OPERATIONS

Data Management Plans

LDC as Data Archive

LDC was founded as a permanent data repository and distribution point for language resources and has fulfilled that role since its inception. The LDC Catalog is a growing digital archive of over 600 holdings that for more than two decades has served as one of the world's major language resource depositories.

Funders like the National Science Foundation (NSF) now require researchers to deposit data in an accessible, trustworthy archive under a data management plan that must be submitted with a research proposal. LDC administers data management plans by providing archiving services and making data widely available under a variety of arrangements that protect intellectual property rights and privacy concerns.

As the first and most active language resource data center, LDC established or adopted many of the practices that the related research communities follow today. LDC's expertise in data curation, distribution and management and its commitment to the broad accessibility of linguistic data make it the repository of choice for researchers.

Examples of NSF-funded data sets distributed through LDC include:

- Grassfields Bantu Fieldwork
- Translanguage English Database
- SLX Corpus of Classic Sociolinguistic Interviews
- Penn Discourse Treebank
- Propbank
- Subglottal Resonances Database

Advantages of Data Center Distribution

The sharing requirement of data management plans may be satisfied in a number of ways: the researcher's website, an institutional website or a data center. Data centers offer numerous advantages because they have in place infrastructures and processes for reviewing, storing and distributing resources over the long-term, a key element for data management plans in general.

Quality: LDC has acquired and institutionalized data management skills that are applied to all holdings. Specifically, LDC performs pre-publication reviews of submitted data and works with contributors to resolve any issues pertaining to content or data integrity.

Data Management Plan Requirements

All applicants for NSF funding must submit a data management plan that describes how a proposal will implement sharing and dissemination of research data and results. A plan includes a description of the:

- Data
- Hosting archive
- Details of access and sharing
- Metadata used
- Pertinent intellectual property rights
- Ethics and privacy issues, if any
- Data, tools and documentation formats
- Plans for archiving and preservation

Stability: LDC has been in continuous operation since 1992, longer than other language-related data centers, with every corpus contributed to the Consortium still available. The steady support of Consortium members assures longevity even when soft-monies are scarce.

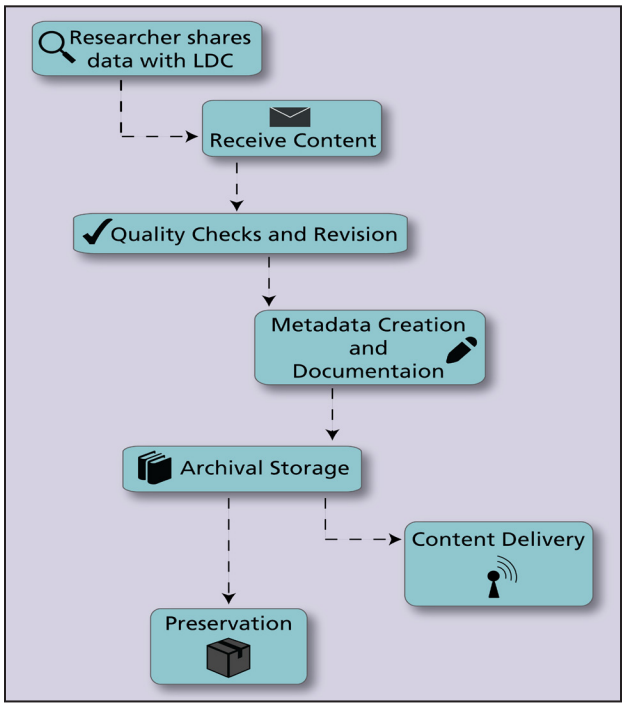
Discoverability: LDC adds all contributions to its Catalog with optimized search capabilities and mirrors the Catalog via OLAC, the Open Language Archives Community, and the ELRA Universal Catalog. Publications and activities of interest are announced on LDC's website, in its monthly newsletter – circulated to more than 8,000 researchers worldwide – and on social media. LDC has identified over 13,000 published papers that rely upon data available in the Catalog.

Expertise and Innovation: As technologies evolve, it is critical for data centers to create new infrastructure in order to maximize quality and effectiveness while minimizing cost and timeline. Recent LDC innovations include infrastructure for data collections from SMS, the WebAnn framework for creating annotation tools, an interface that implements familiar e-commerce concepts, and distribution through the cloud and service grids. Nearly as important as developing the necessary expertise is recognition among user communities. More than half of LDC's 600 titles are contributed and over 3500 organizations worldwide have licensed more than 120,000 copies of LDC's language resources

Linguistic Data Consortium 3600 Market St., Suite 810, University of Pennsylvania, Philadelphia, PA 19104-2653 USA
Telephone: +1.215.898.0464 • Fax: +1.215.573.2175 • ldc@ldc.upenn.edu • www.ldc.upenn.edu

Curation and Distribution Services

LDC offers a range of services that meet the requirements for data management plans and can be customized for a project's particular needs.



Data Curation Process

LDC maintains in-house storage solutions which accommodate over 200TB with the capability to scale to petabytes rapidly and transparently. LDC also leverages commercial cloud storage when appropriate. In addition to a specialized back-up system, LDC ensures that data is migrated to new formats, platforms and storage media as required by best practices in the digital preservation community.

LDC has an established track record for successfully distributing language resources to many users, by numerous methods and under a variety of licensing arrangements. LDC's licenses are compatible with the community's customary uses as well as with intellectual property and human subjects concerns. Comprehensive recordkeeping ensures that users always know their rights to specific data sets. Those practices enhance resource usability, preserve contributor's flexibility and ease the administrative burden.

Distribution services include the following:

Compatibility. LDC offers guidance on file-naming, metadata conventions, corpus structure and format.

Non-exclusive distribution. LDC does not insist on exclusive distribution rights to contributed data. Data creators may deposit their data at LDC and also at their institutional site or by other means.

Management of property rights, privacy and ethical concerns. Principal investigators, their institutions and third-party data providers retain their rights while licensing LDC to process, store and disseminate the resource to the community. In addition, LDC has vast experience in satisfying the legal and regulatory constraints imposed on data collection, annotation and archiving, and its staff includes experts on intellectual property, human subjects protection and export control.

Authorship. Recognizing the need for correct attribution, LDC requires contributors to name resource authors and includes authors as part of the descriptive metadata.

Timed/delayed accessibility. LDC has long experience in holding and protecting data for a timed or delayed release. For example, most NIST (National Institute of Standards and Technology) evaluation campaigns require that data be developed in advance, provided to campaign participants on schedule and then made publicly available only after the evaluation process is completed.

Licensing options. LDC implements procedures to protect the commercial value of language data including research-only licenses and referrals to data owners for commercial licensing.

Costs. LDC works with researchers to develop a funding model that is based on actual costs to assure long-term preservation and advance project goals.

Learn more about how LDC can assist researchers in developing and implementing data management plans from our website: https://www.ldc.upenn.edu/data-management/data-management-plans or contact dmp@ldc.upenn.edu.