DataUp Excel Add-In User's Guide

DataUp: Describe, Manage, and Share Your Data

Version History

Version	Date	Notes
1.0	2/15/13	Initial Draft
1.1	2/25/13	First release
1.2	5/3/13	Minor edits

About the DataUp Project

In today's technologically advanced world, the data generated by researchers is increasingly born digital and subject to intensive transformation and analyses before publication. The various file formats, software, and hardware required to succeed in the modern research landscape can become daunting, especially since education about digital data management has not kept pace with these technological advancements. There is a significant gap between the data management skills needed by modern researchers and their current abilities; the gap is more noticeable given the current increase in funder requirements for data management plans and requests by journal publishers to make supporting data publicly available alongside traditional research articles.

The DataUp project was born out of this need for seamless integration of data management into the researchers' current workflows. We recognized that the large majority of Earth, environmental, and ecological scientists use spreadsheets in the course of their data collection and organization; rather than requiring they learn a new program, we decided to meet them where they already work: Microsoft® Excel and related spreadsheet software.

The DataUp project's goals were to facilitate data management, sharing, and archiving for scientists. The resulting tools from the DataUp project are part of the Investigator Toolkit for DataONE, an NSF DataNet project building cyber-infrastructure that links together existing archives of ecological and environmental scientific data. The OuterCurve Foundation holds the code's copyright and DataUp is part of their Research Accelerators Gallery.

The DataUp project is run through the University of California Curation Center (UC3) at the California Digital Library, a division of the University of California Office of the President. Project funders were the Gordon and Betty Moore Foundation and Microsoft Research Connections.

1 Contents

2 Data Concepts in Excel

This manual relies on the following terminology.

- A *workbook* is an XSLX or XSL file that is the native format for Microsoft Excel.
- A *data set* is the data in a workbook
- Each workbook consists of one or more *worksheets*. Individual worksheets are accessed through the tabs at the bottom of the Excel screen.
- An *entity* is a set of related data within a data set. To be compliant with archival Best Practices, a data set must contain only column labels in the first row and individual data points in each subsequent row. A workbook that contains a data set with multiple entities has each entity in a separate worksheet.

3 Metadata and DataUp

Metadata is key to archival Best Practices. Metadata can be described as data about data. Classical examples include library catalogs and inventories. Examples in modern technology

include the technical and descriptive information digital cameras embed in graphic files and the information (owner, access permissions, creation and modification date/time values) that computer file systems record for all online files.

The DataUp Add-In helps you create two kinds of metadata:

- *General metadata*, describing the worksheet data as a whole. General metadata is described in Table 1.

Column metadata, with individual descriptions for each data column. See "**Error! Reference source not found.**", page **Error! Bookmark not defined.**.

Table 1: General Metadata

Name	Notes
Title of dataset	
Creator: First name	
Creator: Last name	
Creator: Organization	
Creator: Address	
Creator: City	
Creator: State/province	
Creator: Postal code	
Creator: Country	
Creator: Phone	
Creator: Email	
Today's date	
Abstract	
Keyword(s)	
Keyword thesaurus used	
Intellectual rights	
URL for data	
Geographic coverage: Description	
Geographic coverage: West bounding coordinate	
Geographic coverage: East bounding coordinate	
Geographic coverage: North bounding coordinate	
Geographic coverage: South bounding coordinate	
Temporal coverage: Beginning date	
Temporal coverage: Ending date	
Data Contact Person: First name	
Data Contact Person: Last name	

DataUp: Describe, Manage, and Share Your Data

Name	Notes
Data Contact Person: Organization	
Data Contact Person: Address	
Data Contact Person: City	
Data Contact Person: State/province	
Data Contact Person: Postal code	
Data Contact Person: Country	
Data Contact Person: Phone	
Data Contact Person: Email	
Data Publisher: repository name	
Project title	
Project personnel: Last name	
Project personnel role	
Project description	
Funding	
Data table name	
Data table description	Can contain information about the data orientation, number of records, case sensitivity, and temporal, geographic and taxonomic coverage.
Identifier	Unique identifying URL. See "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**.
Formatted citation	Standard citation created by the add-in.

Metadata used by the DataUp project is based on the Ecological Metadata Language (EML). For further information, see the EML web site at http://knb.ecoinformatics.org/software/eml/.

4 Installation

To obtain the add-in, visit http://dataup.cdlib.org/→ Start Using DataUp → Download the add-in. The add-in is packaged as a ZIP archive. To install the add-in, open the archive in Windows Explorer and run the **setup.exe** file.

Installing the add-in creates a program entry named "DataUp.Addin" that appears in the Control Panel list of programs. To remove the add-in, go to the control panel, open "Add and Remove Programs" (Windows XP) or "Programs and Features (Windows 7) and uninstall this program entry.

For detailed installation instructions, refer to the *Release Notes*, available on the web site.

5 DataUp Commands

With the add-in installed, the command ribbon includes a **DataUp** tab (Figure 1).

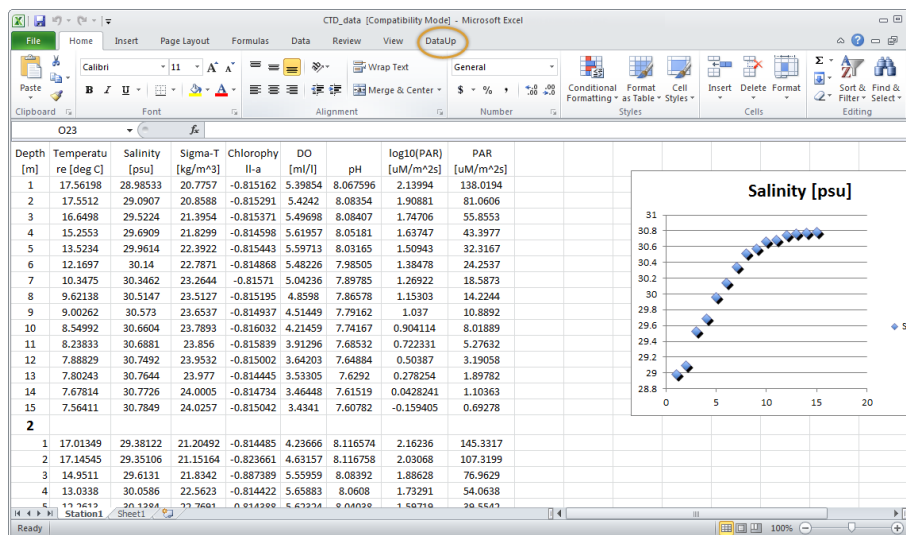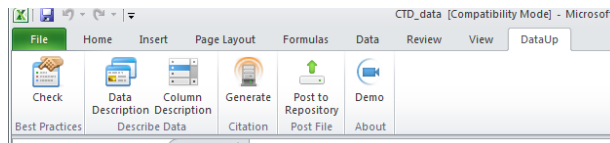*Figure 1: Excel Window with DataUp Installed*

Figure 2: DataUp Commands

The **DataUp** tab contains the following groups and commands (Figure 2):

- Best Practices
  - **Check** – Check for best practices. See "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**.
- Describe Data
  - **Data Description** – Bring up **Create Metadata** dialog. This is used for the following tasks:
    - "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**.
    - "Using the Create Metadata Dialog to Enter New Column Descriptions" on page 10.
    - "Using the Create Metadata Dialog to Edit Existing Column Descriptions" on page 13.

- o **Column Description** – Adds column descriptions directly into the metadata worksheet. See "Using the Column Description Command to Enter New Column Descriptions" on page 14.
- Citation
    - o **Generate** – Generates a unique identifier and citation and adds them to the metadata. See "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**.
- Post File
    - o **Post to Repository** – See "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**.
- About
    - o **Demo**

6 Implementing Best Practices

When archiving scientific data in Excel files, "Best Practices" help you store the data so that it can be retrieved and reused. Best Practices errors can cause data to be lost or misinterpreted. This section describes some common Best Practices errors, most of which can be detected by the add-in **Check** command. It explains why they should be avoided and suggests possible fixes. Finally, the implementation of Best Practices is demonstrated in an example workbook.

Table 2 describes common Best Practices errors.

Table 2: Best Practices Violations

Best Practice Violation	Add-In Detects	Why Avoided	Possible Fix
Descriptive Column Headers	No	Column headers should be simple labels.	Change to label. Move descriptive text to column metadata.
Embedded charts, pictures, shapes, tables	Yes	Not visible when data is exported to a CSV file or opened by program other than Excel	Move to separate tab or file.
Comments	Yes	Not visible when data is exported to a CSV file or opened by program other than Excel	Move to a data column.
Merged cells	Yes	Not maintained when exported to a CSV file. Information may be lost when cells are unmerged upon export.	Unmerge cells and annotate appropriately so that information is not lost.

Best Practice Violation	Add-In Detects	Why Avoided	Possible Fix
Commas	Yes	Commas are often used to separate multiple pieces of information. Cells should have only one piece of information.	Split comma-separated information in separate columns.
Special characters (non-alphanumeric)	Yes	Can cause problems for other programs. May cause problems for other programs. May be modified upon export.	Use alphanumeric characters only. If needed, describe the symbol in a new column
Color-coded text or cell shading.	Yes	Any information conveyed by formatting is lost when data is exported to a CSV file.	Use descriptions or alphanumeric coding schemes in a new column
Mixed data types	Yes	Some programs cannot handle mixed data types.	Use only a single data type in a column. If necessary, split data into multiple columns.
Blank cells within contiguous data	Yes	Might be misinterpreted by some programs.	Define other representation for missing data.
Blank rows or columns within contiguous data	Yes	Typically indicates multiple entities within one worksheet.	Move each entity into its own worksheet.

7 Describing Data

The DataUp Add-in assists you in creating metadata that describes your data. Two kinds of metadata are supported:

- *General metadata*, which describes the whole collection of data. General metadata is described by Table 1 on page 4; procedures entering general metadata are described in "**Error! Reference source not found.**", below.

Column metadata, with individual descriptions for each column. See "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**.

7.1 Entering General Metadata

Use the **Data Description** command to enter general metadata. This brings up the **Create Metadata** dialog (**Error! Reference source not found.**). This dialog has two tabs:

- The **Data Description** tab, which appears initially, is used to enter general metadata.
- The **Column Descriptions** tab, which is used to enter column metadata. See "Using the Create Metadata Dialog to Enter New Column Descriptions" on page 10.

Figure 3: Create Metadata Dialog

Note the following features of **Data Description** tab:

- Mandatory fields are indicated by a red asterisk after the title.
- Fields with data that don't pass validation tests have red outlines and error icons. These include mandatory fields that are empty.
- Any validation error disables the **Save** button.
- The following "Identifier" and "Citation" fields are normally left blank and filled in automatically based on information retrieved from the repository. See "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**.

When the data description is saved, a new worksheet called "Metadata" is created if it does not already exist. The general metadata is embedded in this worksheet (**Error! Reference source not found.**).

Figure 4: General Metadata in "Metadata" Worksheet

	A	B
1	Name	Value
2	Title of dataset	Galapagos Rift water collections June 2008
3	Creator: First name	Jane
4	Creator: Last name	Smith
5	Creator: Organization	San Diego Univesity
6	Creator: Address	415 20th St.
7	Creator: City	San Diego
8	Creator: State/province	CA
9	Creator: Postal code	
10	Creator: Country	USA

Once descriptions are embedded in the "Metadata" worksheet, they can be revised directly in the worksheet. They can also be edited in the **Create Metadata** dialog, which populates its fields with any existing descriptions.

7.2 Entering Column Metadata

There are two ways to enter column metadata:

- Use the **Data Description** command to bring up the **Create Metadata** dialog. The **Column Descriptions** tab on this dialog provides an editor for column descriptions stored in the "Metadata" worksheet. See "Using the Create Metadata Dialog to Enter New Column Descriptions" on page 10 and "Using the Create Metadata Dialog to Edit Existing Column Descriptions" on page 13.
- Use the **Column Description** command. This command simply adds column metadata fields to the "Metadata" worksheet, copying selected column headings for field names. Additional column metadata is entered directly into the worksheet. See "Using the Column Description Command to Enter New Column Descriptions" on page 14.

7.2.1 Using the Create Metadata Dialog to Enter New Column Descriptions

The following steps enter all column metadata at once:

1. Click the **Data Description** command button. The **Create Metadata** dialog appears.
2. Click on the **Column Descriptions** tab header to bring up its tab (Figure 5).

Figure 5: Column Descriptions Tab
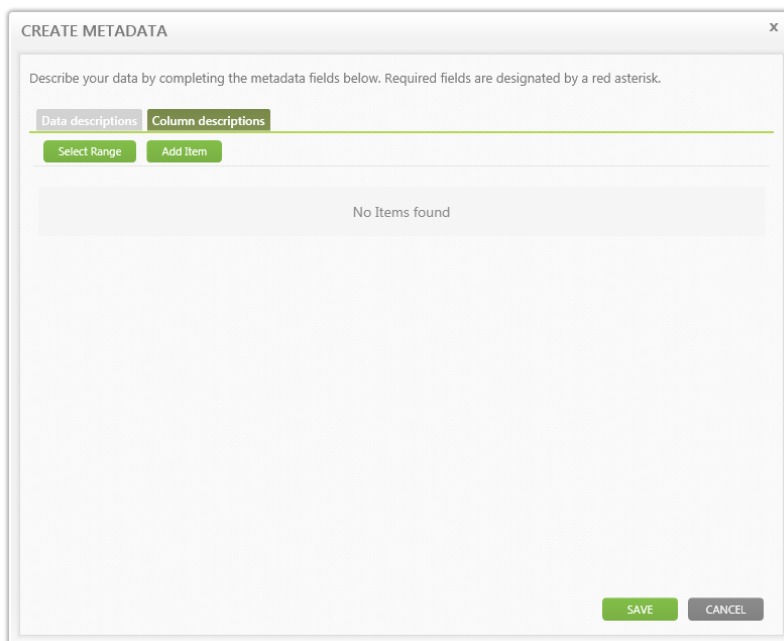
3. Click on the **Select Range** button. The **Parameter Metadata** dialog appears.

Figure 6: Parameter Metadata Dialog
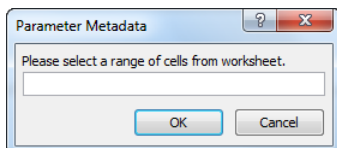
4. With the **Parameter Metadata** dialog still active, select the row of cells containing the column headers. The address of this row appears in the dialog.

Figure 7: Selecting the Row Containing Column Headers
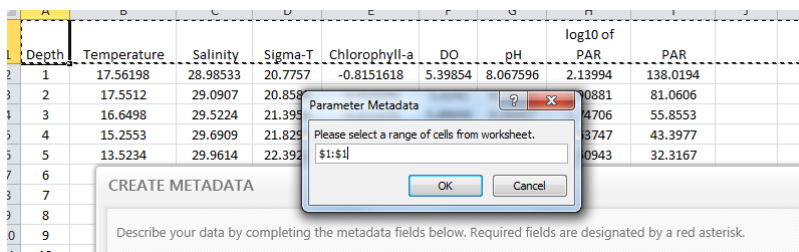
5. On the **Parameter Metadata** dialog, click the **OK** button. On the **Column Descriptions** tab, a metadata item is created for each column, with the column names already filled in (**Error! Reference source not found.**).

Figure 8: New Column Descriptions

6. Fill in the following for each column (**Error! Reference source not found.**):
   • **Entity Name**: A label for the entities. This defaults to, and should be the same as, the Excel worksheet from which the column labels were copied.
   • **Entity Description**: A description of the entity. The Entity Description only needs to be entered once per Entity; the add-in automatically copies this text to the other column metadata.
   • **Name**: The name of the column. This is initially copied from, and should be the same as, one of the column labels you copied in the previous step.
   • **Description:** Description of the column data.
   • **Type:** One of three values, "Numeric", "DateTime" or "Text". A dropdown list is available for quick entry.
   • **Units** (only entered for numeric data)**:** unit of measure for the data. Most common units of measure are available in a drop-down list. More complex or uncommon units are entered by hand.

Figure 9: Complete Column Descriptions
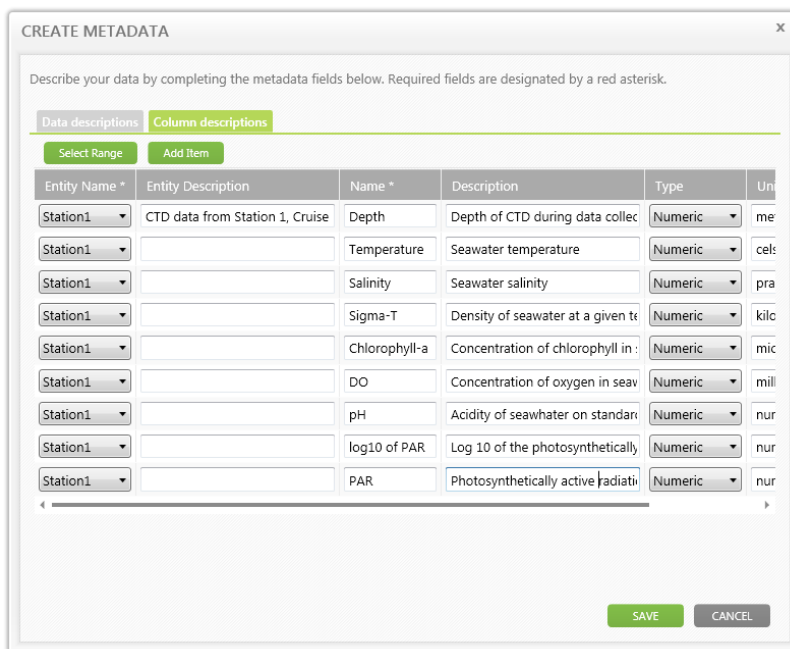
7. Click on the **Save** button. The new column descriptions are embedded in the "Metadata" worksheet.

Figure 10: Column Descriptions Embedded in "Metadata" Worksheet

7.2.2 Using the Create Metadata Dialog to Edit Existing Column Descriptions

If you bring up the **Create Metadata** dialog with column descriptions already embedded in the "Metadata" worksheet, the **Column Descriptions** tab is populated with the existing column metadata. This allows you to edit the column descriptions in the following ways:

- To revise existing column descriptions, manually edit the column fields.
- To add a description for a new column, click the **Add Item** button.
- To remove a column description, click the *x* to the right of the column description.

To save changes back to the "Metadata" worksheet, click on the **Save** button.

7.2.3 Using the Column Description Command to Enter New Column Descriptions

Use these steps to add column descriptions directly to the "Metadata" worksheet:

1. Click on the **Column Description** command button. The **Parameter Metadata** dialog appears (Figure 11).

Figure 11: Parameter Metadata Dialog
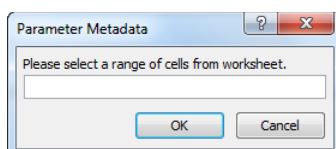
2. With the **Parameter Metadata** dialog still active, select the row of cells containing the column headers. The address of this row appears in the dialog (Figure 12).

Figure 12: Selecting the Row Containing Column Labels

3. Click on the **OK** button. A table for the selected columns is created in the "Metadata" worksheet (Figure 13).

Figure 13: Table Created by Column Description Command

	C	D	E	F	G	H	I	J
1		EntityName ▾	EntityDescription ▾	Name ▾	Description ▾	Type ▾	Units ▾	
2		Station1		Depth				
3		Station1		Temperature				
4		Station1		Salinity				
5		Station1		Sigma-T				
6		Station1		Chlorophyll-a				
7		Station1		DO				
8		Station1		pH				
9		Station1		log10 of PAR				
10		Station1		PAR				
11								
12								

4. Fill in the following for each column (Figure 14):

- **EntityName**: A label for the entity. This defaults to, and should be the same as, the Excel worksheet from which the column labels were copied.
- **EntityDescription**: A description of the entity. This should be the same for each row that has a given **EntityName**. To avoid entering the same text repeatedly, follow this procedure:
    1) Enter the description in the first row.
    2) Right-click the cell and chose **Copy**.
    3) Select the empty cells in the rest of the **EntityName** column.
    4) Right-click the selected cells and choose simple paste (the first icon below **Paste Options**).
- **Name**: The name of the column. This is initially copied from, and should be the same as, one of the column labels you copied in the previous step.
- **Description:** Description of the column data.
- **Type:** One of three values, "numeric", "timedate" or "text".
- **Units** (only entered for numeric data)**:** Unit of measure for the data.

Figure 14: Completed Column Descriptions

EntityName	EntityDescription	Name	Description	Type	Units
Station1	CTD data from Station 1, Cruise 1435	Depth	Depth of CTD during data collection	Numeric	meter
Station1	CTD data from Station 1, Cruise 1435	Temperature	Seawater temperature	Numeric	celsius
Station1	CTD data from Station 1, Cruise 1435	Salinity	Seawater salinity	Numeric	practical salinity units
Station1	CTD data from Station 1, Cruise 1435	Sigma-T	Density of seawater at a given temperature	Numeric	kilogram per meter
Station1	CTD data from Station 1, Cruise 1435	Chlorophyll-a	Concentration of chlorophyll in seawater	Numeric	microgram per liter
Station1	CTD data from Station 1, Cruise 1435	DO	Concentration of oxygen in seawater	Numeric	milligram per liter
Station1	CTD data from Station 1, Cruise 1435	pH	Acidity of seawater on standard 0-14 pH scale	Numeric	number
Station1	CTD data from Station 1, Cruise 1435	log10 of PAR	Log 10 of the photosynthetically active radiation of photos per meter squared per second	Numeric	number
Station1	CTD data from Station 1, Cruise 1435	PAR	Photosynthetically active radiation of photos per meter squared per second	Numeric	number

7.2.4 Column Metadata for Multiple Entities

When an Excel workbook contains multiple entities, Best Practices dictate that each entity get its own worksheet, as shown in "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**. Column Metadata must be provided for each such worksheet.

If column data usage is identical in worksheets, you can copy the metadata for the first entity, paste it back into the "Metadata" worksheet, and modify the **EntityName** and **EntityDescription** cells for the copied metadata. Here we demonstrate this using the example worksheet.

1. Select and copy the existing column metadata for a previous worksheet (Figure 15).

DataUp: Describe, Manage, and Share Your Data

Figure 15: Existing Entity Metadata Selected for Copying

2. Paste the copy directly below the existing column metadata.
3. In the first row of the new column metadata, modify the **EntityName** and **EntityDescription** values to match the next worksheet (Figure 16).

Figure 16: Copied Column Metadata with New Values for EntityName and EntityDescription

4. Select and copy the new **EntityName** and **EntityDescription** values.
5. Select the remaining cells for **EntityName** and **EntityDescription** (Figure 17).

DataUp: Describe, Manage, and Share Your Data

Figure 17: New EntityName and EntityDescription Values Copied with Destination Cells Selected

6. Paste the copied values into the selected cells (Figure 18).

Figure 18: New Column Metadata with Correct Values for EntityName and EntityDescription

8 Creating Citations

Once you've finished describing your data, two final metadata must be provided before the Excel worksheet can be archived.

- An **Archival Resource Key** or **ARK**. This is a unique and persistent URL that supports long-term access to your data. The add-in helps you request your repository to generate an ARK.
- A **Citation** for your data, based on the other metadata, including the ARK.

For more information on ARKs, refer to https://confluence.ucop.edu/display/Curation/ARK.

Follow these steps (some delays might occur while a remote repository is contacted):

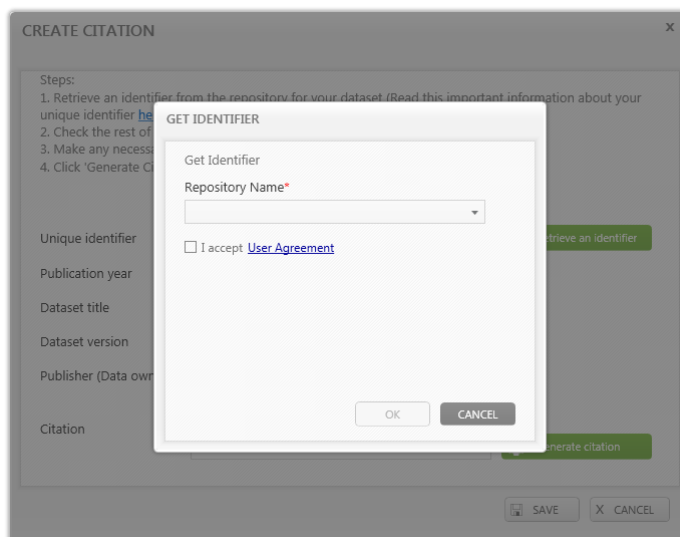1. Click on the **Generate** command. The **Create Citation** dialog appears (**Error! Reference source not found.**).
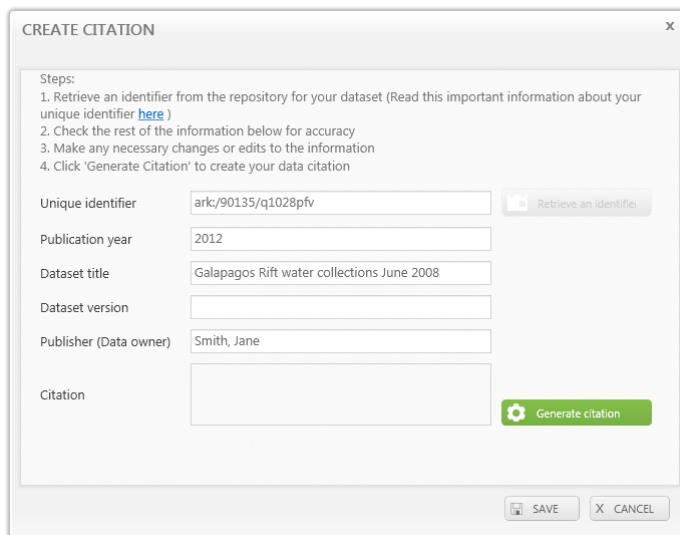
Figure 19: Create Citation Dialog

2. Click on the **Retrieve an identifier** button. The **Get Identifier** dialog appears (**Error! Reference source not found.**).

Figure 20: Get Identifier Dialog

3. Read and accept the User Agreement, then choose a repository from the drop-down list. Click **OK**.
4. The **Create Citation** dialog reappears, with the **Unique Identifier** field filled in (**Error! Reference source not found.**). Click the **Generate citation** button.

Figure 21: Create Citation Dialog with Identifier Filled In

5. The **Citation** field is filled in (**Error! Reference source not found.**). Click the **Save** button.

Figure 22: Create Citation Dialog with Citation Filled In

9 Posting to the Repository

This section describes use of the **Post to Repository** command.

1. Click on the **Post to Repository** command button. The **File Post** dialog appears.
2. Make one of the following choices:
    *   Click on the **Post as CSV** button to post the data as a text file containing comma-separated values. CSV format is accessible to a wide variety of software, but does not support many features of Microsoft Excel. Do not use this format if you need to preserve multiple worksheets, graphics, or embedded objects.
    *   Click on the **Post as XSLX** to post in the XML-based format that is the default for Microsoft Excel starting with the 2007 version. This format supports all Excel features, but may be present problems when read by other software.
3. The "Issues" tab of the dialog appears. This tab provides the same functionality as the **Check** command. If the issues reported preclude posting to the archive, press the **Cancel** button; otherwise press the **Next** button. For more information, refer to "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**.
4. The "Descriptions" tab of the dialog appears. This tab provides the same functionality as the **Data Description** command. Make any final changes to the worksheet metadata and press the **Next** button. For more information, refer to "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**.
5. The "Citations" tab of the dialog appears. This tab provides the same functionality as the **Generate** command. Make any final changes to the worksheet metadata and press the

Next button. For more information, refer to "**Error! Reference source not found.**" on page **Error! Bookmark not defined.**.

6. The "Post" tab of the dialog appears. Choose a repository from the drop-down list, accept the user agreement, and click on the **Post** button.