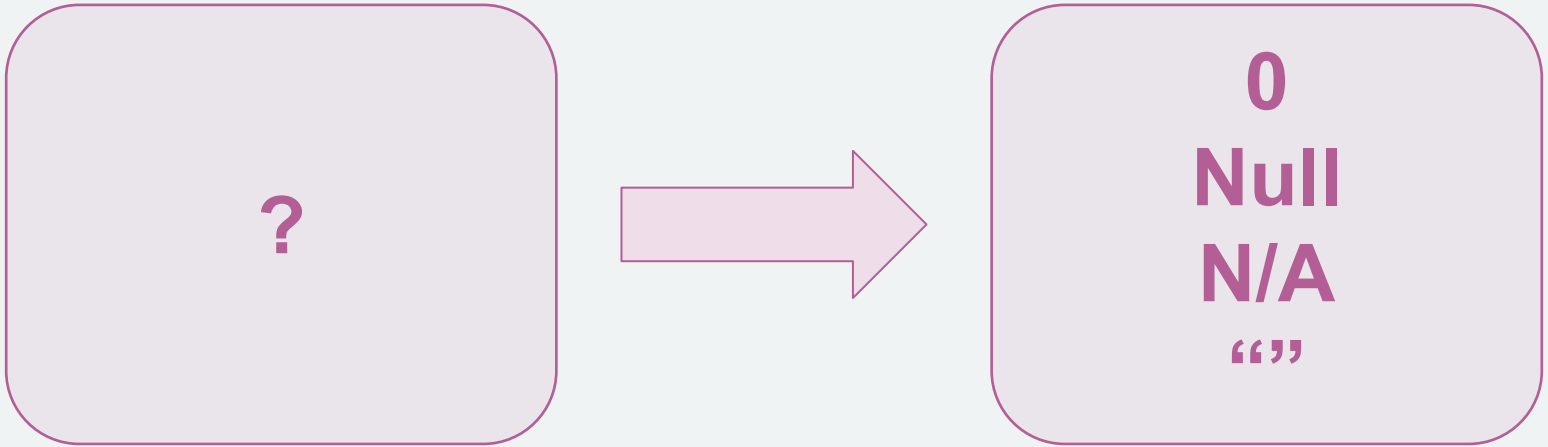
Nothing?

“What happens when there is data missing from a table?”



Nothing?

“An empty cell doesn’t necessarily mean an empty field.”



Containers for Data

"There are different shaped containers for your data. Some are simple, some are crazy. Beware."

1. Field types / Data types
2. Character sets
3. Data structures
4. Functions
5. Structure and Integrity

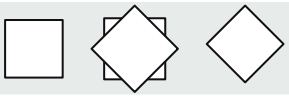
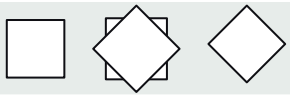
Reference: Trevarthen, L. Feb. 2020, Unsplash License, https://unsplash.com/photos/gWvdUpNQr6g

FIELD MUSEUM

Table

“How do you combine the data in cells to create data in other cells?”

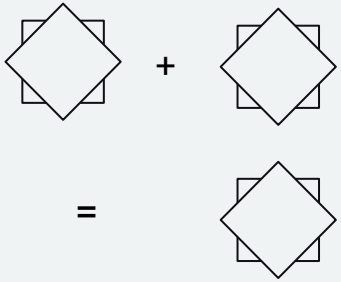
x + y = z

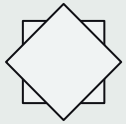
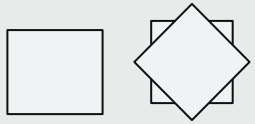
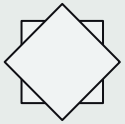
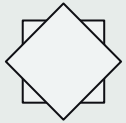
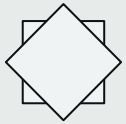
15	6	x+y
		
“15” 	“6” 	Error
15.0 	6.0 	21.0
15 	6 	21

Functions and Formulas

“How do you combine the data in cells to create data in other cells?”

abc *plus* **xyz** =
abcxyz

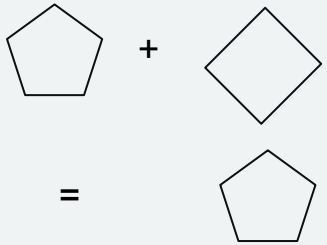


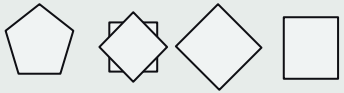
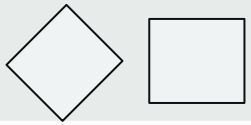
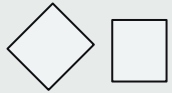
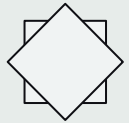
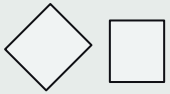
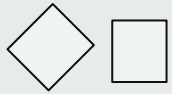
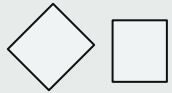
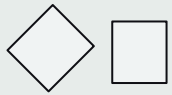
IL	60605	**abc** plus **xyz**
		
“IL” 	“60605” 	“IL60605”
“IL” 	60605 	Expect weird behavior

Functions and Formulas

"How do you combine the data in cells to create data in other cells?"

Date *plus* # of days = new date



21 May 2017	7 or 7.0	Date plus # of days
		
"21 May 2017" 	7 or 7.0 	"Weird behavior" 
42876 	7 or 7.0 	42883  
42876 (21 May 2017) 	**7 or 7.0** 	**28 May 2017** 

Containers for Data

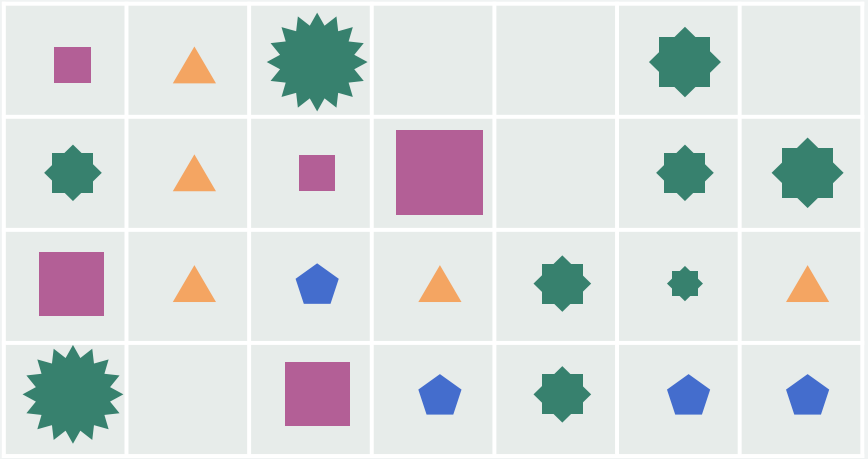
"There are different shaped containers for your data. Some are simple, some are crazy. Beware."

1. Field types / Data types
2. Character sets
3. Data structures
4. Functions
5. Structure and Integrity

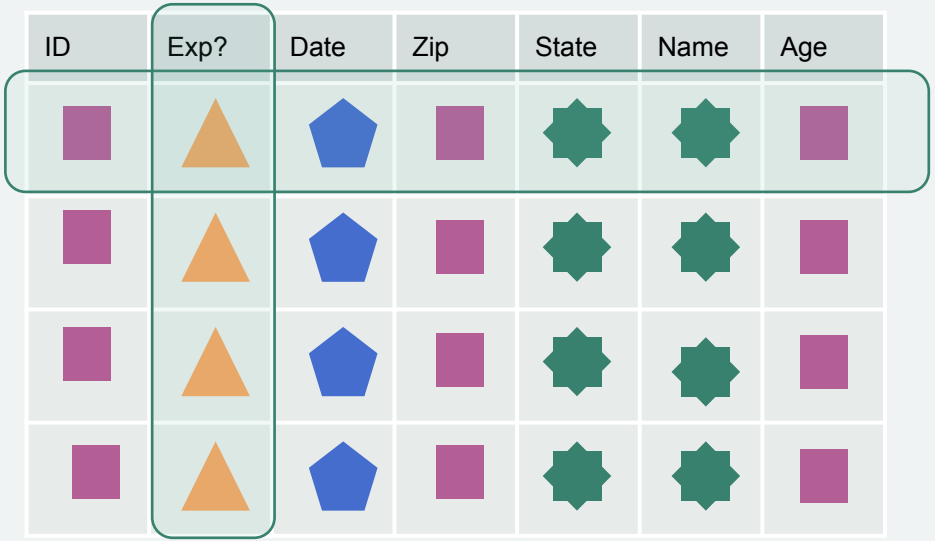
Reference: Trevarthen, L. Feb. 2020, Unsplash License, https://unsplash.com/photos/gWvdUpNQr6g

Grid to Table

“From unstructured to structured.”



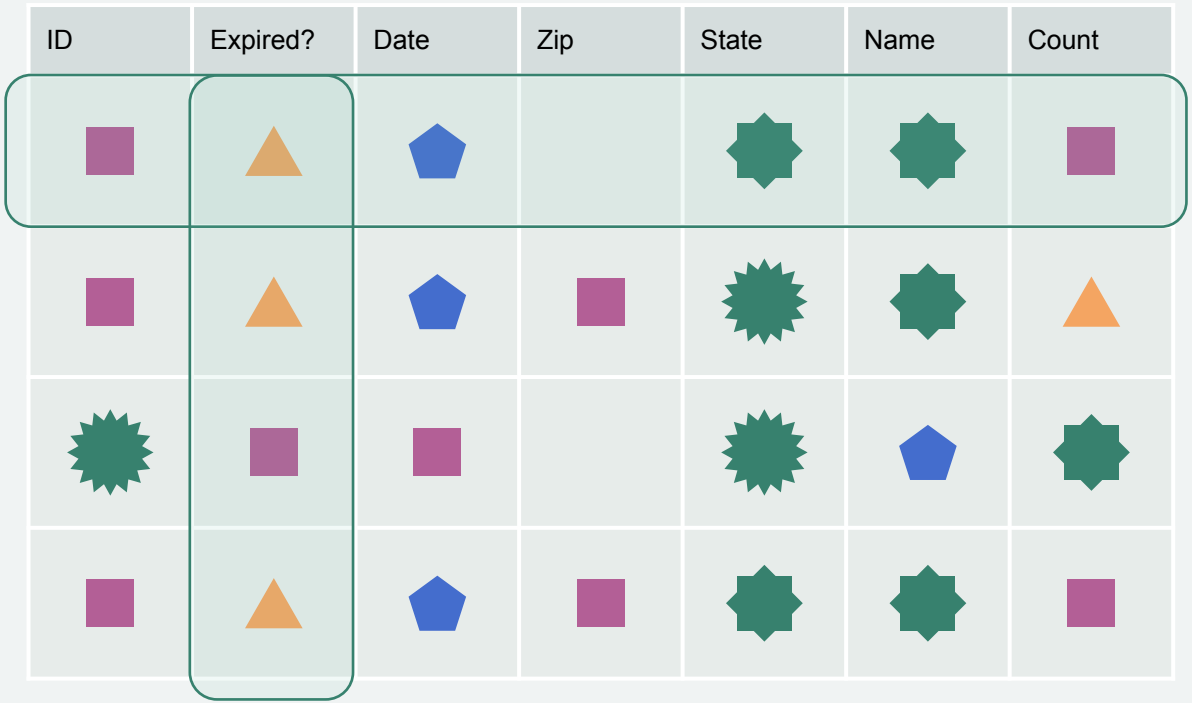
Unstructured Grid



Structured Table

The Problem With Structured Tables

“What happens if you move things around?”



In a structured table there is nothing to stop you rearranging things except your knowledge of the data. If that happens you are right back where you started. So how do you stop that happening?

Data Integrity and Security

“The difference between a spreadsheet and a database table.”

ROW

- Attributes of a record **ALWAYS** stay together.

COLUMN

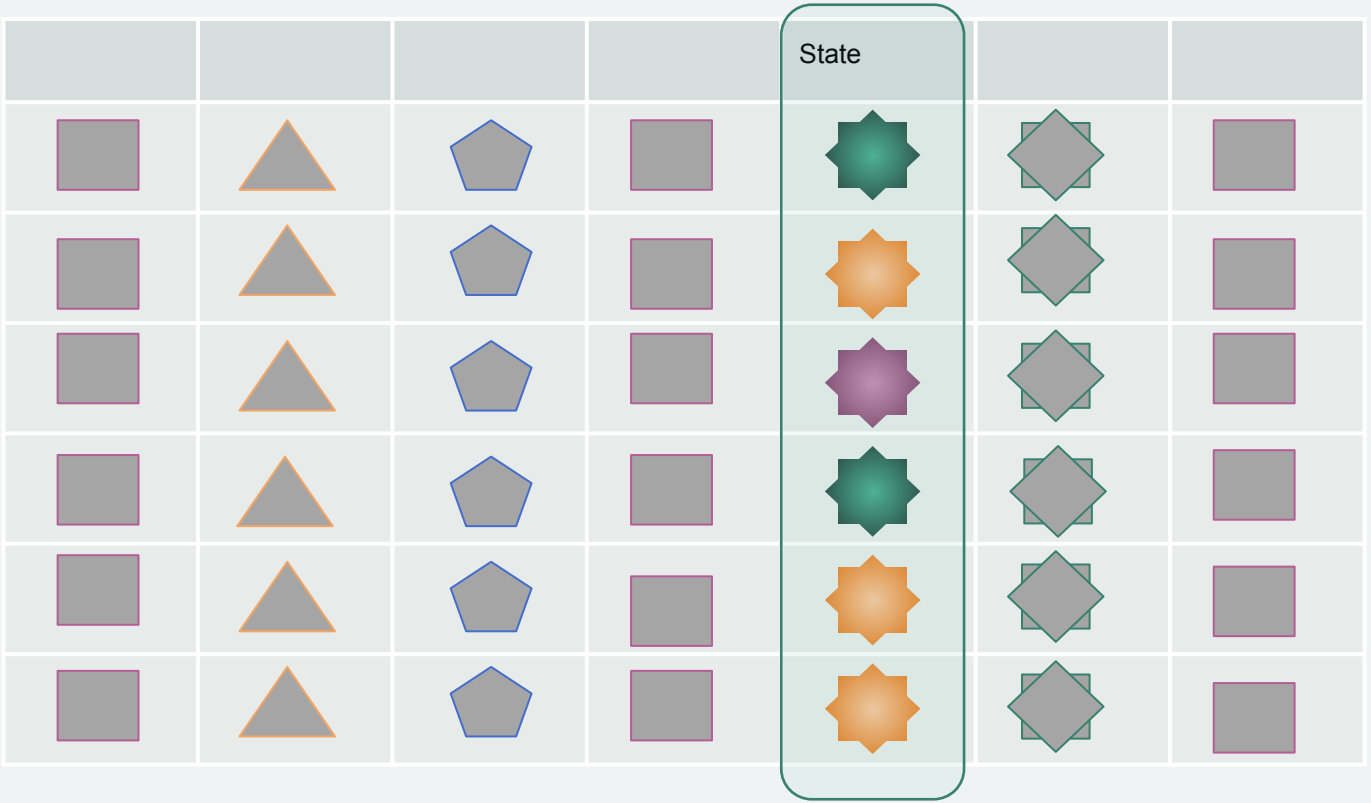
- Any attribute has the **SAME** field/data type for every record.

TABLE

- All data in a table refers to a **SINGLE** concept.

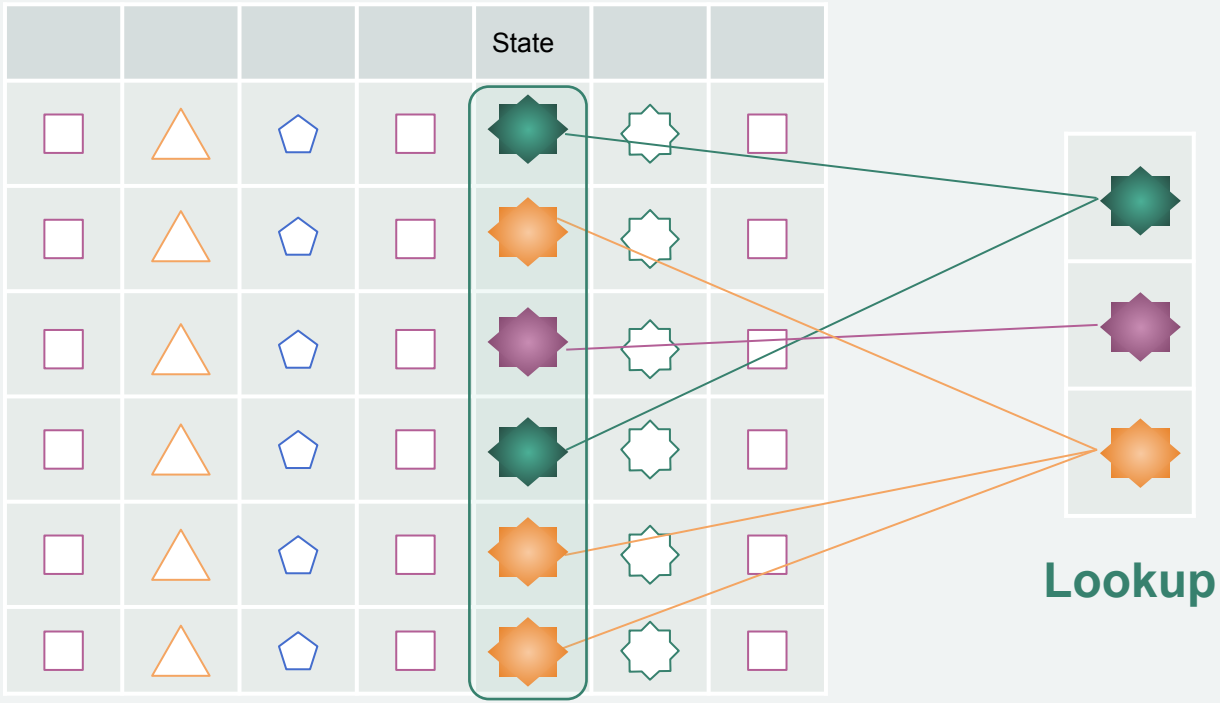
Efficiency and Standardization

“Streamlining your data.”



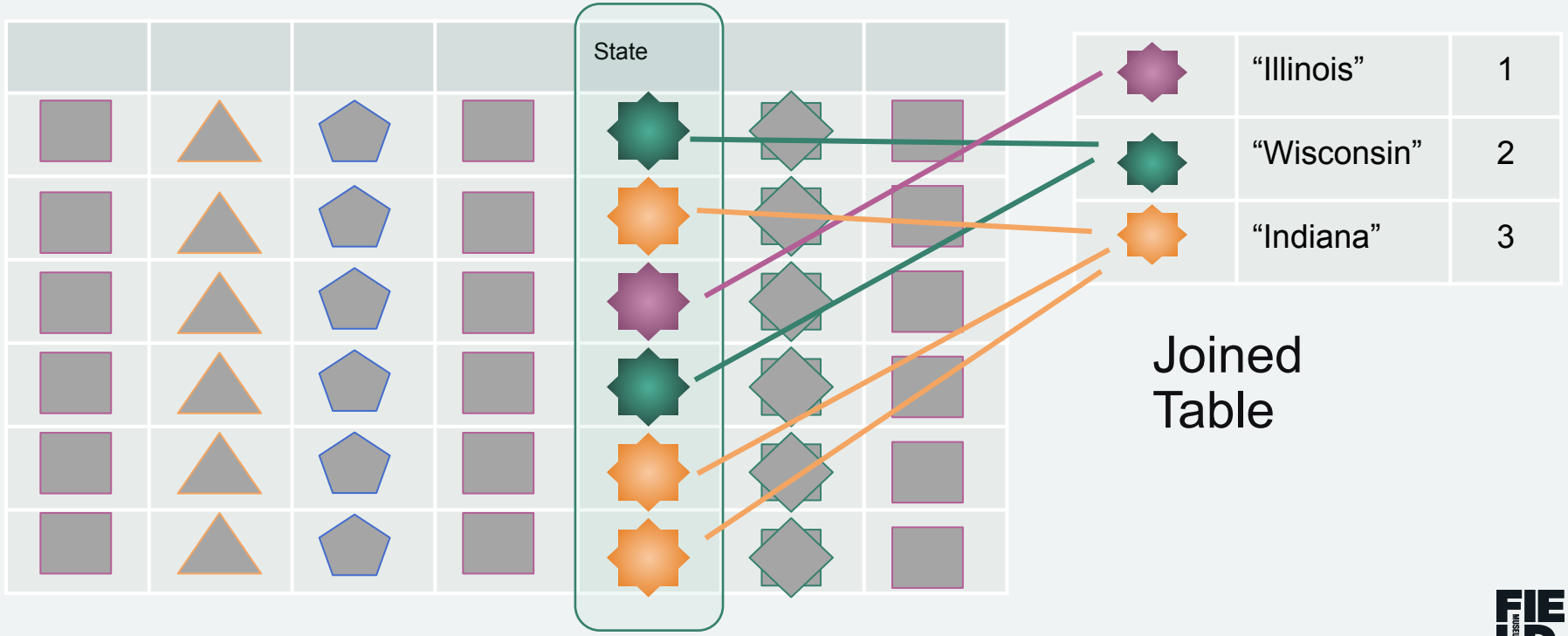
The Problem With Structured Tables

“Streamlining your data.”



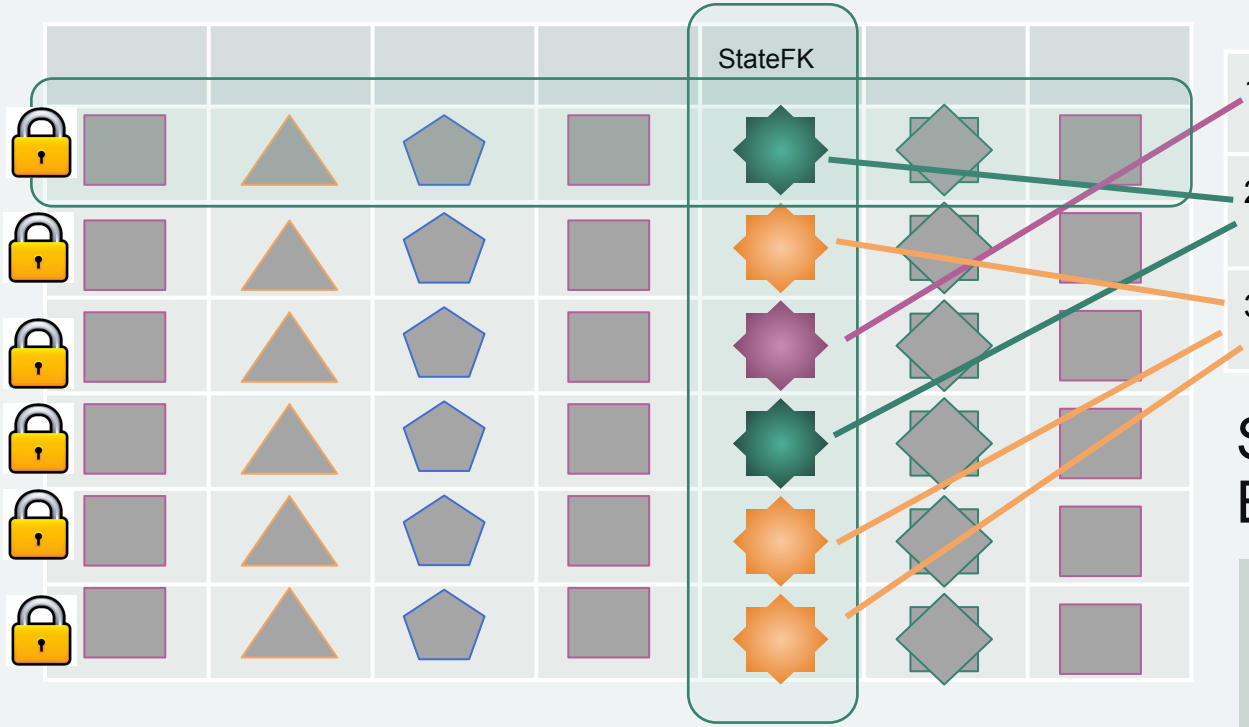
Related Data

“Streamlining your data.”



Relational Data

"What happens when you use more than one table?"



StateID			
1	"Illinois"	"IL"	"Springfield"
2	"Wisconsin"	"WI"	"Madison"
3	"Indiana"	"IN"	"Indianapolis"

Structured Table with Enforced Integrity

This is a very simple relational database. Data about States is held in a separate table to the main records and connected via a Foreign Key (an integer).

FIELD MUSEUM

Normalization

“The process used to organize a database into efficient tables and columns.”

First Normal Form (1NF)

- Remove duplicate columns
- Create separate tables for related data.
- Identify each row with a primary Key

Second Normal Form (2NF)

- Remove subsets of data for multiple rows
- Create relationships with foreign keys

Third Normal Form (3NF)

- Remove columns not dependant on the primary key

Here be Dragons.

Here be Dragons.

Mapping Data	01
Data Relationships	02
Planning…	03

Mapping Data

“Process of Identifying the start field(s) within dataset A, and its(their) corresponding field(s) in dataset B.”

ID 	Exp? 	Date 	Zip 	State 	Name 	Age 	Pink Elephant 	Notes 	Spouse Notes 
1	Y	1/1/67	60607	IL	F. Barrie	76	Blah	dom: 5/9/94	Mary Fry: 80
2	N	8/6/89	56001	WI	C. Walsh	45	Vex		dom: 1/3/08
3	N	4/12/67	53001	WI	Nick Hornby	39	Huh?		Beth Hornby

 Integer  Boolean  Date  Text  Complex

Donor Dataset

FIELD MUSEUM

Mapping Data

HERE BE DRAGONS.

"Process of Identifying the start field(s) within dataset A, and its(their) corresponding field(s) in dataset B."

ID	Exp?	Day	Month	Year	Zip 5-4	Country	State	Pink Elephant	Age	Spouse Age	Spouse Name	First Name	Last Name	Date Married
1:1	1:1	1:∞	1:∞	∞:1	1:0	0:1	1:1	1:0	1:1	∞:∞	∞:∞	1:∞	1:∞	∞:∞

Recipient Dataset

Data Relationships

“What is the relationship between the field in dataset A and dataset B?”

- Some fields will map **one-to-one** (1:1)
- Some fields will map **many-to-one** (∞:1)
- Some fields will map **one-to-many** (1:∞)

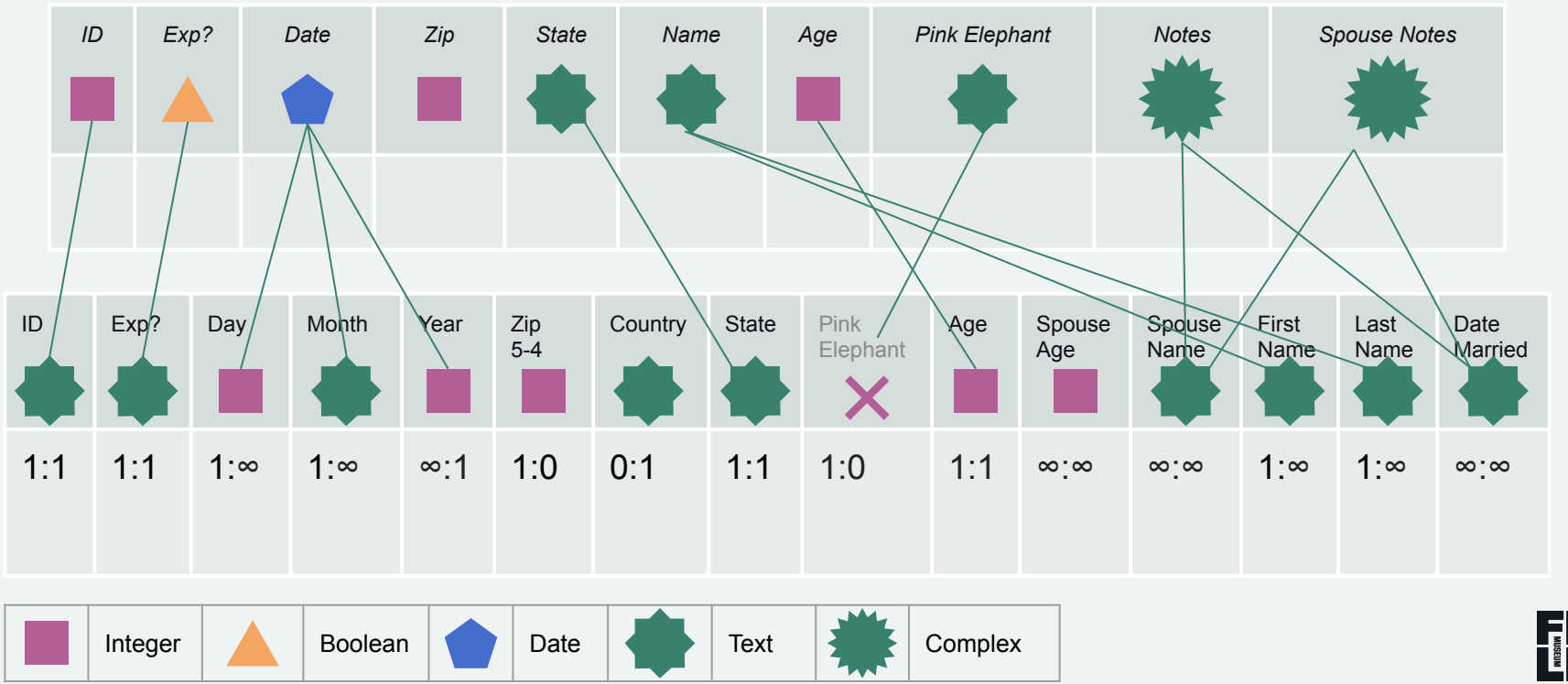
- Some fields might not exist yet: **zero-to-one** (0:1)
- Some fields might not have a place to go: **one-to-zero** (1:0) ?
- Some fields will map **many-to-many** (∞ : ∞)

Reference: https://www.w3schools.com/charsets/ref_html_utf8.asp

Mapping Data

HERE BE DRAGONS.

"Process of Identifying the start field(s) within dataset A, and its(their) corresponding field(s) in dataset B."



Planning

“Where the rubber hits the road.”

- **one-to-one** (1:1)- Beware your field types.
- **many-to-one** (∞:1)- These fields will need to be joined together.
- **one-to-many** (1:∞)- These fields will need to be split apart.

- **zero-to-one** (0:1)- You will have to work out how (or if) you can fill these in.
- **one-to-zero** (1:0)- You will have to either throw away the data or add a field.
- **many-to-many** (∞ : ∞)- “for the love of a higher being!”

Reference: https://www.w3schools.com/charsets/ref_html_utf8.asp

Questions?

Ask now; we might have answers!

Presenter name

Title or credentials

- Bio
- Contact Information

Insert Presenter Image

Thank you!

Sharon Grant
sgrant@fieldmuseum.org
https://www.fieldmuseum.org/

Janeen Jones
jjones@fieldmuseum.org
https://www.fieldmuseum.org/

Kate Webbink
kwebbink@fieldmuseum.org
https://www.fieldmuseum.org/

This project was made possible in part by the **Institute of Museum and Library Services** Grant ME-249136-OMS-21 | **IMLS.gov**