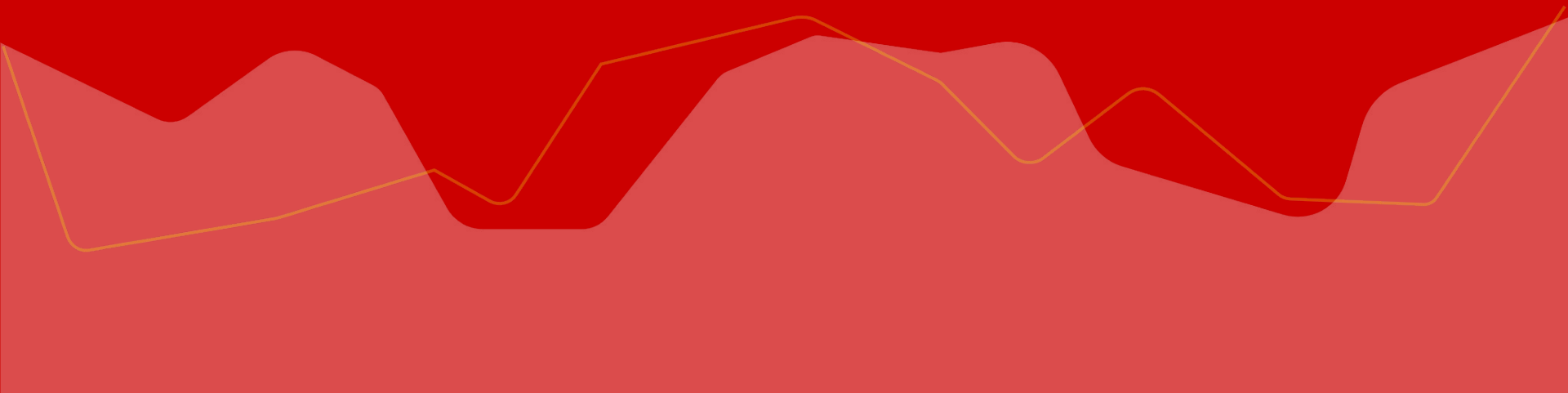


# Building Clean and Archivable Data

*By Jamie Geddes and Nicky Garland*



# 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**

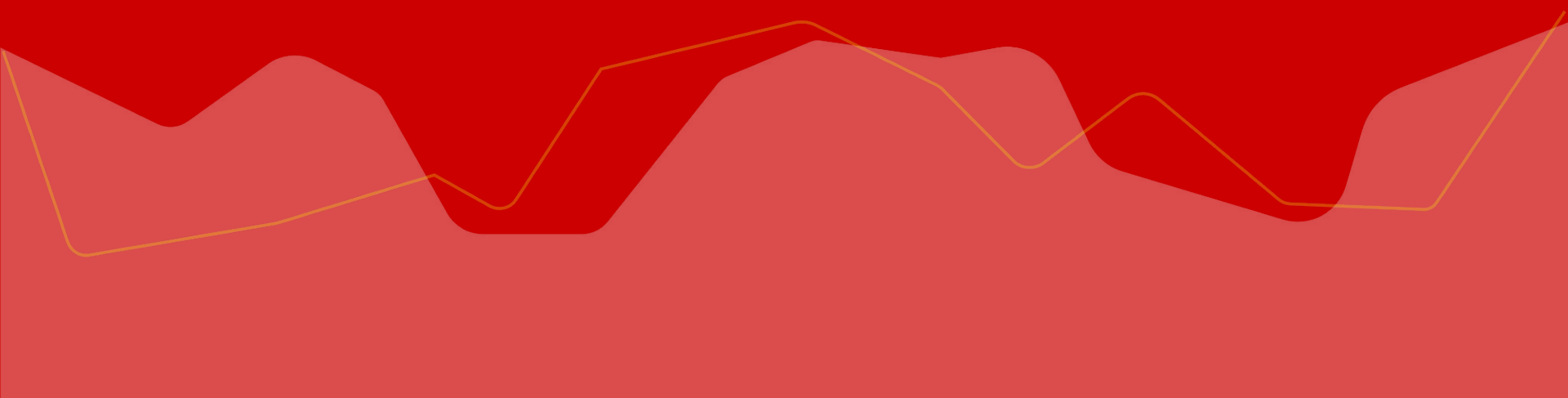
During the session, you can add any questions or comments to the [FAQ document](#).



# Section 1 – Best practice for organising your data

The background of the slide is a solid red color. At the bottom, there is a decorative graphic consisting of a yellow line graph with several peaks and valleys, overlaid on a semi-transparent red area that follows the same jagged shape as the line.

# 1.1 Organising File structure

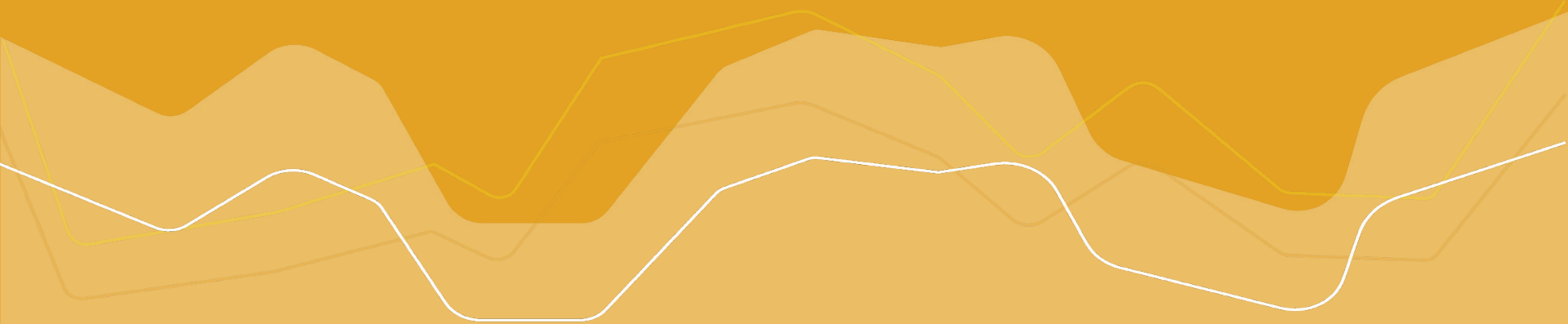


## Why File Structure Matters

- **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.

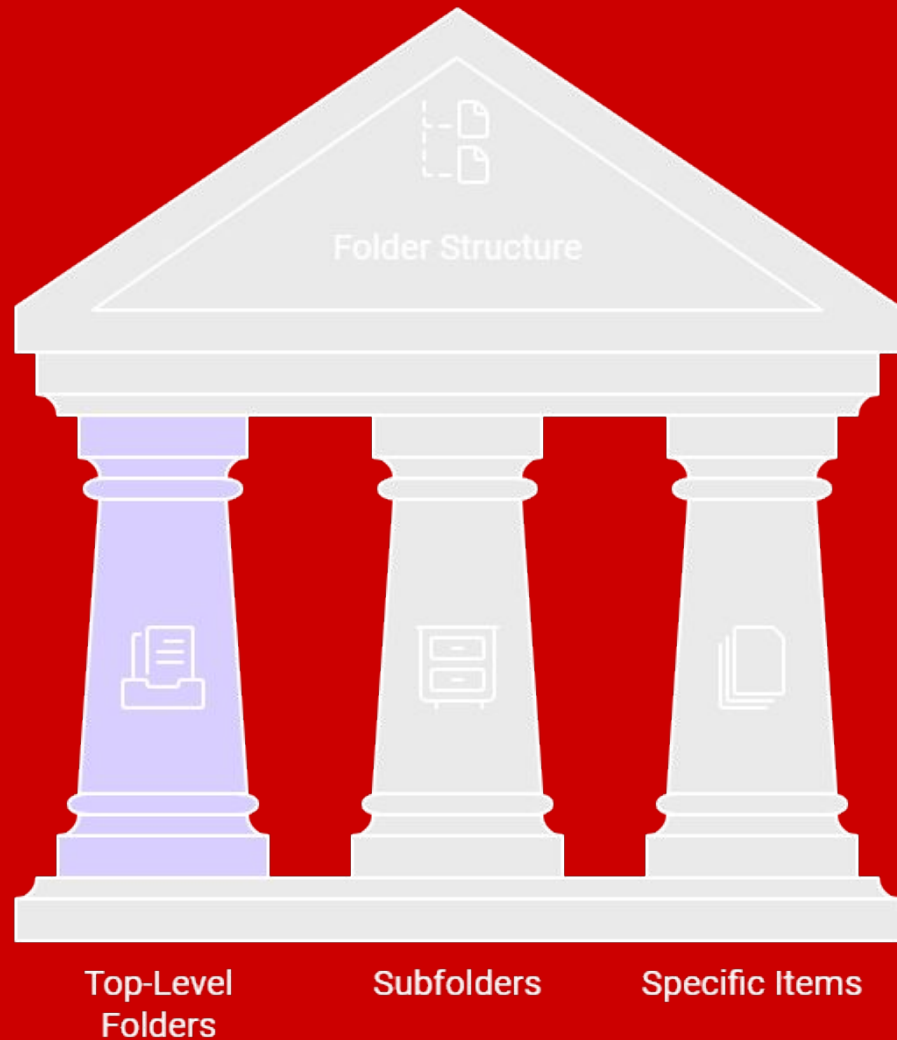


# What are Some Best Practices for Organising File Structures?



## 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

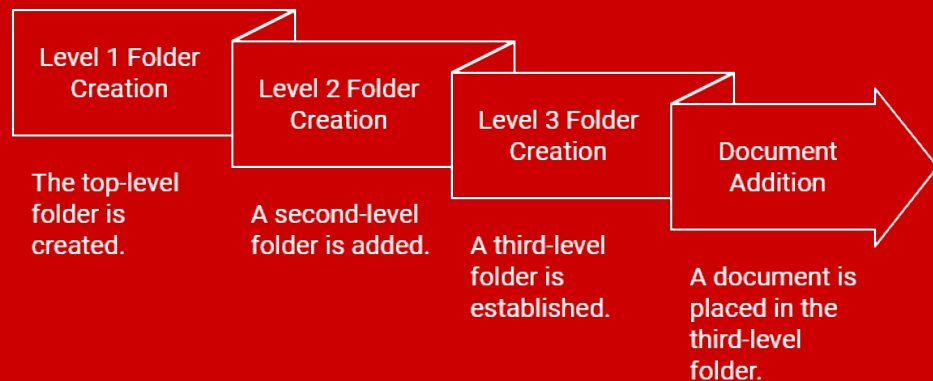




## Implement Reduced File Nesting

- 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

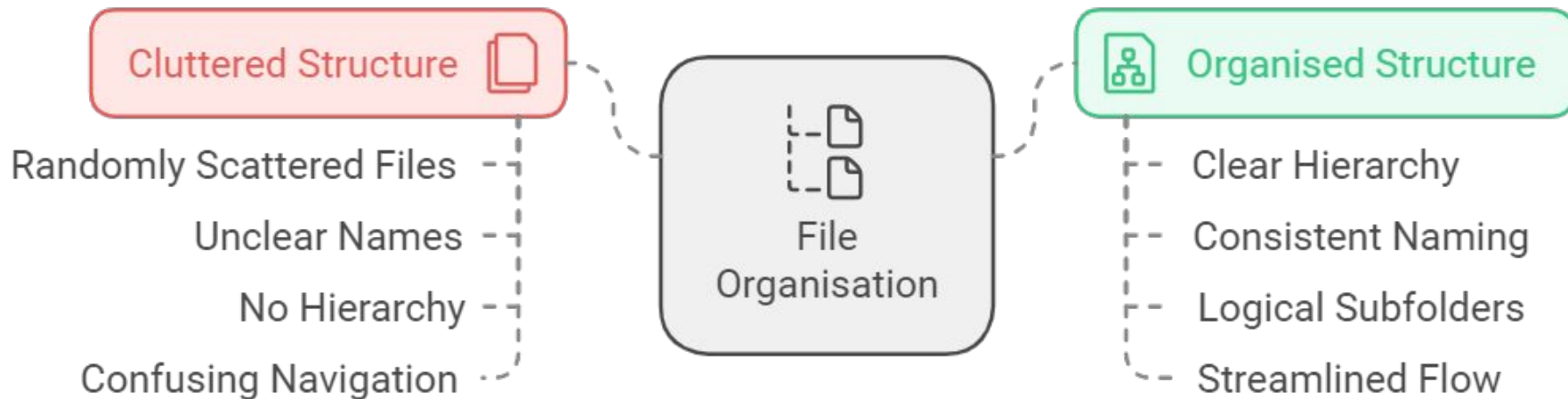
- Avoid **generic folder names** e.g. *“project1” “miscellaneous”*
- Avoid **special characters and spaces** (\* ! ? ( ) \$)
- Create **versioning folders** for structure
- Apply **consistent** naming conventions

data\_final 

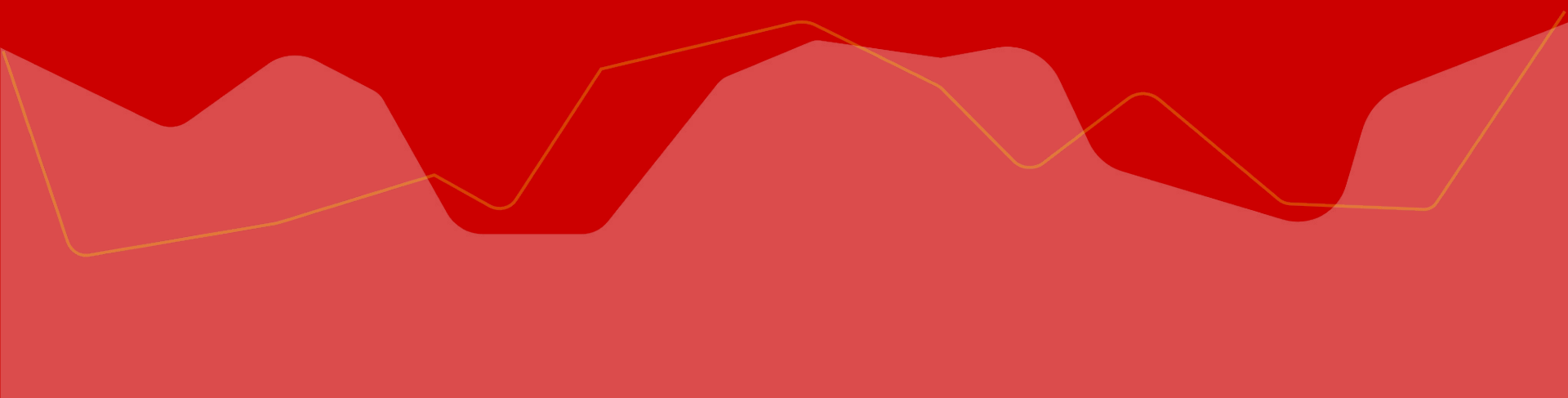
final (1) 

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.



## 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

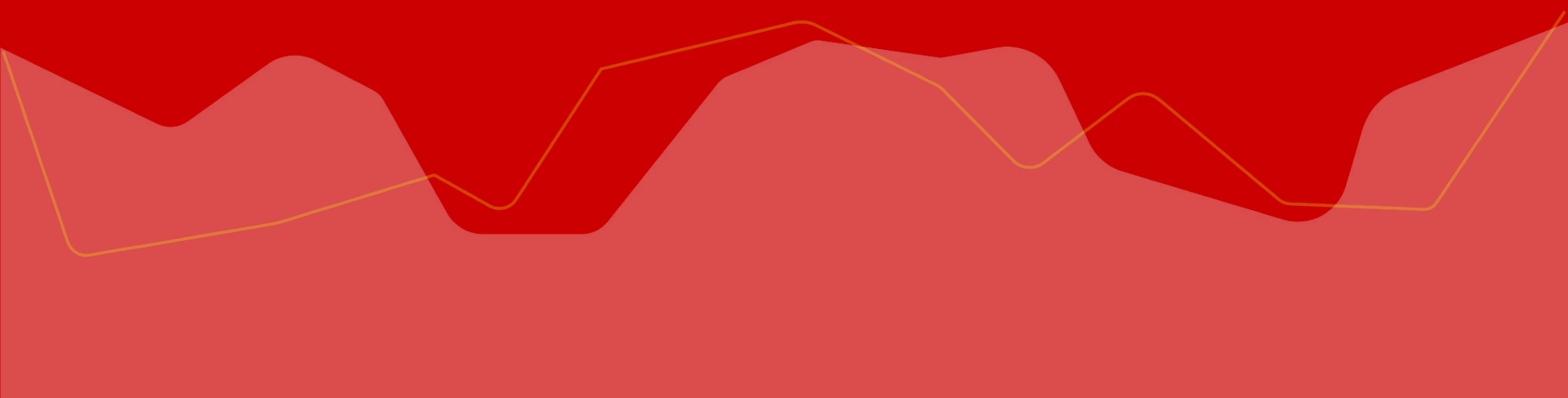
**Good** or **Bad** Organised Data?

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:



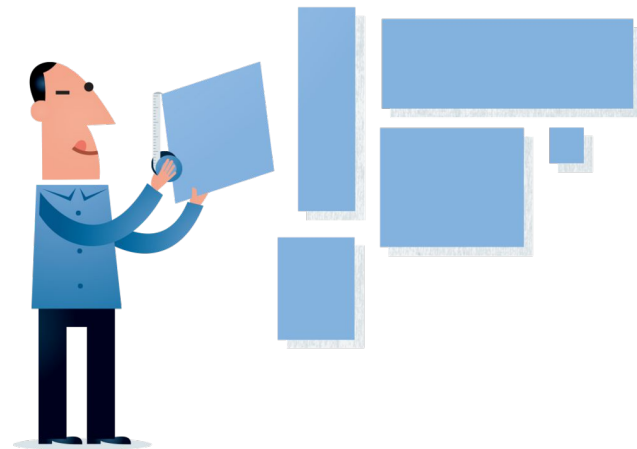


## Selecting Formats

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

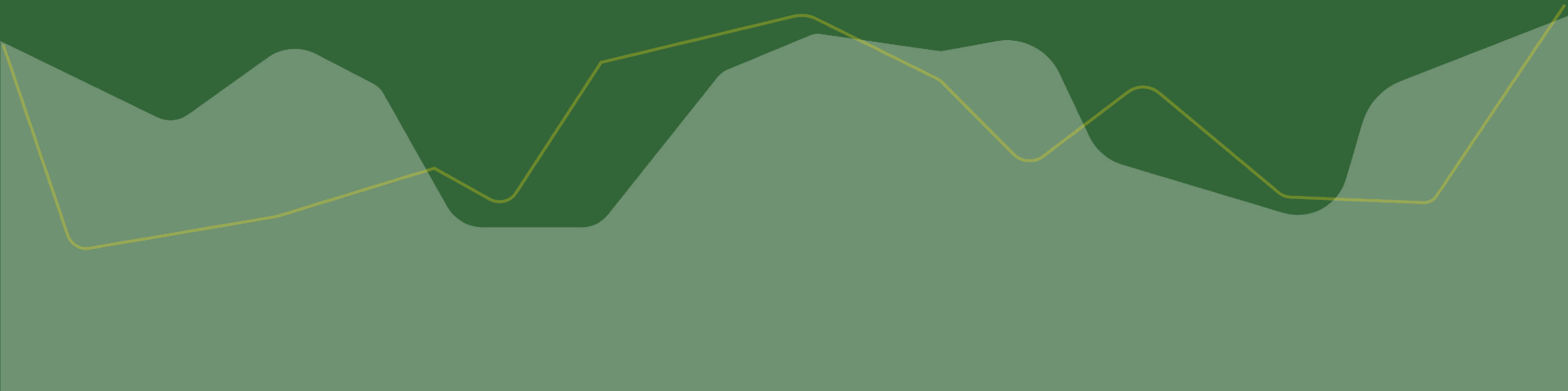
Image Format	Compression	Longevity	Compatibility
<b>JPG/JPEG</b>	Lossy	Low	High (open, widely supported)
<b>TIFF (TIF)</b>	Lossless	High	High (open, widely supported)
<b>PNG</b>	Lossless	High	High (open, widely supported)
<b>GIF</b>	Lossy	Low	High (open, widely supported, limited colors)
<b>BMP</b>	Lossless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



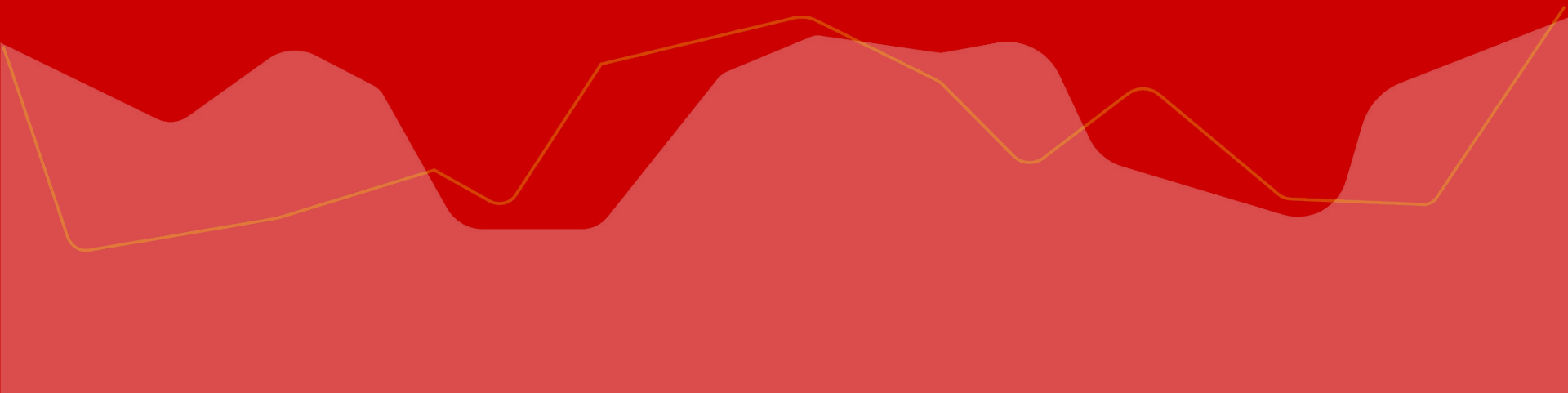
# Break



# Section 2 – How to transform your files in bulk

The bottom of the slide features a decorative graphic consisting of several overlapping, wavy lines in shades of red and orange, creating a layered, mountain-like effect.

# 2.1 Change filenames in bulk





## Bulk rename folders & files:

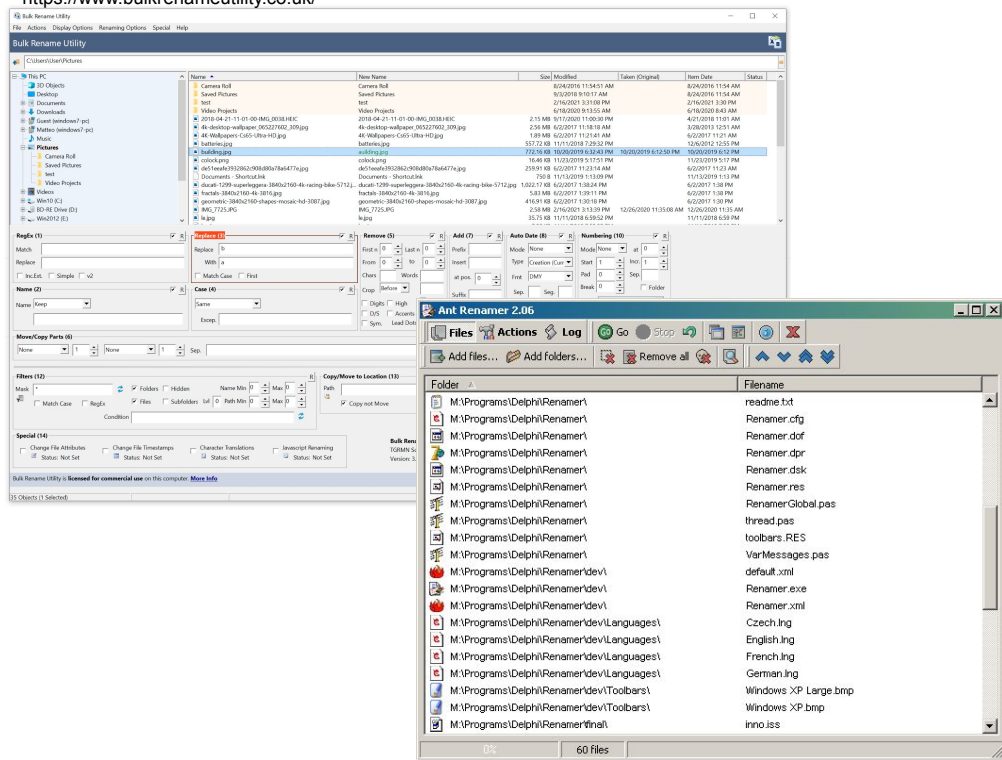
File renaming software:

- 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 (\* ! ? ; ; " \$)

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

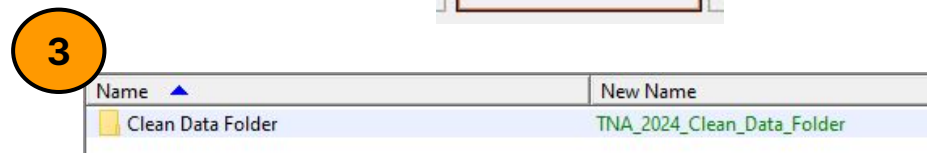
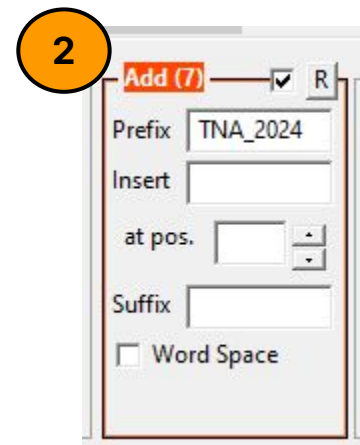
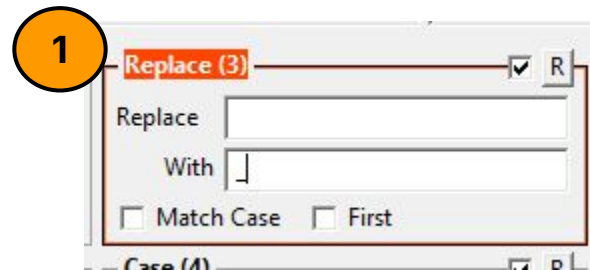


[https://www.antp.be/pic/renamer\\_1.png](https://www.antp.be/pic/renamer_1.png)

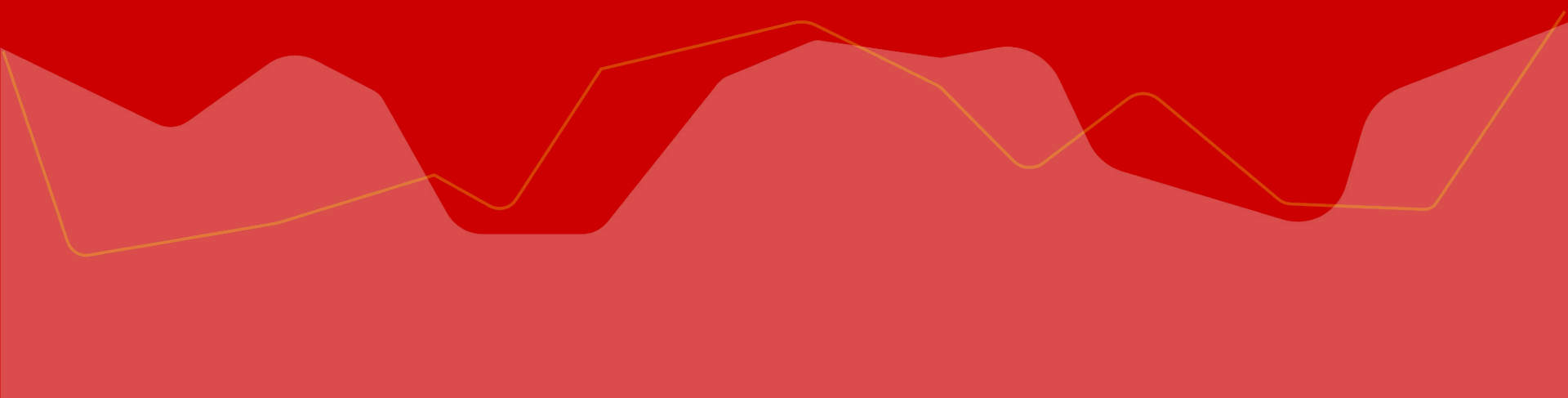
## Explore the bulk rename utility tool

**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



JPG



PNG



BMP



NORMALISATION



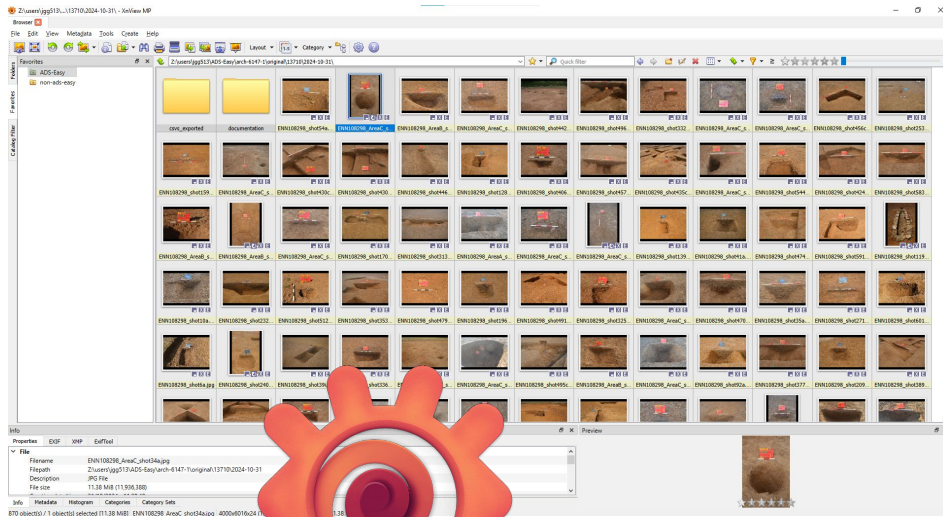
TIFF



TIFF



TIFF

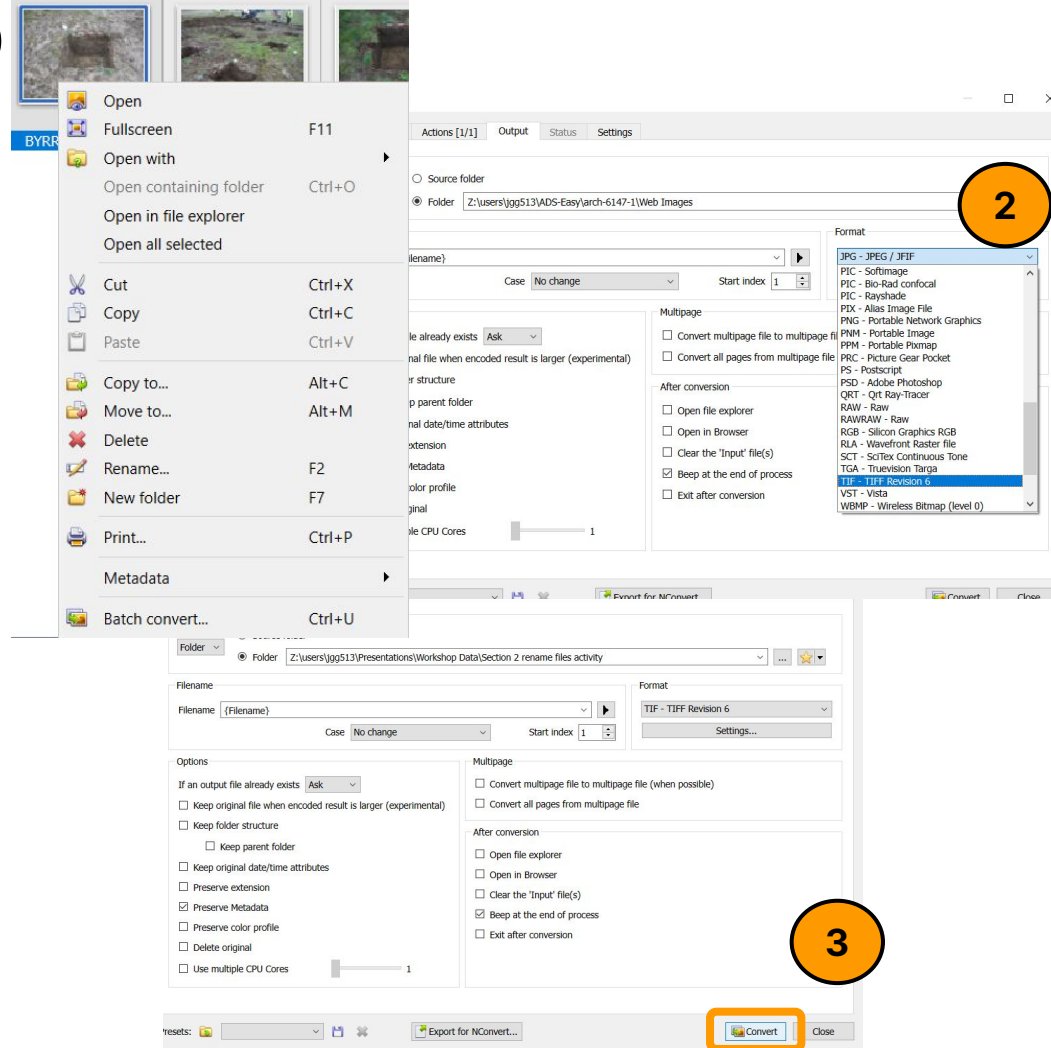


<https://www.xnview.com/en/xnviewmp/>

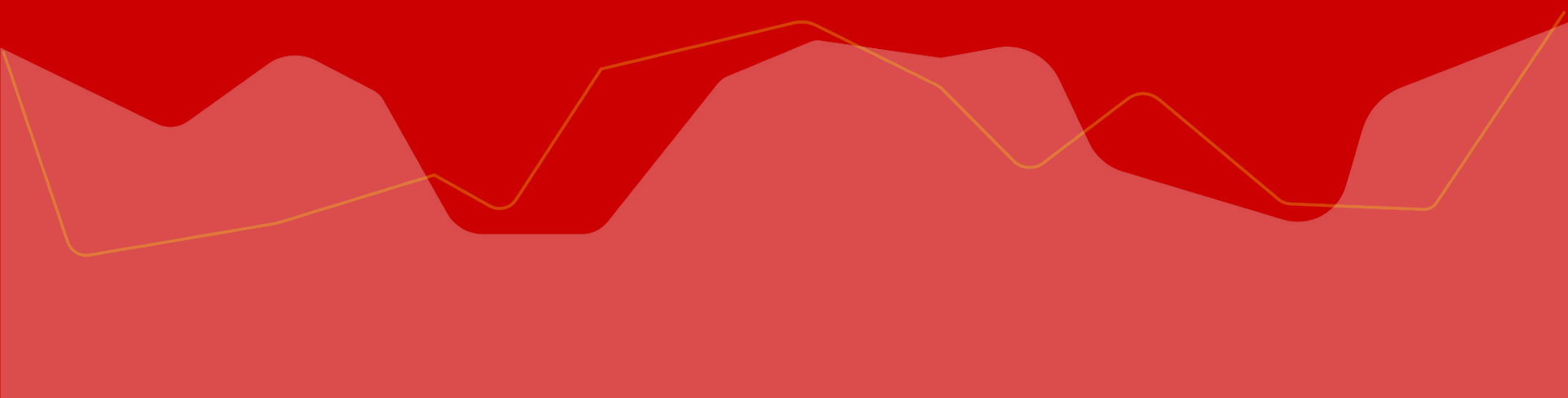
## Explore XNView MP

**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

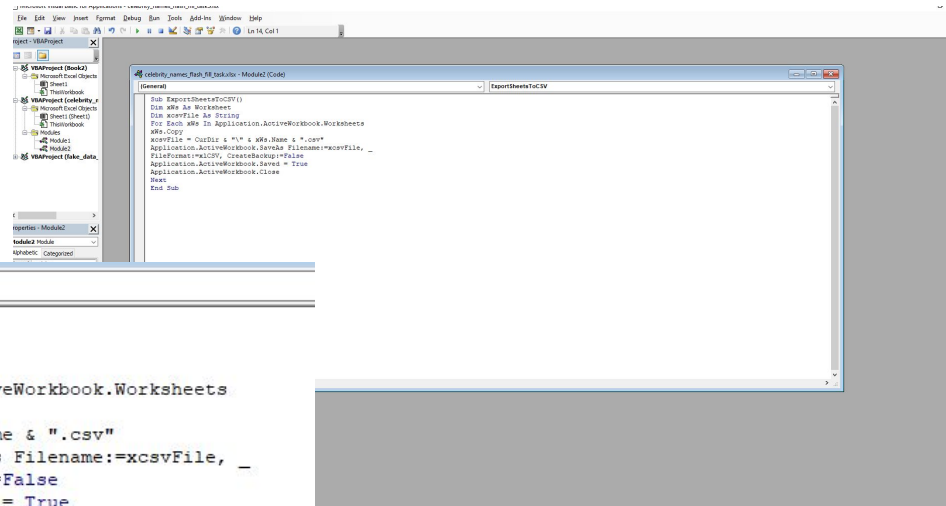


## Bulk Convert Excel sheets to CSV files:

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

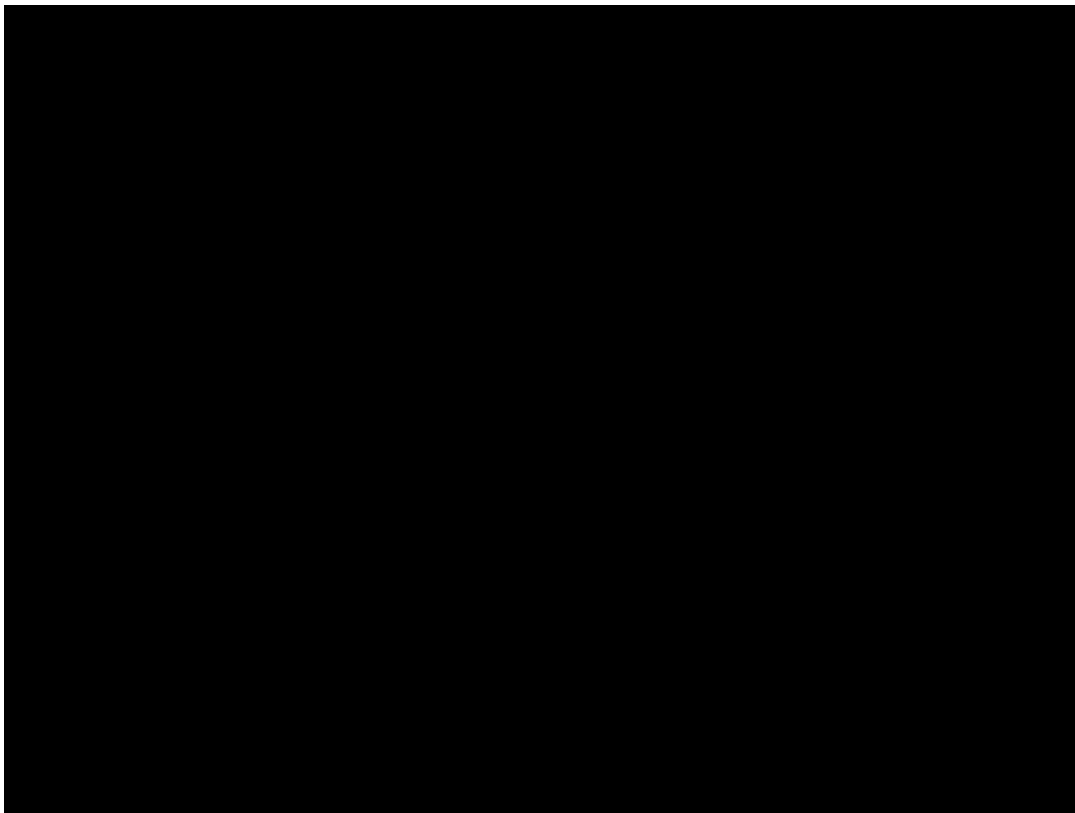
```
(General)

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
```



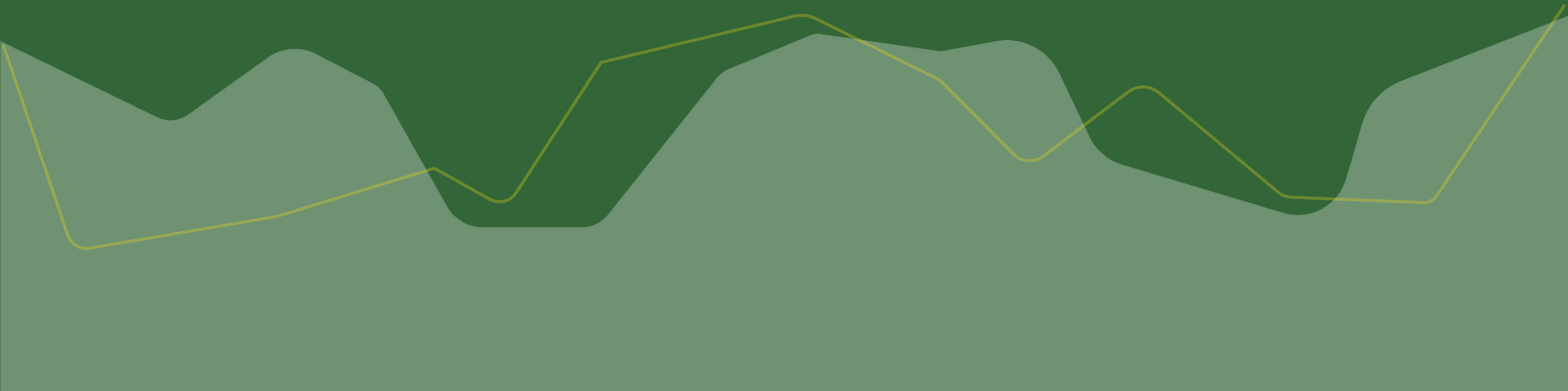
Excel VBA in use

## Bulk Convert Excel sheets to CSV files:





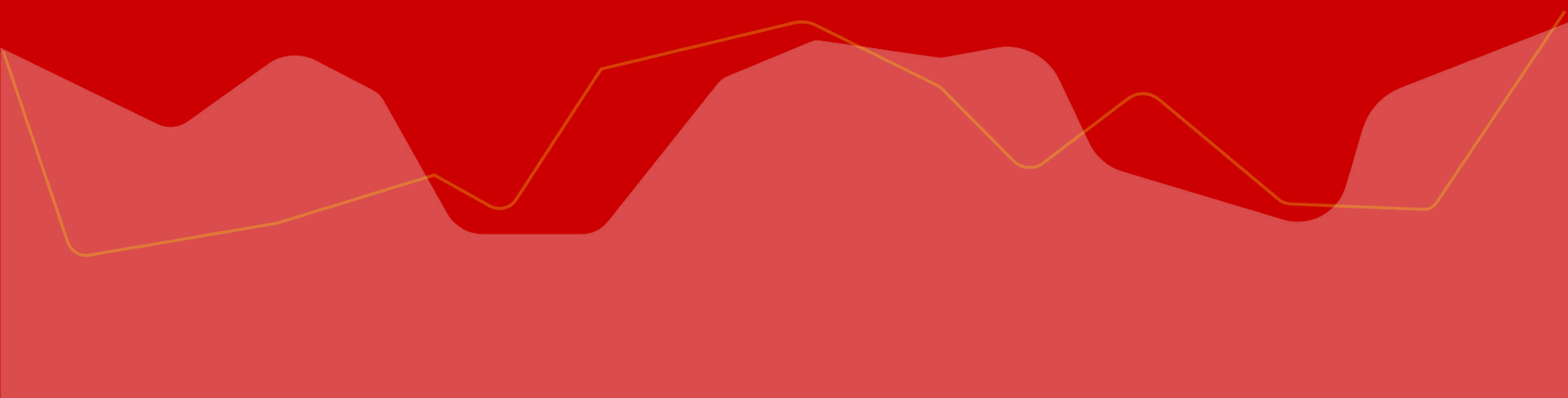
# Break



# Section 3 – How to wrangle spreadsheets in Excel

The background is a solid red color. A decorative yellow line graphic is overlaid on the bottom half of the page, consisting of several connected segments that create a jagged, mountain-like silhouette.

# 3.1 Useful Excel Tools and Formula

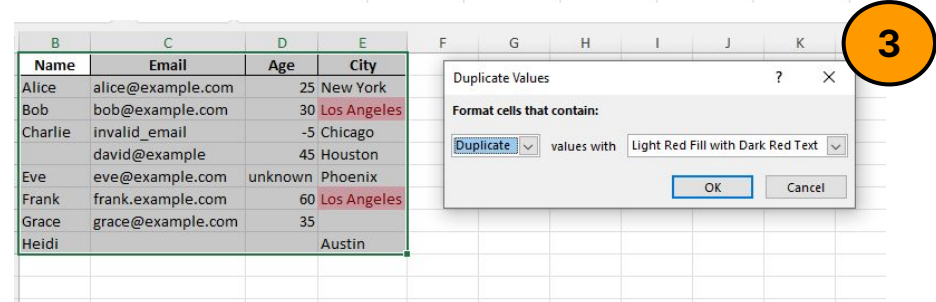
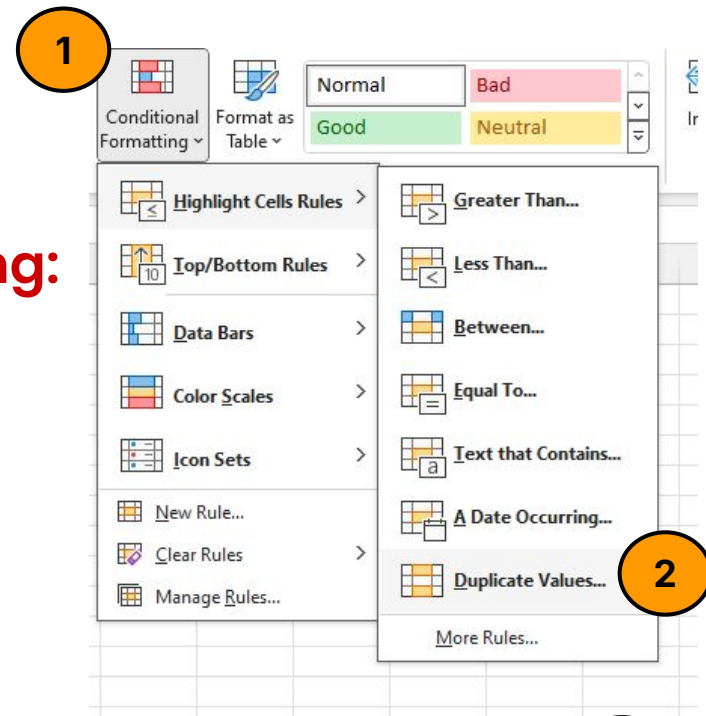


# 1. Conditional Formatting For Errors



## 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



Example of highlighting duplicates with conditional formatting

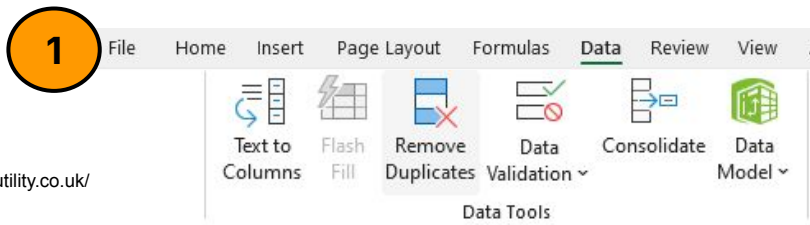
## 2. Remove duplicates in Excel

The bottom of the slide features several overlapping, wavy lines in various shades of orange and yellow, creating a decorative border.

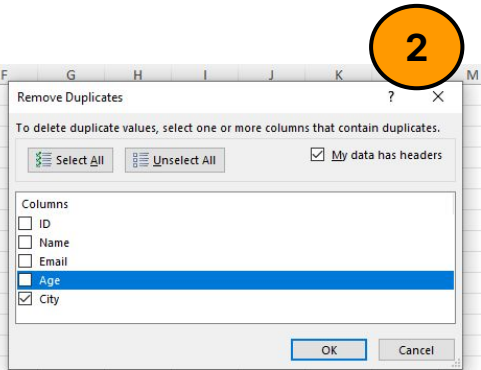
<https://www.bulkrenameutility.co.uk/>

## Remove Duplicate Values:

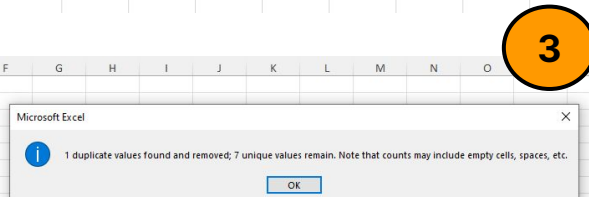
1. In the **Data** tab click on the Remove Duplicates option
  2. **Select the columns** that you want to scan for duplicate entries
  3. View the removed data and check it looks okay
- **Task:** Remove any duplicate values with just the **CITIZAN id** ticked and **Feature name**.



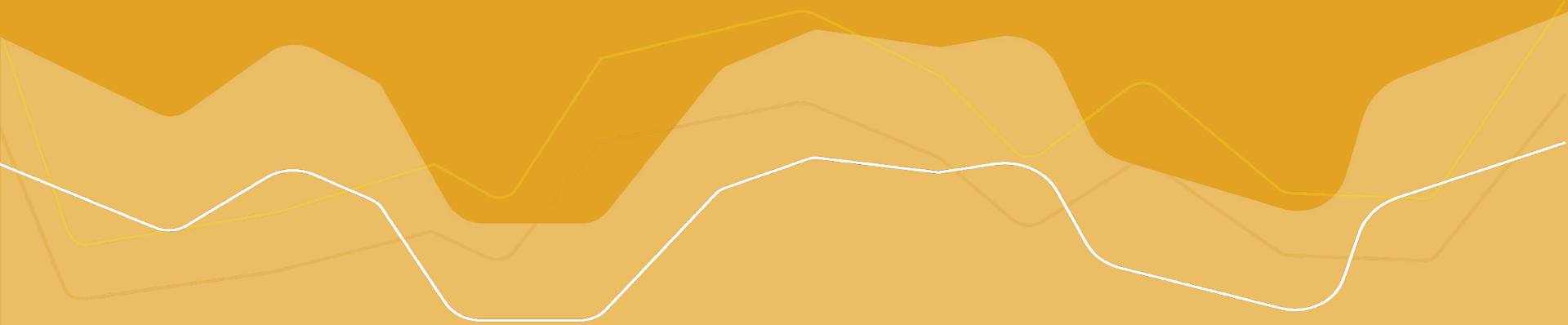
ID	Name	Email	Age	City
1	Alice	alice@example.com	25	New York
2	Bob	bob@example.com	30	Los Angeles
3	Charlie	invalid_email	-5	Chicago
4		david@example	45	Houston
5	Eve	eve@example.com	unknown	Phoenix
6	Frank	frank.example.com	60	Los Angeles
7	Grace	grace@example.com	35	
8	Heidi			Austin



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



### **3. Flash Fill Feature in Excel**

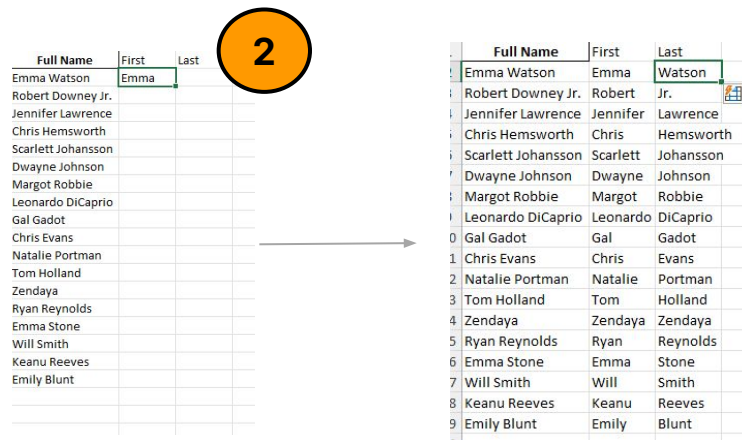
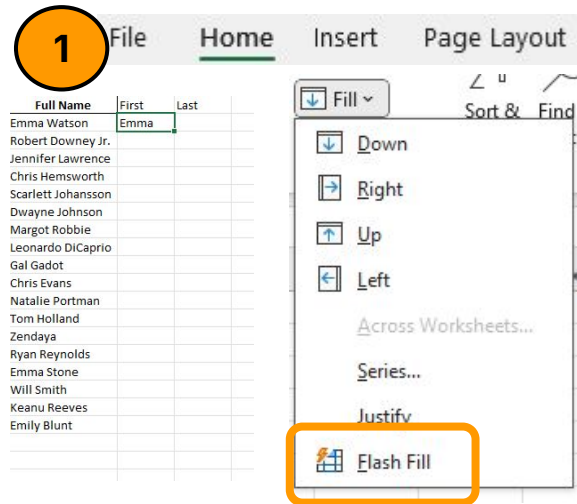




## Flash Fill Feature in Excel

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 |**



## 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 displays the 'Convert Text to Columns Wizard' in Microsoft Excel, showing three steps of the process. Step 1: 'Delimited' is selected as the file type. Step 2: 'Comma' and 'Space' are selected as delimiters. Step 3: 'General' is selected as the column data format, and the destination is set to '\$B\$2'. A data preview window shows the resulting columns.

**Convert Text to Columns Wizard - Step 1 of 3**

The Text Wizard has determined that your data is Delimited. If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

Delimited - Characters such as commas or tabs separate each field.

Fixed width - Fields are aligned in columns with spaces between each field.

**Convert Text to Columns Wizard - Step 2 of 3**

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

Tab

Semicolon

Comma

Space

Other: [ ]

Treat consecutive delimiters as one

Text qualifier: [ ]

**Convert Text to Columns Wizard - Step 1 of 3**

This screen lets you select each column and set the Data Format.

Column data format

General

Text

Date: DMY [ ]

Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Advanced...

Destination: =\$B\$2

Data preview

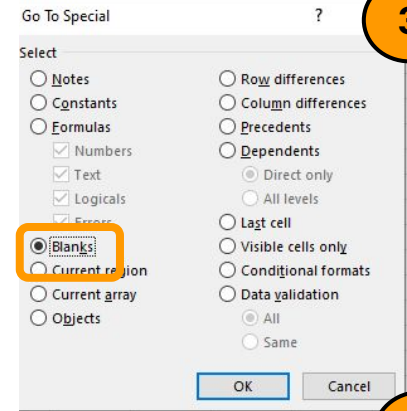
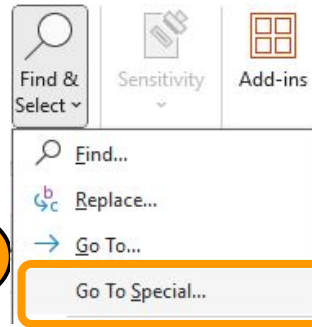
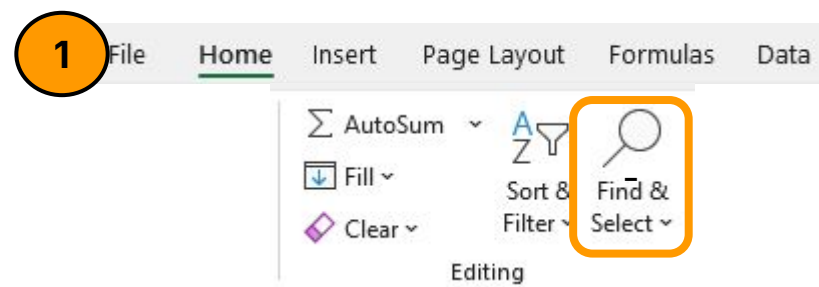
General	General
Emma	Watson
Jennifer	Lawrence
Chris	Hemsworth
Scarlett	Johansson
Dwayne	Johnson

## 5. Quickly find files and replace blank entries

The bottom of the slide features a decorative graphic consisting of several overlapping, wavy lines in various shades of orange and yellow, creating a layered, abstract effect.

## 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**



4 Enter NULL into highlighted box

	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

Ctrl + Enter →

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

5

## 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

ID	Name	Email	Age	City
1	Alice Jone	alice@example.com	25	New York
2	Bob Curtis	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

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

Find and Replace

Find: Find what:

Replace with: NULL

Options >>

Replace All Replace Find All Find Next Close

Book	Sheet	Name	Cell	Value	Formula
fake_data_with_errors.xlsx	Sheet1		SB\$5		
fake_data_with_errors.xlsx	Sheet1		SE\$7		
fake_data_with_errors.xlsx	Sheet1		SC\$8		
fake_data_with_errors.xlsx	Sheet1		SD\$8		

4 cell(s) found

3

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

Find and Replace

Find: Find what:

Replace with: NULL

Options >>

Replace All Replace Find All Find Next Close

Book	Sheet	Name	Cell	Value	Formula
fake_data_with_errors.xlsx	Sheet1		SB\$5	NULL	
fake_data_with_errors.xlsx	Sheet1		SE\$7	NULL	
fake_data_with_errors.xlsx	Sheet1		SC\$8	NULL	
fake_data_with_errors.xlsx	Sheet1		SD\$8	NULL	

4 cell(s) found

Microsoft Excel

1 All done. We made 4 replacements.

OK

Example of removing blank entries with find & replace tool

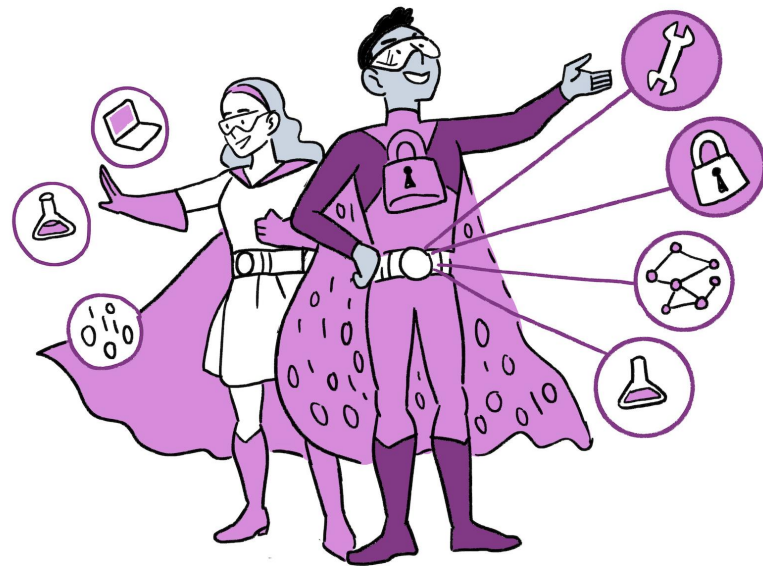
## 6. Useful formula



## 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:

The bottom of the slide features a decorative graphic consisting of several overlapping, wavy lines in shades of red and orange, creating a layered, abstract effect.

## 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

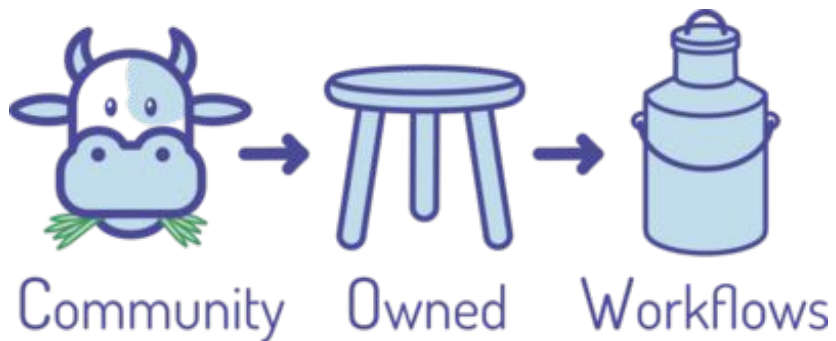
# (BONUS) Use Openrefine to wrangle data

The bottom of the slide features a decorative graphic consisting of several overlapping, wavy lines in a lighter shade of purple, creating a layered, mountain-like effect against the dark purple background.

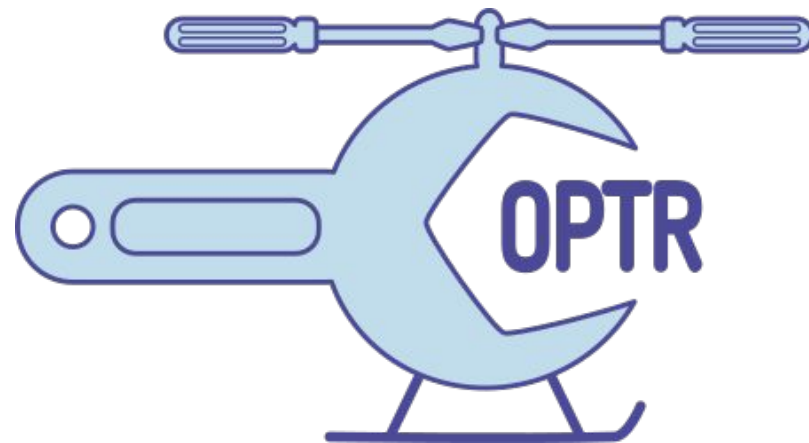
# Q & A/ Project Session



## Useful Resources for software



Community Owned  
Workflows (COW)



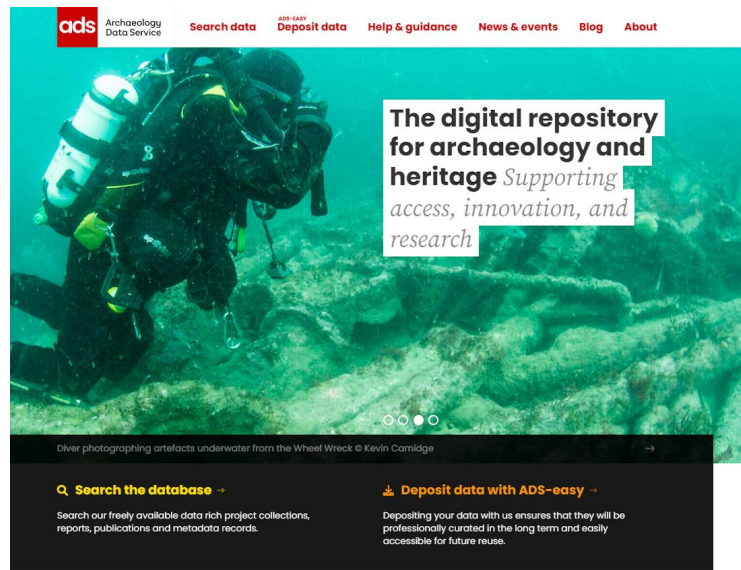
Community Owned Digital  
Preservation Tool Registry (COPTR)

## Keep informed!

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



[nicky.garland@york.ac.uk](mailto:nicky.garland@york.ac.uk)





Archaeology  
Data Service

**Thank you!**

[jamie.geddes@york.ac.uk](mailto:jamie.geddes@york.ac.uk)  
[nicky.garland@york.ac.uk](mailto: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](http://www.archaeologydataservice.ac.uk)



[help@archaeologydataservice.ac.uk](mailto:help@archaeologydataservice.ac.uk)