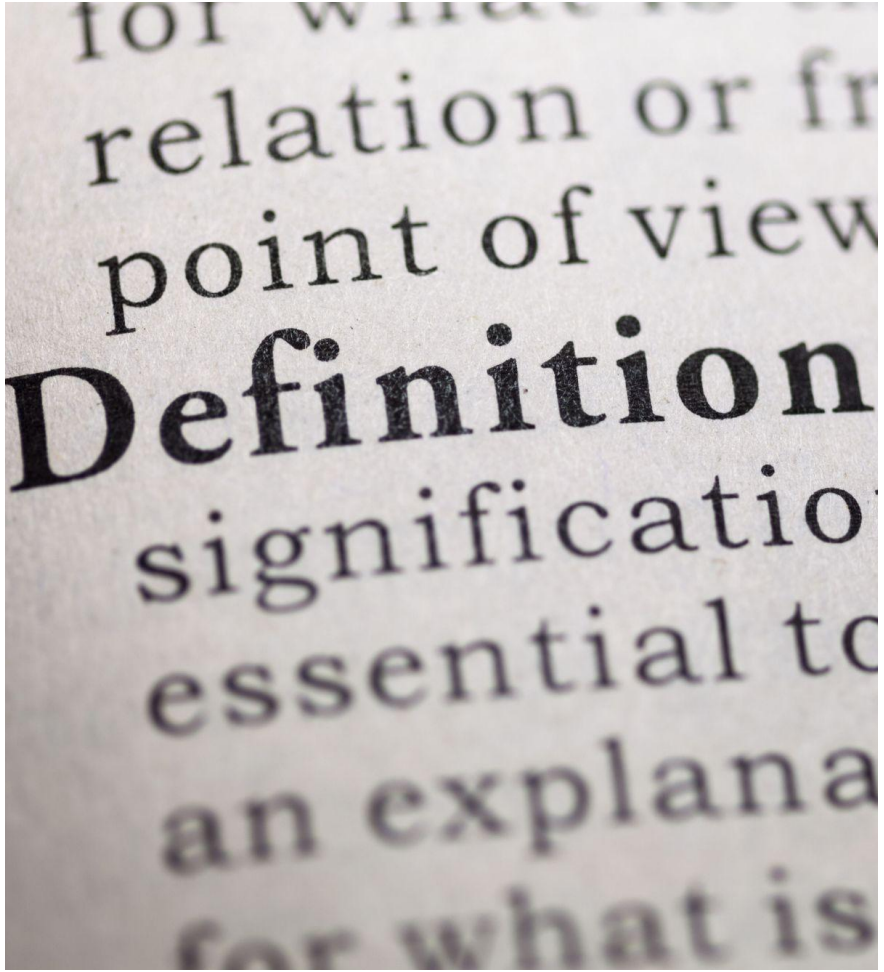
Wrapping Up

You’re almost done!



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

Agenda

Recap	01
Exercises	02
Wrap Up	03

Recap

Data Relationships

- 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 (∞ : ∞)

Field/Data Types

- **Numbers** (numeric) — Two types:
  - integer, long integer
  - float, double
- **Text** (alphanumeric):
  - character, string

- **System Value** (date):
  - date/time, date, time
- **Unstructured Text** (memo):
  - long text, longchar, blob

- **Binary** (boolean):
  - 1/0, 0/1: Yes/No, Y/N, True/False
- **Structured Text** (factor)

Overall Steps

1. Understand/Assess **Donor** (What do you have?)
2. Understand/Assess **Recipient** (What will you have?)
3. Map **Donor** to **Recipient** (How will what go where?)

4. Plan how to execute the mapping
    - Details when/how to clean **donor** fields
        - Wash/Rinse/Repeat…
            - …until each field meets the **Recipient** standards.
5. Import the data to the **Recipient**.

Steps for Planning

What do you need to do:

1. Understand the structure data:
    - Is it a grid? A structured table? Is integrity enforced?
        - What kind(s) of record(s) does a row represent?
        - Do the columns have a consistent field format? (What data-types are in each column?)

2. Understand the data:
    - What do the values themselves look like? How dirty is it?
        - Does the data in the field actually mean what you think it does?

3. Standards
    - Should the data have restrictions?
        - Check controlled vocabs, authority files…

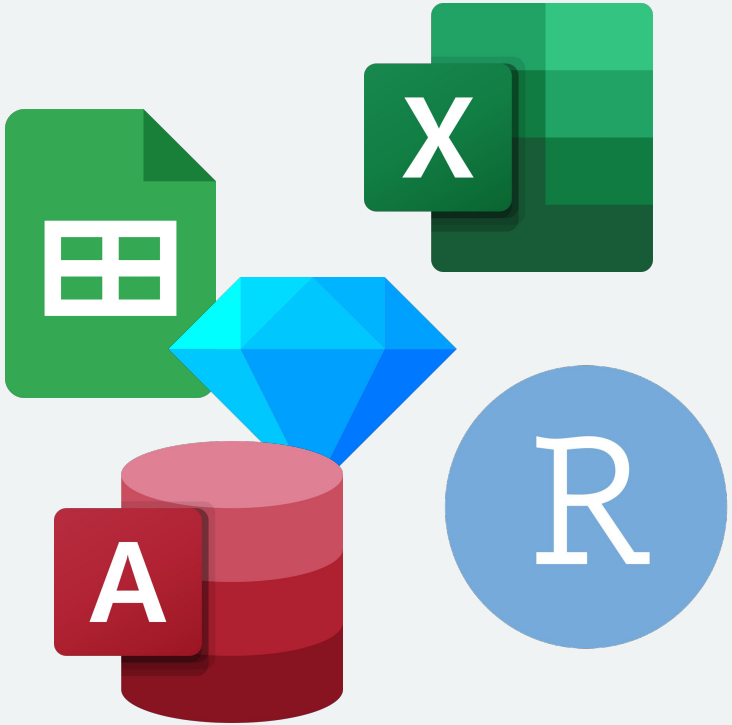
Recap

What you need to do is:

1. Understand the underlying structure of your data
2. Understand your data
3. Standards

Then, given your understanding of those:

- what tools should you use (or **not** use).
- when is a given tool appropriate/most helpful/useful/efficient (or **not**).



Tools: Do’s and Don’ts

Task	Excel	LibreOffice Calc	Google Sheets	OpenRefine	Access	R
Formulas	☑	☑	☑	☑	☑	☑
Charts & Graphs	☑	☑	☑		☑	☑
Auto-formats fields	⚠		⚠			⚠
Allow creation of lookup lists	☑	☑	☑		☑	☑
Allow character sets besides Latin1 (e.g., UTF-8)	⚠	☑	☑	☑	☑	☑
Track changes after closing	⚠	⚠	☑	☑		☑
Maintain integrity between rows and columns	⚠	⚠	⚠	☑	☑	☑
Enforce referential integrity between different sheets	⚠	⚠	⚠		☑	☑
Creates a copy of original data				☑	☑	☑
Allow you to share steps				☑		☑
Directly changes your original data					⚠	
Assists in formula writing					Limited	☑

FIELD MUSEUM

Final Exercises

Exercise 1: Migrating

Migrate the donor dataset `Day3_Dataset1.csv` to the recipient Darwin Core (DwC) spreadsheet.

- Day3_Dataset1: https://docs.google.com/spreadsheets/d/17CkJyWN7I3aF8xH8_d-o-z6kdJX9GpPcQWr5W6UJF3I/edit?usp=sharing
- DarwinCore spreadsheet: https://docs.google.com/spreadsheets/d/1BAoCE3qDgbb-n76w3Hm1ic9y0ksYhMh7MXkwtP-q-Ts/edit#gid=745189688

Preparation:

- Copy the original dataset
- Copy the recipient spreadsheet

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

Exercise 1: Migrating

Migrate the donor dataset `Day3_Dataset1.csv` to the recipient Darwin Core (DwC) spreadsheet.

1.1 Mapping: Map as much of the given donor dataset to recipient DwC sheet as you can.

- Fill in Columns A-E (see next slide)
- Don't clean it yet; just map.
- Think through the "I have this dataset field, but don't know where it goes in DwC" moments…

1.2 Planning: In your mapping table, indicate the following…

- Fill in Columns F-G (as necessary) (see next slide)
- Field types of the "original" data set, and of the final "DwC" data set. (text, numeric, etc.)
- Mapping types between corresponding "original" → "DwC" fields. (1:1, 1:many, etc.)

1.3 Execution: Using the tools and methods you've been shown, execute your plan from exercise 1.2 for the Specimen identifier's name field.

- Take the "`Identified by`" field, and prepare it according to the plan you devised to get that field into its DwC field/s.
- Save the results as a properly formatted .csv file renamed as `Day3_ex13_YOURUSERNAME.csv`

A	B	C	D	E	F	G
Original field name	**Orig. data type**	**DwC field Name**	**DwC field type**	**Relationship** *(e.g., 1:1,etc.)*	**Mapping plan/Steps**	**Questions?** *(e.g., ask the data owner)*
ADP number						
Cat Numb						
count in lot						
Specimen identifier's name						
Kingdom						
Genus						
Species						
Subspecies						
State						
City						
Collectors name						
Images						
Latitude						
Longitude						

Exercise 2: Restructuring Data

Gather and Spread

2.1 Reshaping [R/gather]

- Gather the "`Day3_Dataset1.csv`" dataset into two columns — one for `Cat Numb` and one for `Identifications` (both current and former)
- Export the results as a properly formatted .csv file renamed as `Day3_ex21_YOURUSERNAME.csv`

2.2 Reshaping [R/spread]

- Spread the "`Day3_Dataset2.csv`" (into a dataframe with four columns — one for `Cat Numb` and three for up to three References per `Cat Numb`)
- Export the results as a properly formatted .csv file renamed as `Day3_ex22_YOURUSERNAME.csv`

Find hints on the next two slides.

Exercise 2: Restructuring Data

Hints: 2.1 Reshaping [R/gather]

- Subset the fields you need.
- Make a new single column for `currentID` that concatenates the `Genus` and `Species` fields
  - With R, use the paste() function. If you want, include “Subspecies.”
  - And/or try something with more logic if you want to show “Kingdom” for unidentified specimens, e.g., set `currentID = Kingdom` for rows where `Genus == “Unidentified”`.

- Gather the identification fields (`currentID` and `Former.identification`).
- Drop any unnecessary columns.

Exercise 2: Restructuring Data

Hints: 2.2 Reshaping [R/spread]

- Drop any unnecessary columns.
- Order the data by `Cat Numb`.

- Make a new key column.
  - **Extra hint:** `sequence()` & `rle()` may be useful for this
- Spread the dataframe.

Wrap-up

Comments, Thoughts…

Complete the feedback form:

https://docs.google.com/document/d/1rp5sp7o42Lxm2ZuNLUXj5KymOsPu4mKPpEirTt6S7zs/

Post-Training Assessments

- There are two post-training assessments:
  - Fundamentals of Data Cleaning
  - Tools for Data Cleaning
- The assessments are identical to the pre-training assessments.
  - Compare your scores to measure your progress.

Post-Training Assessments

Fundamentals of Data Cleaning

- Questions in the first assessment are divided into the following two categories:
  - Vocabulary
  - Application of Vocabulary
- Questions in the second assessment are divided into the following two categories:
  - Steps
  - Planning
- Use the chart at right to obtain a proficiency level for each category:

Proficiency Level	Correct Answers
1	0-2
2	3-4
3	5-6
4	7-8

Fundamentals of Data Cleaning

Terminology and Implementation

1

Terminology

Take the Assessment

2

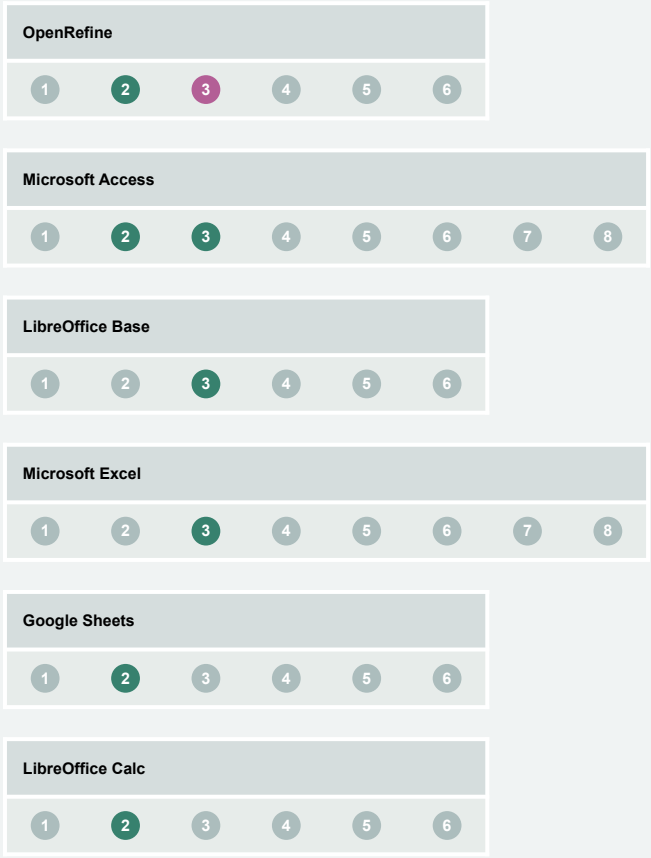
Implementation

Take the Assessment

Post-Training Assessments

Tools for Data Cleaning

- Questions in the first six assessments are divided into five categories:
  - Identifying the Appropriate Tool
  - Viewing Data
  - Cleaning and Standardizing Typos
  - Splitting and Concatenating Columns
  - Importing and Exporting
- The seventh assessment (R) functions as its own sixth category.
- Sum the number of points earned in each category across all seven assessments to obtain an overall score.
  - See the example at right and on the next slide:



7 Viewing Data questions answered correctly

Post-Training Assessments

Tools for Data Cleaning

- Questions in the first six assessments are divided into five categories:
  - Identifying the Appropriate Tool
  - Viewing Data
  - Cleaning and Standardizing Typos
  - Splitting and Concatenating Columns
  - Importing and Exporting
- The seventh assessment (R) functions as its own sixth category.
- Sum the number of points earned in each category across all seven assessments to obtain an overall score.
  - See the example at right:

7 Viewing Data questions answered correctly

Proficiency Level	Correct Answers
1	0-2
2	3-4
3	5-6
4	7-8

Viewing Data proficiency level: **4**

Tools for Data Cleaning

OpenRefine

1

OpenRefine

Take the Assessment

Tools for Data Cleaning

Database Tools

2

Microsoft Access

Take the Assessment

3

LibreOffice Base

Take the Assessment

Tools for Data Cleaning

Spreadsheet Tools

4

Microsoft Excel

Take the Assessment

5

Google Sheets

Take the Assessment

6

LibreOffice Calc

Take the Assessment

Tools for Data Cleaning

R (programming language)

7

R

Take the Assessment

FIELD MUSEUM

Thank you!

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

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

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

Abigail McArthur-Self
amcarthur-self@fieldmusuem.org
https://www.fieldmuseum.org/

Alexis Ramirez
aramirez@fieldmusuem.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**