Packaging, Transforming and Migrating Data From a Scientific Research Project to an Institutional Repository:

The SGS LTER Collection

Nicole E. Kaplan[1], Karen S. Baker[2,] Daniel C. Draper[3,] and Shea Swauger[3]

[1] Natural Resource Ecology Laboratory, Colorado State University
[2]University of Illinois Urbana-Champaign
[3] Colorado State University Libraries

12 December 2014

Photo on cover page: SGS LTER field crew hikes out to check small mammal trapping webs in the grassland uplands of the shortgrass steppe. (Spring 2009).

Executive Summary

This report describes the process of preserving a collection of project-related scientific research materials - data, metadata, and artifacts - produced over 32 years at the Shortgrass Steppe Long Term Ecological Research (SGS LTER) site. The SGS LTER operated out of Colorado State University (CSU), located in Nunn, Colorado and was funded by National Science Foundation (NSF). Preservation plans were motivated by the 2012 decommissioning announcement for this long-term project (1982-2014) and its local data management system. A two-fold strategy was developed to ensure preservation and community access to the entire collection. In addition to satisfying NSF requirements for submission of data to the LTER Network Information System (LTER NIS), the local information manager identified a second task: creation of a collection including data, metadata and a diverse set of materials that together represent the SGS LTER project as a whole. Migration of the SGS LTER data management system was designated a pilot project for curation of research data within the CSU Institutional Repository, as part of Digital Collections of Colorado (DCC).

The SGS LTER collection comprises approximately 5 gigabytes of data and supporting materials. There are close to one hundred datasets produced by SGS LTER that are diverse, small files with extensive metadata, well described using the Ecological Metadata Language (EML). These data are largely field-based, geo-located, time-series measurements, which have been integrated longitudinally. Other series of materials prepared for the collection include over 400 image files, 17 Geographic Information System spatial layers, species lists, and proposals and progress reports to NSF. EML from the SGS LTER data management system was transformed to Dublin Core for discovery through the DCC and was used to implement an expanded set of elements important for research data documentation.

A strategy was developed to meet the requirement for programmatic access by machine to data from the LTER NIS via a landing page created for each data package. In effect, data are publicly available and automatically harvested by other data repositories, transforming the SGS LTER collection from existing independently to contributing as part of a federated network of scholarly research.

Expansion of the notion of curation from submission of research data to that of creating an interoperable SGS LTER project collection within the DCC revealed new issues and activities to consider. Issues that emerged included design of workflows to create and transform metadata, data exchange between source and secondary repositories, versioning and use of persistent identifiers for digital objects, data citation registries for assessing outcomes of research, and the role of a collection-related information manager. This pilot study was made possible by an interdisciplinary, collaborative effort to preserve data and materials from a historical scientific research project.

Report

Packaging, Transforming and Migrating Data From a Scientific Research Project to an Institutional Repository: The SGS LTER Collection

Nicole E. Kaplan, Karen S. Baker, Daniel C. Draper and Shea Swauger

1.0 Introduction

This report describes the process of preserving a 32-year collection of project-related scientific research materials: artifacts, data, and metadata. Studies at this shortgrass steppe (SGS) location began approximately six decades ago with the aim of investigating and documenting the semi-arid grassland ecosystem east of the Rocky Mountains in North America. The SGS research project was funded within the long-term ecological research (LTER) National Science Foundation (NSF) program and based at Colorado State University (CSU). To ensure preservation and community access to the collection at the end of this more than three decade site-based funding period, a twofold strategy was developed to migrate data from a local data management system to two remote repositories, both positioned to provide open access. First, data and its associated metadata have been packaged and staged for ingestion into the recently deployed Network Information System (NIS) of the LTER Program. Second, the collection of all digital and non-digital project materials has been incorporated into the CSU Libraries and the Digital Collections of Colorado (DCC), the institutional repository at CSU Libraries, for preservation and accessibility as a pilot project.

Starting in 1982 until its decommissioning from 2010-2014, the SGS LTER data efforts included archiving and serving data, metadata (documentation describing the data), proposals, reports, and disseminating other information online (e.g. personnel information and citations). The migration of the SGS LTER scientific project collection takes place at a time when development of digital data repositories is at an early stage. Data sharing and open access are new expectations from research sponsors that present both challenges and opportunities for environmental researchers in particular (Reichman et al. 2011). The ARL report (2009) observation that 'repositories are developing rather than developed' holds true today. Good practices for organizing and submitting mixed collections of digital artifacts and data are formative although proactive engagement and design agility seem to be required both now and into the future.

Our aims are to meet the immediate need of preserving the SGS LTER project collection before it ends and to document the process as well. Additional goals include:
- Provide digital *access to all project materials,* including digital research data
- Contribute to development of *information infrastructure at CSU* that supports new approaches, tools, and services for collections of scientific data and related artifacts.
- Use a *collaborative team approach* that includes a variety of information professionals and scientific researchers.
- Explore the assignment of *unique persistent identifiers*, e.g. handles and Digital Object Identifiers (DOIs) to datasets and the implications of such assignments.
- Plan forward for a *web of repositories* by demonstrating interoperability among institutional and domain repositories*.*

Tables of acronyms and applications used are given in the Appendices 10.1 and 10.2. Links to project and application websites (e.g. NREL, EML) are provided in these appendices. After providing background in Section 2 and data-related definitions in Section 3, some of

the issues that arose during migration of a local collection to alternative repositories are summarized in Section 4. Attribution and interoperability needs are discussed in Section 5 followed by some issues requiring local action in Section 6. Concluding thoughts are given in Section 7.

2.0 Background

The SGS LTER information manager approached CSU Libraries personnel in 2012 about planning transfer of the SGS LTER 32-year collection of digital and physical materials to the CSU Libraries and DCC at the end of its funding as an LTER site. Concurrently, a collaboration began with an information scientist interested in the information infrastructure of scientific research projects including data management and curation activities. Background is provided below on the site, the project, and the collection as well as the migration process itself.

Organizations Participating in the Pilot Study
- SGS LTER local information management component in collaboration with the long-term project researchers has made accessible organized and well-described data using the Ecological Metadata Language (EML) in addition to designing and populating a project web site to query and serve data from a backend data management system.
- CSU Libraries is investigating support for preservation and access to scholarly output, including scientific research data in the Digital Collections of Colorado.
- LTER Network has deployed a Network Information System (NIS), a repository that harvests, ingests, and disseminates data from the 26 LTER sites in the network (NIS 2014).
- UIUC Graduate School of Library and Information Science (LIS) has programs to train information professionals (Palmer et al. 2011; Kelly et al. 2013) as well as to carry out LIS research including data practices, data curation, and the development of information infrastructure.

2.1 SGS LTER and Its Local Data Management System

The Shortgrass Steppe was established as an LTER site in 1982. The site has produced knowledge applicable to sustainable land management practices, nutrient cycling, ecosystem science and modeling, landscape and community ecology, global change, infectious disease research, and environmental literacy (Lauenroth and Burke 2008). Measurements have been made at the site starting in 1943, supported over the years by a number of interdisciplinary research programs, including the International Biological Program (1964-1974). A number of participants have contributed to collaborative research and education efforts at the site, including scientists at the CSU Warner College of Natural Resources, College of Agricultural Sciences and College of Natural Sciences, USDA Agricultural Research Service, Rangeland Resources Research Unit, land managers at the USDA Forest Service Pawnee National Grassland Ranger District and ranchers in the Crow Valley Livestock Association. Over the years, these partnerships have been cultivated to

facilitate transfer of ideas, data and technology. The site is administratively anchored at CSU with a 15,000 acre field site 45 miles to the northeast of Fort Collins, Colorado with two sets of field buildings support research and education activities and provide lodging (Figures 1a and 1b). It was an active part of a network of LTER sites coordinated via an LTER Network Office (Figure 1c).

The online SGS LTER Data Management System was maintained during the last decade of the project by 1.5 FTEs. It was built on a local area network and windows-based server, called Ascalon, which was supported at the department level. RAID technology (Redundant Array of Inexpensive Disks) ensured reliability of the system, management of project data, metadata and other information was centralized on the server, and frequent offsite backups were performed. Public access to SGS LTER information was available via a web server maintained by university-wide Academic Computing and Networking Service at CSU. The site information manager served as webmaster as well as liaison to the project regarding content. These arrangements provided the flexibility to relocate data management activities to a series of departments within CSU as project leadership and department administration changed over the life of the project.

Figure 1a. Looking northwest at the row of old IBP buildings. This facility included space for laboratory work, meals and meetings (left), a garage, shop and mobile research unit (middle), and a dormitory (far right). It is still in use today for conducting field work.

Figure 1b. Looking west over the new Semiarid Grassland Research Center, located across the road to the south from the IBP buildings. This facility provides meeting space and housing and is used for retreats, conferences, workshops and educational activities.
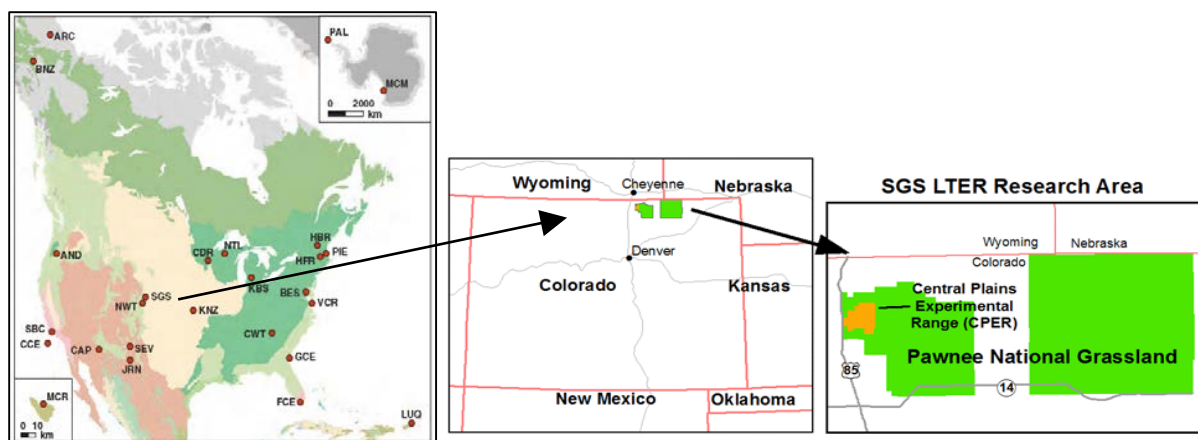
Figure 1c. The network of 26 LTER sites in North America, Antarctica, and French Polynesia (left). The location of the SGS LTER research site and Colorado (middle) and core facilities of the SGS LTER within the Central Plains Experimental Range and Pawnee National Grassland (right).

As an LTER site, scientific research spans a time from 1982, before the internet was in general use, to the digital age in 2014 when support for the concept of open access is developing. A requirement of membership in the LTER is identification of a data manager[1], an approach evident in 1982 (Gorentz 1992) and a role developed over time at annual meetings of an LTER Data Management Committee. Over the years, the role of data management broadened to that of information management in recognition of the spectrum of activities and responsibilities associated with the position (Michener et al. 1994; Baker et al. 2000). The LTER site information managers have developed in time awareness of both data organization and of long-term data care (Karasti and Baker, 2004; Baker and Karasti, 2004). As a result, an SGS LTER information manager was in a position to plan, carry out, and document data work associated with decommissioning of the SGS LTER project despite a lack of materials providing guidance.

The decision to decommission the SGS as a LTER site was made in 2010, with a three-year period to bring both its research and its data management to a close. Its organized, centralized data system supporting aggregation and documentation was well positioned for migration of digital objects and descriptive content. Submission of data packages by 2014 to the LTER NIS was an NSF requirement that accompanied final funding support. In addition, the stories behind the data and the project data management are being collected via transcription of interviews with scientists and staff in collaboration with H. Karasti (Kaplan and Karasti 2013) and K. S. Baker (this report), respectively. The Relational Database Management System (RDBMS) relations are preserved by ingestion of the SGS LTER collection within the DCC to facilitate discovery and navigation among objects in the collection. The aging department server Ascalon and the services it provides will not be supported locally after 2015. Scientists who continue to work with their data collected

---

[1] We use the title 'information manager' in this report hereafter since the LTER Data Management Committee changed the designation of their role from 'data manager' to 'information manager' at the turn of the century (Baker et al, 2000).

during the SGS LTER program, will need access to data and information as well as back-up and exchange services by some other means.
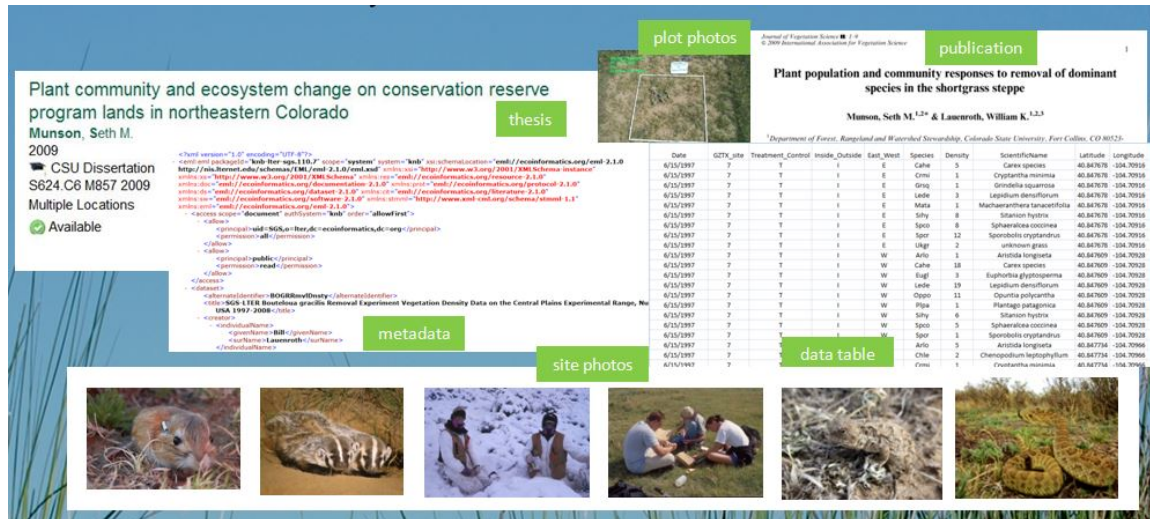

 Figure 2. Data and diverse materials that require managing and have relationships among them

The SGS LTER collection consists of a myriad of diverse, small files with extensive metadata (Figure 2). The data are well described in standardized, structured metadata language while other digital objects are not. The SGS LTER collection is approximately 5GB in size. There are approximately 100 datasets generated and documented by the SGS LTER. The majority of datasets are time-series data that have been integrated longitudinally and documented with EML 2.1.0. Other series of digital objects from the SGS LTER were identified, organized and documented for inclusion in the SGS LTER collection in the DCC. The SGS LTER information manager addressed the need to create metadata both for the various digital materials in the collection that were previously available on the project website and to document the collection itself.

2.2 CSU Institutional Repository and Its Collections

The CSU Digital Institutional Repository (http://lib.colostate.edu/repository) aims to make the research and scholarship of CSU faculty, students, and academic staff publically available now and in the future by offering open access and preservation services that provide guidelines and policies. Starting in 2006, the CSU Libraries began development focusing on digital documents and images as well as audio and video files. For example, digitization projects involving photographic images and water-related materials have contributed to building CSU digital collections (Zimmerman and Paschal 2009; Hunter et al. 2010).

The CSU Libraries is a member of the Digital Collections of Colorado (DCC) partnership that is a shared service initiative of nine university libraries including Colorado State

University-Fort Collins, Colorado State University-Pueblo, University of Colorado at Boulder, University of Colorado at Colorado Springs Kraemer Library, University of Colorado Health Sciences Library, Colorado School of Mines Arthur Lakes Library, Colorado Mesa University Tomlinson Library, and Auraria Library (Figure 3). The partnership established a joint license administered at CSU for the digital asset management system Ex Libris DigiTool to manage and provide access to digital resources. In 2008, the system was implemented by CSU to support the institutional repository and by 2011 was adopted by all member libraries. DigiTool was a relatively mature product with technical support available, so considered to require less investment from local IT staff than other digital asset management applications such as Fedora or Dspace. DigiTool handles both simple and complex objects and has established file formats, metadata standards, some basic web interface customization, and usage statistics. It creates permanent URLs in the form of Handles with unique institution IDs and object IDs. Descriptive metadata incorporated in the application include Dublin Core, MARC, and MODS, although CSU primarily uses a qualified Dublin Core, which is defined in the CSU Core Data Dictionary (Hunter et al. 2008).

This report and the SGS LTER pilot project contribute to experience with organization, migration, and preservation of a project collection that includes datasets and requires expansion of the DigiTool basic metadata (Dublin Core) to incorporate elements important for research data documentation.
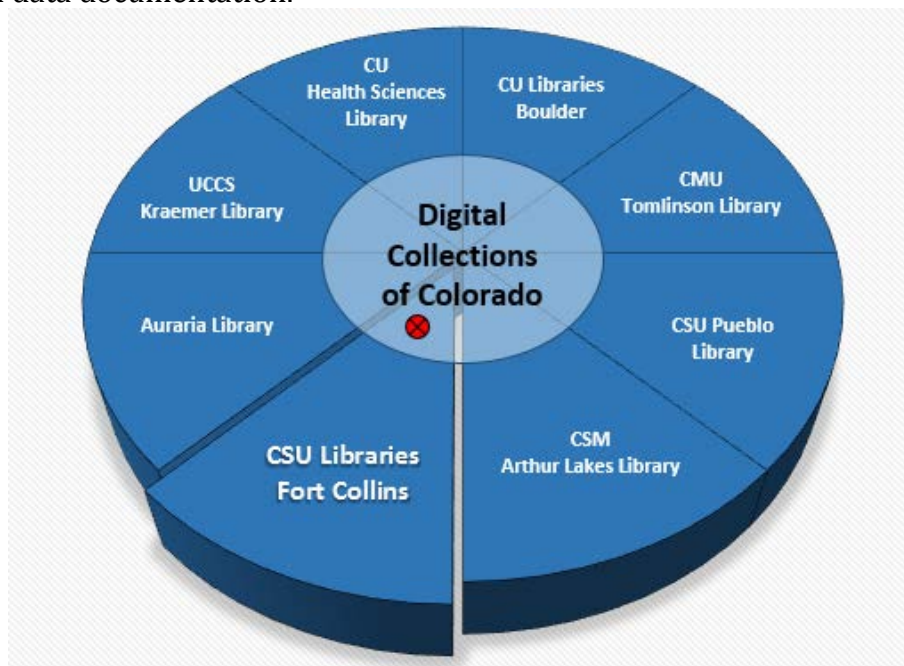

Figure 3. The DCC represents a consortium partnership of member libraries.   Each pie piece represents the collection and services of one member library. The red dot represents the SGS LTER Collection within the CSU institutional repository which resides within the DCC.

2.3 Migrating the Collection

Preservation plans for the SGS LTER collection were motivated by the announcement that SGS local data management would end when the long-term project itself was decommissioned. NSF identified submission of data to the LTER Network Information System (NIS) as a requirement of the final increment of funding. After investigating the possibilities for preserving and ensuring accessibility of the data and metadata at CSU as well as at the LTER Network, the local information manager identified a second task, that of creating a collection of project materials that include not only data and metadata but a diverse set of project materials that together represent the SGS LTER as a whole. The second task was carried out by a four member working group (the authors) in consultation with project scientists and CSU librarians who designated the activity a pilot project. It was recognized the task would provide experience with migrating digital assets from a scientific project to an institutional repository.

2.3.1 LTER NIS data package submission
The process for contributing data to the LTER NIS is initiated by submission of an EML file containing a required link for direct access to data files. The generation of EML for LTER NIS submission starts with the local SGS LTER RDBMS containing all of the metadata content. A local controlled vocabulary of keywords (Appendix 10.3) and units (Appendix 10.4) aligned with the LTER unit registry (Kortz et al. 2009; LTER Unit Registry 2009), are also used to standardize metadata content according to LTER best practices.

 A PERL script together with an XSLT style sheet is used to generate valid EML 2.1.0 in the form of an XML file that can be integrated into the PASTA framework (Provenance Aware Synthesis Tracking System), an application central to the NIS. The LTER Network Data Portal provides a metadata validator (Bohm et al. 2012). The validation criteria and enactment of the validator were the subject of ongoing discussions by the LTER Information Management Committee for a number of years as part of its iterative design. The validator performs a thirty-four point check and reports errors and warnings.

A 'LTER NIS data package' is defined as a valid EML file and the related data files in data table format. Ingestion of this package by PASTA, begins with upload of a metadata file. Once the metadata file is validated, data files are ingested using the metadata link providing machine access to the data. The LTER NIS assigns site-specific package identification and registers the package with EZID for a DOI. Data packages are made available via an online interface providing search by keywords, site, or other metadata entries. LTER data packages are replicated within DataONE and the Knowledge Network for Biocomplexity. In addition, DOIs are being provided to Thomson-Reuters Web of Science as part of a pilot study begun in 2013 for indexing data (James Brunt, personal communication).

2.3.2 Creating a project collection
The process for moving the SGS LTER collection as a whole to the institutional repository was planned as a number of steps by a small working group of participants – SGS LTER information manager, CSU digital librarians, as well as an information science student studying and supporting data practices via action research. They defined the scope of the

collection for transition to ensure persistent, reliable, and interoperable (machine readable and downloadable) access. Organizing the data was a priority along with inventorying project materials (digital and physical). In addition, metadata creation was required for the non-data package objects in the collection that previously were available on the project website. Further, it was necessary to create metadata for the collection itself as one of the collections in the DCC. Other data activities included digitization of datasheets, experimental design and protocol documentation, organization of Geographic Information System (GIS) layers of topographic features, landmarks, and study sites, as well as creation of a series of photographs for the project collection. Digital image data handling was documented in procedural guidelines to address CSU Library Metadata and Selection Criteria for Digital Images (Appendix 10.5).

A 'CSU data package' within the DCC consists of a landing page with data objects. A landing page contains at least one data file, a zipped metadata file containing basic metadata, and relevant supporting objects (e.g. photographs of plots). The landing page may also contain links to referenced objects (e.g. species lists or spatial layers) as well as pertinent materials in other series such as articles, reports, and theses within the DCC. Data and artifacts are organized in series within the SGS LTER project collection, which can be explored in a number of ways: via query within the CSU Libraries Discovery portal, by browsing the DCC hierarchy, and through websites hosted by local CSU units that remain connected to the SGS legacy (e.g. Natural Resource Ecology Lab, USDA-ARS Central Plains Experimental Range, and Semiarid Grassland Research Center). From the landing page of a CSU data package in the DCC, a data package's metadata and related materials can be downloaded as a single zip file as can the data files outside the zip file. The requirement for data tables to be accessed and downloaded programmatically was satisfied by situating the data table outside the metadata and related materials, but within the landing page. In effect, data are openly accessible and machine-accessible by the LTER NIS as well as other data repositories.

Expansion of the notion of curation from submission of research data packages to the LTER NIS to that of creating an SGS LTER project collection within the DCC, including design and development of a method for interoperable data exchange, revealed new activities to consider. Metadata crosswalks, controlled vocabularies, and data citation registries emerged as important topics for discussion. Some questions that arose included:

- What is the mapping across the various layers of metadata (i.e. EML to Dublin Core)?
- What are best practices for forming data citations that will support complete and consistent attribution for the data creator? Today, tracking and attribution is performed by people counting citations or programs (i.e. altmetrics software such as ImpactStory, Plum Analytics, etc.) for calculating index metrics.
- How can repositories document provenance and link to related or sub-sampled collections to acknowledge the point of origin of collections? (i.e. SGS LTER)
- How can a repository stage to be able to meet future needs?
- What additional aspects of interoperability require consideration?

- How can this pilot project inform scientific research projects and/or ending long-term projects that are interested in migrating and preserving their digital assets?

3.0 Data Related Term Definitions

Definitions of what constitutes concepts such as a collection, a landing page, a data package, and a digital object vary depending upon circumstances such as discipline, community, and/or collection. Terms are frequently ambiguous and often used in different ways depending upon practices or perspective presented. There are varied uses of the concept of data package (Kaplan 2014), and below it is defined for two different cases, followed by definitions that illustrate how we use some other terms in this report.

3.1 Terms Relating to Organization of Data

The category 'data organization terms' refers to aggregating and relating individual entities that result in ways of designating or grouping data.

Data Package of the SGS LTER in the CSU Digital Collections of Colorado: a set of two kinds of items: one or more data files (tab delimited text format) and a set of basic metadata materials in a metadata folder zipped into a single file that contains the minimum descriptive information necessary for a basic understanding of the data files. The basic metadata folder contains a Readme file, a metadata file in the EML format, a variable definition file as a tab delimited text file, and related supplementary files such as plot photographs in jpg format.

Data Package for the LTER Network Information System: a set of data and its associated metadata where the metadata is represented in the extensible markup language (XML) that is compliant with the Ecological Metadata Language (EML) schema and is "complete" with regard to metadata quality and content and provides "unfettered" access (in the machine accessible sense) to data, as a direct reference to the data through a unique and persistent identification protocol, such as HTTP or FTP (accessed 4/3/2014 from https://nis.lternet.edu:8443/display/pasta/Data+Package+Background).

Data: one or more related data files with or without metadata.

Digital Object: a single digital file

Dataset: one or more related data tables with one or more contextualizing files

Landing Page: a web page that serves as an online location to 'land' and where information about data or digital objects may be found, including a 'Readme' file and documentation cataloged in the repository; there may be links to one or more data files and supplementary information. A landing page is a complex object that provides context for the data by grouping together files that are related to each other. Some data repositories assign a

persistent identifier (e.g. Handle or DOI) to a landing page. A landing page may serve as a reference webpage for an entity such as an administrative program, a project or, in this case, for a data package. It displays a subset of relevant metadata within the repository. The concept of a landing page is emerging as central to data delivery though it clashes conceptually with the more concrete notion of a data package as an immutable end-product because the information on a landing page may change and any one data object may have updated versions made available over time.

Portal: a web page that services as an entry point, bringing information together in an ordered manner that reflects the hierarchy and relations among aggregated digital materials such as collection(s), sub-collections, series, landing pages, and/or data packages.

Series: many digital entities grouped together because they are related by having similar structure and related content (i.e. photos, published papers, data packages, GIS layers).

Sub-collection: a collection that is nested within another collection and contains digital objects that may not fit within an existing series.

Collection: a number of objects aggregated digitally in some manner that may be grouped into sub-collections or series because of an association such as with a particular creator, program or topic.

3.2 Example of Organization of a Project Collection

The SGS LTER project collection resides within the Digital Collections of Colorado. The organizational scheme is shown by webpage screenshots in Appendix 10.6. The collection is arranged hierarchically as follows:

Level 1. Portal for Digital Collections of Colorado
  Level 2. Portal for Colorado State University
    Level 3. Portal for CSU Colleges
      Level 4. Portal for Warner College of Natural Resources
        Level 5. Portal for Natural Resource Ecology Laboratory (NREL)
              (note: IBP & SGS LTER)
          Level 6. Portal for SGS LTER Project Collection
            Level 7. Data Package Series for SGS LTER
                (note: Two Views: Brief list or Table)
              Level 8. Landing Page for SGS LTER Data Package
                Level 9. Programmatically accessible data files, each with unique Handle

The first six levels of organization are designated as portals, each portal providing links to an array of collections. Level 6 lists categories of materials in the SGS LTER collection. An SGS LTER catalog of data packages is presented at Level 7, where each data package has a landing page (Level 8). The basic metadata on the landing page is in a compressed folder for easy download. The data files are not compressed but are directly viewable within DigitTool as well as downloadable in tab-delimited format for ease of use by humans or

access by other information systems. This means the data are exposed in a manner that makes them programmatically accessible.

The hierarchy of levels provides an example of the ambiguity introduced by term use in the case of a collection of collections. For instance, the SGS LTER Project Collection is recognized as a collection from the perspective of SGS LTER Researchers since it is a stand-alone set of materials that encompasses a body of work associated with the SGS site. Yet, the SGS LTER Project Collection is also a sub-collection from the perspective of the Digital Collections of Colorado. Further, the collection of data packages is a sub-collection of SGS LTER Project Collection. Note also that because of the similarity in structure and content of the data packages, Level 7 labels the set of data packages as a series.

3.3 Terms Relating to a Repository

Source Repository: the repository that data creators work with initially and most closely over time that may be co-located with an administrative site associated with the research project; there is only one source repository.

Secondary Repository: Any repository that is not the source repository that may include all or a subset of data or derived data-products from other repositories, may serve to subset or to aggregate data from multiple sources, and/or may enhance opportunities for discovery by serving domain specific data or a specific research community. Secondary repositories may also make available derived value-added data products or have tools for visualizing and analyzing data (e.g. International Biological Information System at NREL).

3.4 Descriptive Metadata for Datasets and Collections

Metadata provides information about data files. Three kinds of metadata pertinent to the SGS LTER collection migration are given below:

Descriptive Metadata: metadata for discovery and for content documentation are key to automated use and re-use of data. Descriptive information about data or digital objects occurs at many levels. It is recommended that data creators identify the point of data origin and provide enough description to make the data reusable. It is best to provide the 'who, what, where, when and how' of what you are describing as a human-readable document. Data discovery through online browsers and automated access require structured metadata.

Ecological Metadata Language (EML) – a metadata specification in a machine-readable structured format used in the life sciences to describe datasets (Michener et al, 1997; 2011; Jones et al, 2001; see example from SGS LTER data package). The tags for keywordSet, data access policy and intellectual rights are provided below as an example of EML v2.1.0:

```
<keywordSet>
    <keyword keywordType="theme"> biomass</keyword>
    <keyword keywordType="theme"> populations</keyword>
    <keyword keywordType="theme"> plants</keyword>
    <keyword keywordType="theme"> primary production</keyword>
</keywordSet>
<intellectualRights>
    <section>
        <title>URL for Access Policies</title>
        <para>http://www.lternet.edu/policies/data-access</para>
        <section>
            <title>Data Access Policy</title>
            <para> Data sets were provided by the Shortgrass
                    Steppe Long Term Ecological Research (SGS-LTER)
                    Program, a partnership between Colorado State
                    University, United States Department of
                    Agriculture, Agricultural Research Service, and
                    the U.S. Forest Service Pawnee National Grassland.
                    Significant funding for these data was provided by
                    the National Science Foundation Long Term
                    Ecological Research program (NSF Grant Number
                    DEB-1027319). The SGS-LTER project (1980-2014)
                    was established as one of the first sites in the US
                    LTER Network and has produce a rich legacy of
                    digital materials including reports, proposals,
                    images, and data packages. Data, products and
                    other information produced from the SGS-LTER
                    are curated as a collection within the Digital
                    Collections of Colorado
                    (http://digitool.library.colostate.edu/R/
                        ?func=collections&amp;collection_id=3429).
                    Materials can be accessed from the Institutional
                    Digital Repository of Colorado State University or
                    upon request by emailing ecodata_nrel@colostate.edu.
                        All data are open for dissemination and re-use for any purpose,
            but you must attribute credit to the owner and cite use
                    appropriately according to the LTER Data Access
                    Policy. para>
        </section>
    </section>
</intellectualRights>
```

Dublin Core (DC) - a metadata specification used in library sciences that is a basic metadata schema compared with the more complex EML schema. Dublin Core has 15 basic elements that can be expanded. It has been adapted by other communities, e.g. for biodiversity data with Darwin Core. It is used by CSU to describe collections (see example from DCC landing page).

3.5 Terms Relating to Identification and Registration

There are a variety of services that register identifiers for datasets (Duerr et al. 2011).

DataCite: a registry (DataCite.org) that incorporates the 'handle'; it provides a DOI when data citation metadata are completed and a fee paid.

DOI: a digital object identifier issued under the authority of doi.org, the internationally recognized system for registering persistent objects.

Handle: a locally generated unique identifier for locally held digital packages that can be registered via a web-based system handle.net.

EZID: a dataset registration system developed by the California Digital Library that carries out registration of a DOI by working with the registering organization (see http://www.doi.org/registration_agencies.html) that brings together registration agencies (e.g. CrossRef for print; DataCite for data). In this process the use of handles is managed by DataCite via organizational agreement with handle.org.

Thomson Reuters Data Citation Index (Web of Science): an index of particular data repositories and their content, including metadata, which allows users to search for and cite research data across multiple repositories. This service also tracks citations to the data, enabling more reliable bibliometrics and measurement of impact. Currently it is an incomplete index of repositories, providing only a small sample of existing, publically available data repositories; it is one of several emerging citation index and impact factor services (e.g. Google Scholar, Plum Analytics). It has been adopted for use in evaluations of programs and individuals by the colleges at CSU associated with the SGS LTER collection. Impact factors, citations, publications, and other measurements of outcomes contribute to annual reviews for tenure-tracked faculty.

Some information systems designate a landing page as an entity to which one registers a DOI. Others designate a data package or a dataset as the object to which a DOI will be assigned. The discrepancy between these approaches has the potential to create confusion. Furthermore, the practice of assigning DOIs to both datasets and landing pages, in combination with the goals of sharing data and building a web of repositories, results in a dataset potentially being assigned multiple DOIs, one DOI that refers to the dataset and its metadata in one repository, and another DOI that refers to the dataset as part of a landing page in another repository. This scenario may affect usage metrics attached to a DOI, a problem for data creators who rely on those metrics to measure the impact of their data. Communication among those who register DOIs is recommended to develop standards for what should be given a DOI.

The SGS LTER Project work with LTER NIS and the DCC provides examples of identifier use. The LTER NIS assigns a DOI to a landing page that serves a NIS data package. DigiTool, on the other hand, assigns a Handle as a persistent URL made available via the DCC landing

page. DigiTool also creates a system unique identifier defined by DigiTool as a persistent identifier for each digital object.

3.6 Examples of Information Systems

<u>Thompson Reuters' Data Citation Index – a data citation indexing</u> service that captures relationships among citations in publications in a bibliographic database and provides bibliometric services.

<u>DigiTool:</u> commercial software from Ex Libris that manages digital collections and institutional repositories; it is currently in use by a Digital Collections of Colorado consortium of users in Colorado.

<u>LTER NIS:</u> a suite of applications referred to as the Network Information System developed by the Long Term Ecological Research Network Office (LNO); it is currently in use by the LTER network of sites (NIS, 2014).

4.0 Issues When Migrating a Project Collection

Initial design and development requirements recognized for the migration process included:

1. Developing the capacity to create a project collection consisting of scientific data and related digital artifacts (e.g. SGS LTER)
2. Defining a pilot study for designing and developing the migration process
3. Engaging local experts to organize and describe collection related digital artifacts that provide contextual information
4. Developing multiple workflows for migrating data and metadata into the repository
5. Establishing a staging area for hand-off of files from the local system to librarians ingesting objects into the repository (e.g. DropBox)
6. Linking objects in other collections in the institutional repository (e.g. theses, articles) within SGS LTER data landing pages
7. Developing and making visible working standards for each repository (e.g. EML metadata for LTER NIS, qualified Dublin Core for DCC, recommendations for ecological research projects, controlled vocabularies, and a check list for project data migration and preservation (See Section 6.3)
8. Ensuring discoverability and interoperability for searches and harvests (e.g. Google, LTER, OAI-PMH)
9. Establishing relationships among repositories (e.g. LTER NIS and CSU DCC)
10. Working with local research groups to ensure analysis and visualization systems have programmatic access to data (e.g. NREL)
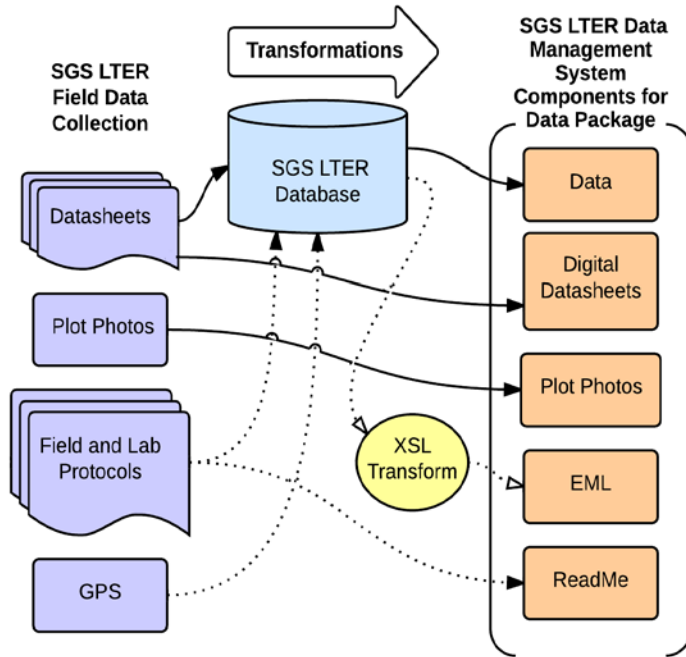11. Planning for versioning and/or appending ongoing scientific dataset series

Figure 4a. SGS LTER data management system workflows begin by collecting data in the field, and then entering and managing data in a relational database management system. Metadata is transformed and packaged along with data and other related materials for migration to the DCC. Solid lines represent digital objects, and dotted lines are metadata. Shown are original field-based data and metadata (purple), data package components (orange), and metadata transformation (yellow).

Figure 4b. Institutional repository workflows for submission, metadata transformation and ingest. Solid lines represent digital objects, and dotted lines are metadata. Shown are data package components and artifacts for submission (orange), metadata transformation (yellow), and collection series in the DCC (green).
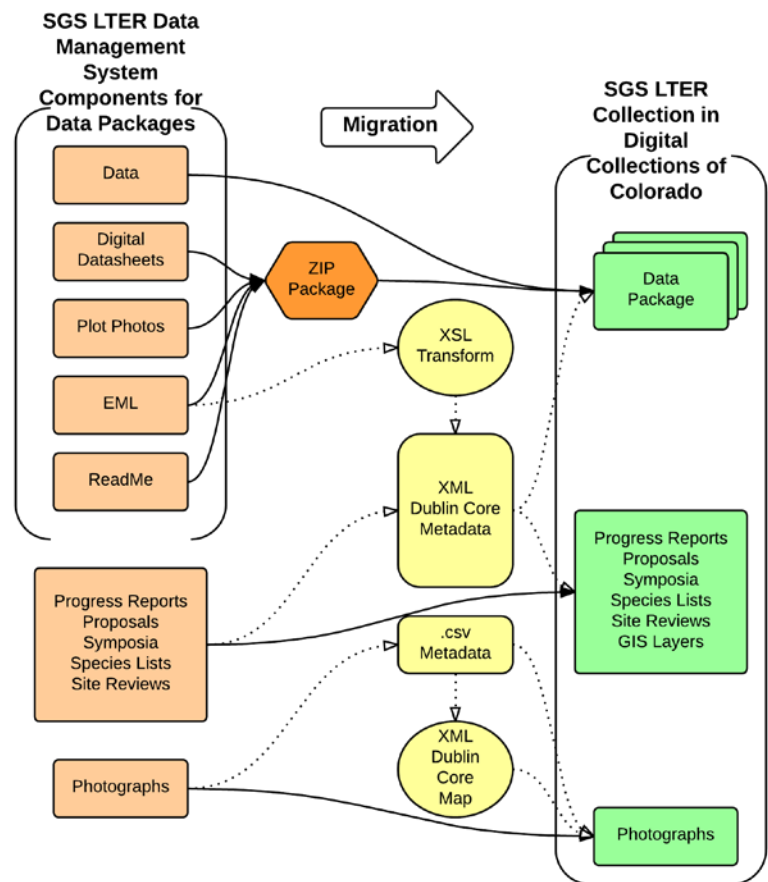
4.1 Digital Workflows for Packaging, Transformation, and Migration

Workflows for transformation and migration of project materials were created to process SGS LTER materials into the DCC. They were developed collaboratively by the project information manager and digital repository staff.

Figure 4a shows the SGS LTER project materials (data sheets, plot photos, protocols, and GPS data). The data and protocols were entered into the project database. The workflow illustrates the staging of data package components in preparation for transfer to the DCC.

Three workflows for migration from the SGS LTER servers to the DDC are shown in Figure 4b. The first workflow pertains to data packages received from the SGS LTER data management system, which include a dataset in tab delimited .txt format and materials directly related to each dataset. Related materials include a README file, variable definition file, digital data sheets, EML metadata file, and plot photographs. The related materials are combined into a zipped file for posting with the data. The zip format ensures ease of download. The EML metadata file is processed using an XSL transformation and converted to XML Dublin Core metadata, which is ingested with the dataset and zip file. Each set of data and zipped content (e.g. data package) comprise the data package sub-collection.

The second workflow includes several SGS LTER related materials, such as progress reports, proposals, symposia, species lists, site reviews, and GIS layers. Dublin Core metadata in XML is created for each object. The metadata references the whole SGS LTER collection and provides direct links to the data records when applicable. It's been crucial for the SGS information manager to designate which materials are related, so that objects can then be related within the DCC. Each object is ingested into separate SGS LTER sub-collections.

The third workflow is used for the processing of SGS digital images. Descriptive metadata developed for the CSU collections is created as a .csv (comma separated value) formatted spreadsheet by SGS staff. The photographs and the .csv file are ingested together with an XML map to generate the Dublin Core elements in DigiTool. Each of the three workflow outputs becomes part of the SGS LTER collection in the DCC.

4.2 Managing Photographs

More than 400 digital images of the shortgrass steppe were captured by various photographers. These include photos of people working and living there, ranching operations, landscape views, weather events, flora and fauna. SGS LTER managed these images and made them available internally for use by community members in posters, books and presentations. A criterion for minimum resolution, size and quality was established for determining which files would be most appropriate for the collection. Prints and negatives were scanned and digitized by SGS staff. All the digital image files that met our criteria were documented with metadata recommended by the Core Data

Dictionary and with keywords applied by the CSU photographers. The photos were prepared as a series in the SGS LTER collection. A few examples are collaged (Figure 5).

Twister June 2010

Prairie Rattlesnake (*Crotalus viridis viridis)* consuming short-horned lizard (*Phrynosoma hernandesi*) 2005

Ord's Kangaroo Rat (*Dipodomys ordii*) September 2009

Flooding July 1997

Bite Count sampling 2002

Grasshopper on *Sphaeralcea coccinea* 2009

Figure 5. Photo collage representative of biotic & abiotic facets of SGS LTER.

The entire collection of photographs can be found within the DCC at
http://digitool.library.colostate.edu/R/?func=collections-result&collection_id=4467.

4.3 Identifying Sub-collections

Data and related digital objects were organized into series within the SGS LTER collection. Together they provide a rich context and descriptive background regarding the history of the SGS LTER project and serve as evidence of how science and education was conducted at the site. The SGS LTER data packages are linked to project artifacts organized in series and to other collections in the DCC (Figure 6).
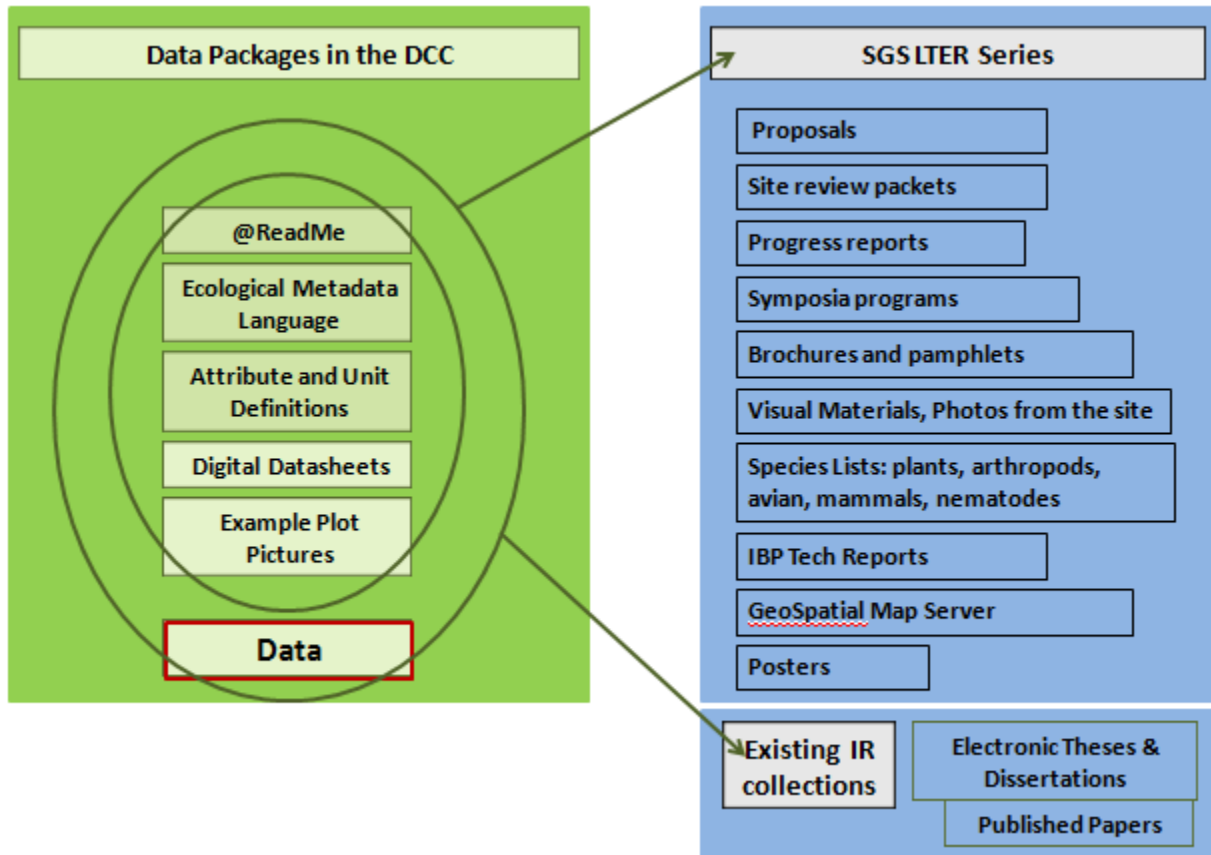
Figure 6. Overview of the SGS LTER collection.

4.4 Mapping EML to Qualified Dublin Core

Structured metadata as EML and Dublin Core are created and stored in an XML format. Additional XML tags were configured in Digitool to accommodate research data (see Appendix 10.7) XML tags in EML were mapped to those in Dublin Core to help streamline the process of developing discovery metadata. The SGS LTER information manager and CSU IR digital services librarian identified EML metadata tags to be used for developing a Dublin Core descriptive metadata template (Figure 7). An XSL transformation was

constructed to map the specified EML tags to Dublin Core and generate other necessary Dublin Core elements before ingest into the DCC (see Appendix 10.8).
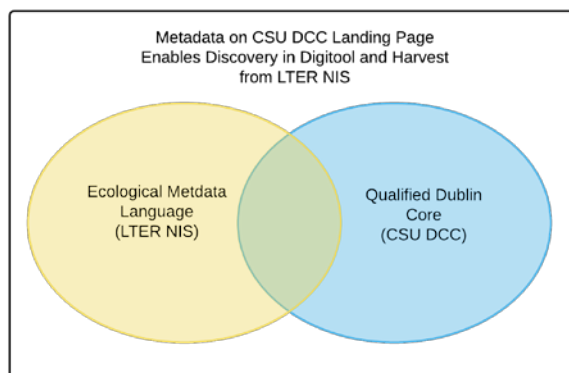
Figure 7. Ecological Metadata Language was the adopted standard of LTER sites, while Dublin Core elements were in use by CSU Libraries. Some EML tags mapped to the Dublin Core while others were generated and enacted for use with ecological research data in the DCC.

5.0 Attribution and Interoperability Needs

5.1 Data Citation Services

The availability of citation indexing or bibliographic citations in a digital database such as Thompson-Reuters makes it possible to track references made to a paper and to tie the metrics to the authors. In a similar manner, when data citations are included in journal article bibliographies, it will be possible to track citations to data. The access to data has been tracked in the past via number of downloads from an electronic posting of the data. The practice of data citation creates a new, location-independent data use metric.

A local Handle has been developed that assigns a unique identifier for entries in the repository. These are registered through Handle.net. Options exist to register Handles (no fee) or using EZID to obtain a DOI (with fee). The DCC effort has not established the assignment of DOIs at this time.

The DataCite Metadata report (http://schema.datacite.org/) is a key document in understanding metadata associated with identifiers. We also found informative two timely reports about how a national research center has begun the data citation process (Mayernik 2012; Mayernik et al. 2012b). Mayernik (2013) highlights unsettled issues that impact data citation arrangements: the newness of DOI use for data, the identification of the citable object, and the long-term prospects for persistence. The report presents five recommendations about how to approach work with DOIs.

5.2 Elements in a Data Citation

As data citation becomes automated, there is a need to consider who will be credited through the mining of data citations as well as a need to develop interoperable standards for forming data citations. At the moment of registration of a dataset, the dataset metadata is mapped to data citation metadata and determines what information will appear in a bibliographic citation and what information is available for mining from the full data citation metadata.

The data citation created by DataCite includes five elements:

> creators, title, date, publisher, DOI

An example DataCite citation from a SGS data package in PASTA:

> Lauenroth, William, 2013, Bouteloua gracilis Removal Experiment Vegetation Density Data on the Central Plains Experimental Range, Nunn, Colorado, USA 1997-2000, ARS #137, SGS LTER, Long Term Ecological Research, [doi: 10.6073/PASTA/D0272AF6E402FE1FE0C5D218B06CFDCB]

An example DataCite citation from an NCAR landing page:

> Mearns, L.O., et al., 2007, updated 2012. *The North American Regional Climate Change Assessment Program dataset*, National Center for Atmospheric Research Earth System Grid data portal, Boulder, CO. Data downloaded 2013-10-15. [doi:10.5065/D6RN35ST]

The title is a free text field so is a candidate for development of local best practice particularly because of its prominence in the data citation provided to dataset users. See following section.

5.3 Dataset Title

Creating dataset titles is important in identification, discovery and citation. There is an LTER Information Management Best Practice (LTER IM, 2011) that states "the dataset <title> should be descriptive and mention the data collected, geographic context and research site (what, where). If the data will not be updated, include the time frame (when)."

An LTER IM example illustrating the title using a template of 'name, location, start data – end date'

```
<title>
        Long-term Ground Arthropod Monitoring Dataset at Ficity, USA from 1998 to 2003
</title>
```

An NCAR dataset title
```
<title>
        North American Regional Climate Change Assessment Program dataset
</title>
```

A terse SGS title:
```
<title>
        Bouteloua gracilis Removal Experiment Vegetation Density Data on the Central
        Plains Experimental Range, Nunn, Colorado, USA
</title>
```

A fuller title may be used in order to have more information appear in a data citation that often use only 4 metadata elements (title, publisher, date, and dataset creator). A fuller title example includes a pertinent identification number such as the permit number issued for sampling and the name of the project/program:

```
<title>
        SGS LTER Bouteloua gracilis Removal Experiment Vegetation Density Data on the
        Central Plains Experimental Range, Nunn, Colorado, USA 1997-2000, ARS #155, SGS
        LTER
</title>
```

5.4 Acknowledgement Information

While a bibliographic citation gives the authors, title, and publication information for printed document and a data citation similarly is the text created to appear in the reference section of a publication, acknowledgement statement gives credit to a number of people and groups associated with a dataset. To date, data mining of the citations and other free text fields has provided information on use metrics.

Information entered into the DataCite metadata may in the long term become an automated approach to gathering information on authorship, groups, and funding agencies for metrics on dataset use. As a result, it becomes important to consider how full acknowledgement information appears in the DataCite information. The contributor element qualifiers allow some but not all of the specificity needed. Using the 'rights' and 'description' elements provide alternatives for preserving the full acknowledgement information that may be parsed at a later date when the metadata requirements, authority schemes, and best practices have been more fully developed.

5.4.1 Using the 'contributor' element with qualifiers
The 'contributor' property in the DataCite metadata schema contains a qualifier with a controlled vocabulary that specifies particular kinds of contributors (DataCite Table 9,

property 7.1, p12):

> ContactPerson, DataCollector, DataManager, Distributor, Editor, Funder,
> HostingInstitution, Producer, ProjectLeader, ProjectManager, RegistrationAgency,
> RegistrationAuthority, RelatedPerson,
> Researcher, ResearchGroup, RightsHolder, Sponsor, Supervisor,
> WorkPackageLeader, Other

DataCite permits some designations such as

> a. authors as creators of type 'Researcher' or 'DataCollector'
> b. funding agency as contributor of type 'Funder' (that is recommended)
> c. local project name as contributor of type 'ResearchGroup' (e.g. SGS LTER)
> d. institution posting the source dataset as contributor of type 'HostingInstitution'
>   (e.g. CSU Libraries)

An example of 'HostingInstitution' from NCAR work shows how a best practice helps to standardize its use and provide organizational information about division and groups involved:

> University Corporation For Atmospheric Research (UCAR):National Center
> for Atmospheric Research (NCAR):Computational and Information Systems
> Laboratory (CISL)

A best practice would be required to define use of the 'producer' element in alignment with the 'HostingInstitution' practice.

The DataCite metadata permits designation of an authority or identifier scheme for these with the 'nameidentifier' and 'nameidentifierScheme'. For an author identified in a Library of Congress authority file (http://authorities.loc.gov/) or who has registered for an ORCID (http://www.orcid.org/), there is an authority control. The Library of Congress approach, however, only targets books and uses an 'after-the-fact' approach of relating name variations to a primary name. Such an approach does not scale readily to the much larger number of individual articles. Similar mechanisms for establishing unique identification of other entities such as institutions and funding agencies have not yet been developed.

There are a number of others to be acknowledged who are associated with collaborative projects for which DataCite does not have an explicit identification property or qualifier, e.g.

> a. local institution with project administration (e.g. CSU)
> b. local institutional department within which individuals work (e.g. NREL)
> c. local site owner where measurements were made (e.g. USDA ARS)
> d. local site collaborating partner (e.g. USFS PNG)

The contributorType 'Other' is available for designating these related partners. It would be possible to establish local best practices for some of these.

An example of the acknowledgement put into the 'Rights' property that is used in the DataCite registry as well as in the EML schema:

> Data sets are open. Please include tag line in report or manuscript: Data sets were provided by the Shortgrass Steppe Long Term Ecological Research group, a partnership between Colorado State University, United States Department of Agriculture, Agricultural Research Service, and the U.S. Forest Service Pawnee National Grassland. Significant funding for these data was provided by the National Science Foundation Long Term Ecological Research program. (SGS LTER example)

5.4.2 Using the 'description' property
Acknowledgement text can also be added to the EML 'abstract' tag, a free text element. It can be mapped to the DataCite 'description' element.

5.5 Designating Relations Between Datasets in Different Repositories
The 'relatedIdentifier' property in the DataCite metadata schema has a qualifier of 'relatedIdentifierType' and 'relationType' with a controlled vocabulary that specifies particular kinds of relations (DataCite Table 9, element 12.2, p28):

> IsCitedBy, Cites, IsSupplementTo, IsSupplementedBy, IsContinuedBy, Continues, HasMetadata, IsMetadataFor, IsNewVersionOf, IsPreviousVersionOf, IsPartOf, HasPart, IsReferencedBy, References, IsDocumentedBy, Documents, IsCompiledBy, Compiles, IsVariantFormOf, IsOriginalFormOf, IsIdenticalTo,

The implications of this element are significant in terms of making clear the relationships among multiple repositories where the dataset and any of its variations appear. The 'relationType' may become in practice a way to link the DOI given a dataset published in a source repository to the DOI of a variation on this dataset appearing in another repository, e.g. using 'relationType' of 'IsVariantFormOf'.

Of import to the data generator in particular as well as to data provenance in general, the content of the 'dataSource' tag is currently not being sent by the LTER NIS to data package indexing systems, such as DataCite. The 'dataSource' element in EML 2.1.0 could be used by the PASTA provenance tracking system for recording the source data when a derived product is created and described with EML (EML Best Practices). It is this tag that would document the source repository (in this case the DCC).

5.6 Packaging and Types

The DataCite element 'resourceType' describes the kind of resource. It has a qualifier with a controlled vocabulary. The language for naming a 'landing page' is not standardized. Two examples are given below:

> XML Examples:
> <resourceType resourceTypeGeneral="Dataset">Reanalysis dataset </resourceType>
> <resourceType resourceTypeGeneral =Dataset">DataPackage</resourceType>

It is recommended that a repository develop a best practice for use of this element.

5.7 Version Documentation

There are a variety of ways to handle new versions of files. The LTER NIS requests a new DOI for each of its new versions. Alternatively, versions of the dataset may be documented using the DataCite version element. This provides a mechanism to avoid obtaining a new DOI for each update. Versioning can also be documented in a Readme file on the landing page of the dataset. This latter method is used by NCAR.

6.0 Key Lessons and Next Steps

6.1 Local Actions as Related to Decommissioning of the SGS LTER project

The following are some of the activities that participants in the SGS LTER scientific project collection pilot anticipate requiring attention at the end of decommissioning work:

1. Develop static and dynamic web content migration and redirection of users to ensure access to information:
   a. Develop new web page(s) with static descriptive content about the project and its history within the Natural Resource Ecology Lab (NREL) at CSU. Visitors will be directed to the SGS LTER collection in the DCC, via a bookmarkable URL from the collections portal. The webpage within NREL will act as a landing page for the SGS LTER collection. (*It is important to note that the collection is managed as a digital object holistically so there is no landing page for the collection portal within Digitool.)
   b. Implement automatic redirects to SGS LTER data where appropriate. The SGS LTER data discovery and download functions built onto the SGS LTER data management system will be removed from the CSU web server supporting SGS LTER.
   c. Create an index inclusive of the project's data packages and other materials with links to handles and/or DOIs for the final report prepared for NSF.

d. Coordinate with information managers at the CSU GeoSpatial Centroid to relate spatial information to non-spatial data packages (i.e. ARS Study Number in data set title maps to attribute in geospatial database).

6.2 Future Actions with Broader Implications

The following summarize some of the important activities that participants in the SGS LTER scientific project collection pilot were able to identify as requiring attention in the future and contributing to emerging recommendations:

1. Design a migration approach
   a. Define organization of data files into collections and/or series (e.g. SGS LTER collection, proposal series).
   b. Consider metadata requirements that meet needs of data users (e.g. species lists and methods for data).
   c. Identify or develop controlled vocabularies for local communities and post them publicly.
      i. LTER Research Themes (http://www.lternet.edu/research/core-areas)
      ii. LTER Unit Registry (http://unit.lternet.edu/unitregistry/)
      iii. SGS LTER recommended keywords (see Appendix 10.3) (Porter 2009)
2. Facilitate migration from local systems to a repository
   a. Develop staging area for transfer of digital materials (e.g. DropBox; project server).
   b. Develop and document work flows for groups of digital files
   c. Create data packages that meet specifications of the repository and/or research community.
3. Take into account external repository requirements
   a. Consider requirements for interoperability.
   b. Implement ways to increase discoverability.
   c. Describe and document relationships among repositories or sources of data or objects.
4. Create a data citations working group with a diverse membership (i.e. cataloger, IT, digital librarian, researcher/data manager as user) to discuss repository-wide implications for data citation requirements. Consider how this group can contribute to a Data Management Standing Committee begin formed upon request of the VPR office.
   a. Investigate membership in EZID and associated costs for registering DOIs.
   b. Develop a citation syntax for CSU Libraries.
   c. Create a complete copy-paste or exportable citation to data within the metadata of each digital entity with a DOI.
   d. Look into what other repositories are doing about DOIs, when they hold various materials in their archive.

e.   Develop discussion on how to manage duplicate data citations or avoid where possible.
f.   Develop vocabularies in coordination with other repositories for specifying names of contributors that document provenance (e.g. hostingInstitution or dataSource).
g.   Consider use of ORCID for identification of institutions and/or data creators.
h.   Discuss use of date field in Dublin Core for data packages. Consider it as published date as timestamp of ingestion versus date which reflects when the analog object was created (i.e. when data were collected in the field).

6.3 Project Data Migration and Preservation Checklist

How to migrate and preserve project data is a key lesson learned from this pilot study. The following checklist summarizes activities important to preserving a field science project's collection of data, metadata, and artifacts. It is inclusive of digital objects considered outcomes that are final products of scientific research.

1. Consult with local researchers, research partners, and organizational departments about perspectives of why and how to preserve data and maintain access.
2. Consult with funding agencies and network or affiliated program offices about preservation requirements and options.
3. Determine if it is necessary to provide opportunities for local data contributors that permits responsiveness to append and/or version the data in the future.
4. Inventory data, metadata, project digital materials (in project directories and on the project website), as well as project-related physical materials (e.g. procedures and protocols, experimental design documents, field books, reports, maps, photographs, file cabinet contents, physical samples).
5. As you make progress on your inventory, periodically review and prioritize the list of data and other materials you plan to preserve and share. Begin packaging, transforming and migrating what you have in hand as a place to start and expect to find additional materials during the process.
6. Identify digital repositories that can accommodate data and materials, as well as meeting any technical needs. Consult with an academic library and/or research repository about its state, scope, and policies as well as the software supporting the repository. Inquire about their experience and capabilities with scientific data and other materials.   If a single repository is not able to accommodate all types of digital data and materials, consider building partnerships with multiple repositories to preserve and serve the collection.
7. If multiple repositories are necessary, review policies regarding linking to other repositories as some may not allow for that functionality. Local institutional repositories currently may have the flexibility and incentive for creating collections and establishing relations to content in other repositories.
8. If multiple repositories are working in partnership, identify and describe relationships between source and secondary repositories in serving project data and materials.

Consider when and where DOIs are assigned and the associated metadata that documents the relationships between source and secondary repository.

9. Develop text for data package metadata that acknowledges the project by name, identifies the funding agency and makes clear the 'primary data repository' as well as links to related repositories.

10. Identify candidate materials for digitization, e.g. procedure manuals, field books, reports, file cabinet contents, slides and photographs. Consult with appropriate repository and/or metadata specialist about procedures for documentation and preservation.

11. Generate metadata for all the objects you plan to preserve in a repository (data packages, image files, reports, proposals, presentations, controlled vocabularies). Consider whether cross-walks can be developed between existing metadata to the formats required by secondary repositories.

12. As a final data product of the project and useful for reporting to your funding agency, create an index inclusive of the project's data packages and other materials with links to handles and/or DOIs.

13. Decide whether the project web site is to be archived and if so, how it will be done.

14. Consider creating a project reference as a historical marker at a local website that links to the project collection in the repository.

7.0 Final Remarks

The local scholarly assets from the SGS LTER project included an array of datasets to preserve and a project information manager able to liaison with digital librarians. A poster given at the International Digital Curation Conference in 2014 provides an early overview of the project (Appendix 10.9) and a Project Data Migration and Preservation Checklist provides a brief summary of tasks that were found useful to consider in bringing a project to a close (Section 6.3).

The SGS LTER collection represents a record of more than thirty years, which is uncommon in ecological research. Although LTER program support is now ending, some scientists plan to continue long-term monitoring efforts on the shortgrass steppe. Researchers may continue updating datasets with field measurements made annually or at a longer frequency. Support for the datasets may well continue in new forms and in coordination with other repositories (e.g. National Agricultural Library). Also, the data will be made available for access by online visualization and analysis tools within the NREL International Biological Information System (http://ibis-live.nrel.colostate.edu/). These tools allow researchers and scientists to download, map and create value added data products to meet their needs for obtaining information regarding species of concern, potential habitat conservation, exotic invasions, and other issues related to biodiversity and natural resource management.

This pilot project focused on all three elements identified by the ARL Report (2009) as central to digital repository service – access, context, and content. Access will be provided by two repositories: the LTER Network Information System and the Digital Collections of Colorado. These represent two different contexts: the LTER arena with a recognized metadata standard and validation checker; the CSU Institutional Repository arena with Digitool's structure for content and standards support. This enhances the opportunity for discovery and preservation. The dual focus contributes to data and system interoperability as well as the vision for a robust knowledge infrastructure of the future (Edwards et al. 2013). The migration to two repositories provides an example of what it means to envision remote repository requirements supporting a web-of-repositories (Baker and Yarmey 2009).

Reflecting on our experiences over the two-year period for migration of the SGS LTER project collection, we recognize the benefits of cultivating a continuing design attitude that acknowledges significant issues must be addressed over time. These include bibliographic rights, data versioning and inter-repository referencing. For the moment, we move forward by documenting our progress in meeting our objectives to capture the rich historical context of a project, gain experience with metadata, workflows, and data citations in order to ensure curation, open access and interoperability for the SGS LTER and CSU digital scholarly assets.

8.0 Acknowledgements

9.0 References

ARL, A. R. L. (2009). The Research Library's Role in Digital Repository Services. In A. D. R. I. T. Force (Ed.).

Baker, K. S., Benson, B. J., Henshaw, D. L., Blodgett, D., Porter, J. H., and Stafford, S. G. (2000). Evolution of a multisite network information system: the LTER information management paradigm. *BioScience, 50*(11), 963-978.

Baker, K. S. and Karasti, H. (2004). The Long-Term Information Management Trajectory: Working to Support Data, Science and Technology: Scripps Institution of Oceanography, http://escholarship.org/uc/item/7d64x0bd.pdf.

Bohm, S., Boose, E., Costa, D., Downing, J., Gastil-Buhl, M., Gries, C., O'Brien, M., and Servilla, M. (2012). Report to the IMC, EML Data Package Checks and the PASTA Quality

Engine. IMC EML Metrics and Congruency Checker Working Group, July 2012, Albuquerque, New Mexico, https://im.lternet.edu/sites/im.lternet.edu/files/Data_package_quality_checks_Report_july2012.pdf.

Baker, K. S. and Yarmey, L. (2009). Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation, 4*(2), 12-27.

CSUL, Colorado State University Library Digital Collections http://lib.colostate.edu/digital-collections/

CSUL-DL, Colorado State University Libraries Digital Repository http://lib.colostate.edu/repository/ http://digitool.library.colostate.edu/R/?func=collections&collection_id=2384

DataCite registry, http://www.datacite.org/

DataCite metadata, http://schema.datacite.org/

DataCite search, http://search.datacite.org/ui

DCI, Dublin Core Metadata Initiative http://dublincore.org/documents/dces/

DigiTool, Digital Collections of Colorado, http://digitool.library.colostate.edu/

Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., Bermudez, B., and Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Information Systems, 4*, 139-160.

Ecological Metadata Language (2014). Accessed Sept. 24, 2014: https://knb.ecoinformatics.org/#external//emlparser/docs/index.html

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., Calvert, S. (2013). Knowledge Infrastructures: Intellectual Frameworks and Research Challenges. Ann Arbor: Report of a workshop sponsored by the National Science Foundation and the Sloan Foundation, Deep Blue.

Gorentz, J. (1992) Data management at biological field stations and coastal marine laboratories. Report of an invitational workshop, W.K. Kellogg Biological Station 22-26 Apr 1990, Lansing (MI): Michigan State University.

Hunter, N.C., S. Liu, P.J. Rettig, and A. Level, 2008. .CSU Core Data Dictionary Version 1.1, Metadata Best Practices Task Force. http://hdl.handle.net/10217/3147d

Hunter, N. C., Legg, K., and Oehlerts, B. (2010). Two Librarians, an Archivist, and 13,000 Images: Collaborating to Build a Digital Collection. *The Library Quarterly, 80*(1), 81-103.

Jones, M. B., Berkley, C., Bojilova, J., & Schildhauer, M. (2001). Managing Scientific Metadata. IEEE Internet Computing, 59-68.

Kaplan, N. (2014). Data Management for Open Access: 5 Questions to ask about your data. http://nrelscience.org/2014/06/10/the-data-lifecycle-part-1-data-management-for-open-access-5-questions-to-ask-about-your-data/

Kaplan, N. and H. Karasti (2013). The Science Stories Behind the Data. LTER Network News Fall 2013, Vol 25(3). http://news.lternet.edu/Article2653.html

Karasti, H. and Baker, K.S. (2004). Infrastructuring for the long-term: Ecological information management. Proceedings of the 37th Annual Hawaii International Conference on System Sciences.

Kelly, K., Marlino, M., Mayernik, M. S., Allard, S., Tenopir, C., Palmer, C. L., and Varvel Jr, V. E. (2013). Model Development for Scientific Data Curation Education. *International*

Journal of Digital Curation, 8(1), 255-264.

Kortz, M., Yarmey, L., Conners, J., Ackerman, T. and Baker, K. (2009). LTER Unit Working Group Projects: Dictionary and Registry, Poster. 2009 LTER All Scientists Meeting 13-17 September, 2009. Estes Park, Colorado. http://oceaninformatics.ucsd.edu/docs/ASM2009_UnitWGPoster.pdf

Lauenroth, W.K. and Burke, I.C. (2008). Ecology of the Shortgrass Steppe. Oxford University Press, New York, New York. 522 p.

Palmer, C. L., Allard, S., and Marlino, M. (2011). *Data curation education in research centers.* Paper presented at the Proceedings of the 2011 iConference.

LTER IM, 2011. LTER IM Best Practices, LTER IM Best Practices Guide: http://im.lternet.edu/node/910 or http://intranet2.lternet.edu/sites/intranet2.lternet.edu/files/documents/Scientific%20Reports/Informatics/emlbestpractices-2.0-FINAL-20110802.pdf)

LTER Example File, http://im.lternet.edu/sites/im.lternet.edu/files/example_full_a.xml

LTER Data Portal, https://portal.lternet.edu/nis/browse.jsp

LTER Unit Registry, http://unit.lternet.edu/unitregistry/

Mayernik, M. S. (2012). Data citation initiatives and issues. *Bulletin of the American Society for Information Science and Technology, 38*(5), 23-28.

Mayernik, M. S. (2013). Bridging data lifecycles: Tracking data use via data citations workshop report. Boulder, CO: National Center for Atmospheric Research.

Mayernik, M. S., Daniels, M. D., Dattore, R. E., Davis, E. R., Ginger, K., Kelly, K. M., M. Marino, M., Middleton, D.E., Phillips, J., Strand, G., Williams, S.F., Worley, S.J., and Wright, M. J. (2012). Data citations within NCAR/UCP. In N. T. Note (Ed.). Boulder, CO: National Center for Atmospheric Research.

Michener, W.K., Brunt, J.W. and S.G. Stafford (1994). Environmental Information Management and Analysis: Ecosystem to Global Scales. CRC Press.

Michener, W. K., Porter, J., Servilla, M., and Vanderbilt, K. (2011). Long term ecological research and information management. *Ecological Informatics, 6*, 13-24.

NIS, 2014. LTER Network Information System Community Website. Accessed September 2014 at: https://nis.lternet.edu:8443/display/NISnew.

Porter, J.H. 2009. Developing a Controlled Vocabulary for LTER Data. LTER Databits Fall 2009. http://databits.lternet.edu/node/70

Reichman, O. J., Jones, M. B. and Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science,* 331(6018), 703-705.

Zimmerman, D. and Paschal, D. B. (2009). An exploratory usability evaluation of Colorado State University Libraries' digital collections and the Western Waters Digital Library web sites. *The Journal of Academic Librarianship, 35*(3), 227-240.

10. Appendices

10.1 Appendix: Table of Acronyms

Acronym	Organizational Name	Link
ACNS	Academic Computing and Networking Services	http://www.acns.colostate.edu
CSU	Colorado State University	http://www.colostate.edu/
DCC	Digital Collections of Colorado	http://digitool.library.colostate.edu/about
DOI	Digital Object Identifier	http://www.doi.org
IBP	International Biological Program	http://www.nasonline.org/about-nas/history/archives/collections/ibp-1964-1974-1.html
IR	Institutional Repository	http://lib.colostate.edu/institutional-repository
LTAR	Long Term Agricultural Research	http://www.ars.usda.gov/Research/docs.htm?docid=21984
LTERNet	Long-Term Ecological Research Network	http://lternet.edu
LNO	LTER Network Office	http://lno.lternet.edu
NREL	Natural Resource Ecology Laboratory	http://www.nrel.colostate.edu/
SGRC	Semi-arid Grassland Research Center	http://sgrc.colostate.edu/
SGS LTER	Shortgrass Steppe Long Term Ecological Research	http://www.sgslter.colostate.edu/
UIUC	University of Illinois Urbana-Champaign	http://illinois.edu
USDA	United States Department of Agriculture	http://www.usda.gov/
USDA ARS	USDA Agricultural Research Service	http://www.ars.usda.gov/
USFS	United States Forest Service	http://www.fs.fed.us/
USFS PNG	USFS Pawnee National Grassland	http://www.fs.usda.gov/arp

10.2 Appendix: Table of Applications

Abbreviation	Name	Application	Link
Centroid	Geospatial Centroid	Geographical visualization at CSU	http://gis.colostate.edu
DCC	Digital Collections of Colorado	Colorado consortium use of DigiTool as part of institutional repositories	http://digitool.library.colostate.edu/R?RN=6077 10997
DigiTool	DigiTool	Digital asset management tool for collections	http://www.exlibrisgroup.com/category/DigiTool Overview
DC	Dublin Core	Metadata standard	http://dublincore.org
Dspace	Dspace	Repository software	http://www.duraspace.org
EML	Ecological Metadata Language	Metadata standard for ecological sciences	http://knb.ecoinformatics.org/software/eml
IBIS	NREL International Biological Information System	Value Added Information System for scientists and citizens	http://www.ibis.colostate.edu/cwis438/websites /IBIS/Home.php?WebSiteID=10
NIS	Network Information System	Information system developed by the LTER	http://nis.lternet.edu
PASTA	Provenance Aware Synthesis Tracking Architecture	Architectural element of the LTER NIS	ttps://nis.lternet.edu:8443/display/NISnew/PAS TA+-+the+NIS+infrastructure

10.3 Appendix: SGS LTER Recommended Keywords

aboveground net primary productivity	density	nitrogen	salinity
animals	disturbance	organic nitrogen	shrublands
arthropods	evaporation	paleoecology	soil moisture
barometric pressure	fertilization	paleopedology	soil survey
belowground temperature	fire ecology	paleosol	soil temperature
biomass	genetics	plant/animal interactions	soils
blue grama	geology	plants	soils series
carbon dioxide	grasslands	playa	stream terrace
catena	hillslope	Pleistocene	swift fox
chronosequence	holocene	population dynamics	temperature
cloud cover	invertebrates	precipitation	toposequence
cover	landscape	radiation	vegetation structure
coyote	mammals	relative humidity	water
	meteorology	rodents	weather

10.4 Appendix: SGS LTER Controlled Vocabulary: Units

name	unitType	abbreviation
acre	area	a
area	area	a
atmosphere	pressure	atm
bar	pressure	bar
becquerel	radionucleotideActivity	Bq
britishThermalUnit	energy	btu
bushel	volume	b
bushelsPerAcre	volumetricArea	b/a
calorie	energy	cal
celsius	degrees	C
centigram	mass	cg
centimeter	length	cm
centimeterPerYear	speed	cm/year
centimetersPerSecond	speed	cm/s
cubicCentimetersPerCubicCentimeters	volumePerVolume	cm^3/cm^3
cubicInch	volume	in^3
cubicMeter	volume	m^3

decigram	mass	dg
decimeter	length	dm
degree	angle	'I
dekagram	mass	dag
dekameter	length	dam
dimensionless	dimensionless	
fahrenheit	degrees	F
farad	capacitance	F
foot	length	ft
footPound	energy	ft lb
gallon	volume	gal
grad	angle	grad
gram	mass	g
gramcaloriePerCentimeterSquared	solar radiation (Langley)	LY
gramsPerCentimeterSquaredPerSecond	arealMassDensityRate	$g/cm^2/sec$
gramsPerCubicCentimeter	massDensity	g/cm^3
gramsPerGram	massPerMass	g/g
gramsPerHectarePerDay	arealMassDensityRate	g/ha/day
gramsPerMeterSquaredPerYear	arealMassDensityRate	$g/m^2/yr$
gramsPerSquareMeter	arealMassDensity	g/m^2
gramsPerYear	massFlux	g/yr
gray	specificEnergy	Gy
hectare	area	ha
hectogram	mass	hg
hectometer	area	hm
henry	inductance	H
hertz	frequency	Hz
hour	Time	hr
inch	length	in
joule	energy	J
katal	catalyticActivity	kat
kilogram	mass	kg
kilogramPerCubicMeter	massDensity	kg/cm^3
kilogramsPerHectare	arealMassDensity	kg/ha
kilogramsPerHectarePerYear	arealMassDensityRate	kg/ha/yr
kilogramsPerMeterSquaredPerSecond	arealMassDensityRate	$kg/m^2/s$
kilogramsPerMeterSquaredPerYear	arealMassDensityRate	$kg/m^2/yr$
kilogramsPerSecond	massFlux	kg/s
kilogramsPerSquareMeter	arealMassDensity	kg/m^2
kilohertz	frequency	KHz

kiloliter	volume	kL
kilometer	length	km
kilopascal	pressure	kPa
kilovolt	potentialDifference	kV
kilowatt	power	kW
liter	volume	L
litersPerSquareMeter	volumetricArea	l/m^2
lumen	luminosity	lm
lux	illuminance	lx
megagram	mass	Mg
megahertz	frequency	MHz
megapascal	pressure	MPa
megavolt	potentialDifference	MV
megawatt	power	MW
meter	length	m
metersPerGram	massSpecificLength	m/g
microgram	mass	_g
microgramsPerGram	massPerMass	_
microliter	volume	_l
micrometer	length	_m
micron	length	_
mile	length	mile
millibar	pressure	mbar
milligram	mass	mg
milligramsPerCubicMeter	massDensity	mg/m^3
milligramsPerSquareMeter	arealMassDensity	mg/m^2
millihertz	frequency	mHz
milliliter	volume	ml
millimeter	length	mm
millimetersPerSecond	speed	mm/s
millimolesPerGram	amountOfSubstanceWeight	_/g
millivolt	potentialDifference	mV
milliwatt	power	mW
minute	time	min
molesPerGram	amountOfSubstanceWeight	_/g
molesPerKilogram	amountOfSubstanceWeight	_/kg
molesPerKilogramPerSecond	amountOfSubstanceWeightFlux	_/kg/s
nanogram	mass	ng
nanomolesPerGramPerSecond	amountOfSubstanceWeightFlux	_/g/s
newton	force	N
nominalDay	time	dd

nominalHour	time	hhhh
nominalLeapYear	time	
nominalMinute	time	mm
nominalMonth	time	mm
nominalWeek	time	
nominalYear	time	yyyy
number	dimensionless	
numberPerGram	massSpecificCount	
numberPerKilometerSquared	arealDensity	
numberPerMeterCubed	volumetricDensity	
numberPerMeterSquared	arealDensity	
ohm	resistance	_
ohmMeter	resistivity	_m
pascal	pressure	Pa
pint	volume	pint
pound	mass	lbs
quart	volume	qt
radian	angle	rad
second	time	sec
second	time	Msec
siemen	conductance	S
sievert	doseEquivalent	Sv
squareCentimeters	area	cm^2
squareFoot	area	ft^2
squareKilometers	area	km^2
squareMeter	area	m^2
squareMile	area	$mile^2$
squareMillimeters	area	mm^2
squareYard	area	yd^2
tesla	magneticFluxDensity	T
ton	mass	ton
tonne	mass	T
tonnePerHectare	arealMassDensity	T/ha
tonnesPerYear	massFlux	T/yr
volt	potentialDifference	V
watt	power	W
weber	magneticFlux	Wb

10.5 Appendix: CSU Library Metadata and Selection Criteria for Digital Images

Information on the Metadata Spread Sheet for image files:

1. Some background information:
   The spreadsheet header names in templatePhotoLog&Metadata_06_Warner_Natural_Resources_Ecology_Lab_DigiTool.csv are used for staff working on the project. Each spreadsheet column is mapped in ingest with the attached xml file, UDPC_mapping.xml, to the corresponding Dublin Core (DC) field in DigiTool. Each DC field is defined in the CSU Core Data Dictionary, CSU-Core-1.1.pdf.

2. The file: SGSLTERphotoTemplateSallieV2.xlsx should be used to inventory and document digital images of the natural world contains headers and rows as an example.

3. Fill out the included columns and leave any cells blank you don't know or are missing the information.

4. All the filenames shall begin with "NRELSGSL", and append the current filename with an underscore. Each image filename will be the same as the master file name in the spreadsheet. I simply prepend "NRELSGSL" to each filename to ensure all images are identified as part of the SGS collection. Also, we need to have the filenames with either no spaces or with underscores, e.g. NRELSGSLBlueHeronMLindquist.jpg or NRELSGSL_Blue_Heron_MLindquist.jpg; however, all dashes and spaces need to be changed to align with the preceding examples. Each exact photograph filename needs to be mirrored in the spreadsheet's first column.

5. The purpose for adding a logical filename in the spreadsheet is to give a unique number to each SGS photograph. Example entries are in the spreadsheet to help exemplify how it works. It may seem redundant, but it is an important internal number for the IR.

6. Do we want the photographs to be separated into groups (arthropods, aves, flora, etc.) or all placed in one collection. If you want them separated, it will be easiest to place each group on different spreadsheets and annotate the group in the spreadsheet filename.

7. Criteria for Image Files:
   a. Each should be at least 800 on the long side; however, if you have important photos under 800, those could be added as well.
   b. Save as something about 1 – 1.5 Meg at 200 or 300 dpi for 4" x6" to 8" x 12" prints.

10.6 Appendix: SGS LTER Project Collection in CSU Digital Collections of Colorado

Level 1. Portal for the CSU DCC.

Colorado State University
Libraries

Digital Collections

DCC Home   Search / Browse   Results   Previous Searches   My Account   About   Login   End Session   Help

Simple Search   Advanced Search

A word or phrase: ⦿ Contains ○ Exact ○ Starts With

[   ]   [Search]

Select collection: [Colorado State University, Fort Collins ▲▼]

Browse All Collections > **Colorado State University, Fort Collins**

CSU Departments and Schools (4430)
Agricultural and Resource Economics
Animal Sciences
Anthropology ...

CSU Colleges (2290)
University Libraries
Veterinary Medicine and Biomedical Sciences
Warner College of Natural Resources

CSU Centers and Research Institutes (2232)
Center for Collaborative Conservation
Colorado Institute of Public Policy
Colorado Water Institute ...

CSU Theses and Dissertations (2453)
2011-
1980-2010
1950-1979 ...

CSU Archives and Special Collections (38536)
Agricultural and Natural Resources Archive
Manuscripts
Special Collections ...

CSU Conference Proceedings and Events (44)
Bridging the Gap Conference
CI Days: Cyberinfrastructure 2010 in the Rockies
Forest Biomass Conference

CSU Journal Publications (29)
Furthering Perspectives
Journal of Student Affairs
Journal of Undergraduate Research and Scholarly Excellence

CSU Student Research Projects (43)
Celebrate Undergraduate Research and Creativity (CURC) Showcase

Books (12)

The Digital Collections of Colorado include publications, theses and dissertations, presentations, historical materials and other scholarly works in various formats.
Share • Bookmarkable URL • Statistics for this collection • About the Digital Collections of Colorado • Contact Us • Home
© Copyright 2014 Colorado State University, the University of Colorado, Colorado School of Mines and Colorado Mesa University

Level 2. Portal for Colorado State University.

Level 3. Portal for CSU Colleges.

Level 4. Portal for Warner College of Natural Resources; Note: Natural Resource Ecology Laboratory (NREL) and International Biome Program (IBP) in addition to SGS LTER.

Level 5. Portal for NREL including the SGS LTER.

Level 6. Portal to SGS LTER Project Collection.

Level 7. Two views of the SGS LTER Project Series of Data.

Level 8. Landing page for a SGS LTER data package.

10.7 Appendix: Dublin Core Elements Configured for Data

Label	Dublin Core Element	Notes
Award	<dcterms:award>	
Referenced by	<dcterms:isReferencedBy/>	
References	<dcterms:references/>	
Source	<dc:source/>	Amended for data with "Data source:" in value.

10.8 Appendix Metadata Mapping from EML to CSU Qualified Dublin Core for Data Packages and Related Materials

Element	EML	Dublin Core
Abstract		<dcterms:abstract>
Author	<dataset> <Creator> <Individual name> <givenName> <surName>	<dc:author
Award	<funding> <section> <para>	<dcterms:award>
Date	<PubDate>	<dc:date>
Geographic Coverage	<coverage> <geographicCoverage> <geographicDescription>	<dcterms:spatial>
Keywords	<keywordSet>	<dcterms:keywords>
Rights	<intellectualRights> <section> <para>	<dc:rights/>
Temporal Coverage	<temporalCoverage> <rangeOfDates> <beginDate> <calendarDate> <endDate>	<dcterms:temporal>
Title	<dataset> <title>	<dc:title>
Publisher	<publisher>	<dc:publisher>
Contributor		<dc:contributor>
Distribution	<distribution> <online> <url function>	

10.9 Appendix: SGS LTER Project Data and Materials Migration Poster

The SGS LTER project was described in a poster presented at the International Digital Curation Conference (IDCC, 24-27 February, 2014; San Francisco). The poster file can be found in the DCC here: http://hdl.handle.net/10217/81458.