Ag Data Commons Data Submission Manual

User Edition v1.3

Knowledge Services Division | National Agricultural Library | 10301 Baltimore Ave | Room 207
Beltsville, Maryland 20705 | Prepared February 2016

Table of Contents

I. Principles
II. Steps in the data submission process
    Is the Ag Data Commons right for your data?
    Register for an Ag Data Commons account
    Submit your data
    Clone an existing dataset
III. Description of fields on “Edit Dataset” page
    Primary fields
    Purpose & Methods
    Geographic Information
    Citations
    Related Content
    Contact
    Keywords
    Administrative
    Highlight
IV. Description of fields on “Submit Data” page
V. Guidelines: Formatting Your Data
VII. DOI Opt-out
VIII. Review by ADC curator
Appendix 1: Data Dictionary Guidelines
Purpose
Automatically generating a data dictionary
Manually creating a data dictionary
    For a spreadsheet
Manually creating a data dictionary
    For a database
Blank Template
Examples
Questions/Contact

I. Principles

Research datasets and tools that relate to those datasets have value of their own, regardless of whether they are associated with published peer-reviewed literature. This means they can have a different set of authors or publishers, their own digital object identifier (DOI), and their own unique description. The presence or absence of peer-reviewed papers that describe them does not affect a dataset's qualification to be hosted on Ag Data Commons (henceforth referred to as ADC).

On the ADC, a "dataset" refers to the entire metadata record for one or more resources on the ADC; the dataset refers to information entered on the "Edit dataset" page. A "resource" refers to the data files you upload or link to on the "Submit data" page. This may be a spreadsheet, database, photographic collection, software tool, or other resource that you wish to make available for the purpose of advancing scientific knowledge. Again, multiple resources may be included as part of a single dataset, depending on how you wish to define and describe your data.

II. Steps in the data submission process

Is the Ag Data Commons right for your data?

Before proceeding, determine whether your data is suitable for inclusion on the Ag Data Commons.

1. Was the research behind your data all or partially funded by the USDA?
    a. If YES, continue.
    b. If NO, your data is unsuitable for the ADC at this time. As the ADC grows, we may consider expanding to include data with other sources of funding.

2. Is your data able to be freely distributed and re-used by others?
    a. If YES, continue.
    b. If NO, your data is unsuitable for the ADC at this time. We host and support open data to facilitate sharing and reuse.
    c. If you are unsure, please see our section on determining licenses.

Register for an Ag Data Commons account

3. Select "Register" at the upper right corner of the ADC home page.

4. Enter your credentials, and select "Create new account" at the bottom of the page. An ADC curator will be notified of your account request.

5. Once your account has been approved (which will usually be within 5 business days), you will receive an email with details on how to complete your account registration. You must complete the registration process before you can submit data.

Submit your data

6. Now, decide how you want to organize your data: Do you have many datasets or only one with multiple files? As the data submitter, you are best suited to make this decision. Try to imagine yourself as a user of your data, and ask yourself what the most useful grouping would be. Are there enough differences among your data files that they require unique descriptions of methods and data dictionaries (lists of the measurements included in the files)? Will they likely be reused independently, and should have unique citations? If so, you should create distinct ADC records.

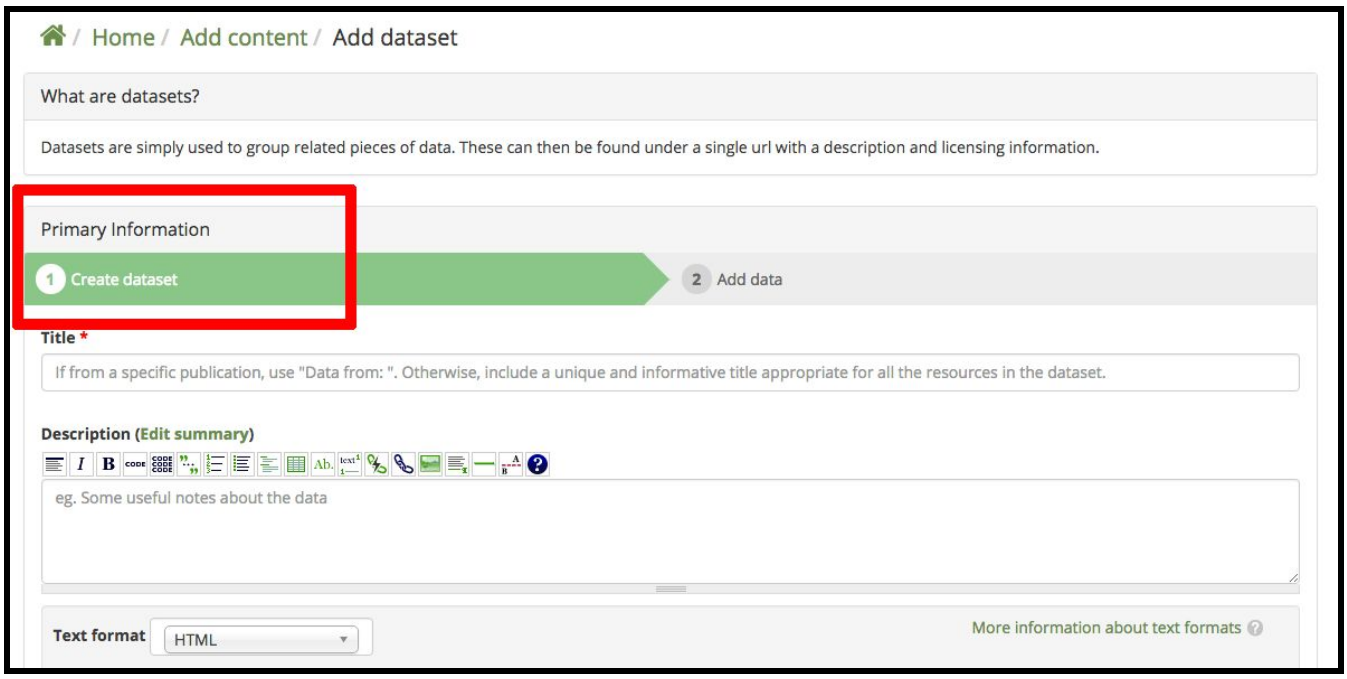
    On the other hand, it would be confusing and redundant to generate unique ADC records where one will suffice. Remember that you can denote relationships among datasets in the ADC like "Is Part of" and "Cites other datasets." In this way it is possible to have separate but related records for your data if you feel that is the most appropriate way to represent it.

7. At this point, you should also decide whether you want your data resource(s) to be published publicly, or if you would like there to be an embargo. We allow for

resources to be embargoed for up to 3 years. See “Scheduling Options” under [Description of fields on “Submit Data” page]() for more information.

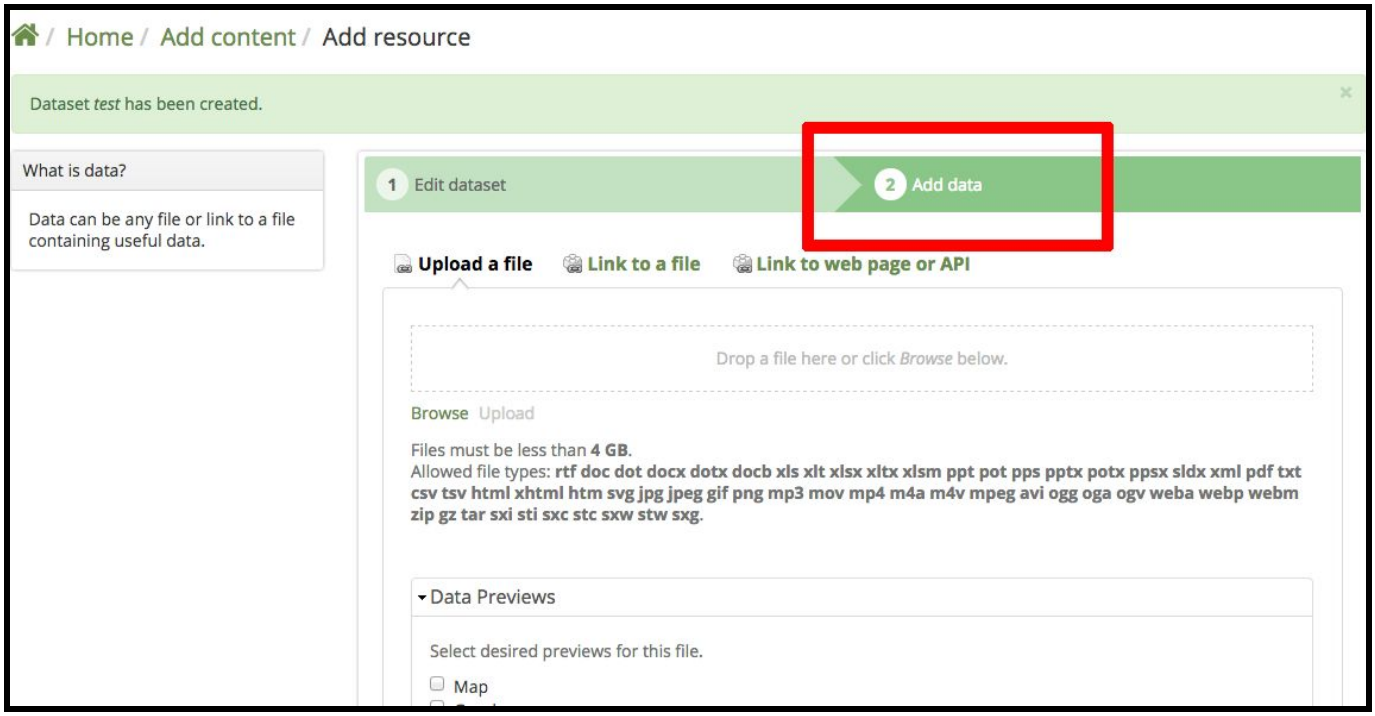
8. Once you have made these decisions, select “Log in” at the upper right corner of the ADC home page.

9. Select “Submit Dataset” at the upper right of the screen. Now you should be on the “Create Dataset” form.



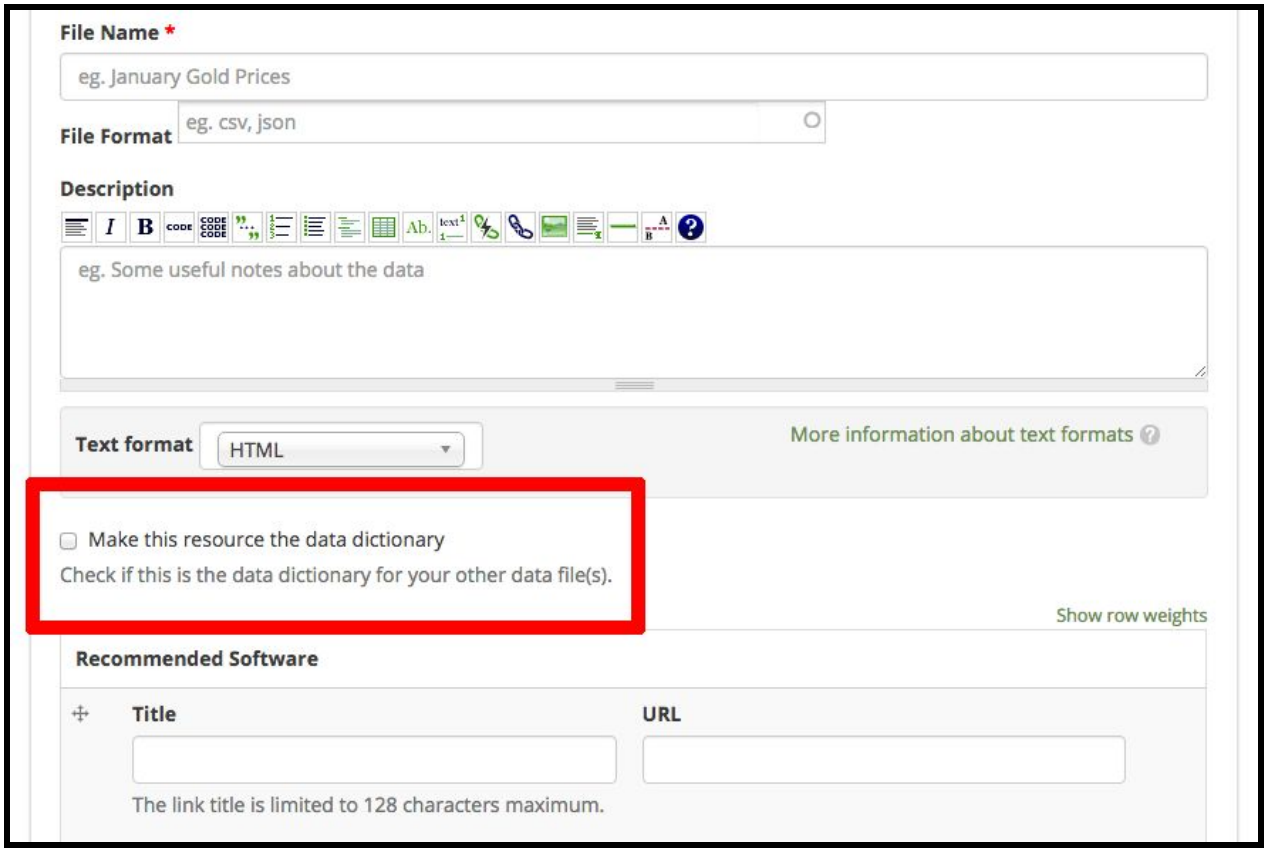
10. Proceed to populate the fields shown. See our [Description of fields on “Edit Dataset” page]() for detailed explanations of each field.
    a. Note that the gray rows toward the end of the page (such as “Purpose and Methods” and “Temporal Information”) are expandable. Select each header to view the full range of fields ADC has for describing your data.

11. After you have finished filling out the dataset creation form, select “Next: Add data” at the bottom of the screen. Now you should be on the “Add data” form.



12. Upload or link to your data files and any related content. Make sure they are named and formatted the way in which you want them to appear. See our Description of fields on “Submit Dataset” page for detailed explanations of each field.

13. Remember to upload your data dictionary as a resource here, and check the box titled “Make this resource the data dictionary” near the center of the page. See our Data Dictionary Guidelines for more information on preparing data dictionaries.



14. If you need to add resources to your dataset record at a later date, you can return to your record in the future and select “Add Resource” at the top of the Dataset page. To edit resources at a later date, select on the resource title, then select “Edit” at the top of the page.

15. If navigating from the “Edit Dataset” page, select “Save” at the bottom of the page to save changes. If navigating from the “Submit Data” page, select “Save” or “Save and add another” at the bottom of the page to save changes, or to save and add an additional resource.

16. When you view your saved datasets/resources on ADC, they will be highlighted in pink. This indicates they are unpublished, nothing else.

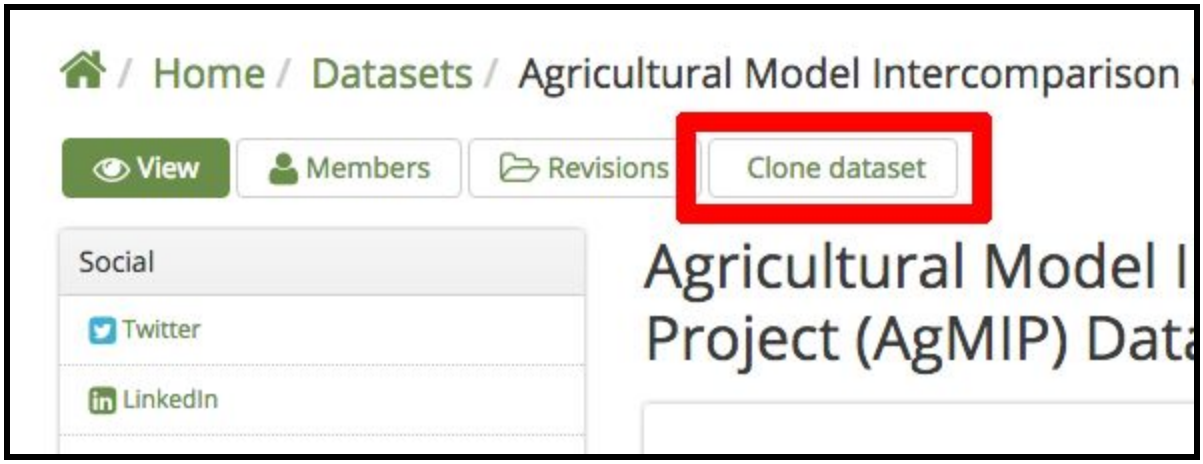
17. An ADC curator has now been automatically notified of your submission. We will review your dataset and contact you if we need more information to ensure the data is adequately described.

18. An ADC curator will notify you when your dataset is published. Note that we reserve the right to refuse publication for any reason.

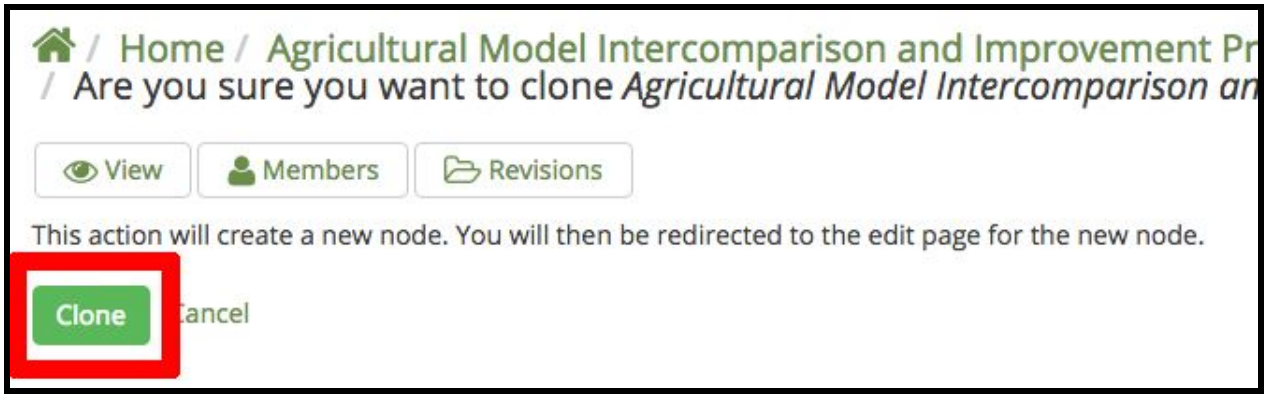
Clone an existing dataset

There are many cases where a content author wants to submit an updated version of a dataset already in the Ag Data Commons, or wants to upload a similar record to one that has already been created. If many similarities exist between an existing record and one not yet created, content authors can choose to clone one of their existing datasets as opposed to creating the dataset from scratch. As a new version, the cloned dataset has slight differences in metadata and will receive a new DOI if applicable, but most of the metadata between the original and new datasets is the same.

1. Log in with your user account
2. Navigate to an existing dataset you have created whose metadata is almost exactly the same as the dataset you want to create
3. Click on the “Clone Dataset” button at the top of the record



4. You are now on the confirmation page. Click on the “Clone” button to create the new dataset



5. Create a clone: a new dataset is created prepopulated with the metadata from the original dataset, but not including any resources or DOI or landing fields
6. You can then change any of the fields for the newly created dataset
7. The new dataset automatically includes a pointer to the old dataset

III. Description of fields on "Edit Dataset" page

*** Required** **Note that all fields without an asterisk are optional. If N/A, leave blank.**

Primary fields

- **Title*** : Enter a descriptive dataset title as you want it to appear; include dates, locations, and specific metrics that make your dataset unique. If the data is from

a Primary Article Citation (see below), use the naming convention "Data from: title of article"

- **Description*** : A rich free text description that provides as much explanation as possible about the dataset: how and why it was generated, and how it should (or should not) be used. This can be modified from article text (e.g. Abstract, Methods, Objectives), but should focus on characterizing the data, not the journal article. **Please provide explanations for all acronyms and abbreviations.**

- **Author**:
  - Name* : Enter the Last name, First name (e.g. Doe, John) of all persons involved in the data collection. Authors can be different from those listed in a primary/related article (or presented in a different order); multiple allowed.

  - Identifier Kind: Select from the drop-down list to select the unique author identifier kind, if applicable. For example, ORCID, ResearcherID, etc.

  - Identifier: Enter the unique author identifier here, if applicable. For example, an ORCID identifier should be entered as 1234-5678-9123-4567.

- **Dataset DOI (digital object identifier)**: DOI for the dataset, not the journal article that may be referencing it. If a DOI does not already exist, ADC will obtain one for you. See our section on opting out of this option if you do not want a DOI created for your dataset.

- **ISO Topic**: High-level subject categorization, also referred to as ISO Topic Categories. Select one or more from the drop-down list; multiple allowed.

- **Product Type**: This field automatically defaults to Dataset. However, it may be changed to better describe the main or most important part of the dataset's actual content (e.g. Computer model, Photograph collection, Database, Presentation, Software Tool, etc.).

Purpose & Methods

- **Intended Use**: Explain the intended use and benefits of the dataset. What purpose do you expect the data to serve? For example, precipitation data may be collected to study patterns of groundwater recharge, to validate watershed models, etc; Life-cycle assessment (LCA) data may be intended for a wide range of impacts in private or public use and/or for product comparisons, etc.

- **Use Limitations**: Explain the limitations regarding the dataset's usability. For example, estimates may be biased over water, equipment may have malfunctioned during a specified time, granularity may mean it is unsuitable for certain kinds of analysis.

- **Equipment or software used:** Name the equipment and software used to collect and process the data. Provide make and model, name and version number, and a stable URL for each tool used to collect and process the data.

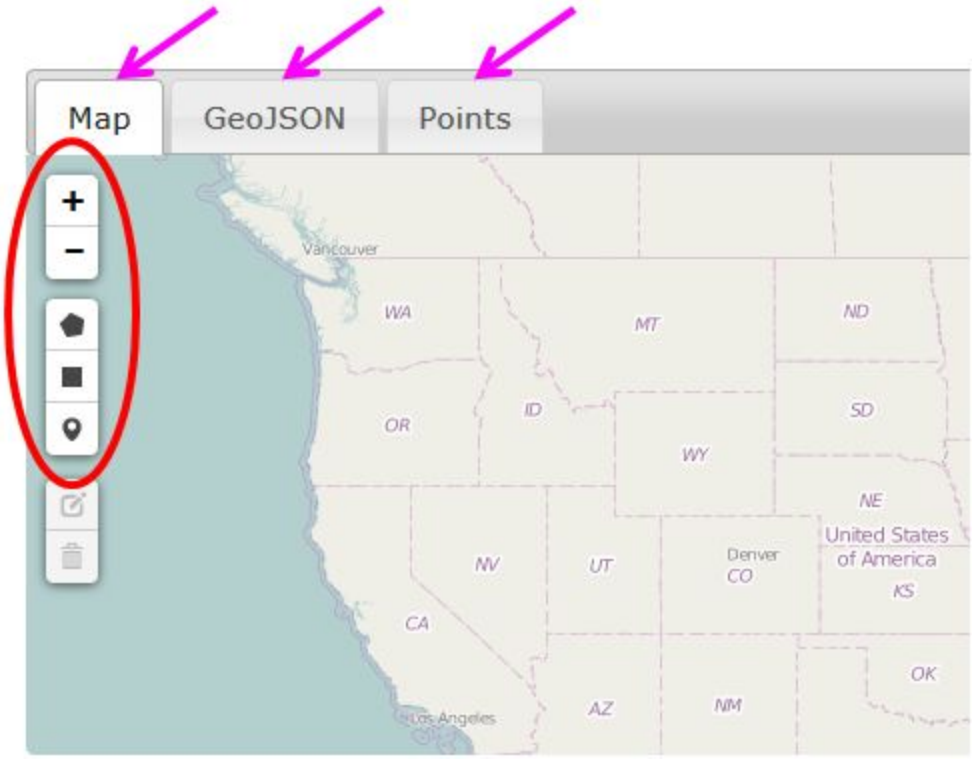
Geographic Information

- **State or Territory**: Select as many as are applicable from the drop-down list of states; multiple allowed.

- **Spatial Description**: This free text can be an address, city, state, region, or other spatial description. Geonames are recommended but not required.

- **Global Map**: You may also use the interactive map (projected in WGS84) to indicate where your data were collected. See the left side of the map for buttons to manipulate data input.
    - The + and - enable you to zoom in or out, depending on the level of geographic detail you wish to represent. If you have global data, you may zoom out as much as necessary to indicate data were collected across multiple countries.
    - Select the third button on the left to enter one or more polygons
    - Select the fourth button to enter one or more bounding boxes

- Select the fifth button to drop one or more points on the map



Select the tabs at the top of the map to enter data in other ways. The GeoJSON tab enables you to enter raw GeoJSON data. GeoJSON is a technical standard that will appeal to GIS enthusiasts. See DCAT spatial/geographical coverage for more info.

Select the last tab on the right, Points, to manually input one or more points in decimal degrees. This is recommended if you know the exact point where data were collected.

These features can be used exclusively or in combination to enter multiple points, polygons, or bounding boxes to a single record if applicable.

Temporal Information

- **Temporal Coverage**: The span of time during which data were collected. Add temporal coverage in one of the following formats (start date/end date; if data collection is ongoing, leave the end date blank):

    - YYYY-MM-DD/YYYY-MM-DD
    - YYYY-MM/YYYY-MM
    - YYYY/YYYY

  See DCAT spatial/geographical coverage for more info.

- **Frequency**: The frequency with which dataset is published e.g. None, Daily, Weekly, Monthly, Annually, Continuously, Irregularly, Decennial - R/P10Y, Quadrennial - R/P4Y, Bimonthly - R/P2M, etc. For example, data with a publish frequency repeating once every 10 years would be designated Decennial - R/P10Y. See DCAT frequency for more info.
    - See the Project Open Data Standard definition and usage of this field at the following link: https://project-open-data.cio.gov/v1.1/schema/#accrualPeriodicity

Citations

- **Primary Article**

    - **Full Citation:** Enter the full bibliographic citation, in APA, to a published article that directly describes this dataset (i.e. a data paper). Leave blank if there is no primary article connected to this dataset.

    - **Article DOI:** Enter the article's DOI as a full URL. For example, "10.123/ABC/123"

    - **PubAg AGID:** If the article is in PubAg, enter the AGID here. When you navigate to an article in PubAg, the URL will look like this: http://pubag.nal.usda.gov/pubag/article.xhtml?id=61025
      The AGID is the string of numbers following "id="
      Leave blank if not applicable.

- **Methods Citation**

    - **Full Citation:** Enter the full bibliographic citation, in APA, to a published article that describes the procedures for data assembly in greater detail; multiple allowed.

- **Article DOI:** Enter the article's DOI as a full URL. For example, "10.123/ABC/123"

- **PubAg AGID:** If the article is in PubAg, enter the AGID here. When you navigate to an article in PubAg, the URL will look like this: http://pubag.nal.usda.gov/pubag/article.xhtml?id=61025
  The AGID is the string of numbers following "id="
  Leave blank if not applicable.

- **Related Article**

  - **Full Citation:** Enter the full bibliographic citation, in APA, to a published article that is related to this dataset. Use your discretion to list only the most useful and relevant one(s); multiple allowed.

  - **Article DOI:** Enter the article's DOI as a full URL. For example, "10.123/ABC/123"

  - **PubAg AGID:** If the article is in PubAg, enter the AGID here. When you navigate to an article in PubAg, the URL will look like this: http://pubag.nal.usda.gov/pubag/article.xhtml?id=61025
    The AGID is the string of numbers following "id="
    Leave blank if not applicable.

- **Preferred Dataset Citation**: If you have a specifically formatted citation for this dataset that you want others to use, enter it here. This will ensure your dataset is cited to your satisfaction if/when it is re-used. If this field is left blank, a standard dataset citation will be constructed for you.

- **Metadata Sources:** If you are using metadata that has already been prepared for this dataset, indicate your sources of metadata here.

Related Content

- **Is Part Of**: Is this dataset part of a larger dataset on the ADC? If so, type the first letters of the container dataset's title, then select from the autocomplete suggestions.

- **Cites other datasets**

    - **Full Citation**: Provide the full citation(s) for datasets that were used to construct this one. This is particularly relevant for datasets that are aggregations or adaptations of other data.

    - **DOI or other link**: Provide the cited dataset's DOI as a full URL, e.g. "http://dx.doi.org/10.123/ABC/123". If there is no DOI, provide another link to the dataset.

- **Related To**: Provide a name and URL to a resource that provides additional context to the dataset, e.g. a project website, manual, or data documentation. This field is not to denote relationship among datasets, but relationships among the dataset and non-dataset resources; multiple allowed.

Contact

- **Contact Name**: Enter a long-term contact person for the dataset. This can be different from the dataset author/originator e.g. a data curator or administrator. See Project Open Data for more info. Format as Last name, First name.

- **Contact Email**: Contact person's email address.

- **Reviewer**: Name of person other than author who has reviewed and approved the data (not the metadata); multiple allowed.

- **Publisher**: Publisher of the dataset. This will be displayed in the dataset citation ADC suggests to users of your data and is required to obtain a DOI. The field will default to "Ag Data Commons" unless you change it.

Keywords

- **User-supplied Tags**: Free text keywords to facilitate discovery of the dataset. This field is not part of a controlled vocabulary, and is intended to capture meaningful tags to support search and browse; multiple allowed.

- **Project**: Choose from the hierarchy to select a project designation, if applicable. These projects are specific to the National Agricultural Library.

Administrative

- **License**: Enter the license assigned to this dataset; for federally generated data, this should be a Public Domain dedication such as CC Zero or US Federal Public Domain. Unsure which license to choose? Walk through the following questions to help determine the right one. The following is based on the latest guidance from the Federal CIO council.

1. Are all the data owners federal employees who generated the data as part of their work?
    a. If YES, then the license recommendation is CC Zero.
    b. If YES, but you are concerned about international public domain, then the license recommendation is US Federal Public Domain.
    c. If NO, see #2.

2. If all data owners are not federal government employees, were the data generated under a collaborative grant, or other licensing agreement?
    a. If YES, check the terms of that agreement for a “rights in data” or other related clause. Use the licensing information specified there.
    b. If YES, but no specific license is mentioned in your agreement, see #3.

3. If your licensing agreement allows non-federal data owners to choose a license, the license recommendation is CC Zero (preferred), CC BY, or CC BY SA.

License definitions and additional information can be found at http://opendefinition.orglicenses/ or https://creativecommons.org/licenses/

   a. If more restrictive licensing is requested by non-federal data owners, then the data are unsuitable for the ADC. We host and support open data to facilitate sharing and reuse.

4. If your data was generated with no US federal funding, then they are unsuitable for the ADC at this time. As the ADC grows, we may consider expanding to include data from externally funded researchers. If you have questions or need assistance, email us at NAL_ADC_Curator@nal.usda.gov.

- **Public Access Level:** Enter the degree to which this dataset *could* be made publicly available, regardless of whether it has been. See Project Open Data for more info.

- **Bureau Code**: Please specify one or more bureau codes. For example USDA-Agricultural Research Service (ARS) funded research should be “005:18”. See Project Open Data for more info.

- **Program Code**: Please specify one or more program codes. ARS and National Institute of Food and Agriculture (NIFA) funded research should be “005:037 - Research and Education”. See Project Open Data for more info.

- **Funding Source(s)**: Indicate all sources of funding for the dataset. Include the institution name and project number, if available; multiple allowed.

- **ARIS Log Number:** This only applies to datasets created by Agricultural Research Service employees. Enter the ARIS log number, if applicable.

Highlight

- **Highlight Image**: This image should be an informative and attractive picture or graphic, and may be used to accompany your dataset. Files must be less than 4 GB, and in one of the following file types: .png .gif .jpg or .jpeg. ADC curators reserve final approval for all image selections.

IV. Description of fields on “Submit Data” page

*** Required Note that all fields without an asterisk are optional. If N/A, leave blank.**

- **File Name***: Enter a descriptive resource title as you want it to appear; if uploading multiple resources, include information that distinguishes this one from the others you will provide. It is particularly helpful to include the format of the resource in the title and it is not necessary to repeat details that are in the title of the dataset.

- **Option to Link to a file, Link to an API, or Upload a File**:

  - **Upload a file:** If you are uploading a file, multiple can be dragged into the interface window at once. Or, select “+ Add Files” in the lower left corner of the dialog box. Remember to select “Start upload.”
    - If your file is in csv or text format, select the delimiter used, if applicable.

  - **Link to a file:** Enter a link to one of the following file formats: csv, html, xls, json, xlsx, doc, docx, rdf, txt, jpg, png, gif, tiff, pdf, odf, ods, odt, tsv, geojson, or xml.

  - **Link to an API:** Link to an API if you want to link to a web page.

- **File Format**: Specify format of the resource (e.g. CSV, HTML, XML, etc.). All resources linked to from an API should be designated as HTML regardless of the type of resource found at that link.

- **Description**: Provide a detailed description of your resource. For example, if it is an Excel file with multiple tables, describe what the different tables contain.

Imagine that you are a novice user: what information would you need to make sense of this resource?

  - **Text Format**: This refers to the text format of the description block. Although it should not be necessary, you can change the format to suit your preference.

- **"Make this resource the data dictionary" checkbox**: Check this if the resource you are uploading is the data dictionary. See Appendix 1 of this document for more information on data dictionaries.

- **Recommended Software:** Provide a name, version number, and stable URL for software tools recommended to view or run this resource.

- **Dataset:** This field will auto-populate with the title of the dataset associated with this resource. Do not edit or add to this field.

- **Weight**: This field determines the order in which the resources will be displayed on the dataset page. The lowest value will be displayed first.

- **Described by**: This field is restricted to ADC curators. Submitters should leave this field blank.

- **URL Path Settings**: A URL will automatically be generated for this metadata record unless you uncheck the "Generate automatic URL alias" box; once it's unchecked, you may enter a customized URL if you wish.

- **Revision Information**: If necessary, use this feature to revise your data resource, and provide an explanation as to why it is being revised.

- **Scheduling Options**: If you would like to place an embargo on the publication of your resource, enter the date you would like it to be published here. Format is YYYY-MM-DD. You may enter a date up to three years in the future. If you would like your data resource to be published immediately upon curator approval, with no embargo, leave this field blank.

V. Guidelines: Formatting Your Data

- Provide data in csv (comma-separated values) wherever possible for tabular data.

- Include meaningful column headers in the spreadsheet: This will allow the data to be converted into other formats and display successfully.

- Author your spreadsheet one table per tab, as opposed to more than one table in a tab. This ensures the data will be machine readable.

- No blank rows or columns between data.

- Do not use zeroes or leave a cell blank. Select a code to identify missing data; using -999 or -9999 is a common convention. Indicate the code for missing data in the data dictionary.

- When exporting data to another format, check to ensure that no cells with missing data have zeros, or are blank in the resulting file.

- Use standard terminology wherever possible (include both the Latin and common names for plants and animals). As linked data becomes the norm, this will increase the impact of your dataset by making it easier to find.

    - As a scientific researcher, you may have access to disciplinary thesauri or ontologies that you wish to use. If not, a couple suggestions are the National Agricultural Library Thesaurus (a broad thesaurus of Agricultural Terms), and the Integrated Taxonomic Information System (which provides standardized names for plants, animals, fungi, and microbes).

VII. DOI Opt-out

By default, ADC will mint a new DOI for your dataset upon submission. If you already have a DOI, or otherwise wish to opt-out of this service, please e-mail: NAL-ADC-Curator@ars.usda.gov

In the body of your email, include the title and author(s) of the dataset that should *not* receive a DOI.

VIII. Review by ADC curator

Once your data has been submitted, an ADC curator will be notified that your record is ready for review. She will:

1. Review your metadata and contact you if further information is needed to ensure the data is sufficiently described.

2. Add ADC and NALT keywords to enhance retrievability, making your dataset easier to find in the repository.

3. Obtain a DOI if your dataset does not already have one.

4. Contact you before your data goes live on ADC so that you can review and approve the final record.
    a. Note that ADC reserves the right to deny publishing your dataset for any reason.

Appendix 1: Data Dictionary Guidelines

Table of Contents

Purpose

Automatically generating a data dictionary

Manually creating a data dictionary

- For a spreadsheet
- For a database

Blank Template

Examples

Questions/Contact

Purpose

Data dictionaries are used to provide detailed information about the contents of a dataset or database, such as the names of measured variables, their data types or formats, and text descriptions. A data dictionary provides a concise guide to understanding and using the data. Ideally, all Ag Data Commons (ADC) records for datasets and databases should include or point to a data dictionary. It is preferred that these data dictionaries be machine readable, in csv format.

If your data are managed in a standard relational database you will likely be able to generate a data dictionary through your software. This will provide a document that is consistently formatted and contains what is needed for others to understand your data. See the following section for more information.

If your data are managed in spreadsheets, text files, or comma separated values, you will need to manually prepare a data dictionary. To support machine-readability, we recommend preparing your data dictionary as a spreadsheet. If you prefer to prepare it as a .doc or .pdf, we recommend embedding a data dictionary table in your document that can be easily extracted. A data dictionary template and examples can be found toward the end of this document.

The following are recommended guidelines for data dictionaries; not requirements. These guidelines are subject to change, as best practices are evolving.

Automatically generating a data dictionary

- If your data is stored in a relational database, it may be able to generate a data dictionary for you.
- If your data is stored in a spreadsheet, you will need to manually create a data dictionary.

Enterprise-level databases often contain built-in tools for automatically generating data dictionaries. Consult your database administrator or software documentation for instructions specific to your system.

To generate a data dictionary from MS Access, select the "Database Tools" tab, then select "Database Documenter" (under “Analyze”). Now you should see the Documenter dialog box. If you have entered a description of your fields in Design View, they will carry over to the generated data dictionary here. It should resemble the following:

	ColumnWidth:	Default	
	DataUpdatable:	False	
	DecimalPlaces:	Auto	
	Description:	(g/cm3), Particle density as mass of solids per volume of solids.	
	DisplayControl:	Text Box	
	Format:	@;"No data"	
	OrdinalPosition:	3	
	Required:	False	
	SourceField:	particle_density	
	SourceTable:	soil_properties	
porosity		Double	8
	AllowZeroLength:	False	
	AppendOnly:	False	
	Attributes:	Fixed Size	
	CollatingOrder:	General	
	ColumnHidden:	False	
	ColumnOrder:	Default	
	ColumnWidth:	Default	
	DataUpdatable:	False	
	DecimalPlaces:	Auto	
	Description:	(cm3/cm3), Measured porosity of soil as volume of voids per bulk volume. Please provide measured values only, do not provide a value that is calculated from the bulk and particle densities.	
	DisplayControl:	Text Box	
	Format:	@;"No data"	
	OrdinalPosition:	4	
	Required:	False	
	SourceField:	porosity	
	SourceTable:	soil_properties	
OM_content		Double	8
	AllowZeroLength:	False	
	AppendOnly:	False	
	Attributes:	Fixed Size	
	CollatingOrder:	General	
	ColumnHidden:	False	
	ColumnOrder:	Default	
	ColumnWidth:	Default	
	DataUpdatable:	False	
	DecimalPlaces:	Auto	
	Description:	(%), Mass of organic matter content as a percentage of the total solid mass. If necessary, estimate organic matter content based on carbon content and explain in table [comment_soil_properties] field	

Manually creating a data dictionary

For a spreadsheet

Submit a spreadsheet (.XLS or .XLSX) with one tab for introductory information, and separate data dictionary tabs that correspond to each existing tab in your dataset. For example, if your dataset consists of three tabs, your data dictionary will have four tabs: the first for introductory, background information, and three more to correspond to the three tabs of data. Consider using our data dictionary template to get started.

Best Practices:

- ★ Submit your data as a spreadsheet or csv
- ★ One table per tab
- ★ No extraneous comments
- ★ No empty cells, columns, or rows (enter n/a if nothing applies)
- ★ Spell out all abbreviations
- ★ Element definitions should be stated in the singular, be succinct, and be able to stand alone from other element definitions
    - ○ This complies with recent recommendations to data stewards in the ISDA, and is based in part on the ISO/IEC 11179, Information Technology - Metadata Registries (MDR) standard (http://metadata-standards.org/11179/)

If you would rather submit a DOCX or PDF, embed tables in your document so they will be exportable. Submit a .DOCX or .PDF with the following:

- Introductory and explanatory text
  - Explain context: is if from a singular research article, or a larger project?
  - Provide a URI (Uniform Resource Locator), which will usually be a URL or a DOI (Digital Object Identifier) for the dataset or related journal article.
  - Other pertinent information such as version, date released, etc.)

- A listing of elements (fields), in addition to the following:
  - Element source table
  - Element definition
  - Element variables
  - Element data type
  - Element field length
  - Required y/n and/or null value note

Manually creating a data dictionary

For a database

Use this option only if you are unable to automatically generate a machine-readable data dictionary. Guidelines for automatically generating a data dictionary from a database are in the previous section of this document.

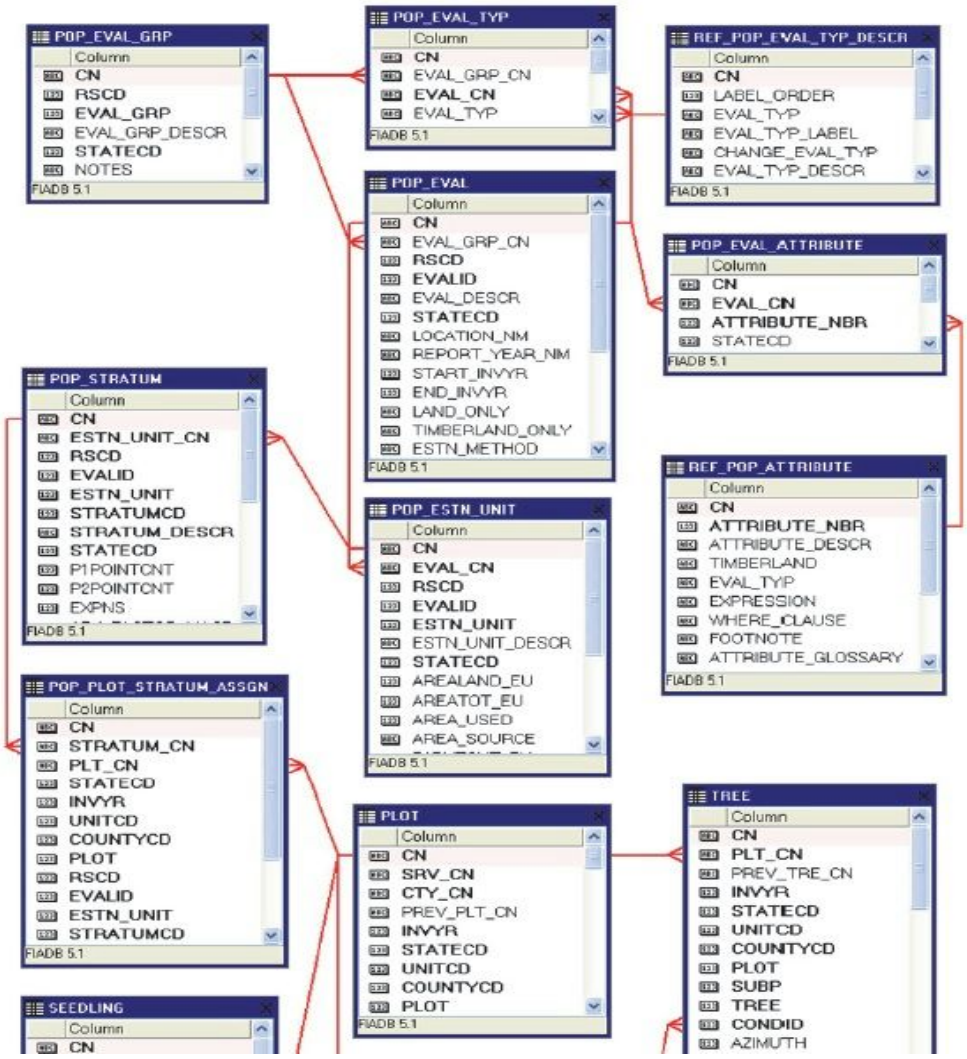
We suggest submitting your data dictionary as a spreadsheet (consider using our blank template below). If you would rather submit a DOCX or PDF, embed tables in your document so they will be exportable. Submit a DOCX or text searchable PDF with the following:

- Introductory and explanatory text
  - Explain context: is if from a singular research article, or a larger project?
  - Provide a URI (Uniform Resource Locator), which will usually be a URL or a DOI (Digital Object Identifier) for the dataset or related journal article.
  - Other pertinent information such as version, date released, etc.)

- A listing of elements (fields), in addition to the following:
  - Element source table
  - Element definition
  - Element variables
  - Element data type
  - Element field length
  - Required y/n and/or null value note

- If possible, a data diagram or data model showing the relationships among tables. For example:

Figure 4. National FIADB linkages -- Relationships among phase 1 and population estimation tables to the phase 2 plot and other frequently used tables.



Source:[The Pacific Northwest Forest Inventory and Analysis Database]

Blank Template

This can be used to manually create a data dictionary. Use one row for each data element, and do not leave rows, columns, or cells blank. Add rows and columns as necessary, and enter n/a if nothing applies. See below for an explanation of column headers.

Spreadsheet tab	Element or value display name	Description	Data type	Character length	Acceptable Values	Required?	Accepts null value?

Explanation of column headers:

- **Spreadsheet tab**: If your spreadsheet has multiple tabs, identify the tab you are describing.
- **Element or value display name**: What is the name used in your data file?
- **Description**: Write a brief definition, stated in the singular, that could stand alone from other element definitions.
- **Data type**: For example, indicate varchar, integer, date, etc.
- **Character length**: For example, the maximum length for Excel is 255, so indicate 255 or less.
- **Acceptable values**: List all acceptable values, separated by pipes (|). This may be a field name or a range of values.
- **Required?**: Enter y/n to indicate whether this field is required.
- **Accepts null value?**: This required to run calculations on your data. Indicate y/n if null value is allowable.

  A note on null values: Null is the absence of a recorded value for a field. A null value differs from a value of zero in that zero may represent the measure of an attribute, while a null value indicates that no measurement has been taken.

Examples

Examples 1 & 2: Created by the ADC team using the source dataset, [Data from: Enabling proteomic studies with RNA-Seq: the proteome of tomato pollen as a test case:]

Spreadsheet tab	Element or value display name	Description	Data type	Character Length	Acceptable Values	Required?	Accepts null value?
S1	454db1	454db1 source Database searched for RNA sequences	varchar	255	n/a	y	n
S1	SGN	SGN source Database searched for RNA sequences	varchar	255	n/a	y	n
S1	All unigenes	Total number of unigenes from database searched	integer	255	All unigenes	y	n
S1	Matched Unigenes	Number of matched unigenes from database searched	integer	255	Matched Unigenes	y	n
S1	Average Length	Average length of RNA sequences obtained from 454 sequencing experiments databases	integer	255	0.0-9999	y	n
S1	Date	date	date	10	YYYY-MM-DD	y	n
S1	Standard Deviation	Standard Deviation of length of RNA sequences	integer	20	0.0-9999	y	n

Example 1

Spreadsheet tab	Element or value display name	Description	Data type	Character Length	Acceptable Values	Required?	Accepts null value?
S5	SGN unigene	Sol Genomics Network unigene build identifier	varchar	255	n/a	y	n
S5	Protein Function	The protein or a complex assembly	varchar	255	n/a	y	n
S5	Closer Organism	Latin name of plant identified as "closer organism" to SGN unigene	varchar	255	n/a	y	n
S5.	Functional Classification	The functional classification of the unigene	varchar	255	Cell growth/division | Cell Structure | Disease/defense/ Stress | Metabolism | Protein Destination and Storage | Protein synthesis | Signal transduction | Transcription | Transport | Other functions |	y	n
S5	SP-NN value	Signal peptide neural network numeric value	integer	20	0.00-0.999	y	n
S5	SP-NN result	Signal peptide neural network result	Char(2)	4	SP: Signal Peptide NN: Neural Network algorithm N: No secreted	y	n

					Y: Secreted		

Example 2

Example 3: Selected columns from the USDA National Finance Center Insight Training Participant Guide Version 1.0, June 2014. Source: [USDA National Finance Center Insight Training Participant Guide Version 1.0 June 2014]

National Finance Center - *Insight* Data Dictionary

			PPS Mapping			Description
Data Element	Data Type	Size	COBOL Data Element	FOCUS Name	ALIAS	Description
APPOINT_TYPE_DESCR	VARCHAR2	50	NA	NA	NA	Insight data elements to describe TYPE-APPOINTMENT-CODE
BI_WEEKLY_END_DATE	DATE	7	N/A	N/A	N/A	End date of pay period.
BI_WEEKLY_START_DATE	DATE	7	N/A	N/A	N/A	Start date of pay period
DAY_END_DATE	DATE	7	N/A	N/A	N/A	Field not intended for ad hoc reporting
DAY_ID	NUMBER	22	N/A	N/A	N/A	Field not intended for ad hoc reporting
DAY_NUM_OF_MONTH	NUMBER	22	N/A	N/A	N/A	A sequential number of the day within the month. Use to indicate the day number of month. Example 26th day of month
DAY_NUM_OF_WEEK	NUMBER	22	N/A	N/A	N/A	A sequential number of the day within the Week. Use to indicate the day number of week. Example 3rd day of
	NUMBER	22	N/A	N/A	N/A	A sequential number of the day within the year. Use to indicate the day number of Year. Example, 239th day of year
	DATE	7	N/A	N/A	N/A	Field not intended for ad hoc reporting
	NUMBER	22	N/A	N/A	N/A	Field not intended for ad hoc reporting
	NUMBER	22	N/A	N/A	N/A	Full name of the fiscal month. (January, February, ..., December)

Example 3

The following are more USDA, ecological, and agricultural data dictionary examples:

- [The Pacific Northwest Forest Inventory and Analysis Database]
- [NRCS New Hampshire Soil Data Dictionary]
- [ICASA Version 2.0 Data Standards for Agricultural Field Experiments and Production]

Questions/Contact

If you still have questions regarding data dictionaries, e-mail us at

NAL-ADC-Curator@ars.usda.gov