# LONG-TERM AGROECOSYSTEM RESEARCH (LTAR) NETWORK DATA SHARING PRINCIPLES AND GUIDELINES

*Section 9 of the LTAR Network Data Management Guidelines*

Version 1.0

11 May 2020

Contributors: Prepared by Nicole Kaplan[1], Dan K. Arthur[2], and Alisa Coffin[3]. Input provided by Joe Alfieri[4], Erin Antognoli[5], David Archer[6], Jennifer Carter[6], Justin Derner[1], David Huggins[7], Earl Keel[3], Tom Moorman[8], Pat Nash[8*], Glenn Moglen[4], Susan McCarthy[5], Eli Moore[5*], Cynthia Parr[5], John Sadler[8*], Tim Strickland[3], as well as LTAR data managers, LTAR site directors, USDA Agricultural Research Service Research Leaders, and others from the USDA National Agricultural Library.

# Contents


Contributor Affiliations ..... 1
Purpose ..... 2
Background ..... 2
Definitions and descriptions ..... 3
Data sharing principles ..... 6
Data sharing expectations ..... 6
Data sharing embargos ..... 7
Terms of co-authorship and attribution ..... 8
Public access and reuse ..... 8
Disclaimer of liability ..... 9
References ..... 9


# Contributor Affiliations

| | *Affiliation* |
|---|---|
| 1 | Central Plains Experimental Range LTAR site |
| 2 | Upper Chesapeake Bay LTAR site |
| 3 | Gulf Atlantic Coastal Plain LTAR site |
| 4 | Lower Chesapeake Bay LTAR site |
| 5 | USDA National Agricultural Library, Beltsville, MD |
| 6 | Norther Plains LTAR site |
| 7 | Cook Agronomy Farm LTAR site |
| 8 | Central Mississippi River Basin LTAR site |
| * | Former affiliation |

## Purpose

The Long-Term Agroecosystem Research (LTAR) network is a research network grounded in experimentation and observation charged with developing a national roadmap for the sustainable intensification of agricultural production. Thus, this document describes LTAR network data sharing principles and guidelines with the intent that all LTAR data will be available for research collaboration and the development of agroecosystem management recommendations and education. The results produced by the LTAR network will be used to inform policy, serving as a legacy of long-term agroecosystem observations for future generations.

The principles and guidelines are intended to:

1. Establish data sharing principles for the LTAR network;
2. Facilitate data sharing among LTAR network and non-network scientists, supporting cross-site and multi-site scientific investigations, as well as the LTAR common experiments;
3. Satisfy USDA policy to make available to the scientific community and to the public all digital scientific data arising from unclassified research and programs funded wholly or in part by USDA;
4. Clarify expectations for authorship and intellectual property rights; and
5. Facilitate development of enhanced data products and data user interface technologies that automate data management processes as well as facilitate comparison of decision management outcomes.

## Background

Consistent with the draft USDA Public Access Policy for Digital Scientific Research (May 17, 2019), the LTAR Data Sharing Principles and Guidelines applies to "digitally formatted scientific data assets resulting from unclassified intramural or extramural scientific research that is supported wholly or in part by USDA, for which the funds were obligated after the approval of this policy" (lines 29-31).

Specifically, the LTAR Data Sharing Principles and Guidelines applies to:

- Digital information, measurements, or statistics that are collected for the purpose of study, analysis, calculation, and/or decision making for LTAR experiments and observations;
- Processes (data management plans, publications, programming scripts);
- Digital scientific data arising from any USDA LTAR related grant or cooperative agreement to awardees including but not limited to states, localities, regulated parties, volunteer organizations, contractors, cooperative agreement holders, grantees, cooperating federal government agencies, intergovernmental organizations, and educational institutions, obligated in fiscal year 2014 or later;
- LTAR digital scientific data potentially within the scope of this policy at the discretion of the dataset creator/steward, and subject to other statutory requirements, including:
    - Models and model-related content, including programming code, parameters, input data, simulated outputs and data products;
    - Data from secondary sources (typically referred to as secondary or outside data).

## Definitions and descriptions

***Common Experiment Data:*** These data are characterized as being collected within areas applied with specific experimental treatments. These data may be collected for a collaboratively designed common experiment with participation from multiple LTAR site scientists and can be made available for research collaborations by contacting the LTAR site data creator or data steward.

***Data Creators:*** People responsible for experimental and dataset design, collection, analysis and quality assurance to ensure data are fit for use. Dataset creators are analogous to authors of a publication, and datasets should be cited in an analogous manner following guidelines described at http://www.datacite.org/whycitedata. Researchers should include dataset creators as collaborators and co-authors on publications for which they provide data.

***Data Embargo:*** A ban on the public release of data for a specified time period.

***Data Products:*** Data and metadata collected, managed and transformed into useful and usable digital information. Knowledge products use data products to create tools, such as a decision support application.

***Data Publications:*** Archived and accessible in a network data management system, data publications are published and curated by a web-based data repository. A Digital Object Identifier (DOI) is assigned for citation.

***Data Sharing:*** Distributing scientific output within and amongst the LTAR network and its collaborators. Sharing data requires development of metadata, including data dictionaries to define variables and units. This approach in turn creates the potential for broader data sharing via mapping to standard network exchange formats (Servilla et al. 2016). Standard network exchange formats are expected in order to facilitate data integration, analysis, and publication, as well as to inform design and development of innovative network data systems.

***Data Stewards:*** People, usually LTAR site leaders or data managers, responsible for managing individual datasets to promote data accessibility and reusability across the network, and to meet data content and data quality requirements of all users. Act as liaison if the data contains personally identifiable information (PII) or intellectual property (IP), and in the case where the data creators are no longer engaged in LTAR network research.

***Fitness for Use:*** Data products are determined as fit for use by data creators in the context of the peer reviewed project research design and data management plans for which they were collected. It is accomplished by establishing well-documented quality assurance and quality control (QA/QC) procedures, and/or information on the stage or level of (QA/QC) checking, and/or production (e.g., provisional and certified or final product). Fitness for use ensures researchers can acquire valid

scientific data with high integrity to answer their questions. Researchers should also consider their tolerance for error when including provisional data products in analysis activities (Johnson 2017).

***LTAR Data Inventory:*** An all-site data inventory created for LTAR scientists and prospective collaborators to discover data that LTAR sites have within their local site data management systems. The data inventory includes information such as variable names, units, frequency of measurement, temporal extent, experimental design, and creator and/or steward contact information. The data inventory is being used to create a controlled vocabulary of variable names to facilitate cross-site comparisons and keywords to enhance data discovery in the future. The data inventory will include data under embargo but which have not yet been released to the public. Available data may include legacy data available only through direct contact with the LTAR site. The inventory may also include summaries of data that are subject to privacy concerns.

***Metadata:*** Metadata is information about data. According to Michener et al. (1997), "Metadata represent the set of instructions or documentation that describe the content, context, quality, structure, and accessibility of a data set." All metadata should be ISO 19115 compliant where possible (see FGDC metadata standards, https://www.fgdc.gov/metadata/iso-standards). The National Agricultural Library provides data management policy and planning, repository management, data and metadata curation and consultation, and preservation. (https://www.nal.usda.gov/main/data). NOAA provides an example of the minimum elements required for ISO 19115 (https://geo-ide.noaa.gov/metadata-standards) and USGS provides an example of ways to create metadata for various types of data (https://www.usgs.gov/products/data-and-tools/data-management/metadata). At a minimum, metadata should contain contact information, a meaningful title, information on the experimental design, methods including instrumentation types, temporal and spatial coverage, and definitions of variables with units (i.e. a data dictionary). Inclusion of QA/QC procedures and file format descriptions is encouraged. Geospatial datasets should also include geoprocessing steps. Synthesized data products, which are derived from other data sources, should also be documented with complete metadata from primary sources. Such metadata will help users understand in detail how the data were collected and determine appropriate applications for future re-use. Archived data should use non-proprietary formats.

***Model Input Data and Model Outputs:*** Computer models typically receive input in the form of a file that contains information of various types: control data which communicates to the computer model how to operate, model parameters that relate to governing equations executed within the computer model, input parameters and input data that describe the biophysical characteristics of the modeled system, and observational data used for model calibration, validation, and performance assessment. Highly processed input and observational data for models will be made available to the LTAR working groups and collaborators associated with modeling activities. Computer models produce output data, sometimes also referred to as "pseudo-data". Output data may be a processed version of observational data or may represent simulations of hypothetical conditions. In some cases, model input data are outputs of antecedent models, for example, weather and climate models provide inputs to economic models. The specific model and version should be documented in the metadata (i.e., via link to

information) on specific executable and parameter file, or within program packages with which these data were compiled and/or run.

***Network Data Management System:*** A centralized system, which is co-designed by people in the community with expertise in hardware, software, standards, workflows, practices, and policies, to enable data management for a network of sites. A network data management system provides data management services, organization of data, searching and browsing for data, integrating data and information from multiple sites for cross-site queries, and downloading data and metadata for collaborators. The Agricultural Collaborative Research Outcomes System (AgCROS; https://agcros-usdaars.opendata.arcgis.com/) is the primary LTAR network data management system.

***New Data Collection Co-located with LTAR:*** Data collection efforts co-located with an LTAR site should follow the guidelines from the LTAR site research sponsor. Visiting scientists and students should follow the guidelines from their LTAR site research sponsor for sharing data, but contact information as well as descriptions of the purpose, methods, and locations of their studies must be collected and stored within the LTAR site data management system.

***Privacy Restricted Data:*** Data that can be used to easily identify individual LTAR collaborators, including specific geolocated information about production capacity and producer identity. While geospatial data alone (e.g., field boundaries) may not rise to the level of privacy restricted data, care will be taken, at the discretion of LTAR agreement holders, to mask the data prior to publication or public release, to prevent the discovery of personally identifiable information (PII). Masking procedures may follow similar practices used by other US government agencies and masked records will include a flag denoting the type of mask.

***Public data:*** Distributed scientific output in public data repositories with open access.

***Public Data Repository:*** A publicly accessible web-based data system usually hosted by an institution, library, or center in which metadata and/or final data products are curated, published, and discoverable. Repositories may also allow interoperability via machine-to-machine access between systems and repositories, such as with an Application Programming Interface (API). Public data repositories may contain metadata that links to another repository curating the data and can place an embargo and/or restriction on public access to data for a specified period of time to address any concerns related to timing of publication. A repository can provide long-term preservation and curation of data, ensuring data availability and integrity in the future, as well as DOIs for use in citations. Data repositories created by USDA can be accessed via the National Agricultural Library's Ag Data Commons (https://data.nal.usda.gov/) data access system.

***Quasi Real-Time Data:*** These data are sensor-based, collected frequently, and are large homogeneous files. They are most often compiled as a time-series and are appended regularly to the local long-term data record. They may include driving variables that characterize systems, such as weather data,

stream gage data, and may not be collected exclusively as part of an experimental or observational research design.

***Site Data Management System*** – A local system for a site, which is co-designed by dedicated data scientists working with researchers, for managing the workflow of data, storing data, tracking versions of data (e.g., provisional or final), implementing quality control procedures, documenting metadata, and preparing data for sharing and/or public access.

***Working Data*** – Data that may be provisional and have not yet been published or determined as "fit for use".

## Data sharing principles

The LTAR scientific community recognizes that:

1. Data produced by the LTAR network must follow USDA policies for sharing data and public access;
2. The LTAR network recognizes the value of allowing free interchange of data and working data will be made available through collaboration;
3. The production of scientific data represents extensive, concerted expertise, time, effort and financial investment;
4. Data in the LTAR network are produced in response to scientific research questions. As such, they are intellectual property and are not intended to be provided as a service in and of itself;
5. The LTAR network welcomes internal and external research collaborations wherein network scientists are fully involved in the data collection, analysis, and publication;
6. Data creators/stewards are responsible for negotiating data access and establishment of collaborations, as described in "Data Sharing Expectations", below.

## Data sharing expectations

LTAR experiments are designed to identify phenomena and correlate mechanisms in agroecosystems that act over longer time scales, so the temporality influences how the scientific data will be interpreted. Therefore, in many cases the common experiment must run its course for the data to be meaningful. After termination of the period of data collection, these data will be packaged into products and are expected to be released to the public within 30 months of the completion of data collection, as defined in the project plan.

Sharing of working data is expected and will be conducted in collaboration with LTAR scientists. The success of network research projects depends upon efficient re-use of shared data from sites and from collaborations with non-LTAR sites and/or researchers who can provide expanded context for interpretation of LTAR data, data products and technology transfer information. In addition, as funders and publishers require authors to share data as supplemental materials, expectations for public access to data will increase.

All LTAR investigators and collaborators who receive LTAR support[1] agree to:

1. Share experimental project plans and designs with the LTAR network, prior to conducting experiments or observational studies, in a central repository like Basecamp. Project plans should include a data management plan. See examples for the level of detail expected.

2. Manage data and metadata within a site data management system or network data management system in accordance with up to date LTAR data management guidelines. All metadata will outline any requirements for data access. Metadata may be released through a network data management system or public data repository.

3. Share initial metadata with the LTAR network within one year of collection to facilitate collaborations in scientific and data management activities, such as integration, analysis, reporting, and product development (e.g., publications[2], technical bulletins, applications, decision support systems, risk assessments, etc.).

4. Share data with collaborators as soon as possible. Share LTAR data products with the public within 30 months of the completion of data collection, as defined in the project plan. Release data and metadata through a network data management system or public data repository.

5. Publish data and data products in AgCROS[3], or other stable scientific data repositories (e.g., Ameriflux, UNH PhenoCam Network, Nature's Scientific data, Mendeley, etc.), with links provided in AgCROS and Ag Data Commons.

## Data sharing embargos

All data will be shared in accordance with the LTAR network and research project data management plan as described in the USDA ARS Policies and Procedures (P&P) document, "Data Management & Public Access Requirements for USDA ARS"[4]. Procedures for waivers for extending the embargo period on working data are described in the aforementioned P&P. Some data will be embargoed as described below. However, all data will include metadata records as previously noted.

[1] To be consistent with the draft USDA Public Access to Results of USDA-funded Scientific Research https://www.usda.gov/sites/default/files/documents/USDA-Public-Access-Implementation-Plan.pdf
[2] USDA-ARS, Office of Communications. Feb. 2, 2020. "Publishing (Print and Electronic Material)", P&P 113.1 ARS v.2. https://axon.ars.usda.gov/Employee%20Tools/REEAdminIssuances/Documents/113.1-ARS.v2.pdf
[3] AgCROS: https://agcros-usdaars.opendata.arcgis.com/ (Delgado et al. 2018)
[4] USDA-ARS, Office of Communications. Forthcoming. "Data Management & Public Access Requirements for USDA ARS", P&P (TBD).

Exclusions can occur requiring data privacy when appropriate with no time limit. Metadata may also be exempt from publication when appropriate. Time limits for exemptions and exclusions, if applicable, should be described in the project plan. Situations that require data and metadata remain private may include instances when the following are included:

- Personally Identifiable Information (PII)
- Proprietary trade data (as codified in formal agreements)
- Public release of information would prevent obtaining a patent or licensing agreement
- Data related to protecting critical infrastructure
- Data related to the physical location of threatened or endangered species
- Data from tribal lands
- Data from culturally sensitive sites
- Other data whose release is limited by law, regulation, security requirements, policy (e.g., classified data, dual use research of concern), or non-ARS funded sources.

## Terms of co-authorship and attribution

Data users are expected to consult data creators/stewards when developing their research plans. This is a good practice to address appropriate applications and analyses of data, as well as encourage collaboration. Terms of co-authorship and acknowledgement and/or data product citation shall be discussed at the onset of a collaborative research project and documented as part of the research plan. The USDA ARS P&P document "Authorship of Research and Technical Reports and Publications"[5] should be used as a reference.

## Public access and reuse

1. All LTAR network data products except those excluded for reasons described above are released to the public and may be freely copied, distributed, edited, remixed, and built-upon with the expectation that acknowledgement of LTAR data creators is given as described below. Publications, models, and data products that make use of these datasets should include proper acknowledgement, including citing datasets, in a similar way to citing a journal article (i.e., author, title, year of publication, name of LTAR publisher, edition/version, and URL/DOI information). See http://www.datacite.org/whycitedata.

   The following example text should be included in any acknowledgement of use of LTAR data:

   *The data used in this analysis were provided by the Long-Term Agroecosystem Research (LTAR) network. LTAR is supported by the United States Department of Agriculture*

[5] USDA-ARS, Office of Communications. Feb. 27, 2018. "Authorship of Research and Technical Reports and Publications", P&P 152.2 v.2.
https://axon.ars.usda.gov/Employee%20Tools/REEAdminIssuances/Documents/152.2.pdf

2. Communication with data creators is encouraged to ensure appropriate reuse of the publicly available data according to its fitness for use, and appropriate authorship on studies.

3. As indicated above, working data are available for consultation, collaboration, and co-authorship with dataset creators or stewards.

4. Data users are encouraged to notify the dataset creator/steward or ARS National Program Office when any derivative work or publication based on or derived from the dataset is distributed.

## Disclaimer of liability

Neither the United States Government, nor any of its employees, makes any warranty, express or implied, including the warranties of merchantability and fitness for a particular purpose, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. (Accessed 2/18/2020: https://www.ocio.usda.gov/policy-directives-records-forms/information-quality-activities)

## References

Delgado, J. A., Vandenberg, B., Kaplan, N., Neer, D., Wilson, G., D'Adamo, R., ... & Arthur, D. (2018). Agricultural Collaborative Research Outcomes System (AgCROS): A network of networks connecting food security, the environment, and human health. Journal of Soil and Water Conservation, 73(6), 158A-164A.

Johnson, Peter A. (2017) Models of direct editing of government spatial data: Challenges and constraints to the acceptance of contributed data." Cartography and Geographic Information Science 44.2: 128-138.

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. Ecological Applications, 7(1), 330-342.

Servilla, M., Brunt, J., Costa, D., McGann, J., & Waide, R. (2016). The contribution and reuse of LTER data in the Provenance Aware Synthesis Tracking Architecture (PASTA) data repository. Ecological Informatics, *36*, 247-258.