

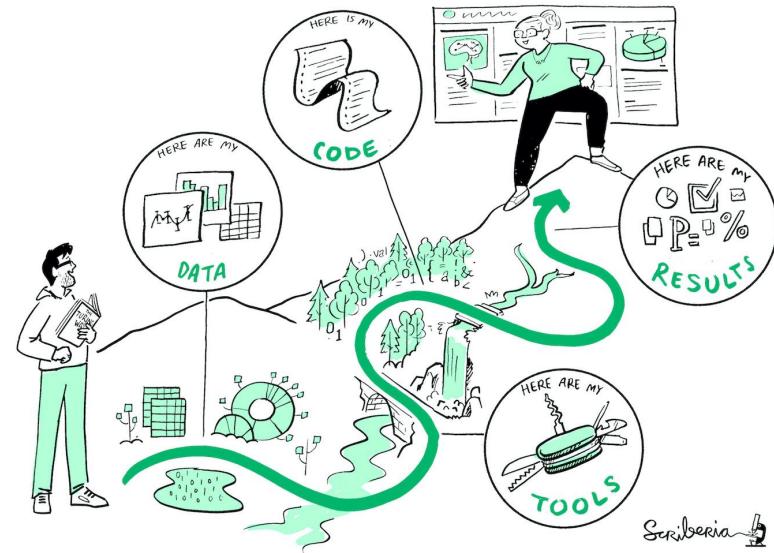


Archaeology
Data Service

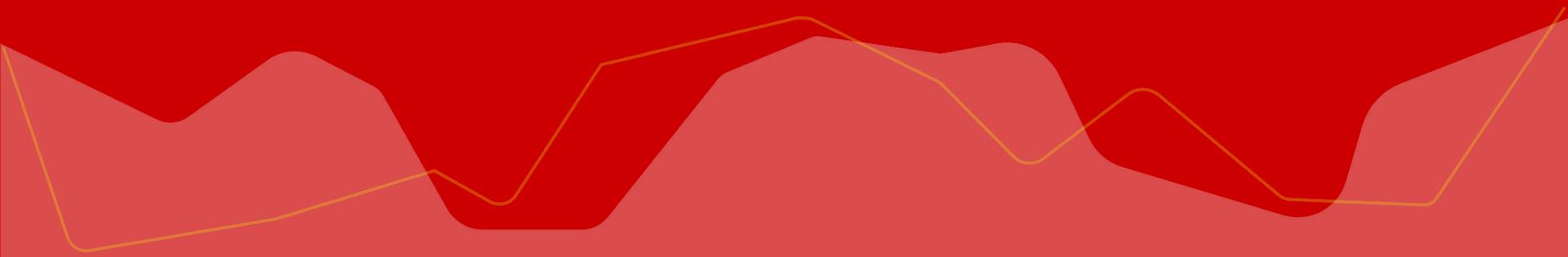
Building Clean and Archivable Data

By Jamie Geddes and Nicky Garland

The Turing Way Community. (2022). The Turing Way: A handbook for reproducible, ethical and collaborative research (1.0.2). Zenodo. <https://doi.org/10.5281/zenodo.7625728>



Today's workshop



Outline - Today's Workshop

10:10-10:35: **Section 1 - Best practice for organising your data**

10:35-10:40: Comfort Break (5 mins)

10:40-11: **Section 2 - How to transform your files in bulk**

11:00-11:10: Coffee Break (10 mins)

11:10-11:30: **Section 3 - How to wrangle your datasets (focusing on spreadsheets)**

11:30-12:00: **Work on Project/Q & A Session**

12:00: **Workshop ends**

During the session, you can add any questions or comments to the [FAQ document](#).

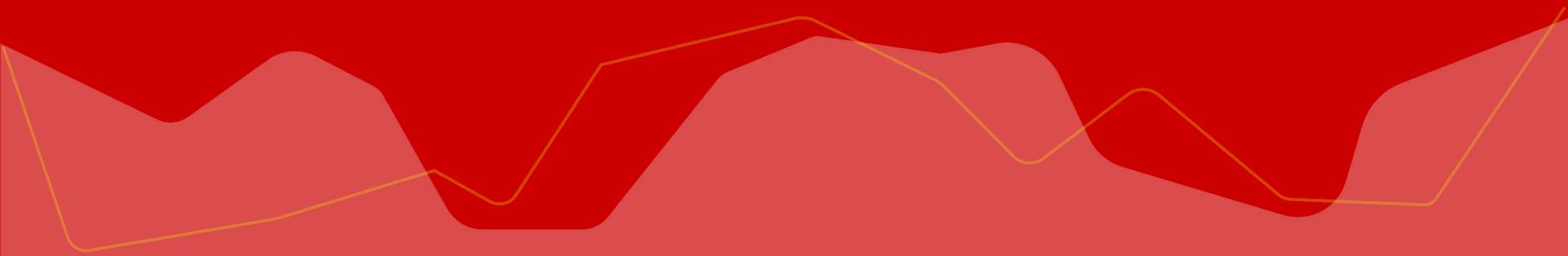


"York, King's Manor - geograph.org.uk - 5689199" by Mike Faherty is licensed under CC BY-SA 2.0.

Section 1 - Best practice for organising your data



1.1 Organising File structure



Why File Structure Matters

- **Accessibility:** A clear structure helps quickly locate files.
- **Collaboration:** Enables efficient teamwork with clear, intuitive file paths.
- **Organisation:** Organised files reduce confusion over versions and updates.



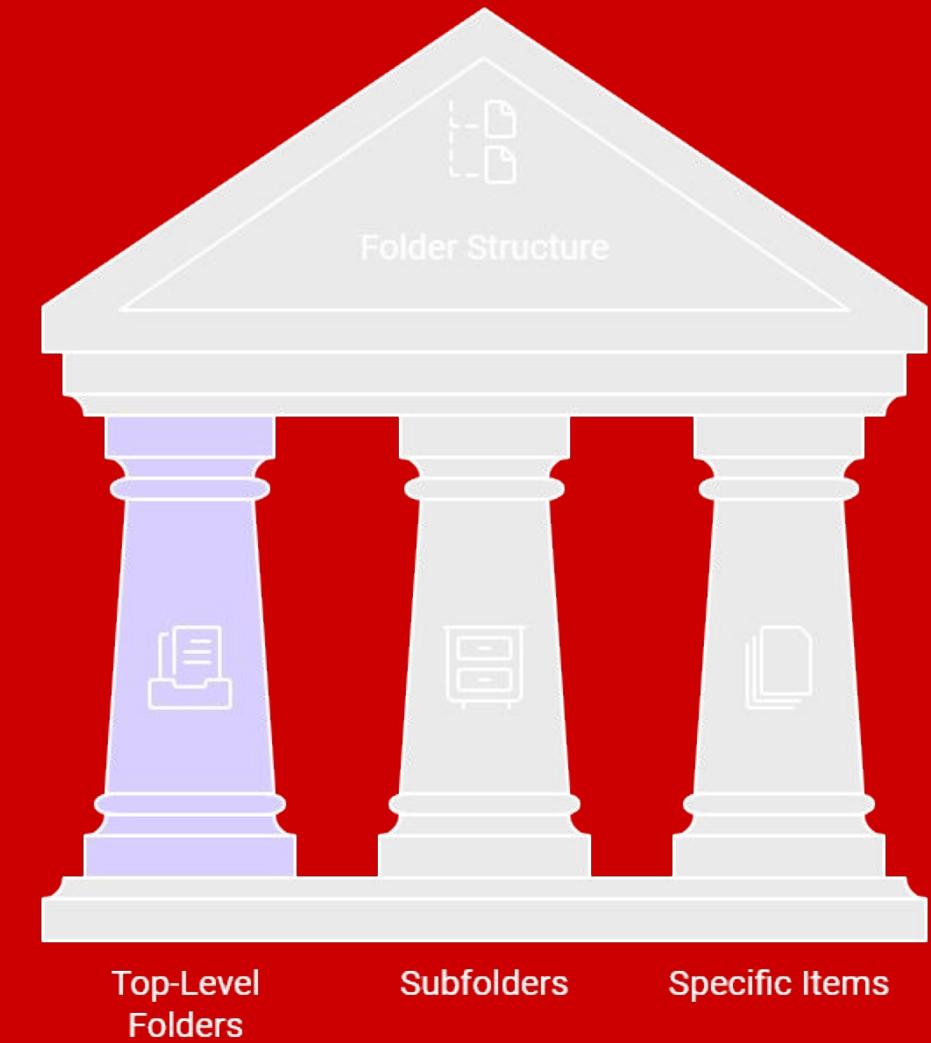
<https://doi.org/10.5281/zenodo.1114788>

What are Some Best Practices for Organising File Structures?



Define a clear Hierarchy

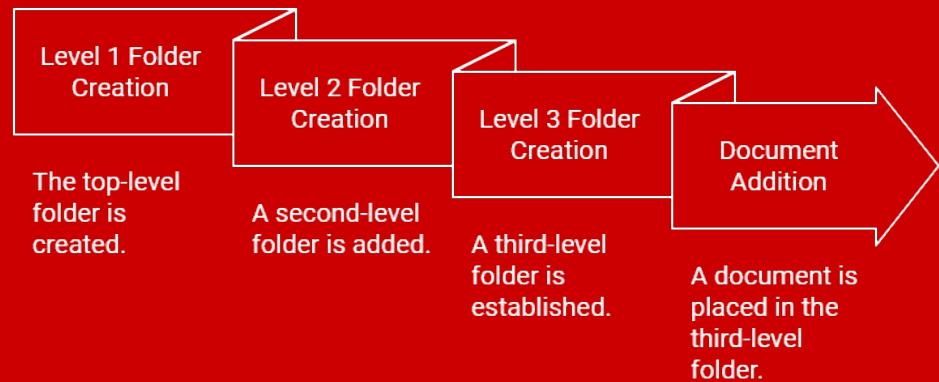
- **Top-level Folders** e.g. Project name, year, category
- **Subfolders** e.g. Images, Reports, Geospatial Data
- **Specific Items** e.g. shapefiles, pdf, xlsx



Implement Reduced File Nesting

- Use subfolders as needed, but **sparingly.**
- **Limit to three levels** of containers where possible.
- helps users **navigate to file paths easily** and find the materials they need.

Nested File Structure Sequence





Use descriptive folder names

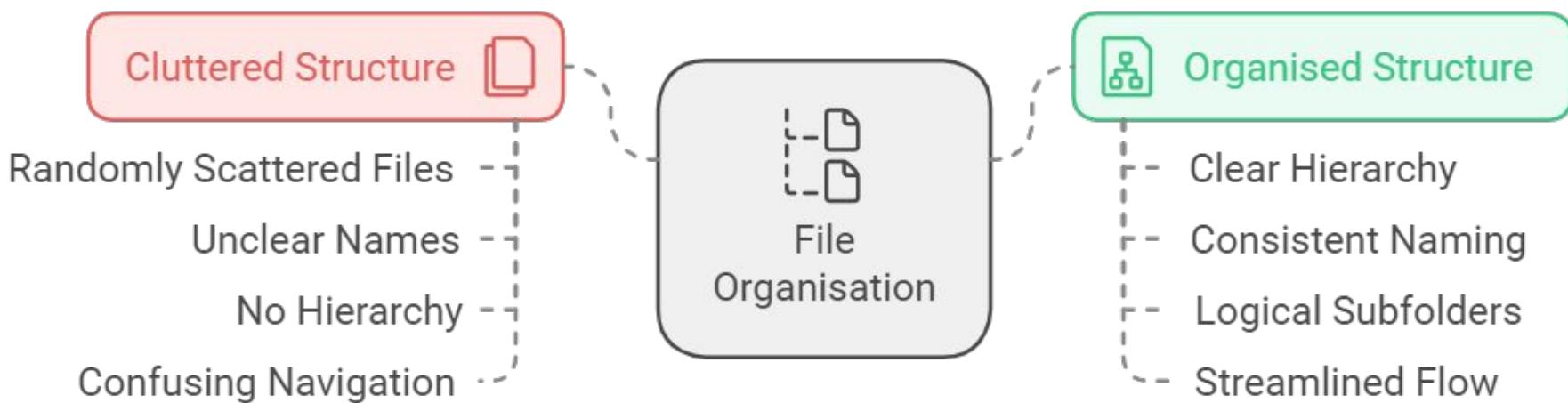
- Avoid **generic folder names** e.g.
"project1" "miscellaneous"
- Avoid **special characters and spaces** (* ! ? () \$)
- Create **versioning folders** for structure
- Apply **consistent** naming conventions

data_final 

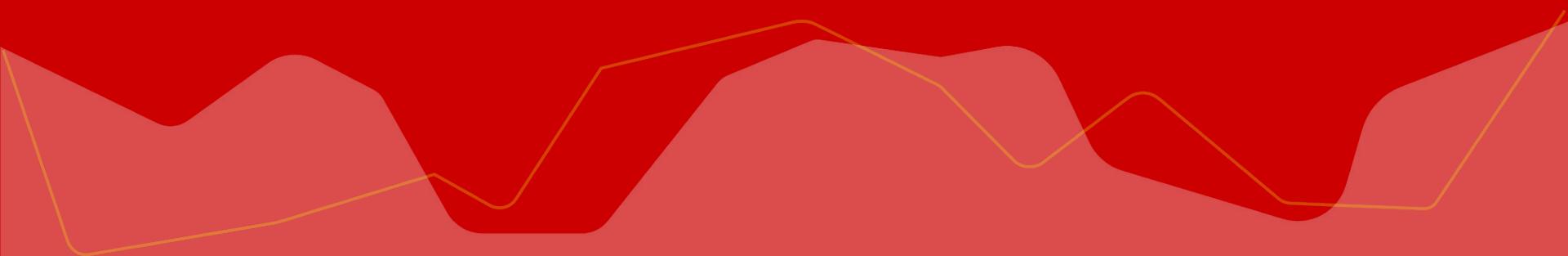
final (1) 

ProjectName_V2_2024-11-02 

Best Practices for Organising File Structures



1.2 Organising File Names



Why File Names Matter

- Main Identifier users see before access resource
- Help users quickly understand file contents, status or version
- Consistent names improve the organisation of the directory tree.



<https://doi.org/10.5281/zenodo.11147887>

File Naming Guidelines

Characters and Formatting:

- Use **alpha-numeric characters (a-z, 0-9)**, hyphens (-), and underscores (_).
- Avoid special characters and other punctuation (e.g. #, %, &).

Upper vs. Lower Case:

- Use one format consistently (e.g. all lowercase).
- Avoid mixing upper and lower case within the same name for readability.

Spaces: Use underscores instead of spaces (e.g. Project_Report_2024).

Unique and Descriptive:

- Choose names that clearly describe the file's content (e.g. Budget_Report_V1).



Good

Bad

Activity: Chaos or Clarity –File Naming Challenge

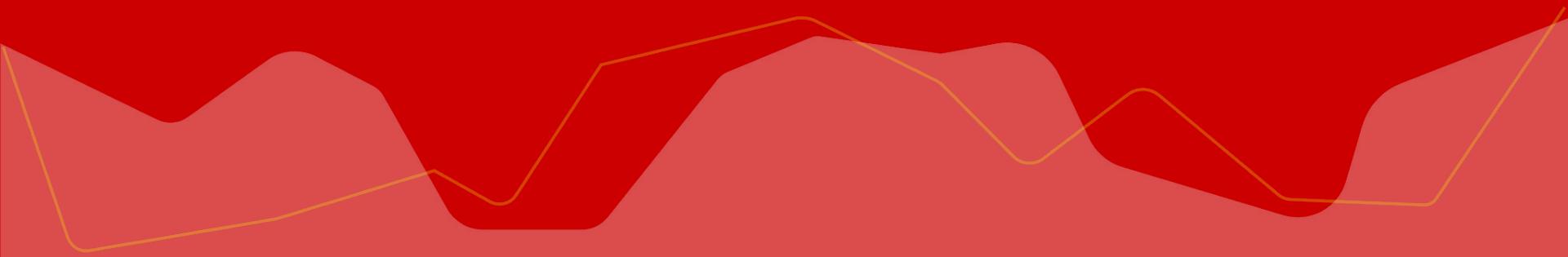
Good or **Bad** Organised Data?

Displayed will be an example directory

- Discuss if you think each folder or file is named “**Good**” or “**Bad**.”
- Suggest **Improvements** to make any “**Bad**” names clearer.



1.3 Selecting how to save your data:

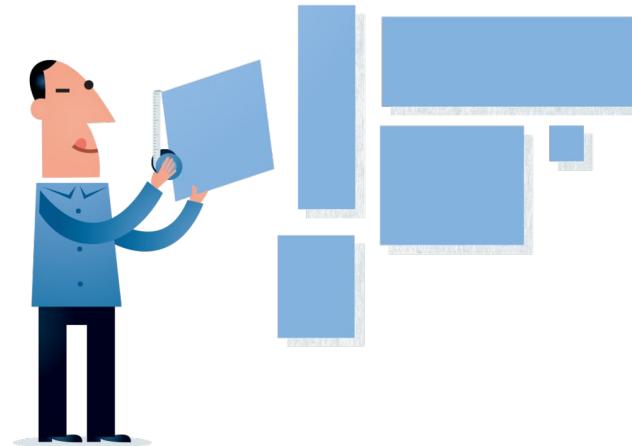


Selecting Formats

Formats selected should best **preserve the qualities of the content.**

A few things to consider:

- Open source vs proprietary
- Ubiquity (how widely used)
- Compression vs uncompressed
- Documentation and standards
- Lossless vs Lossy
- What are other similar people/orgs doing?



Proprietary Vs Open Formats

Proprietary Definition: *Formats that require specific software to open and may limit long-term access to your data.*

Risks:

- **Limited access** if software becomes outdated or unsupported.
- **Restricted sharing with collaborators** who don't have access to the required software.

Example:

- SAS7BDAT: A statistical data format only readable by SAS software, posing challenges for future data access.



Proprietary Vs Open Formats

Open Format Definition: *Non-proprietary, widely supported formats designed for accessibility and longevity.*

Benefits:

- **Long-Term Usability:** Open formats are less likely to become obsolete, ensuring that data remains accessible.

- **Cross-Compatibility:** Often compatible with multiple software options, making collaboration easier.

Example:

- TIFF: Open image format that support lossless compression, preserving data quality.





Lossy vs Lossless Data Formats Example

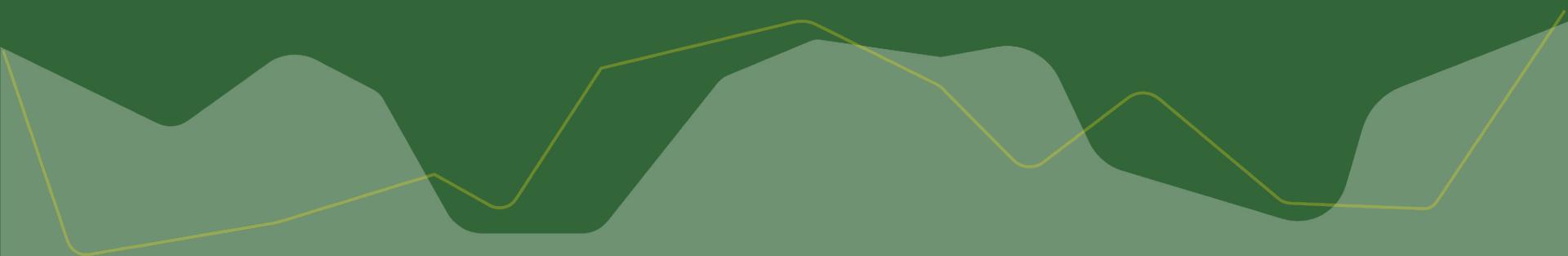
Image Format	Compression	Longevity	Compatibility
JPG/JPEG	Lossy	Low	High (open, widely supported)
TIFF (TIF)	Lossless	High	High (open, widely supported)
PNG	Lossless	High	High (open, widely supported)
GIF	Lossy	Low	High (open, widely supported, limited colors)
BMP	Lossless	Low	High (Windows-compatible, less common on Mac/Linux)
DNG	Lossless	High	Moderate (Adobe-supported, open but less common)

Resources for Selecting Formats

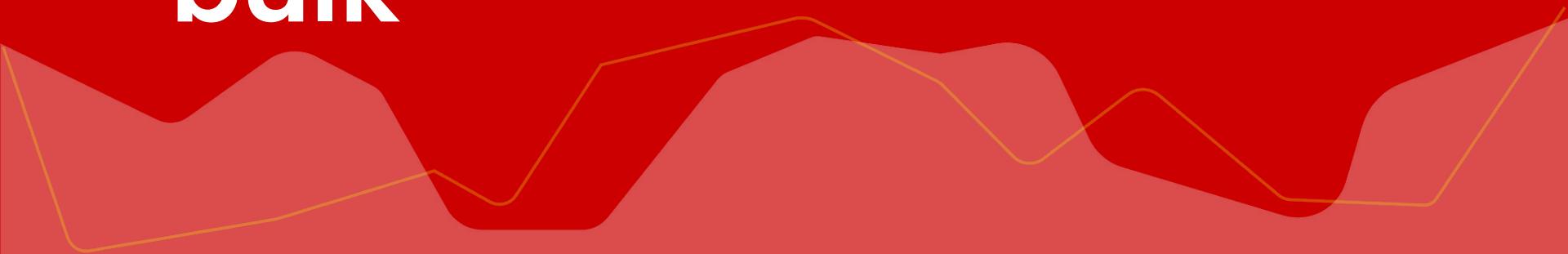
- DPC's 'Bit List' of Endangered Digital Species
- Library of Congress recommended format specifications
- OPF File Format Risk Registry
- PRONOM



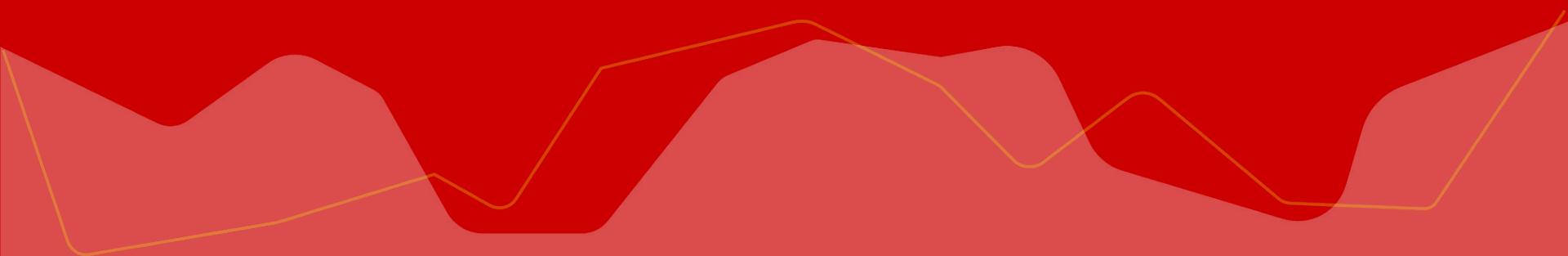
Break



Section 2 - How to transform your files in bulk



2.1 Change filenames in bulk





Bulk rename folders & files:

File renaming software:

- Windows **Bulk Rename Utility**
- Mac **Ant Renamer**

Use these tools to create a consistent and clear series of filenames

Remove special characters and spaces from filenames (* ! ? : ; " \$)

The image displays two screenshots of file renaming software. On the left is the 'Bulk Rename Utility' interface for Windows, showing a file list with columns for Name, New Name, Size, Modified, Taken (Original), Item Date, and Status. It includes a search bar, a 'Replace' dialog, and various configuration options like Match Case, Case (4), and Auto Date (8). On the right is the 'Ant Renamer 2.06' interface for Mac, showing a file list with columns for Name, Type, Creation (Crea), Mod. Date, and Mod. Time. It features a toolbar with actions like Copy, Paste, Cut, and Delete, and a sidebar for folder management.

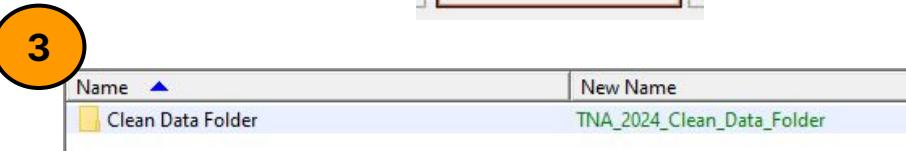
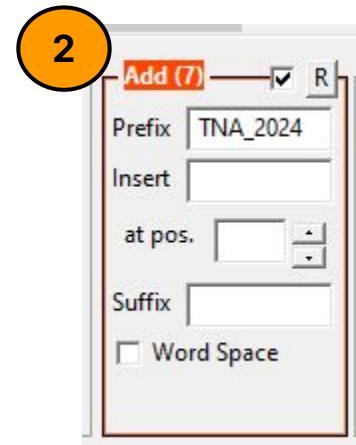
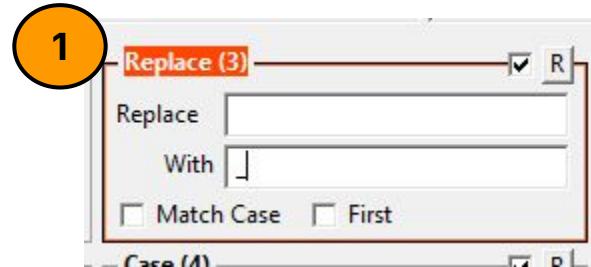
<https://www.bulkrenameutility.co.uk/>

https://www.antp.be/pic/renamer_1.png

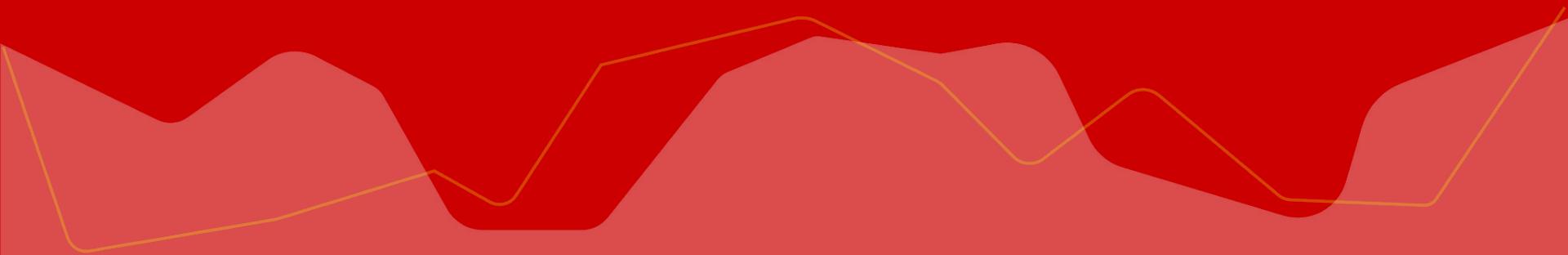
Explore the bulk rename utility tool

Task: In the windows search bar type
“Bulk Rename Utility”

- Replace Spaces in filenames
with underscores (_)
Or
- Add a prefix of “TNA_2024” to the
excel file or folder the data is
stored in



2.2 How to Convert Images





Bulk Convert Images files to a lossless open format:

Image conversion software:

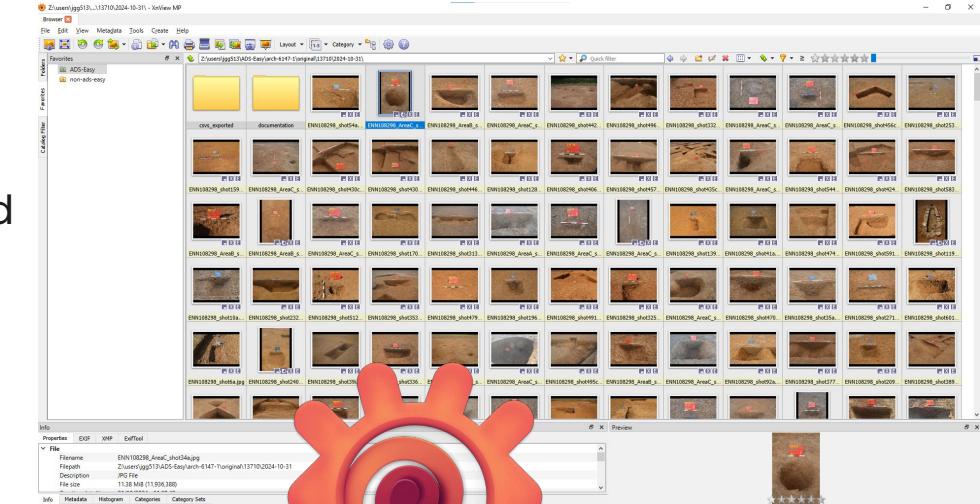
- Xnview MP

Normalise images. Migrating to a standardized format (e.g images to uncompressed TIFF)

A Persistent file format is needed to preserve data because it is expected to remain usable, reliable, and accessible over a long period of time



NORMALISATION

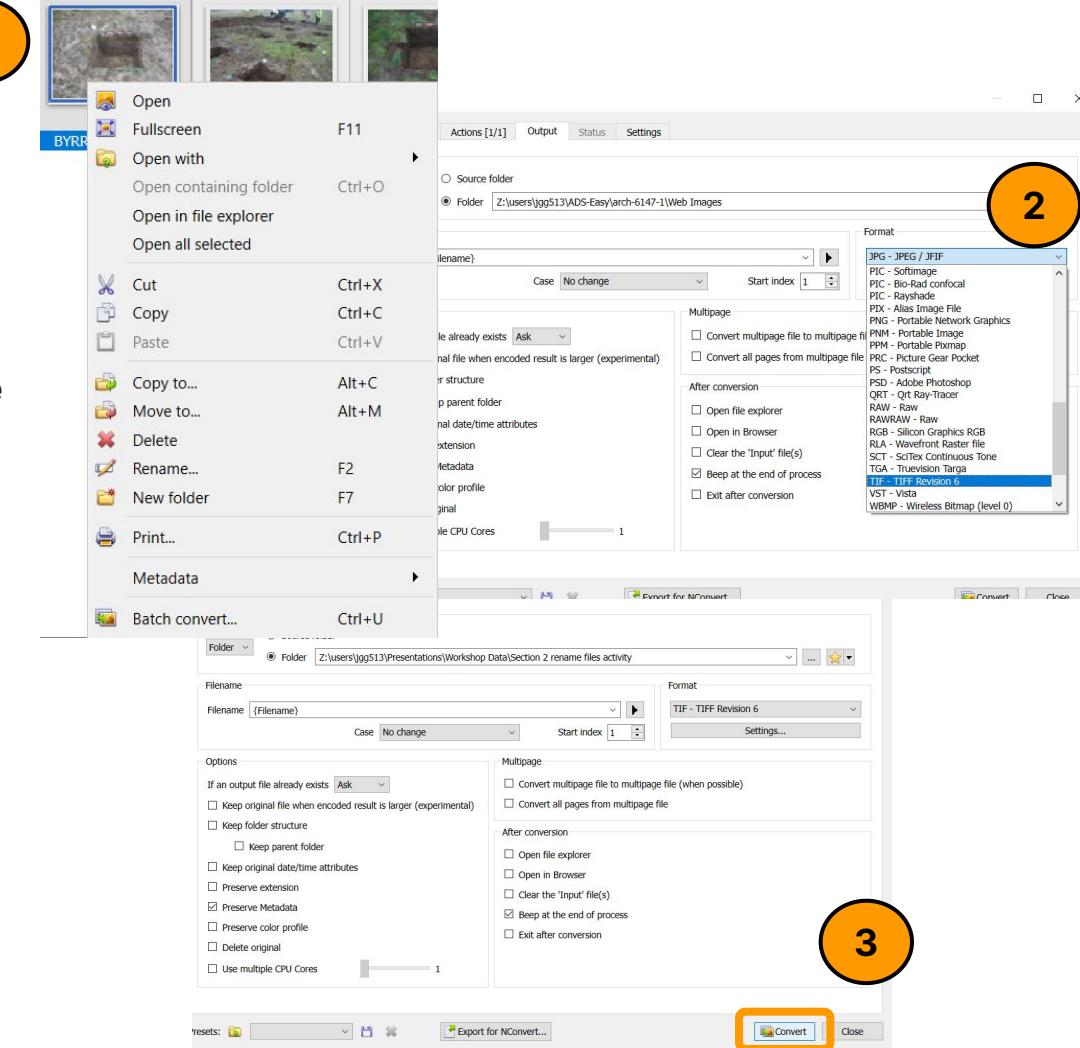


<https://www.xnview.com/en/xnviewmp/>

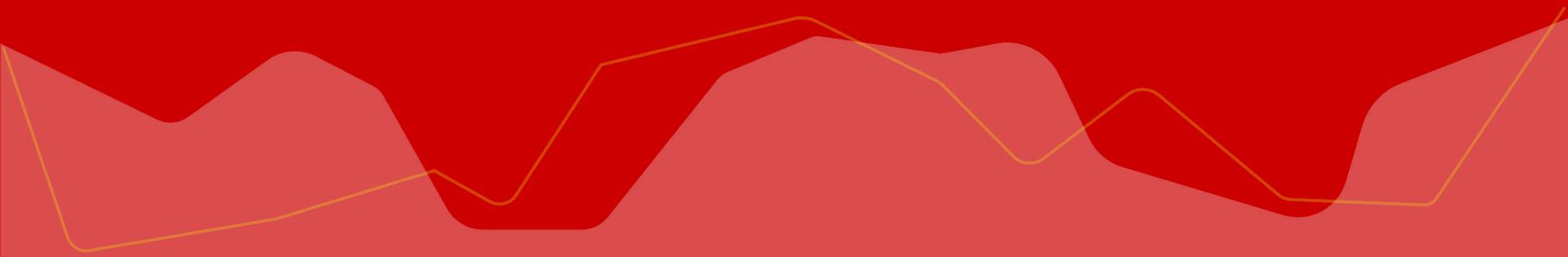
Explore XNView MP

Task: In the windows search bar type "XNView MP"

- Find the folder with test images
- Convert jpg files into TIF format
- Add a prefix "TNA" to the {filename} field see what happens when files are converted

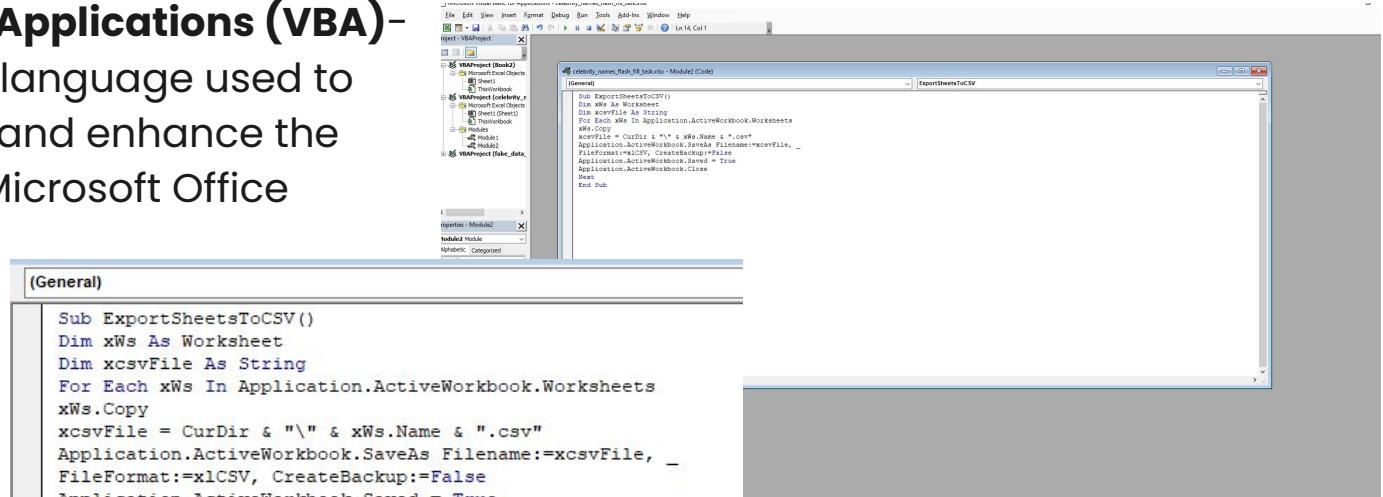


2.3 Bulk Export Spreadsheets to csv



Bulk Convert Excel sheets to CSV files:

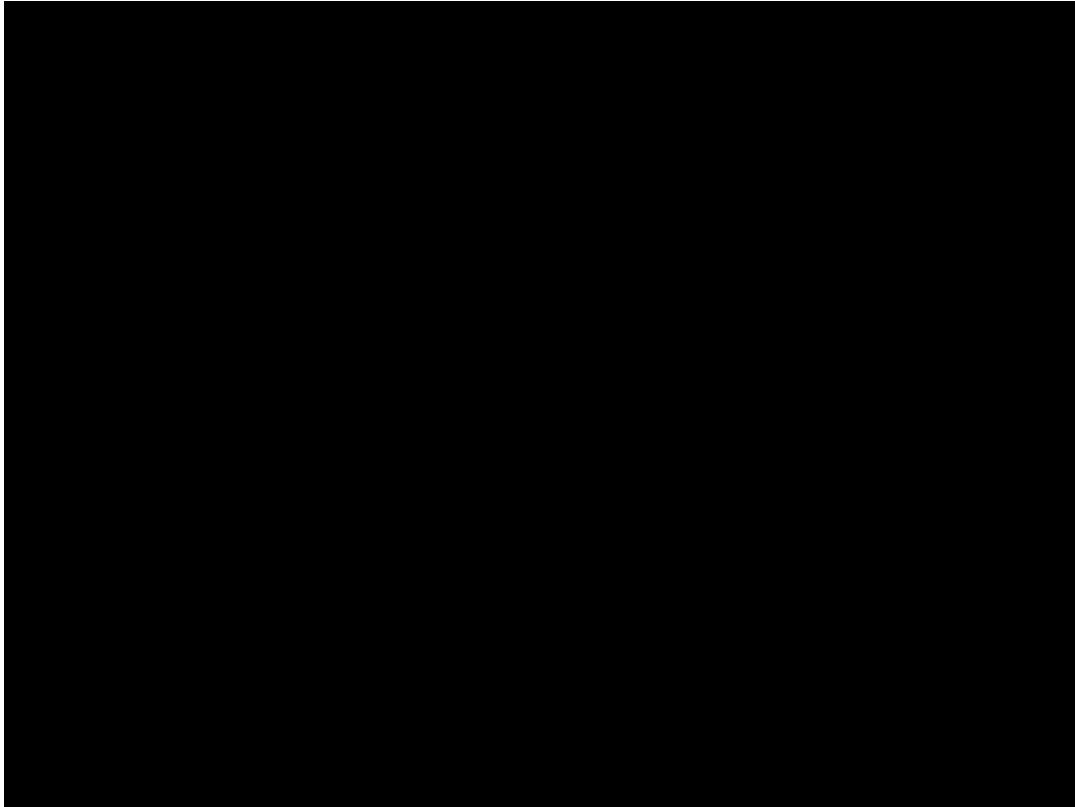
- **Visual Basic for Applications (VBA)**—
a programming language used to
automate tasks and enhance the
functionality of Microsoft Office
applications.



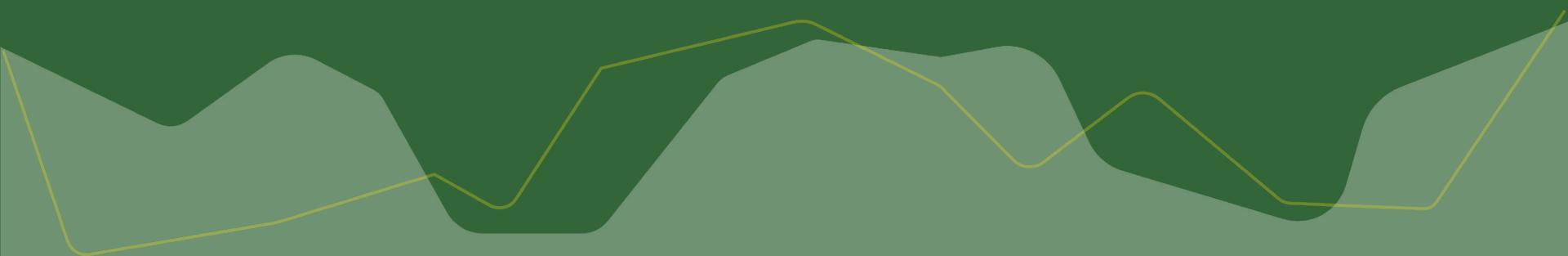
```
Sub ExportSheetsToCSV()
    Dim xWs As Worksheet
    Dim xcsvFile As String
    For Each xWs In Application.ActiveWorkbook.Worksheets
        xWs.Copy
        xcsvFile = CurDir & "\" & xWs.Name & ".csv"
        Application.ActiveWorkbook.SaveAs Filename:=xcsvFile, _
        FileFormat:=xlCSV, CreateBackup:=False
        Application.ActiveWorkbook.Saved = True
        Application.ActiveWorkbook.Close
    Next
End Sub
```

Excel VBA in use

Bulk Convert Excel sheets to CSV files:

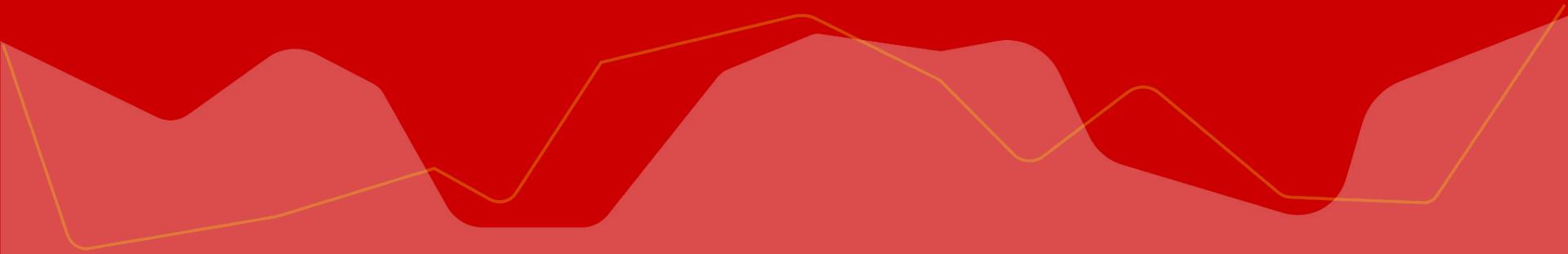


Break



Section 3 – How to wrangle spreadsheets in Excel

3.1 Useful Excel Tools and Formula



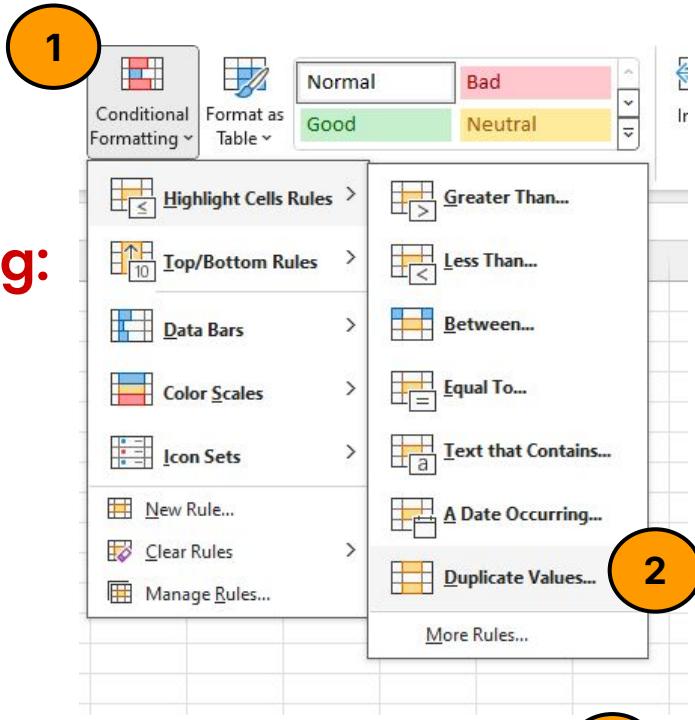
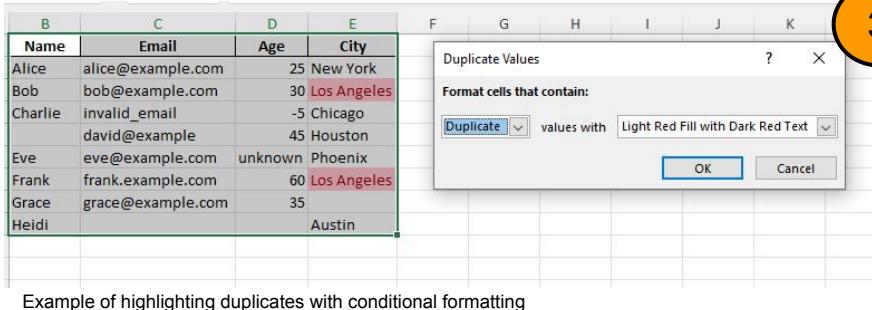
1. Conditional Formatting For Errors



<https://www.bulkrenameutility.co.uk/>

Error Detection with Conditional Formatting:

1. In the **home** tab click on the conditional formatting option
 2. Hover over highlight cell rules & Select **Duplicate Values**
 3. **Select the format** for the cells that contain duplicate
- **Task:** Highlight the CITiZAN id column and view duplicate values in the dataset with conditional formatting

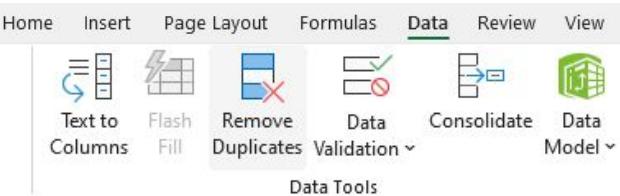
The screenshot shows a Microsoft Excel spreadsheet with a table of data. The columns are labeled B, C, D, E, F, G, H, I, J, K. The data includes columns for Name, Email, Age, and City. The 'City' column contains values like 'New York', 'Los Angeles', 'Chicago', 'Houston', 'Phoenix', and 'Austin'. A 'Duplicate Values' dialog box is open, showing the 'Duplicate' option selected under 'Format cells that contain:' and 'Light Red Fill with Dark Red Text' selected under 'values with:'. The table has a row selection for the row containing 'Los Angeles' in the City column.

Example of highlighting duplicates with conditional formatting

2. Remove duplicates in Excel



1

<https://www.bulkrenameutility.co.uk/>

Remove Duplicate Values:

1. In the Data tab click on the Remove Duplicates option
 2. Select the columns that you want to scan for duplicate entries
 3. View the removed data and check it looks okay
-
- **Task:** Remove any duplicate values with just the CITiZAN id ticked and Feature name.

The screenshot illustrates the process of removing duplicate values in Microsoft Excel. It shows three stages: 1. A table of data with 8 rows. 2. The 'Remove Duplicates' dialog box where 'Age' and 'City' columns are selected. 3. The final result showing one duplicate entry removed.

A	B	C	D	E
ID	Name	Email	Age	City
1	Alice	alice@example.com	25	New York
2	Bob	bob@example.com	30	Los Angeles
3	Charlie	invalid_email	-5	Chicago
4		david@example	45	Houston
5	Eve	eve@example.com	unknown	Phoenix
6	Frank	frank.example.com	60	Los Angeles
7	Grace	grace@example.com	35	
8	Heidi			Austin

Remove Duplicates

To delete duplicate values, select one or more columns that contain duplicates.

Select All Unselect All My data has headers

Columns

ID

Name

Email

Age

City

OK Cancel

A	B	C	D	E
ID	Name	Email	Age	City
1	Alice	alice@example.com	25	New York
2	Bob	bob@example.com	30	Los Angeles
3	Charlie	invalid_email	-5	Chicago
4		david@example	45	Houston
5	Eve	eve@example.com	unknown	Phoenix
7	Grace	grace@example.com	35	
8	Heidi			Austin

Microsoft Excel

1 duplicate values found and removed; 7 unique values remain. Note that counts may include empty cells, spaces, etc.

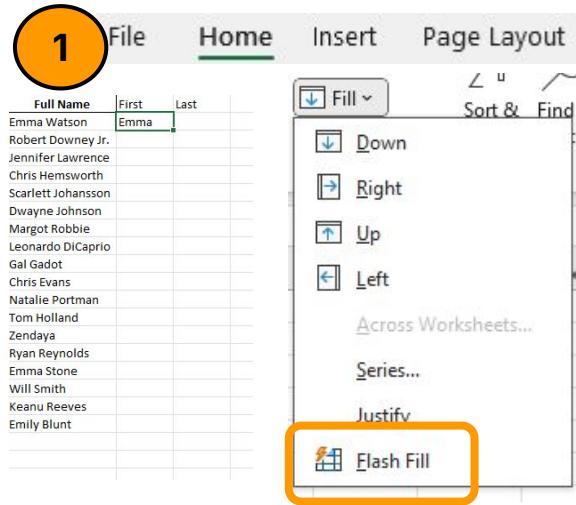
OK

3. Flash Fill Feature in Excel

Flash Fill Feature in Excel

1. Highlight your first cell entry in a column left to original data column you want to split
2. In the **home** tab click on the fill drop down and then **flash fill** (or press **ctrl+e**)
3. Check the patterned entries that appear look correct

Task: Use flash fill feature to enter TDP number into the TDP Number column from feature name e.g. **FCY02 | A102 |**



2

Full Name	First	Last
Emma Watson	Emma	
Robert Downey Jr.	Robert	Jr.
Jennifer Lawrence	Jennifer	Lawrence
Chris Hemsworth	Chris	Hemsworth
Scarlett Johansson	Scarlett	Johansson
Dwayne Johnson	Dwayne	Johnson
Margot Robbie	Margot	Robbie
Leonardo DiCaprio	Leonardo	DiCaprio
Gal Gadot	Gal	Gadot
Chris Evans	Chris	Evans
Natalie Portman	Natalie	Portman
Tom Holland	Tom	Holland
Zendaya	Zendaya	Zendaya
Ryan Reynolds	Ryan	Reynolds
Emma Stone	Emma	Stone
Will Smith	Will	Smith
Keanu Reeves	Keanu	Reeves
Emily Blunt	Emily	Blunt

An arrow points from the last row of the first table to the first row of the second table, indicating the flow of data.

Full Name	First	Last
Emma Watson	Emma	Watson
Robert Downey Jr.	Robert	Jr.
Jennifer Lawrence	Jennifer	Lawrence
Chris Hemsworth	Chris	Hemsworth
Scarlett Johansson	Scarlett	Johansson
Dwayne Johnson	Dwayne	Johnson
Margot Robbie	Margot	Robbie
Leonardo DiCaprio	Leonardo	DiCaprio
Gal Gadot	Gal	Gadot
Chris Evans	Chris	Evans
Natalie Portman	Natalie	Portman
Tom Holland	Tom	Holland
Zendaya	Zendaya	Zendaya
Ryan Reynolds	Ryan	Reynolds
Emma Stone	Emma	Stone
Will Smith	Will	Smith
Keanu Reeves	Keanu	Reeves
Emily Blunt	Emily	Blunt

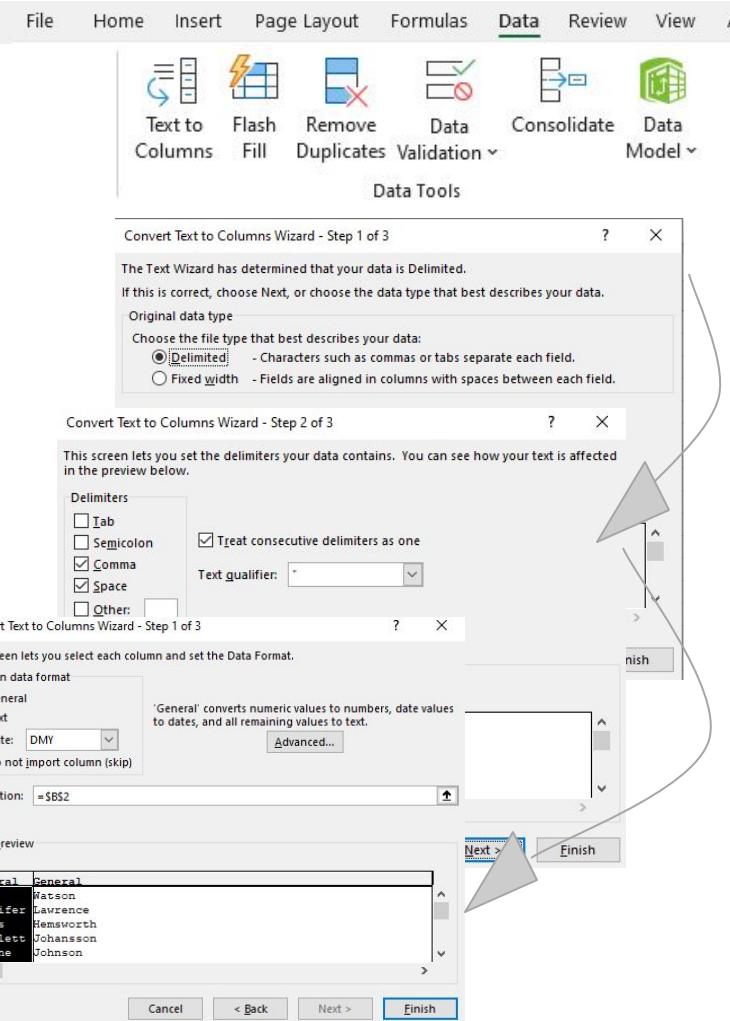
4. Text to column feature



How to convert text to column

1. Select the cell or column that contains the text you want to split.
2. Select **Data > Text to Columns**.
3. In the Convert Text to Columns Wizard, select **Delimited > Next**.
4. Select the **Delimiters** for your data. For example, Comma and Space. You can see a preview of your data in the Data preview window.
5. Select **Next**.
6. Select the **Destination** in your worksheet which is where you want the split data to appear.
7. Select **Finish**.

Task: Use text to column to remove the TDP number from the feature names



The screenshot shows the 'Convert Text to Columns Wizard' process in Microsoft Excel, spanning three windows:

- Step 1 of 3:** Shows the 'Original data type' as 'Delimited' (selected) and the 'File type' as 'Delimited'. It notes that characters like commas or tabs separate each field.
- Step 2 of 3:** Shows the 'Delimiters' section with 'Comma' and 'Space' selected. The 'Text qualifier' is set to '''. A preview of the data shows it correctly splitting the names into first and last names.
- Step 3 of 3:** Shows the 'Column data format' section with 'General' selected for all columns. The 'Destination' is set to '= \$B\$2'. A preview window shows the original names in the first column and the split names in the second column. Buttons for 'Cancel', '< Back', 'Next >', and 'Finish' are at the bottom.

5. Quickly find files and replace blank entries



Find and Replace Blank Entries

Method 1:

1. In the **home** tab click on the **find & select** option (or press **ctrl+g**)
2. Select “**go to special**”
3. Select the “**Blanks**” option. All blanks will be highlighted in the sheet
4. Type **NULL** into highlighted cell & press **ctrl + enter**
5. All blank fields will be updated to **NULL**

The image shows a Microsoft Excel interface with several numbered callouts:

- 1**: A callout pointing to the 'File' tab in the top ribbon.
- 2**: A callout pointing to the 'Find & Select' dropdown in the 'Editing' group of the ribbon.
- 3**: A callout pointing to the 'Blanks' option in the 'Select' section of the 'Go To Special' dialog.
- 4**: A callout pointing to a table where the fourth row's 'Name' cell contains 'NULL'.
- 5**: A callout pointing to the bottom right corner of the table, indicating the keyboard shortcut 'Ctrl + Enter'.

Find & Select Dialog:

- Find & Select
- Sensitivity
- Add-ins
- Find...
- Replace...
- Go To...
- Go To Special...

Go To Special Dialog (Callout 3):

Select

- Notes
- Constants
- Formulas
- Numbers
- Text
- Logical
- Errors
- Blanks
- Current region
- Current array
- Objects

? OK Cancel

Table Data (Callout 4):

ID	Name	Email	Age	City
1	Alice	alice@example.com	25	New York
2	Bob	bob@example.com	30	Los Angeles
3	Charlie	invalid_email	-5	Chicago
4	NULL	david@example	45	Houston
5	Eve	eve@example.com	unknown	Phoenix
7	Grace	grace@example.com	35	
8	Heidi			Austin

Enter NULL into highlighted box

Ctrl + Enter →

Find and Replace Blank Entries

Method 2:

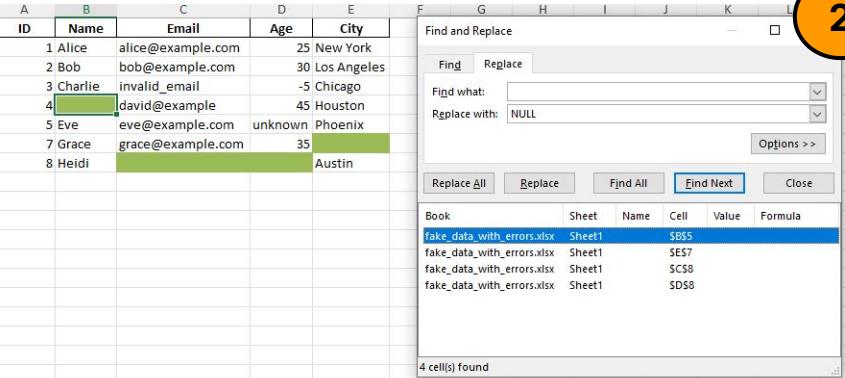
1. Highlight the data and press **ctrl + H**
2. Enter into the replace with box **NULL**
3. Select **Replace All** button
4. A pop up box will appear with the amount of blank entry replacements made

Task: Using either method 1 or method 2, find all blank entries and replace them with “**NULL**” or “**Blank**” values.

1

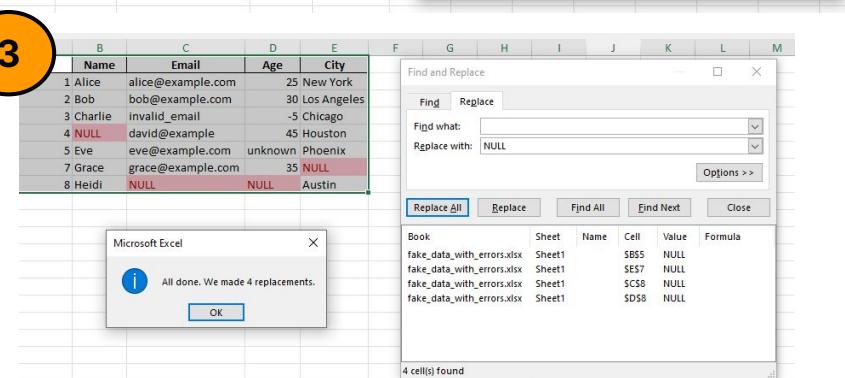
A	B	C	D	E
ID	Name	Email	Age	City
1	Alice	alice@example.com	25	New York
2	Bob	bob@example.com	30	Los Angeles
3	Charlie	invalid_email	-5	Chicago
4		david@example	45	Houston
5	Eve	eve@example.com	unknown	Phoenix
7	Grace	grace@example.com	35	
8	Heidi		Austin	

2



3

A	B	C	D	E
ID	Name	Email	Age	City
1	Alice	alice@example.com	25	New York
2	Bob	bob@example.com	30	Los Angeles
3	Charlie	invalid_email	-5	Chicago
4	NULL	david@example	45	Houston
5	Eve	eve@example.com	unknown	Phoenix
7	Grace	grace@example.com	35	NULL
8	Heidi	NULL	NULL	Austin



Example of removing blank entries with find & replace tool

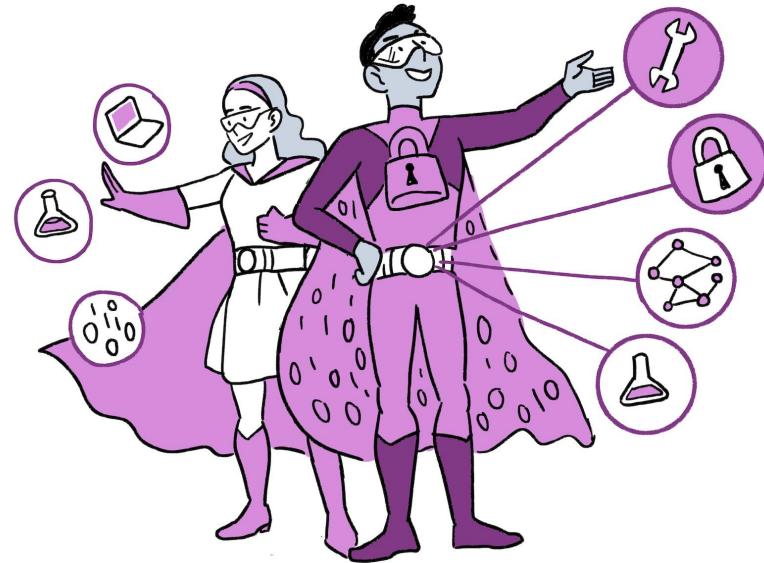
6. Useful formula



Useful formula

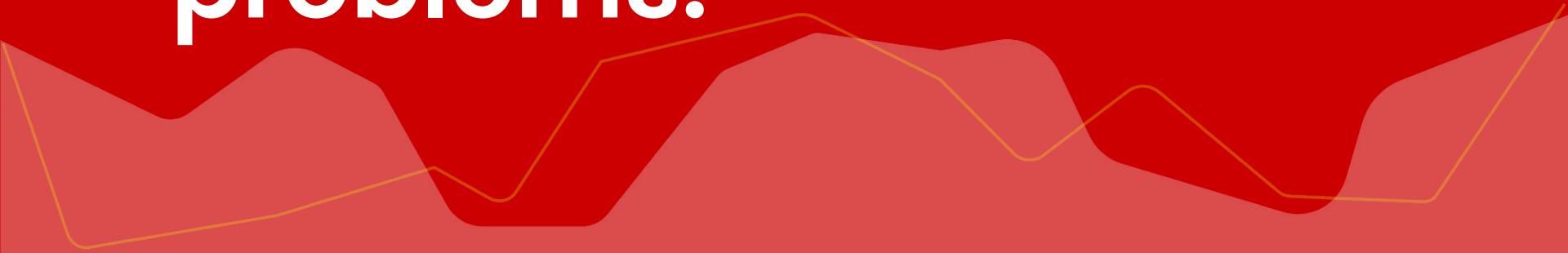
Extra formula to look into for your projects datasets :

- =TRIM(A1) removes extra spaces.
- =CLEAN(A1) removes non-printable characters.
- =SUBSTITUTE(A1, "old_text", "new_text") replaces specific characters or words.



Scriberia

3.2 Data Format problems:





Automatic Type Casting in Excel

Excel automatically guesses the data type (e.g., text, number, date) for cells, known as **automatic type casting**.

Common Issues:

- **Date Misinterpretation:** Numbers or codes can unintentionally become dates, altering data accuracy.
- **Loss of Leading Zeros:** IDs or postal codes may lose leading zeros if interpreted as numbers.

Best Practice:

- **Pre-Set Cell Formats:** Before entering data, specify the correct format (Text, Number, Date) to prevent Excel from changing data types.
- **Verify Data:** Regularly review data entries for unexpected conversions.

Entered Value	Auto-Cast by Excel	Intended Data Type	Issue
2023-10	2023-10-01	Text	Auto-cast to date
ID01234	ID01234	Text	No change - as intended
00345	345	Text	Leading zeros removed
05-20	2020-05-20	Text	Auto-cast to date



Managing Missing Data

The Issue:

- Placeholder values like `-999`, `999`, or `0` are sometimes used to represent missing data.
- Software may interpret these values as real data rather than nulls, leading to inaccurate analyses.

Solution:

- **Use Consistent Null Indicators:** Choose a standardised, easily identifiable placeholder, or use explicit null markers if your software supports them.
- **Verify with Software Settings:** Check if the analysis software you're using interprets your null indicator correctly.

Best Practice:

- Document and clearly define your null indicator for team members and future data use.

Null Values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		NEVER use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL, Excel	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software.		Avoid
NULL	Can cause problems with data type.	SQL	Good option
None	Uncommon. Can cause problems with data type.	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space.		Avoid
Missing	Uncommon. Can cause problems with data type.		Avoid
-, +, ..	Uncommon. Can cause problems with data type.		Avoid

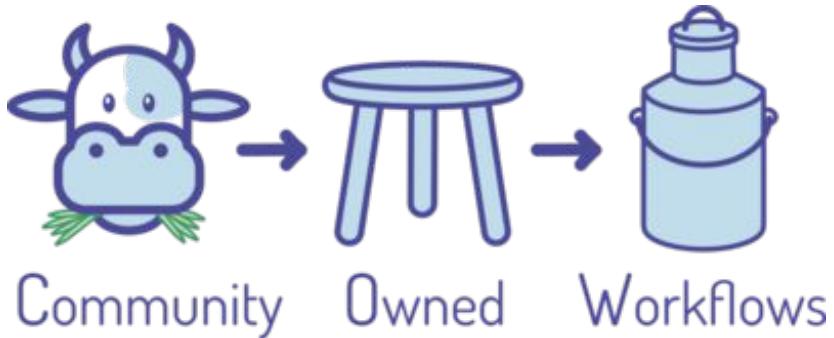
<https://datacarpentry.org/spreadsheets-socialsci/aio.html#null>

(BONUS) Use Openrefine
to wrangle data

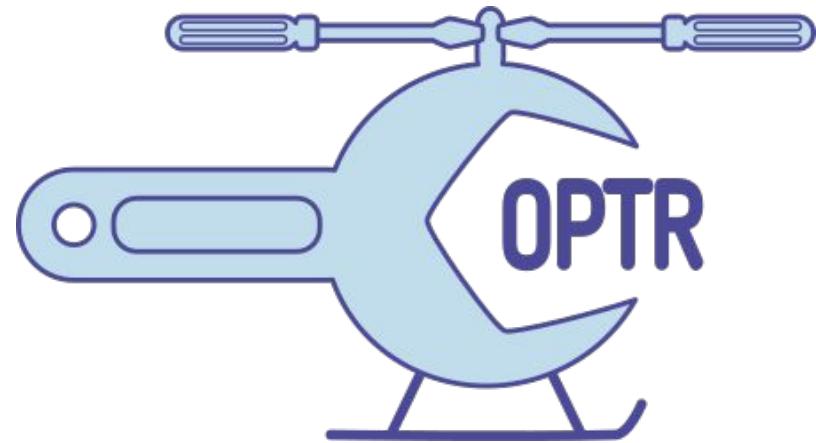
Q & A/ Project Session



Useful Resources for software



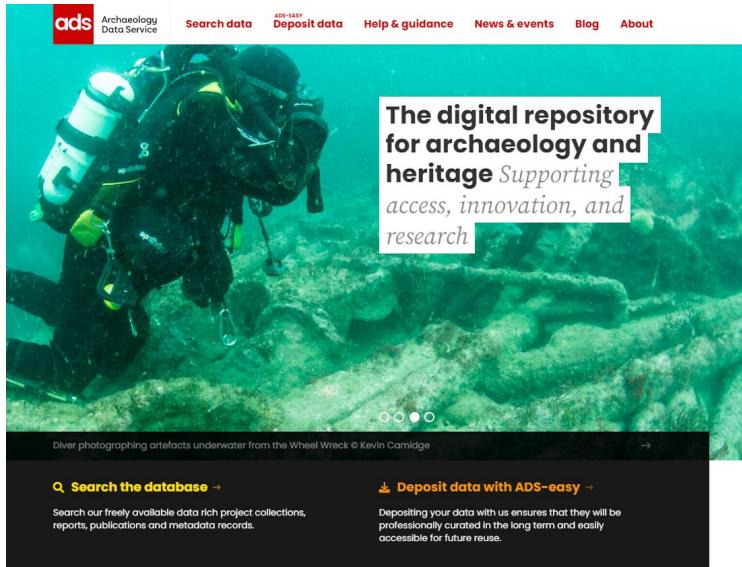
Community Owned
Workflows (COW)



Community Owned Digital
Preservation Tool Registry (COPTR)

Keep informed!

- ADS Website
 - [News and Events](#), [Blog](#)
- ADS Newsletter ([info here](#))
- Social media



The screenshot shows the homepage of the Archaeology Data Service (ADS) website. At the top, there's a red header bar with the 'ads' logo, the text 'Archaeology Data Service', and links for 'Search data', 'ADS-EASY Deposit data', 'Help & guidance', 'News & events', 'Blog', and 'About'. Below the header is a large image of a scuba diver underwater, examining artifacts on a shipwreck. To the right of the image is a white box containing the text: 'The digital repository for archaeology and heritage Supporting access, innovation, and research'. Below the image, there's a caption: 'Diver photographing artefacts underwater from the Wheel Wreck © Kevin Camidge'. At the bottom of the page, there are two columns of links: 'Search the database' and 'Deposit data with ADS-easy'.

nicky.garland@york.ac.uk



Archaeology
Data Service

Thank you!

jamie.geddes@york.ac.uk
nicky.garland@york.ac.uk



@Nicky_Garland

@ADS_Update



Archaeology Data Service
Department of Archaeology
University of York
The King's Manor
Exhibition Square
York, YO1 7EP



www.archaeologydataservice.ac.uk



help@archaeologydataservice.ac.uk