OPERATIONS

Share Data through LDC

Data Sharing and LDC's Mission

The mission of the Linguistic Data Consortium (LDC) 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 academic, governmental and private organizations.

These communities require data across languages, genres and formats. Thus, another key aspect of LDC's mission is to expand and diversify its Catalog of resources. 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 monthly 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 (HLT) and related fields.

Data Sources

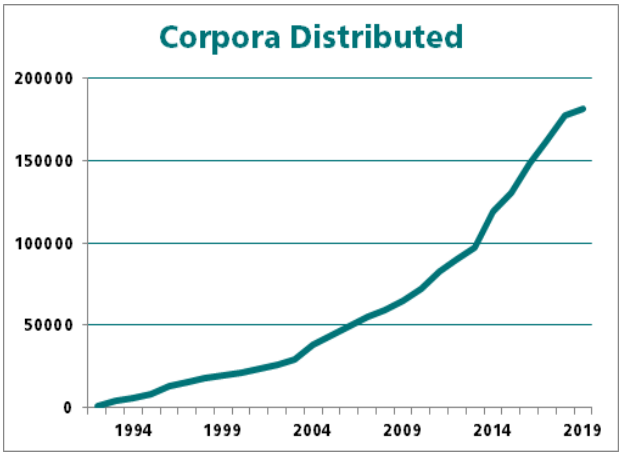
LDC's Catalog contains data developed or contributed by information publishers, researchers, universities and private organizations, including data sets deposited under data management plans. These multilingual language resources run the gamut - text, speech, video and lexicons - and are collected and 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 and distributing 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 25 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 monthly newsletter announces new publications and is circulated to approximately 8,000 readers.
- Since 1992, LDC has distributed more than 175,000 copies of corpora to over 6,000 organizations in 100 countries.
- LDC's Catalog contains over 800 titles in more than 90 languages and holdings increase monthly (36 titles per year).



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

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

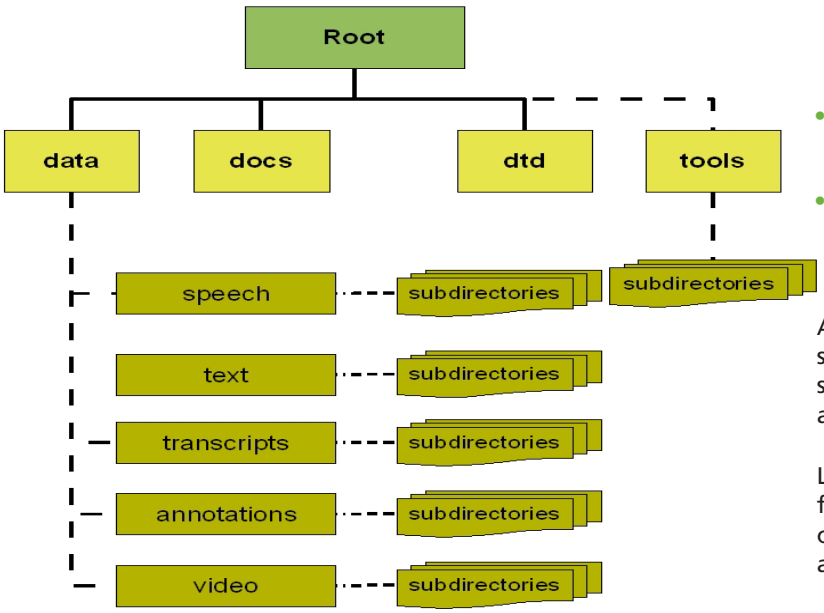
Video

- File containers: avi, dv, mpeg-ps (vob), mpeg-ts, mov, mp4
- Codecs: mpeg1, mpeg2, mpeg4, DivX, Mjpeg, Dv, Xvid

Text

- UTF-8 or UTF-16 encoding
- XML, SGML markup

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, complete the LDC submissions form at https://www.ldc.upenn.edu/data-management/providing/submission and provide a data sample. The following information is required:

- Primary contact person: name, email address and telephone number
- 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: a rough estimate for hours of speech, words of text or tokens as appropriate. At a minimum, provide the data size 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 according to NTSC or PAL standards
  - Text: encoding and markup
- Origin and genre of data used, collection methodologies and copyright information
- A brief narrative description or abstract detailing the nature of the publication, its applications and other relevant information

All publications should be "camera ready." LDC staff performs quality reviews of formats, directory structure and documentation to assess file integrity and data quality.

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