Linking data repositories - an illustration of agile data curation principles through robust documentation and multiple application programming interfaces

IN13A-1821

Karl Benedict[1], Mark Stephen Servilla[2], Kristin Vanderbilt[3], Jonathan Wheeler[4]

Problem

The replication of data and associated documentation between systems for data security, enhanced discovery and access, or use on the secondary system is increasingly difficult as the number and variety of data objects increases. The development of data management and archiving systems and workflows that lower the barriers to automated movement of data and documentation between systems can enable efficient and scalable replication/migration of data assets between systems and between systems and users. Conceptually, this model of streamlined data management and curation based on sound system design principles may be described as agile data curation.

When confronted with the planned discontinuation of funding for the Sevilleta LTER research site, the LTER Network Infomation Information System designer (Servilla) and the Sevilleta Information Manager (Vanderbilt) approached the research data services team in the UNM Libraries (Benedict & Wheeler) to discuss the options for replication of the Sevilleta data and metadata assets currently managed within the LTER data management system (known as PASTA) into UNM's institutional repository - LoboVault. The workflow that was developed to complete this replication is an excellent example of how efficient data transfer, management and reuse can be facilitated through the development of independent systems that are based upon design principles that enable well-structured machine-to-machine communication (via well documented Application Programming Interfaces - APIs), support of standards based documentation models, and effective communication between project participants throughout the development, testing, and implementation process. While not specifically intended to do so, the process used (and described in this poster) illustrates a number of core principles of agile data curation.

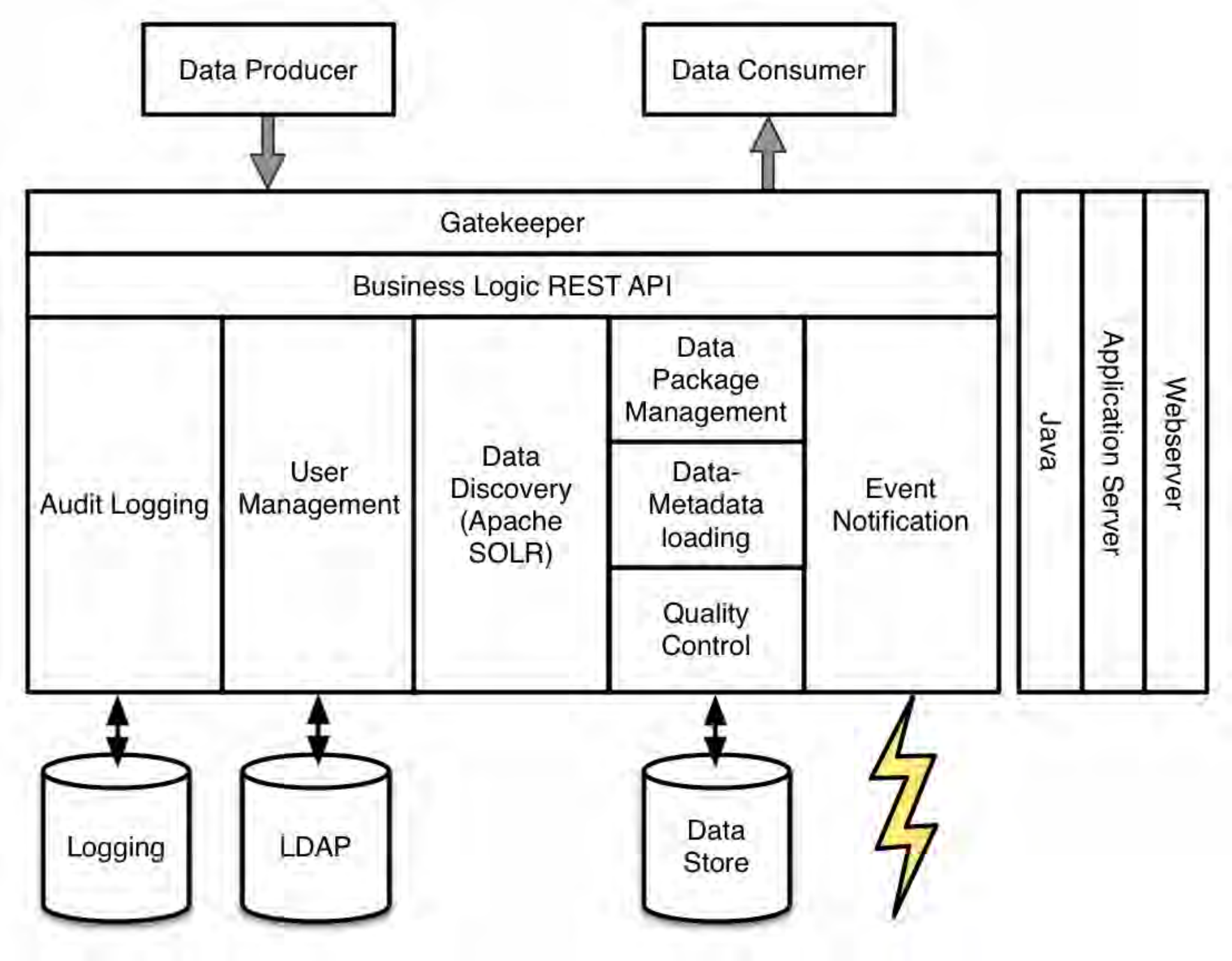


Figure 1. PASTA Conceptual Diagram

Design Goals and Attributes of the Provenance Aware Synthesis Tracking Architecture (PASTA)

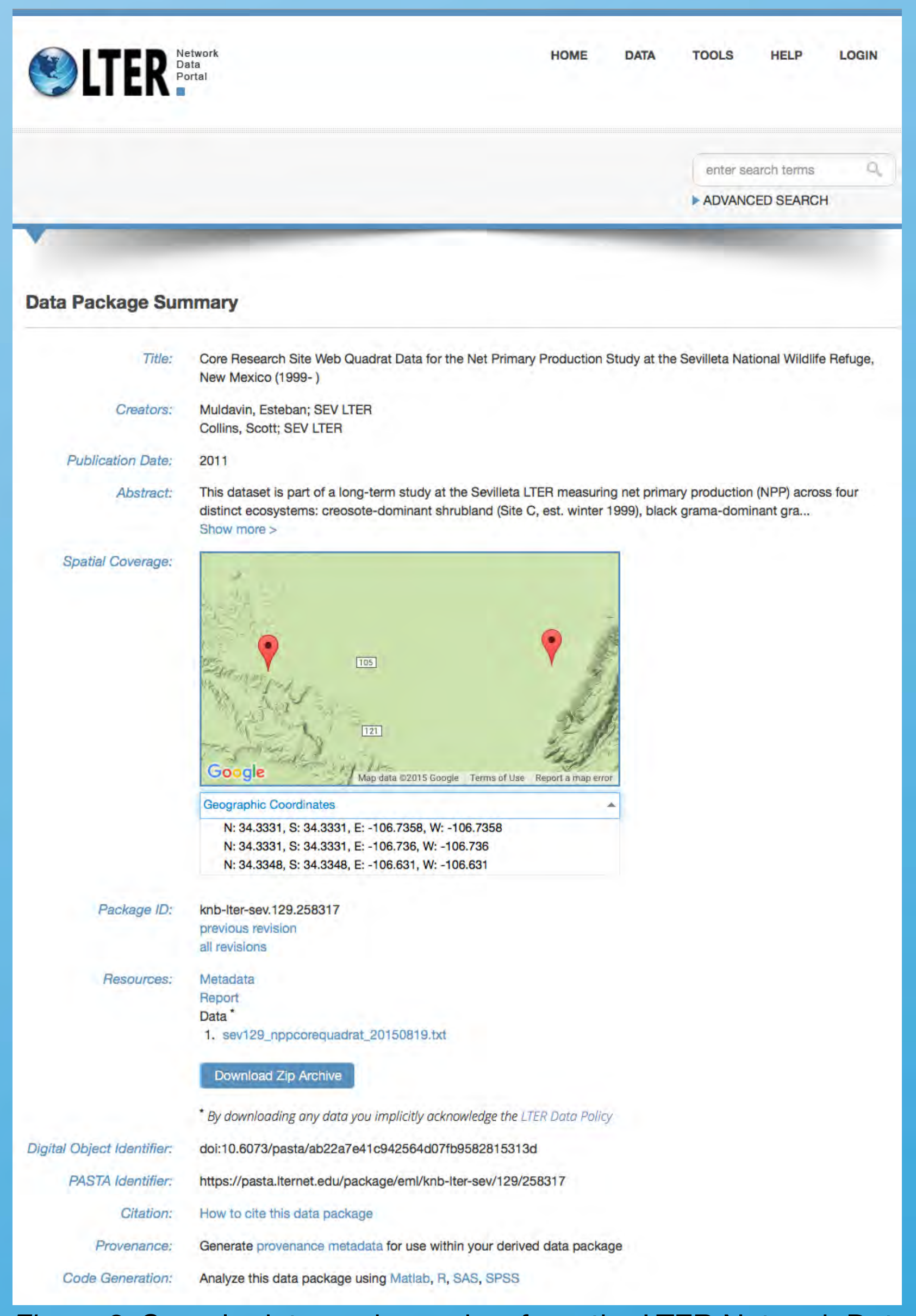
The PASTA platform was designed as part of the LTER Network Information System[6] to support the data discovery, access and use needs of the community of Long Term Ecological Research sites and the users of the data that they produce. It has been developed with a well documented Application Programming Interface (API[7]) that enables create, evaluate, read, update, delete, list, and search for data package resources in the PASTA system. In particular it has the following characteristics:

- Use of the Ecological Metadata Language[8] as its metadata standard
- Employs the Metacat XML database to support metadata search
- Uses data warehousing methods to provide a uniform representation of the heterogeneous data managed within the system
- Provides a web interface[9] for browsing, searching, and viewing the contents of the PASTA system

Integration Workflow Development Process

The integration of the PASTA data packages for the Sevilleta LTER site leveraged the API's of both PASTA and DSpace to automate the process in a way that would meet the immediate needs for the Sevelleta collection while being extensible to all of the collections within PASTA. The specific capabilities of the developed workflow include:

- Leveraging the extensibility of the DSpace metadata model to integrate location data within the PASTA EML metadata into the structured metadata within LoboVault using the Darwin Core extension to the Dublin Core standard
- Maintaining dataset version information provided by PASTA within the LoboVault records
- Conversion of the EML metadata content into corresponding DSpace Dublin Core or Darwin Core elements for search and presentation
- Providing access to the source EML metadata and data objects housed within LoboVault
- Providing an interactive data location map based on information extracted from the source EML metadata



Data Package Summary

Title: Core Research Site Web Quadrat Data for the Net Primary Production Study at the Sevilleta National Wildlife Refuge, New Mexico (1999-)

Creators: Muldavin, Esteban; SEV LTER
Collins, Scott; SEV LTER

Publication Date: 2011

Abstract: This dataset is part of a long-term study at the Sevilleta LTER measuring net primary production (NPP) across four distinct ecosystems: creosote-dominant shrubland (Site C, est. winter 1999), black grama-dominant gra...
Show more >

Spatial Coverage:
Geographic Coordinates
N: 34.3331, S: 34.3331, E: -106.7358, W: -106.7358
N: 34.3331, S: 34.3331, E: -106.736, W: -106.736
N: 34.3348, S: 34.3348, E: -106.631, W: -106.631

Package ID: knb-lter-sev.129.258317
previous revision
all revisions

Resources: Metadata
Report
Data *
1. sev129_nppcorequadrat_20150819.txt

* *By downloading any data you implicitly acknowledge the LTER Data Policy*

Digital Object Identifier: doi:10.6073/pasta/ab22a7e41c942564d07fb9582815313d

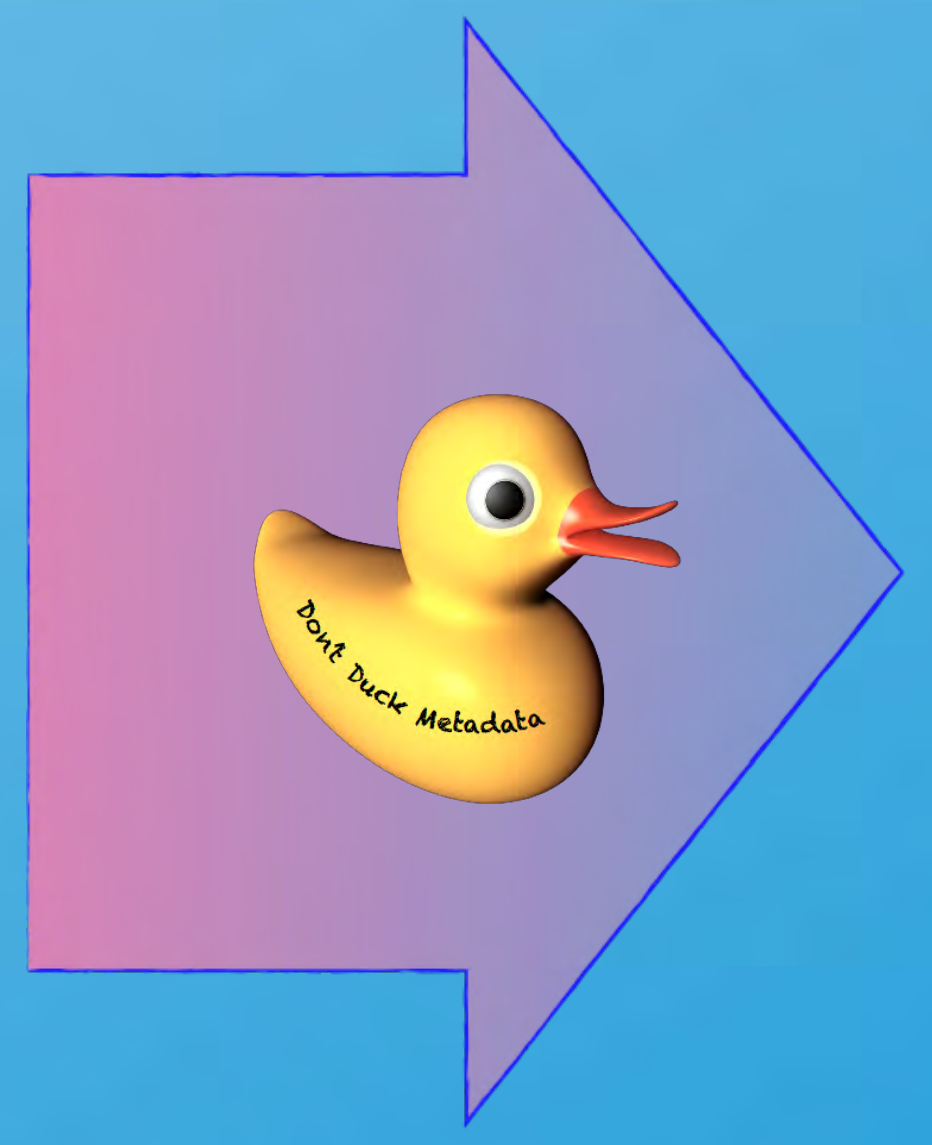
PASTA Identifier: https://pasta.lternet.edu/package/eml/knb-lter-sev/129/258317

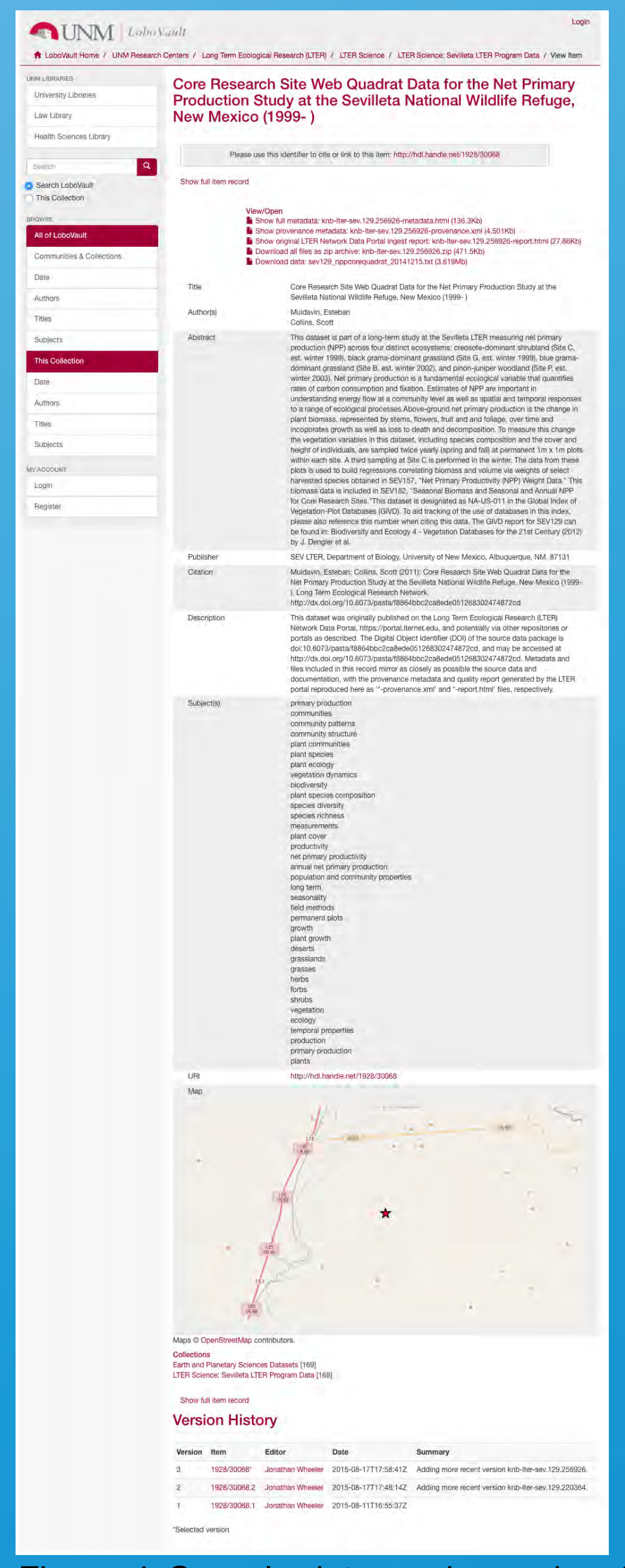
Citation: How to cite this data package

Provenance: Generate provenance metadata for use within your derived data package

Code Generation: Analyze this data package using Matlab, R, SAS, SPSS

Figure 3. Sample data package view from the LTER Network Data Portal[14]





Core Research Site Web Quadrat Data for the Net Primary Production Study at the Sevilleta National Wildlife Refuge, New Mexico (1999-)

Please use this identifier to cite or link to this item: http://hdl.handle.net/1928/30068

Show full item record

View/Open
- Show full metadata: knb-lter-sev.129.256926-metadata.html (136.3Kb)
- Show provenance metadata: knb-lter-sev.129.256926-provenance.xml (4.501Kb)
- Show original LTER Network Data Portal ingest report: knb-lter-sev.129.256926-report.html (27.86Kb)
- Download all files as zip archive: knb-lter-sev.129.256926.zip (471.5Kb)
- Download data: sev129_nppcorequadrat_20141215.txt (3.619Mb)

Title: Core Research Site Web Quadrat Data for the Net Primary Production Study at the Sevilleta National Wildlife Refuge, New Mexico (1999-)

Author(s): Muldavin, Esteban
Collins, Scott

Abstract: This dataset is part of a long-term study at the Sevilleta LTER measuring net primary production (NPP) across four distinct ecosystems: creosote-dominant shrubland (Site C, est. winter 1999), black grama-dominant grassland (Site G, est. winter 1999), blue grama-dominant grassland (Site B, est. winter 2002), and pinon-juniper woodland (Site P, est. winter 2003). Net primary production is a fundamental ecological variable that quantifies rates of carbon consumption and fixation. Estimates of NPP are important in understanding energy flow at a community level as well as spatial and temporal responses to a range of ecological processes.Above-ground net primary production is the change in plant biomass, represented by stems, flowers, fruit and and foliage, over time and incoporates growth as well as loss to death and decomposition. To measure this change the vegetation variables in this dataset, including species composition and the cover and height of individuals, are sampled twice yearly (spring and fall) at permanent 1m x 1m plots within each site. A third sampling at Site C is performed in the winter. The data from these plots is used to build regressions correlating biomass and volume via weights of select harvested species obtained in SEV157, "Net Primary Productivity (NPP) Weight Data." This biomass data is included in SEV182, "Seasonal Biomass and Seasonal and Annual NPP for Core Research Sites."This dataset is designated as NA-US-011 in the Global Index of Vegetation-Plot Databases (GIVD). To aid tracking of the use of databases in this index, please also reference this number when citing this data. The GIVD report for SEV129 can be found in: Biodiversity and Ecology 4 - Vegetation Databases for the 21st Century (2012) by J. Dengler et al.

Publisher: SEV LTER, Department of Biology, University of New Mexico, Albuquerque, NM. 87131

Citation: Muldavin, Esteban; Collins, Scott (2011): Core Research Site Web Quadrat Data for the Net Primary Production Study at the Sevilleta National Wildlife Refuge, New Mexico (1999-). Long Term Ecological Research Network. http://dx.doi.org/10.6073/pasta/f8864bbc2ca8ede051268302474872cd

Description: This dataset was originally published on the Long Term Ecological Research (LTER) Network Data Portal, https://portal.lternet.edu, and potentially via other repositories or portals as described. The Digital Object Identifier (DOI) of the source data package is doi:10.6073/pasta/f8864bbc2ca8ede051268302474872cd, and may be accessed at http://dx.doi.org/10.6073/pasta/f8864bbc2ca8ede051268302474872cd. Metadata and files included in this record mirror as closely as possible the source data and documentation, with the provenance metadata and quality report generated by the LTER portal reproduced here as '*-provenance.xml' and '*-report.html' files, respectively.

Subject(s): primary production
communities
community patterns
community structure
plant communities
plant species
plant ecology
vegetation dynamics
biodiversity
plant species composition
species diversity
species richness
measurements
plant cover
productivity
net primary productivity
annual net primary production
population and community properties
long term
seasonality
field methods
permanent plots
growth
plant growth
deserts
grasslands
grasses
herbs
forbs
shrubs
vegetation
ecology
temporal properties
production
primary production
plants

URI: http://hdl.handle.net/1928/30068

Map

Collections
Earth and Planetary Sciences Datasets [169]
LTER Science: Sevilleta LTER Program Data [168]

Show full item record

Version History

Version	Item	Editor	Date	Summary
3	1928/30068*	Jonathan Wheeler	2015-08-17T17:58:41Z	Adding more recent version knb-lter-sev.129.256926.
2	1928/30068.2	Jonathan Wheeler	2015-08-17T17:48:14Z	Adding more recent version knb-lter-sev.129.220364.
1	1928/30068.1	Jonathan Wheeler	2015-08-11T16:55:37Z	

*Selected version

Figure 4. Sample data package view from the LoboVault LTER collection[15]

Mapping of Agile Software Development Principles into Data Curation

Agile Software Development Principles[5]	Agile Data Curation Principles
Our highest priority is to satisfy the customer through early and continuous delivery of valuable software.	Maximize the impact of research data through accelerated capacity for discovery, access and use of valuable data
Welcome changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage.	Expect unanticipated needs for and uses of research data (and documentation) and develop flexible systems to support new uses and users without significant modifications
Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.	Facilitate automated interaction with data and metadata assets through well documented public web services that enable disintermediated use and reuse of research data
Business people and developers must work together daily throughout the project.	Data creators and data curators should work closely throughout planning, research and preservation activities to ensure the most efficient and streamlined process
Build projects around motivated individuals. Give them the environment and support they need, and trust them to get the job done.	Identify key individuals in a data curation project that have the requisite knowledge and motivation to do the job and get out of their way
The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.	Identify the most effective method(s) for maintaining close communication and *use* them
Working software is the primary measure of progress.	Delivery, access, use and citation of research data are the primary measures of success
Agile processes promote sustainable development. The sponsors, developers, and users should be able to maintain a constant pace indefinitely.	Design principles that enable steady delivery of incremental improvements to research data discovery, access and use should be consistent with a sustainable level of effort and funding from sponsors, data creators and curators, and users
Continuous attention to technical excellence and good design enhances agility.	Continuous attention to technical excellence and good design enhances agility
Simplicity--the art of maximizing the amount of work not done--is essential.	Start with the basics and only make systems more complex as needed, while maintaining a low bar to entry
The best architectures, requirements, and designs emerge from self-organizing teams. At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behavior accordingly.	Continuously work to develop and evolve a community of data providers, curators and users that all participate in the ongoing evolution of the research data systems that they interact with

Lessons Learned

This data and metadata integration project provides both an illustration of many of the agile data curation principles outlined below, but it also highlighted a number of ongoing challenges in addition to the benefits:

Benefits:

- The availability of robust APIs in both PASTA and LoboVault enabled a high degree of automation that provides the promise of significantly streamlined ingest of additional PASTA collections if/when needed
- The detailed and consistent EML metadata standard used in PASTA provided a rich resource of machine-readable documentation that could easily be mapped into LoboVault's metadata model
- The project provided a concrete project around which the LTER and Library data management teams could collaborate - reinforcing the development of a growing research data management community of practice at UNM
- Core PASTA data products are now safely replicated in a complementary system for long-term preservation and access

Challenges:

- The different versioning models implemented within PASTA and LoboVault introduce a step in the processing of materials that hasn't been fully automated yet. This poses a challenge in maintaining synchronization between the systems.
- The communication model based primarily on email exchanges introduced some delays in developing shared understanding that could have been achieved more quickly through in-person interactions.
- Both PASTA and LoboVault support permanent identifiers but the identifiers used in the systems differ in that PASTA uses DOIs and LoboVault (currently) uses Handles. The specification of identifying the canonical identifier remains to be done.

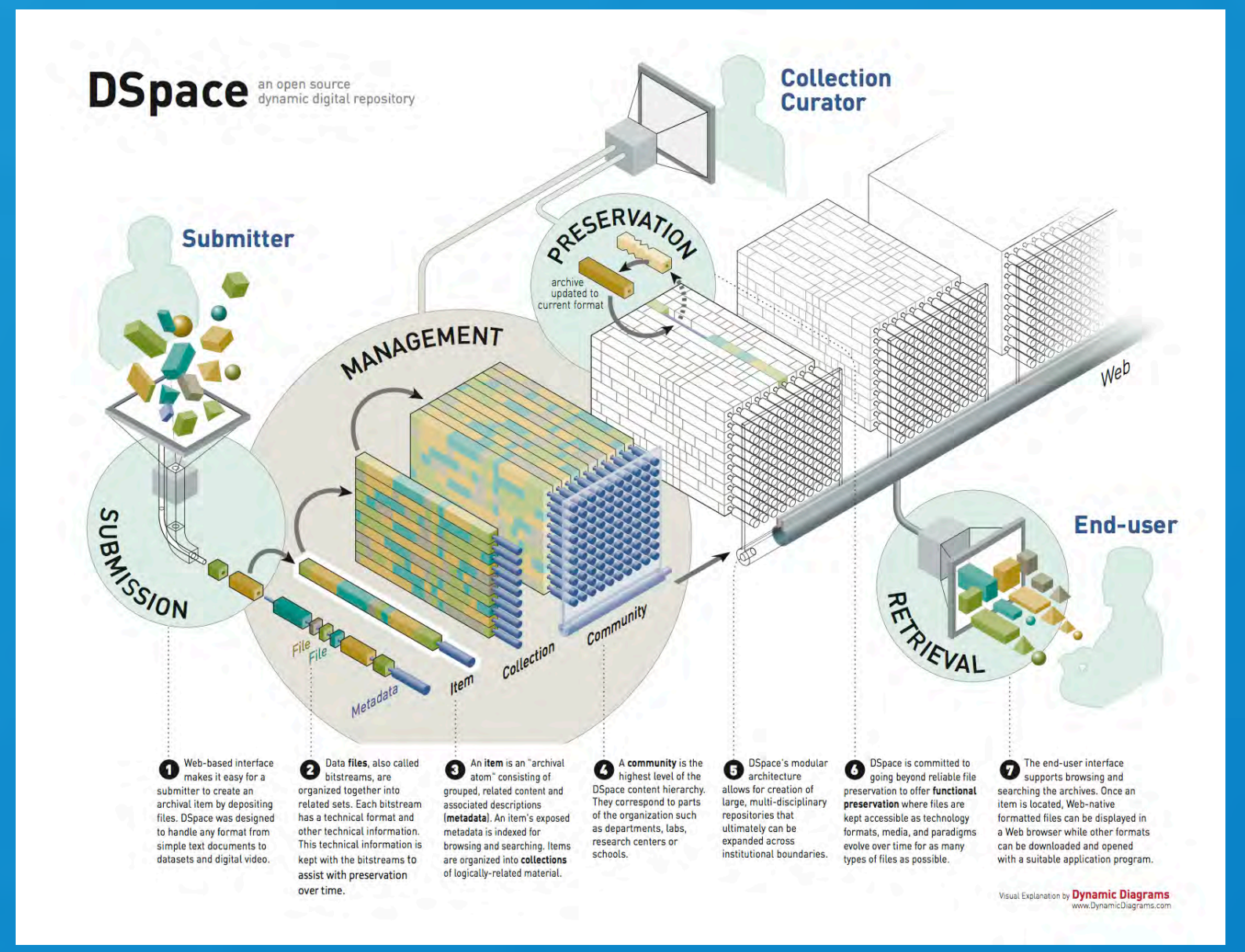

DSpace
an open source dynamic digital repository
Collection Curator
Submitter
PRESERVATION
archive updated to current format
MANAGEMENT
SUBMISSION
File
File
Metadata
Item
Collection
Community
Web
End-user
RETRIEVAL
1 Web-based interface makes it easy for a submitter to create an archival item by depositing files. DSpace was designed to handle any format from simple text documents to datasets and digital video.
2 Data files, also called bitstreams, are organized together into related sets. Each bitstream has a technical format and other technical information. This technical information is kept with the bitstreams to assist with preservation over time.
3 An item is an "archival atom" consisting of grouped, related content and associated descriptions (metadata). An item's exposed metadata is indexed for browsing and searching. Items are organized into collections of logically-related material.
4 A community is the highest level of the DSpace content hierarchy. They correspond to parts of the organization such as departments, labs, research centers or schools.
5 DSpace's modular architecture allows for creation of large, multi-disciplinary repositories that ultimately can be expanded across institutional boundaries.
6 DSpace is committed to going beyond reliable file preservation to offer functional preservation where files are kept accessible as technology formats, media, and paradigms evolve over time for as many types of files as possible.
7 The end-user interface supports browsing and searching the archives. Once an item is located, Web-native formatted files can be displayed in a Web browser while other formats can be downloaded and opened with a suitable application program.
Visual Explanation by Dynamic Diagrams www.DynamicDiagrams.com

Figure 2. DSpace Conceptual Diagram[13]

Design Goals and Attributes of the DSpace Digital Repository Platform (LoboVault)

The University of New Mexico Libraries' DSpace[10] based institutional repository, LoboVault, is an open access portal to the scholarly publications and research data of UNM faculty, graduate student theses and dissertations, and university administrative records. As a general purpose repository, LoboVault has been developed to provide an easy to use access and discovery resource for the full array of scholarly content and data types produced at UNM. The recent implementation of collection-scale batch ingest procedures is further designed to facilitate efficient publication of research output subject to federal publication and data sharing requirements. Relevant features include:

- An extensible metadata model based on qualified Dublin Core[11]
- Metadata federation and integration with external systems via OAI-PMH[12]
- Batch ingest, editing, and cross-registration of content and metadata
- As of DSpace v5, a RESTful API supporting search, create, read, update, and delete functions.
- The development of a *sandbox* version of the repository for development and testing of workflows before execution of bulk ingest processes

Acknowledgements

The research reported here has been supported through funding from the National Science Foundation Cooperative Agreement DEB#0832652 (5/1/09 - 4/30/15), DEB#1440478 (4/1/2015 – 3/31/2017: Sevilleta LTER grant) and the College of Univerity Libraries and Learning Sciences at the University of New Mexico

[1] University of New Mexico, University Libraries - kbene@unm.edu↩
[2] University of New Mexico, LTER Network Office↩
[3] University of New Mexico, Sevilleta LTER Office↩
[4] University of New Mexico, University Libraries↩
[5] http://agilemanifesto.org/principles.html↩
[6] http://lno.lternet.edu/content/network-information-system↩
[7] https://pasta.lternet.edu/package/docs/api↩
[8] https://knb.ecoinformatics.org/#external//emlparser/docs/index.html↩
[9] https://portal.lternet.edu/nis/home.jsp↩
[10] http://www.dspace.org/↩
[11] http://dublincore.org/↩
[12] http://www.openarchives.org/OAI/openarchivesprotocol.html↩
[13] http://www.dspace.org/sites/dspace.org/files/media/DSpace%20Diagram_0.pdf - published by dspace.org and shared under a Creative Commons BY-SA license↩
[14] https://portal.lternet.edu/nis/mapbrowse?packageid=knb-lter-sev.129.258317↩
[15] https://repository.unm.edu/handle/1928/30068↩