



OPERATIONS

Share Data through LDC

Data Sharing and LDC's Mission

LDC's mission is to support language-related education, research and technology development by creating and sharing language resources including data, tools and standards. An important aspect of that mission is ensuring that LDC's published resources reach a broad spectrum of users – students, scholars, researchers, developers – in academia, government and industry.

Because these communities require data across languages, genres and formats, another key aspect of LDC's mission is to continuously expand and diversify its language resource catalog. LDC encourages you to share your data with the Consortium to further that mission.

A CoreTrustSeal trustworthy data repository

The LDC Catalog meets high standards for...



- Data access
- Rights management
- Curation
- Data integrity and authenticity
- Archival storage
- Security

LDC's Global Network

Data distributed through LDC becomes part of a strong global network. All published resources appear in LDC's online Catalog, accessed daily by users worldwide. LDC's newsletter keeps the community abreast of all new publications, and its reach ensures the attention of interested researchers.

LDC members receive free copies of corpora published in their membership year, guaranteeing exposure of your data to major organizations working in human language technologies and related fields.

Data Sources

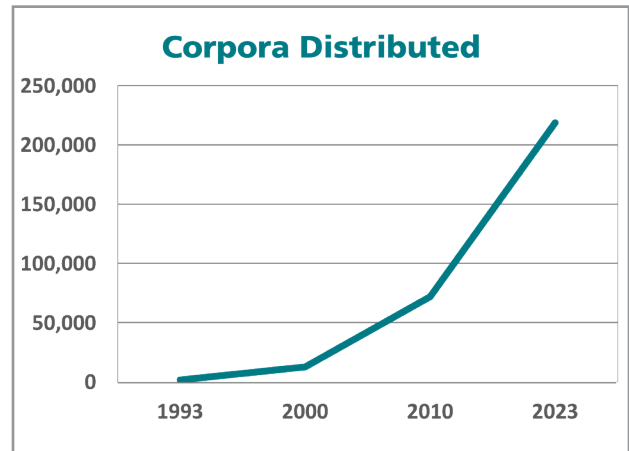
LDC's Catalog contains data developed or contributed by information publishers, researchers, universities and private organizations, including data sets deposited under data management plans. These multi-language resources run the gamut - text, speech, video, image and lexicons - and are collected and annotated under a multitude of conditions.

Why You Should Share Your Data through LDC

LDC is a permanent data archive that has in place infrastructures and processes for reviewing, storing, distributing and preserving resources over the long-term. All data sets deposited at the Consortium are accessible and discoverable through optimized search capabilities. Data storage and delivery infrastructure are developed and enhanced to reflect best practices for digital repositories.

Here are some additional points to consider:

- LDC has over 30 years of experience in publishing data sets and negotiating complex user license agreements with data providers on behalf of the user community.
- LDC does not insist on exclusive distribution rights so you can make your data available through additional channels.
- LDC's newsletter announces new publications and is circulated to more than 8,000 readers.
- Since its founding in 1992, LDC has distributed more than 200,000 copies of corpora to over 6,000 organizations in 100 countries.
- LDC's Catalog contains over 900 titles in more than 90 languages and holdings increase annually.



Corpus Integrity and Structure

LDC is committed to publishing high quality, broadly accessible resources. Because each resource is used by researchers working on a variety of platforms, LDC seeks to standardize data submissions to the extent possible.

Preferred Submission Formats

Audio

- FLAC, MS WAV, MP3, RIFF

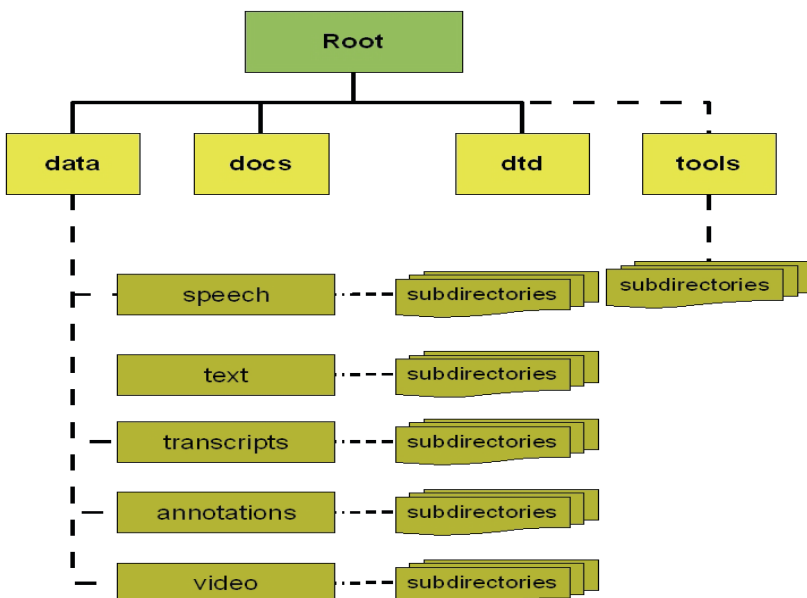
Video

- AVI, MP4

Text

- UTF-8 encoding
- XML markup with DTD for validation

A consistent corpus directory structure promotes easy reference by data users. Although data providers have leeway to organize their data as they see fit, LDC prefers the directory hierarchy below for published corpora:



Need to Know

To submit a corpus for publication, register for an LDC Submissions account: <https://submissions ldc.upenn.edu>, create and edit your submission and provide a data sample. The following is among the required information:

- Title of publication including version number if applicable
- Contributing author(s)
- A list of language and dialect names in the corpus using the ISO 639-3 standard
- Data type (speech, text, video, lexicon, other)
- Estimated delivery date
- Size of publication: estimated hours of speech, words of text or tokens as appropriate. Data size should be provided in bytes
- Format: audio, video, text and markup schemes
 - Audio: sample rate, format and size, compression type and recording environment
 - Video: frame rate and size and aspect ratio as determined by the data stream
 - Text: encoding and markup
- Origin and genre of source data, collection and annotation methodologies, data ownership and copyright information
- A readme document that describes the publication, its applications and other relevant information, aimed at a generic user who does not have the specialized knowledge of the corpus developer/author

LDC Submissions provides infrastructure and resources for describing and uploading data sets, communicating with LDC's publications team, and submitting release-ready corpora. LDC staff performs quality reviews of formats, directory structure and documentation as part of the submissions review process to assess file integrity and data quality.

LDC will assist you through each step from the initial inquiry to publication so that the data will be complete, error free and ready to use.