

# DataUp Features

## Background

The DataUp tool began life as a project led by the California Digital Library (CDL), working in conjunction with Microsoft Research. CDL developed version 1.0 of dataup as an Addin to Excel, and which had the basic functionality. Subsequently CDL decided to switch DataUp 2.0 to a web app (so you loaded your excel file to the web for QA, recording metadata, and posting to a repository) – Landcare felt this was a backward step, requiring internet access to use, and adding additional steps outside of the Excel software.

CDL ceased their work on the DataUp webapp, and gave permission for Landcare to take over development of the Excel addin. We have since greatly improved on their prototype: refining the metadata and its handling; adding the ability to work with multiple worksheets; providing the ability to store both the xlsx and a tsv/csv (tab or comma separated values) version of the data; the ability to request a DOI for the data; automating deposit to our DataStore repository; and generally enhancing the user experience.

The Landcare version of DataUp is 3.x

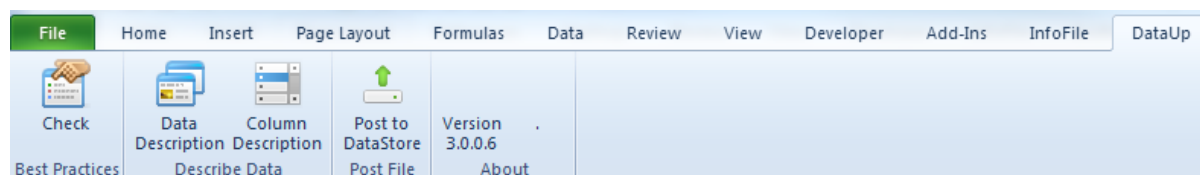
## Summary

The components of DataUp are:

- Recording basic metadata about the data
- QA data: check for best practices/common errors
- Generation of a preferred citation for your data (including a DOI if requested)
- Automation of deposit to the Landcare DataStore repository<sup>1</sup>

DataUp can be used offline (e.g. in a hut in the field) and you can use the relevant parts whilst you are still working on your data e.g. record the metadata early on, and later you can clean up the data and deposit to the DataStore.

The tab or comma separated value (tsv/csv) text file format is recommended as a more robust long term storage format (e.g. if Microsoft change the file format of excel in future), but DataUp enables easy depositing of both tsv/csv and/or xlsx.



Below is a more detailed breakdown of what DataUp does (based on the original CDL documentation, updated to reflect the Landcare Research changes).

---

<sup>1</sup> For Landcare staff use. DataUp could be modified to point to alternate users repositories.

## The DataUp tool has four main features:

### 1. Create metadata

Metadata is “data about data”. It is the who, what, when, where, why, and how for your dataset. Often researchers do not explicitly record these details; instead this information is kept in lab notebooks, on spreadsheet tabs, or in their heads. Among the most challenging aspects of being a good data steward is creating quality standard metadata to accompany your dataset – this ensures the data is best able to be interpreted and possibly reused in future.

DataUp will walk you through creating standard metadata using a form that becomes part of your spreadsheet, allowing for future understanding and reuse. When the data file is uploaded to the DataStore repository, the metadata is used to create the Dataset record along with information about each of the data files (worksheets) uploaded. Landcare chose to align the metadata with the Datacite metadata standard - a domain agnostic, simple metadata schema, which can also be used if we issue a DOI for your data (so you, and others, can cite it in associated papers).

Metadata is recorded at both the file level and the column (data field) level. File level metadata describes the project and includes a dataset title, author names, a contact person and email, and geographic location (to enable spatial searching). A few key elements of the file level metadata are mandatory, the remainder are optional. Column level metadata (i.e. attribute metadata) includes information about the variables in your dataset, the units of measure, and descriptions of the worksheet and each column of data.

The screenshot displays the 'CREATE METADATA' form in two overlapping windows. The top window shows the 'Data descriptions' tab with file-level metadata fields. The bottom window shows the 'Column descriptions' tab with a table for column-level metadata.

**File Level Metadata (Data descriptions tab):**

- Project title (job-code) \*: Impact of possums in Northland (221001-1234)
- Title of dataset \*: Puketi Forest Foliage browse survey 2012
- Funding Source(s): MBIE, DOC
- Data Publisher: Organisation name: Landcare Research
- Data Publisher: Author(s) \*: Jones, Sue; Smith, Bob
- Data Contact Person: First name \*: Bob
- Data Contact Person: Last name \*: Smith
- Data Contact Person: Email \*: smithb@LandcareResearch.co.nz
- Publication date \*: [Empty]
- Abstract: Abstract would go in here...
- keyword(s): Vegetation Survey, Foliar Browse
- Keyword thesaurus used: [Empty]

**Column Level Metadata (Column descriptions tab):**

Worksheet Name *	Worksheet Description *	Name *	Description *	Type *	Units	Delete
Plots	Plot attributes	Transect	Transect number	Numeric	number	X
Plots	Plot attributes	Plot	Plot number	Numeric	number	X
Plots	Plot attributes	Altitude	Altitude above sea level	Numeric	metre	X
Plots	Plot attributes	Aspect	Plot Aspect	Numeric	degree	X
Plots	Plot attributes	Slope	Hill Slope	Numeric	degree	X
Plots	Plot attributes	Drainage	(Good, (M)edium, (P)oor)	Text		X
Plots	Plot attributes	Can_Ht	Canopy Height	Numeric	metre	X
Samples	Foliage browse survey data	Distance	Distance from plot centre	Numeric	metre	X
Samples	Foliage browse survey data	Species	Tree Species NVS code	Text		X
Samples	Foliage browse survey data	Tag	Tree ID tag number	Numeric	number	X
Samples	Foliage browse survey data	Sdiam	Stem diameter	Numeric		X

## 2. *Check for best practices*

Excel is very flexible in how it allows you to arrange and annotate your data. This can be useful while you are working on the data, but when analysis is complete and you are thinking about depositing your dataset for 'safe keeping' in the data repository consideration of best practice and the datasets future come into play. The features that Excel has such as embedded graphs, coloured text or shading, embedded comments etc work fine in excel (.xls, .xlsx), but if over time the version of excel is updated and no longer opens the old versions then not only are those features inaccessible, but the data may be too. Even if the excel file is able to be opened in the future, will others know what the coloured shading means (will you even remember)?

Part of best practice is where possible storing the data in a non-proprietary format such as comma (or tab) delimited text files (CSV/TSV). To enable this it is necessary to avoid (or strip out) things which could affect the ability to read back and interpret the text files in the future. DataUp checks for a number of issues, and in many cases can help fix or provide tips on how to fix:

- Embedded charts, tables, pictures
- Embedded comments
- Commas
- Special characters
- Color coded text or cell shading
- Columns have mixed data types
- Non-contiguous data
- Merged cells
- Blank cells
- Header row absent or more than one header row
- Duplicate field names

In addition to identifying the locations of these problems, DataUp explains why they are potentially problematic, and offers suggested alternatives or the ability to remove embedded charts, comments, and colour coded cells in bulk. Users also have the ability to ignore these suggestions and continue without addressing issues (e.g., if you plan to archive the data in .xlsx format embedded charts will work).

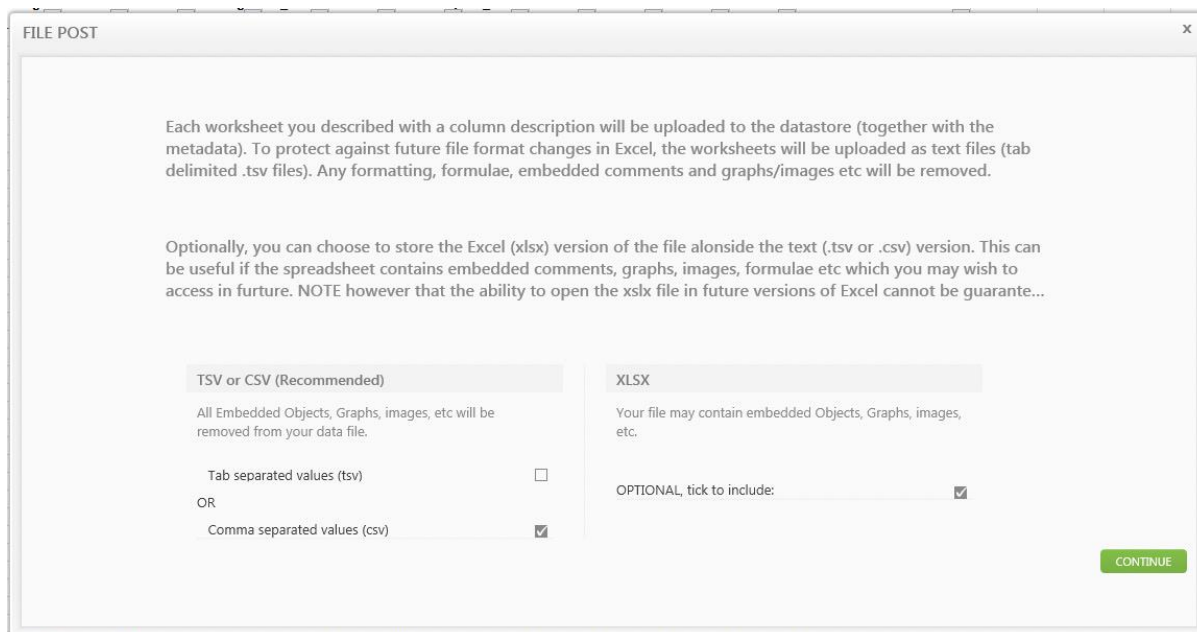
DataUp allows storing of the fully formatted excel file in the DataStore repository, **alongside** the best practice more 'future-proofed' CSV or TSV (or similar) file. Thus you can store a copy of the file with any graphs, comments etc AND a robust CSV/TSV version.

## 3. *Get credit for data: obtain an identifier*

Valuing and incentivizing the time and effort required to manage data well is an important factor in fostering data sharing and reuse. One way to allow data producers to get credit for this is to enable data citation. Rather than citing papers that summarize results from a data set, researchers can also cite data sets themselves. For this to be possible, the data must be well documented, archived, and have a persistent, unique identifier. Landcare Research is registered to issue Digital Object Identifiers (DOIs). DataUp will assist you by generating the citation you wish to have associated with the dataset and requesting a DOI which can be included in associated papers and reports, allowing you and others to cite your data directly, put it on your CV, and help determine its impact in your research community.

#### 4. Archive & share data

Once you have created metadata, you can deposit the data directly to a repository via DataUp for archiving/publishing. Landcare have connected DataUp to our CKAN repository [DataStore](#), which together with our ability to issue DOIs meets journal requirements for publishing data associated with your paper (e.g. [Nature - institutional repositories](#)). DataStore is also indexed by Google, and available for discovery and use by the public.



FILE POST

Each worksheet you described with a column description will be uploaded to the datastore (together with the metadata). To protect against future file format changes in Excel, the worksheets will be uploaded as text files (tab delimited .tsv files). Any formatting, formulae, embedded comments and graphs/images etc will be removed.

Optionally, you can choose to store the Excel (xlsx) version of the file alongside the text (.tsv or .csv) version. This can be useful if the spreadsheet contains embedded comments, graphs, images, formulae etc which you may wish to access in future. NOTE however that the ability to open the xlsx file in future versions of Excel cannot be guaranteed.

**TSV or CSV (Recommended)**

All Embedded Objects, Graphs, images, etc will be removed from your data file.

Tab separated values (tsv)

OR

Comma separated values (csv)

**XLSX**

Your file may contain embedded Objects, Graphs, images, etc.

OPTIONAL, tick to include:

CONTINUE

#### Why?

Good data management practices ensure others can use and reuse your data well into the future. In addition, DataUp can help you meet newly implemented funder requirements for data management. By documenting your data, making your data publicly available, and providing a persistent identifier for citation, you are also contributing to open science, and transparent research processes.