



OPERATIONS

Share Data through LDC

Share Your Data through LDC

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 requires data across languages, genres and formats. Therefore, a second aspect of LDC's mission is to expand and diversify its Catalog of resources. LDC encourages you to share your data and help broaden the network of available resources.

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 greater exposure to major organizations working in human language technologies (HLT) and related fields.

Data Sources

LDC's Catalog contains data from information publishers, researchers, universities and private organizations. These multilingual language resources run the gamut – written, spoken and signed languages encoded in text, speech and video – and are collected and annotated under a multitude of conditions.

Speech data may include materials from interviews and meetings and from broadcast programming and telephone conversations. Text materials come from many sources, including transcripts, newswires, websites, books and periodicals.

LDC also publishes lexicons and dictionaries in common and less commonly resourced languages.

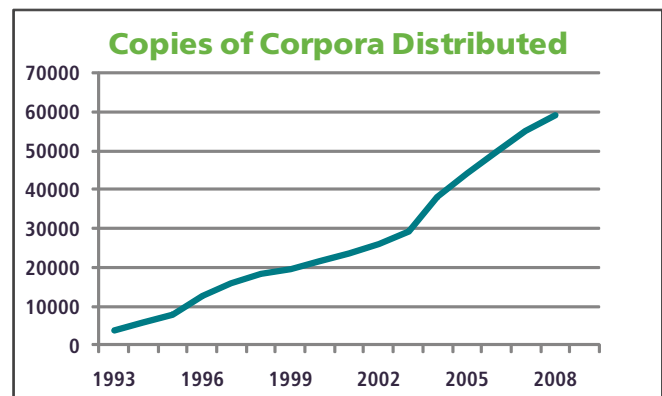
Why Should You Share Data through LDC?

LDC is a recognized leader in language resource creation and distribution. Project sponsors rely on LDC data to support research and technology development in cutting-edge applications. Data providers benefit from having their materials used in research programs because their work is further tested for its validity and utility.

LDC is a data archive. Some information providers do not archive their materials or else store them in an inadequate format. In such cases, LDC becomes the best, or only, source for the data. LDC provides free copies of digital data archives to data providers.

Here are some additional points to consider:

- LDC has 17 years of experience in publishing data sets and negotiating complex user license agreements with data providers on behalf of the user community.
- LDC has distributed nearly 65,000 corpora since its founding in 1992.
- LDC's monthly newsletter announces new publications and is circulated to approximately 8,000 readers.
- LDC's Catalog has over 500 titles and new titles are added monthly (30-36 titles per year).
- LDC corpora are published in over 60 languages and new languages are added to the Catalog every year.



Publication Variety

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

Preferred Submission Formats

Audio

- NIST SPHERE, WAV/RIFF, AIFF

Video

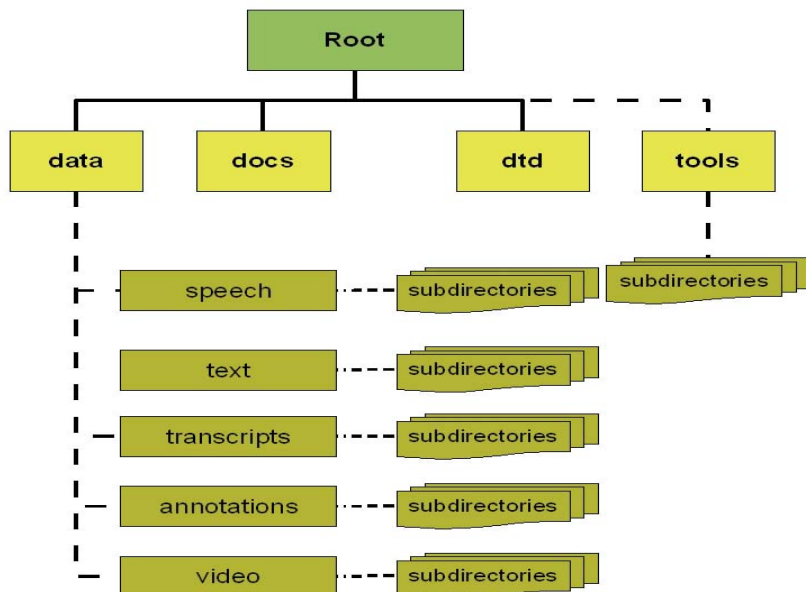
- File containers: avi, dv, mpeg-ps (vob), mpeg-ts, mov, mp4
- Codecs: mpeg1, mpeg2, mpeg4, DivX, Mjpeg, Dv, Xvid

Text

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

Corpus Integrity and Structure

LDC seeks to standardize the corpus directory structure to promote 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, write to ldc@ldc.upenn.edu with the following information:

- Corpus title: should be descriptive and inform potential users about the publication
- Primary contact person: name, email address and telephone number
- Contributing author(s)
- Data type: speech, text, video, lexicon, other
- Size of publication: provide estimates in bytes and hours of speech or video, words or tokens of text as appropriate
- Publication format: audio, video and text formats 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
- Data origin and genre, including copyright information
- Name of the project (if any) for which the corpus was developed
- Language: language and dialect names in ISO 639-3 standard
- Corpus description: nature of the publication, intended application(s) and other relevant information (2-3 paragraphs)
- Data sample

LDC will assist you through each step of the process, from the initial inquiry to the final publication.