

# Solve for X(ML): Transforming metadata to transform data access

John Huck (Metadata Librarian)  
University of Alberta Libraries

Presentation at NADDI 2019, April 25, Ottawa

# A story...

...about a project to improve  
user access to linguistics  
datasets at U of A Libraries

1 - Grappling with the status quo

2 - A simplified access model

3 - Metadata as the key

---

# 1 - Grappling with the status quo

# Linguistic Data Consortium (LDC)

- Based at the University of Pennsylvania
- Publishes 40-50 linguistic datasets (corpora) a year
- Has released close to 800 since 1993
- Text, sound, and video data types
- Data created by different research programs and initiatives
- All corpora are released on DVDs or hard drives
- Most corpora can be downloaded from the **LDC Catalog**



# Example of a corpora

- BOLT Arabic Discussion Forums (LDC2018T10)
  - **Description:** "BOLT Arabic Discussion Forums was developed by the Linguistic Data Consortium (LDC) and consists of 813,080 discussion forum threads in Egyptian Arabic harvested from the Internet using a combination of manual and automatic processes."
  - **Creators:** Jennifer Tracey, Haejoong Lee, Stephanie Strassel, Safa Ismael
  - **Program:** The DARPA BOLT (Broad Operational Language Translation) Program developed genre-independent machine translation and information retrieval systems.
  - **Languages:** Egyptian Arabic
  - **Extent:** ~30 GB of zipped HTML and XML files
  - **Released:** March 15, 2018



# LDC business model


- Organizations purchase annual memberships
- Members gain perpetual access to corpora released that year
- Members grant download privileges to affiliated individuals
- Corpora can also be purchased individually
- A longstanding membership gives U of A Libraries access to about 570 corpora



# User access in 2017 at U of A

- Downloads from LDC Catalog (primary mode of access)
- Discs catalogued, with in-person access through Data Team
- A common arrangement among academic libraries with LDC subscriptions

*However...*

- This meant searching in two places
  - Larger corpora were not available for download
  - The cataloguing workflow had been stopped for two years
- 

# Initial goals for improving access

*Could we...*

- Catalogue the download access?
- Host the larger corpora locally for download?
- Restart the cataloguing workflow?



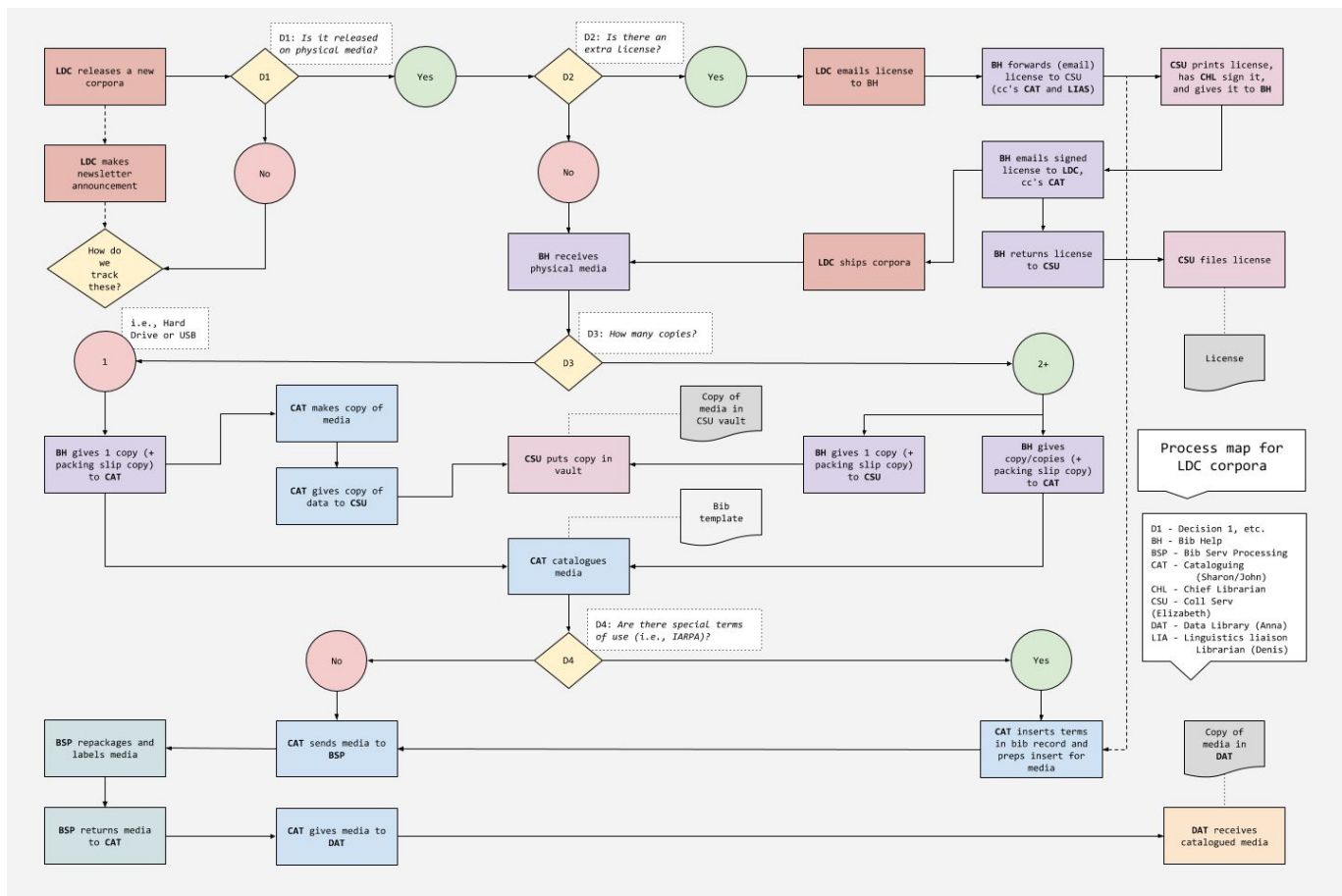


# Obstacles

- Cataloguing download access was not practical
  - Landing pages did not lead directly to the download
  - Users needed to create account and have it approved
- Local hosting was not feasible
  - Large corpora exceeded Dataverse file-size cap
- Could we justify cataloguing only the discs?
  - Non-preferred format
  - Were we misleading users?
  - Many people involved in library process



# Library processing for discs



# Re-evaluating the LDC Catalog

- Holdings information not clearly indicated
- No obvious path to download from landing page
- Difficult to write instructions for users



# LDC Catalog



- ABOUT
- MEMBERS
- COMMUNICATIONS
- LANGUAGE RESOURCES
- Data
- Obtaining Data
- Catalog
- By Year
- Top Ten Corpora
- Projects
- Search
- Memberships
- LDC Online
- Data Scholarships
- Tools
- Papers
- LIR Wiki
- DATA MANAGEMENT
- COLLABORATIONS

Home > Language Resources > Data

My Account Logout Bin: (Empty)

## 2015-2016 CoNLL Shared Task

**Item Name:** 2015-2016 CoNLL Shared Task  
**Author(s):** Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol T. Rutherford, Bonnie Webber, Chuan Wang, Hong Min Wang, Rashmi Prasad

**LDC Catalog No.:** LDC2017T13  
**ISBN:** 1-58563-812-9  
**ISLRN:** 228-559-981-287-1  
**Release Date:** September 14, 2017  
**Member Year(s):** 2017  
**DCMI Type(s):** Text  
**Data Source(s):** newswire  
**Project(s):** CoNLL  
**Application(s):** discourse parsing  
**Language(s):** English, Chinese, Mandarin Chinese  
**Language ID(s):** eng, zho, cmn  
**License(s):** LDC User Agreement for Non-Members

**Online Documentation:** LDC2017T13 Documents

**Licensing Instructions:** Subscription & Standard Members, and Non-Members

**Citation:** Xue, Nianwen, et al. 2015-2016 CoNLL Shared Task LDC2017T13. Web Download. Philadelphia: Linguistic Data Consortium, 2017.

### Introduction

2015-2016 CoNLL Shared Task, LDC Catalog Number LDC2017T13 and ISBN 1-58563-812-9, contains the Chinese and English training, development and test data for the 2015 and 2016 CoNLL (Conference on Computational Natural Language Learning) Shared Task Evaluation which focused on shallow discourse parsing.

The Conference on Computational Natural Language Learning (CoNLL) is accompanied every year by a shared task intended to promote natural language processing applications and evaluate them in a standard setting. Shallow discourse parsing is the task of parsing a piece of text into a set of discourse relations between two adjacent or non-adjacent discourse units. This task is called shallow discourse parsing because the relations in a text are not connected to one another to form a connected structure in the form of a tree or graph.

LDC has also released the following CoNLL Shared Task data sets:

- 2006 CoNLL Shared Task - Ten Languages (LDC2015T11)
- 2006 CoNLL Shared Task - Arabic & Czech (LDC2015T12)
- 2008 CoNLL Shared Task Data (LDC2010T12)
- 2009 CoNLL Shared Task Part 1 (LDC2012T03)
- 2009 CoNLL Shared Task Part 2 (LDC2012T04)

### Data

This release consists of the tokenized, tagged, and parsed tags in English and Chinese. The English train, dev and test data are from Wall Street Journal material in Penn Discourse Treebank Version 2.0 (LDC2008T05); English blind test data are from kinwines. Chinese train, dev and test data are news material from Chinese Discourse Treebank 0.5 (LDC2014T21); Chinese blind test data are from kinwines.

### Samples

Please view this source sample and annotation sample.

### Updates

None at this time.

### Copyright

Portions © 1987-1999 Dow Jones & Company, Inc., © 1994-1998, 2006, Xinhua News Agency, © 2008, 2012, 2012, 2013, 2014, 2017 Trustees of the University of Pennsylvania

### Available Media

Web Download

### Fee

Extra Copy

Request Data



- ABOUT
- MEMBERS
- COMMUNICATIONS
- LANGUAGE RESOURCES
- Data
- Obtaining Data
- Catalog
- By Year
- Top Ten Corpora
- Projects
- Search
- Memberships
- LDC Online
- Data Scholarships
- Tools
- Papers
- LIR Wiki
- DATA MANAGEMENT
- COLLABORATIONS

Home > Language Resources > Data

My Account Logout Bin: (Empty)

## My Account

### University of Alberta

**My Address: (Edit)**  
John Huck  
Libraries  
5-25E, Cameron Library  
University of Alberta  
Edmonton, Alberta T6G 2J8  
Canada

john.huck@ualberta.ca  
(Edit e-mail or password)

**Organization Contact:**  
Bib Help  
Acquisitions & General  
Collections  
5th Floor, Cameron Library  
Edmonton, T6G 2J8  
Canada  
780 492 7017  
libldc@ualberta.ca

**Account Options**  
Corpora Invoiced  
Agreements  
Downloads  
LDC Online  
Receive Newsletter

### Membership Years

2003 (Not-for-Profit, Standard)	2004 (Not-for-Profit, Subscription)
2006 (Not-for-Profit, Subscription)	2007 (Not-for-Profit, Subscription)
2008 (Not-for-Profit, Subscription)	2009 (Not-for-Profit, Subscription)
2010 (Not-for-Profit, Subscription)	2011 (Not-for-Profit, Subscription)
2012 (Not-for-Profit, Subscription)	2013 (Not-for-Profit, Subscription)
2014 (Not-for-Profit, Subscription)	2015 (Not-for-Profit, Subscription)
2016 (Not-for-Profit, Subscription)	2017 (Not-for-Profit, Subscription)
2018 (Not-for-Profit, Subscription)	Non Membership Years



- ABOUT
- MEMBERS
- COMMUNICATIONS
- LANGUAGE RESOURCES
- Data
- Obtaining Data
- Catalog
- By Year
- Top Ten Corpora
- Projects
- Search
- Memberships
- LDC Online
- Data Scholarships
- Tools
- Papers
- LIR Wiki
- DATA MANAGEMENT
- COLLABORATIONS

Home > Language Resources > Data

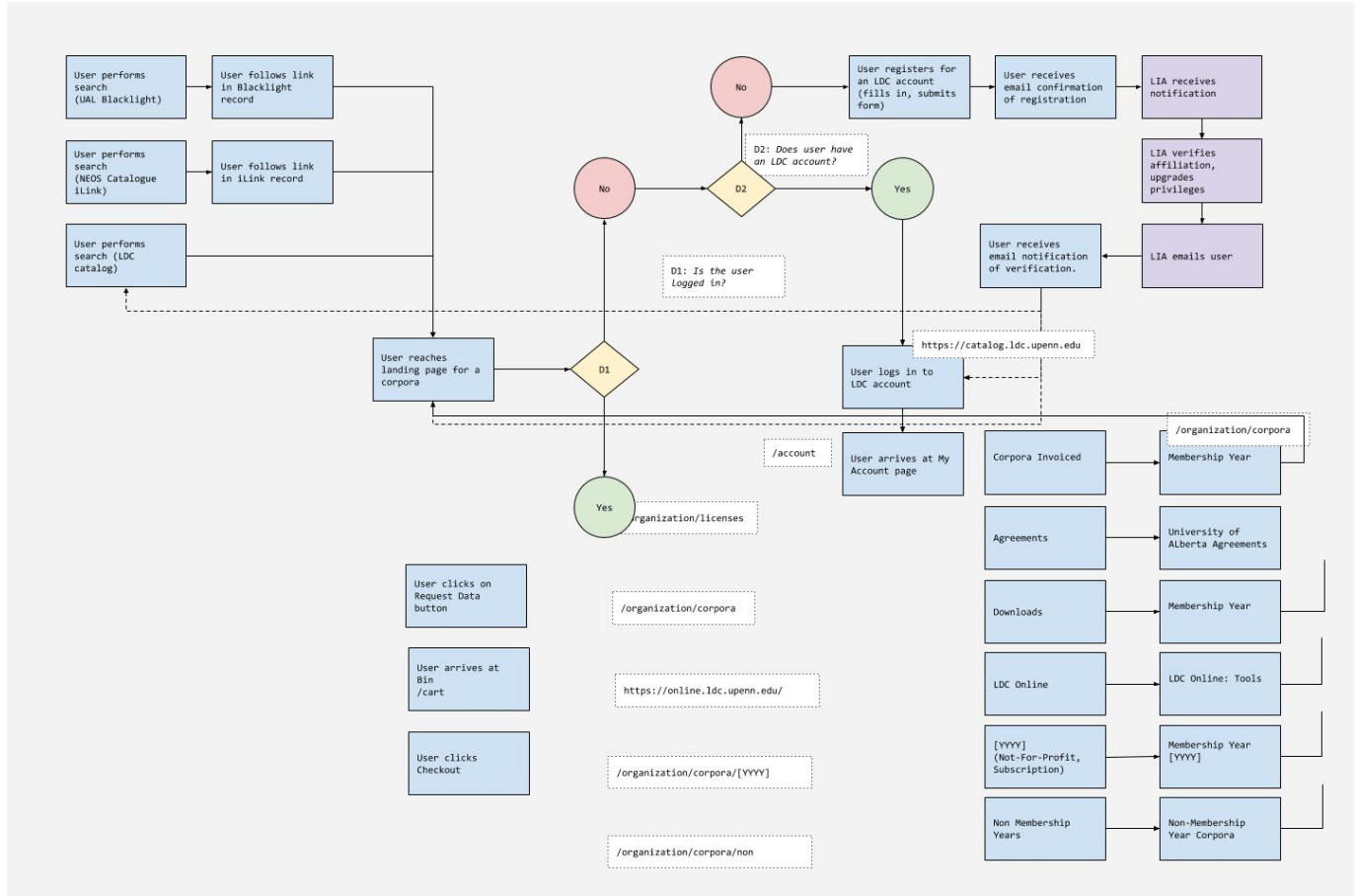
My Account Logout Bin: (Empty)

## Corpora Available for Download

LDC Catalog ID	Corpus Name	Invoice Date	Download	File
LDC2015T07	GALE Phase 3 and 4 Arabic Broadcast News Parallel Text	2018-01-18	(D)	gale_p3_4_arab_bn_parallel_LDC2015T07 File Size: 2.68 MB MD5 Checksum: 0c6552505884f19502502c006e72c206
LDC2015S09	LDC Spoken Language Sampler - Third Release	2018-01-18	(D)	spoken_lang_samp_3 File Size: 87.1 MB MD5 Checksum: 4003388971e71105100716e8d00f
LDC2015S01	GALE Phase 2 Arabic Broadcast News Speech Part 2	2018-01-18	(D)	gale_p2_arab_bn_speech_p2_LDC2015S01 File Size: 3.64 GB MD5 Checksum: f692b330006b0778be239e87935e9
LDC2015S04	Mandarin-English Code-Switching in South-East Asia	2018-01-18	(D)	seasia_LDC2015S04 File Size: 2.84 MB MD5 Checksum: 19950a013958464ef0e1fcd30e483
LDC2018T01	DEFT Spanish Treebank	2018-01-18	(D)	DEFT_Spanish_Treebank_LDC2018T01 File Size: 1.87 MB MD5 Checksum: 8840f0ff00d74070e0e12e4001082
LDC2017T18	GALE Phase 4 Chinese Broadcast News Transcripts	2017-12-15	(D)	gale_p4_chinese_bn_transcripts_LDC2017T18 File Size: 3.37 MB MD5 Checksum: 4356809094e40634e35c104146725
LDC2017S25	GALE Phase 4 Chinese Broadcast News Speech	2017-12-15	(D)	gale_p4_chinese_bn_speech_LDC2017S25 File Size: 3.3 GB MD5 Checksum: 5a611604011704e03053005047e4
LDC2017S21	ASPIRE Development and Development Test Sets	2017-12-08	(D)	ASPIRE_Development_and_Development_Test_Sets_v2.0_LDC2017S21 File Size: 17.8 GB MD5 Checksum: 83a16b0a0e7150276452182373167
LDC2017S22	ASPIRE Babel Kurnia Kurdish Language Pack	2017-12-08	(D)	ASPIRE_Babel_Kurnia_Kurdish_Language_Pack_LDC2017S22 File Size: 1.98 MB MD5 Checksum: 03653771810040c0c01780c304025

# Attempt to map the user path

I gave up ...



# Other concerns

*Users...*

- Could apparently order and pay for data
- Were sometimes compelled to sign electronic agreements
- Could see electronic agreements signed by others



# Conclusions

- LDC Catalog was not appropriate for end-users in current form
- Site architecture was unlikely to be redesigned in near future

*Back to the drawing board!*



## 2 - A simplified access model



# Goals for a user-centred access model

*Make it simple!*

- One place to search everything
- Digital delivery whenever possible
- Don't force users to care about the format

*with some constraints...*

- Couldn't host corpora locally
- Couldn't let users access LDC Catalog




# The project team

*Project had accumulated several stakeholders*

- Bibliographic Services (Cataloguing)
- Data Team (Public Services)
- Linguistics liaison librarian (Public Services)
- Collections Strategies Unit (Acquisitions)



# Factors in our favour

- LDC metadata available from an OAI-PMH service
  - **Metadata Team** - Proof-of-concept transformation to MARC
  - **Data Team** - Already used Google Drive to deliver data
  - **Cataloguing Coordinator** - Willing to load unorthodox records
  - **Collection Services** - Willing to allow "purchase-on-demand" requests
- 

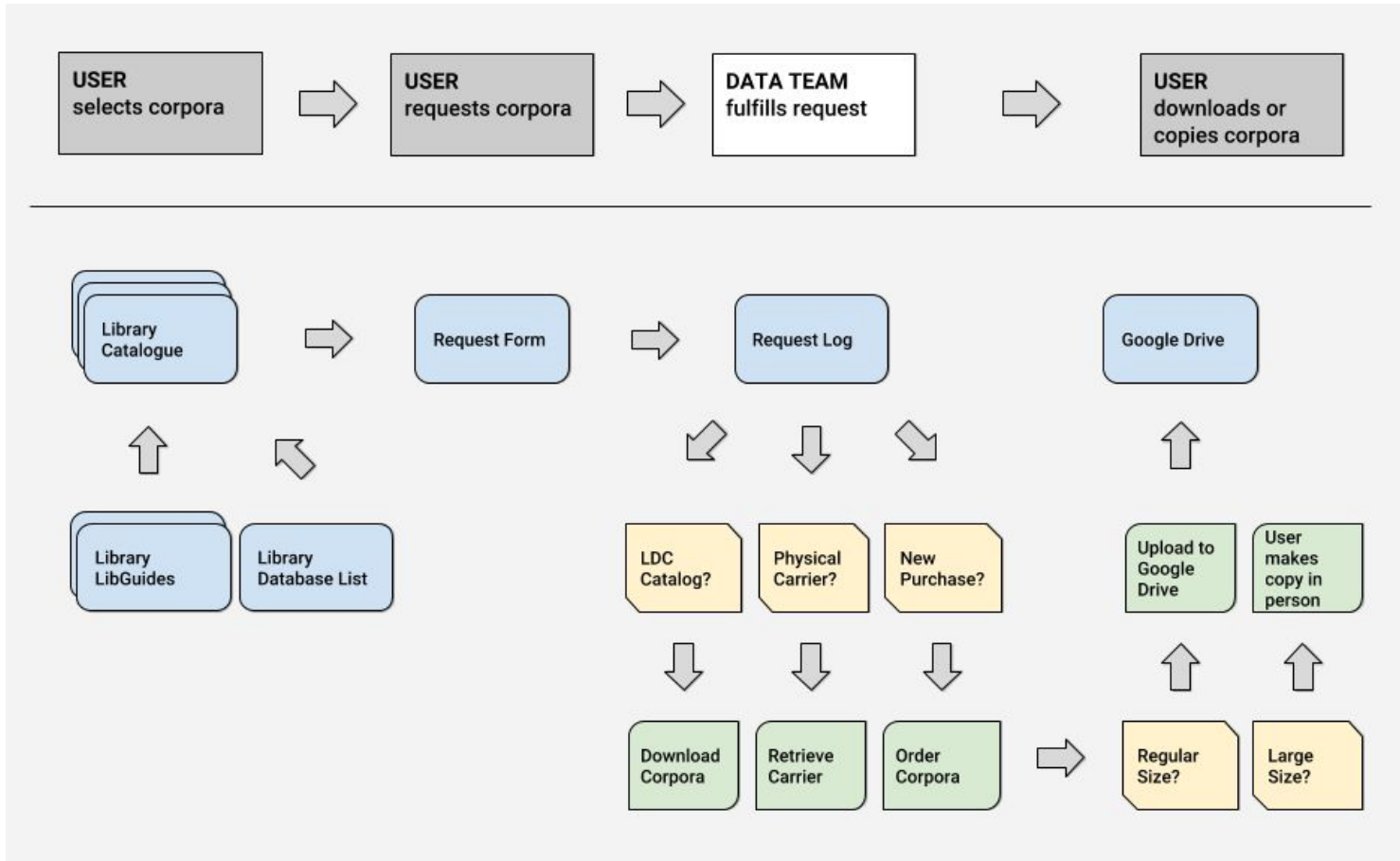
# A new plan for improved access

*We would...*

- Create and load MARC records for all LDC corpora
- Include a link to a request form in each record
- Deliver data to users with Google Drive



# LDC mediated access model



# A simplified model for the user

*Users could...*

- search across all published LDC corpora
- request access to datasets the library didn't own
- would not need to know about available formats
- would only need to know the title or LDC number

## Linguistic Data Consortium (LDC) Corpora Retrieval Request

Use this form to request Linguistic Data Consortium (LDC) linguistic corpora.

Requested items will be made available to you via the University of Alberta Google Share Drive for a fixed period of time so you may make your own copy. Google Drive shared items will be permitted to your UA CCID. You will receive an e-mail with instructions when your requested item/s are ready for access.

This form is monitored M-F, 8:30-4:30

Protection of Privacy - The information requested on this form is collected under the authority of Section 33 (c) of the Alberta Freedom of Information and Protection of Privacy Act and will be protected under Part 2 of that Act.

Your email address ([jhuck@ualberta.ca](mailto:jhuck@ualberta.ca)) will be recorded when you submit this form. Not you? [Switch account](#)

**\*Required**

What is the LDC# and/or title of the LDC corpus you wish to request? (E.g.: LDC2012T02 or "English translation treebank") \*

Your answer

Any comments or further information?

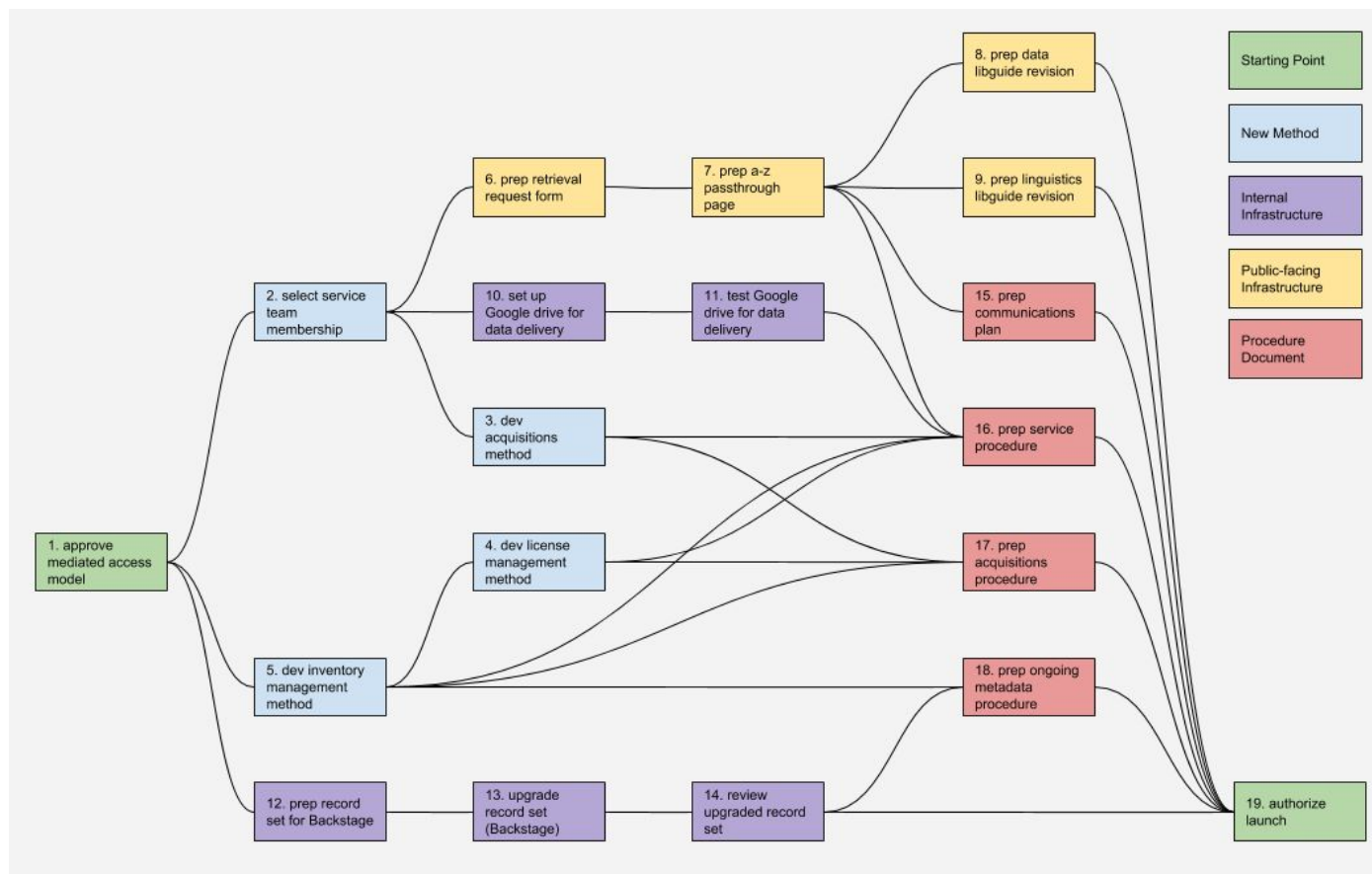
Your answer

**SUBMIT**

Never submit passwords through Google Forms.

*Request Form*

# Planning for project launch



# 3 - Metadata as the key



# Characteristics of the LDC metadata

- **High-quality**

```
<olac:olac xmlns:olac="http://www.language-archives.org/OLAC/1.1/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <dc:contributor>Du Bois, John W.</dc:contributor>
  <dc:contributor>Chafe, Wallace L.</dc:contributor>
  <dc:contributor>Meyer, Charles</dc:contributor>
  <dc:contributor>Thompson, Sandra A.</dc:contributor>
  <dc:date xsi:type="dcterms:W3CDTF">2000</dc:date>
  <dcterms:issued xsi:type="dcterms:W3CDTF"
    >2000-01-01</dcterms:issued>
  <dc:description>*Introduction* The Santa Barbara Corpus of
    Spoken American English is based on hundreds of
    recordings of natural speech from all over the
    United States, representing a wide variety of
```

*OLAC Metadata*

# Characteristics of the LDC metadata

- High-quality
- **Consistent**

```
<xsl:for-each select="dc:publisher[contains(., 'Linguistic')]":  
  <marc:datafield tag="264" ind1=" " ind2="1">  
    <marc:subfield code="a">  
      <xsl:text>[Philadelphia, Pennsylvania]: </xsl:text>  
    </marc:subfield>  
    <marc:subfield code="b">  
      <xsl:text>Linguistic Data Consortium, </xsl:text>  
    </marc:subfield>  
    <marc:subfield code="c">  
      <xsl:text>[</xsl:text>  
      <xsl:value-of select="$year"/>  
      <xsl:text>]</xsl:text>  
    </marc:subfield>  
  </marc:datafield>  
</xsl:for-each>
```

*XSLT Stylesheet*



# Characteristics of the LDC metadata

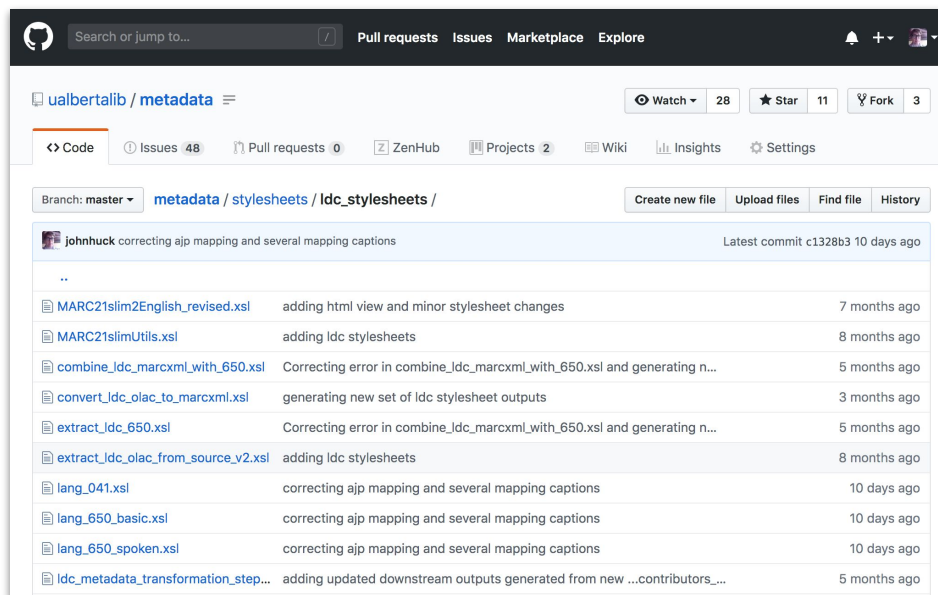
- High-quality
- Consistent
- **Comprehensive**

```
LEADER 03337nmm a22005173i 4500
001 8587995
006 m o u
007 cu |||||u|||
008 190313s2000 pau u eng d
020 a| 1585631647
020 a| 9781585631643
024 8 a| LDC2000S85
024 8 a| 4077318196684 q| ISLRN
035 a| on1090038764
039 a| exclude
040 a| AEU b| eng e| rda c| AEU d| AEU
042 a| dc
043 a| n-us---
050 4 a| PE2808.8 b|.S26 2000
090 a| Internet Access b| AEU
245 0 0 a| Santa Barbara Corpus of Spoken American English Part I.
264 1 a| [Philadelphia, Pennsylvania] : b| Linguistic Data Consortium, c| [2000]
300 a| 1 online resource.
336 a| computer dataset b| cod 2| rdacontent
```

*MARC record*

# Characteristics of the LDC metadata

- High-quality
- Consistent
- Comprehensive
- **Current**



The screenshot shows the GitHub interface for the repository 'ualbertalib / metadata'. The page displays the commit history for the 'master' branch, specifically for the subdirectory 'metadata / stylesheets / ldc\_stylesheets /'. The table lists recent commits with their descriptions and timestamps.

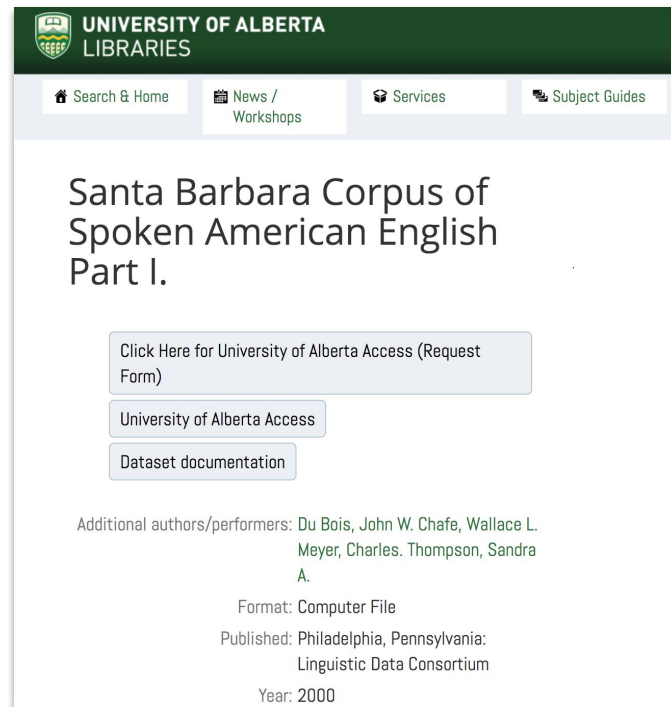
Commit	Description	Time
johnhuck	correcting ajp mapping and several mapping captions	Latest commit c1328b3 10 days ago
MARC21slim2English_revised.xml	adding html view and minor stylesheet changes	7 months ago
MARC21slimUtils.xml	adding ldc stylesheets	8 months ago
combine_ldc_marxml_with_650.xml	Correcting error in combine_ldc_marxml_with_650.xml and generating n...	5 months ago
convert_ldc_olac_to_marxml.xml	generating new set of ldc stylesheet outputs	3 months ago
extract_ldc_650.xml	Correcting error in combine_ldc_marxml_with_650.xml and generating n...	5 months ago
extract_ldc_olac_from_source_v2.xml	adding ldc stylesheets	8 months ago
lang_041.xml	correcting ajp mapping and several mapping captions	10 days ago
lang_650_basic.xml	correcting ajp mapping and several mapping captions	10 days ago
lang_650_spoken.xml	correcting ajp mapping and several mapping captions	10 days ago
ldc_metadata_transformation_step...	adding updated downstream outputs generated from new ...contributors_...	5 months ago

GitHub repo

# Unlocking a solution through metadata

## *Metadata...*

- transcended format
- independent information space
- smooth layer over contingent processes



The screenshot shows the University of Alberta Libraries website. The header includes the university logo and name. Below the header is a navigation bar with links for Search & Home, News / Workshops, Services, and Subject Guides. The main content area displays the title "Santa Barbara Corpus of Spoken American English Part I." and three buttons: "Click Here for University of Alberta Access (Request Form)", "University of Alberta Access", and "Dataset documentation". Below the buttons, there is a list of additional authors/performers: Du Bois, John W. Chafe, Wallace L. Meyer, Charles. Thompson, Sandra A. The format is listed as "Computer File", published by "Linguistic Data Consortium" in Philadelphia, Pennsylvania, in the year 2000.

UNIVERSITY OF ALBERTA  
LIBRARIES

Search & Home News / Workshops Services Subject Guides

Santa Barbara Corpus of Spoken American English Part I.

Click Here for University of Alberta Access (Request Form)

University of Alberta Access

Dataset documentation

Additional authors/performers: Du Bois, John W. Chafe, Wallace L. Meyer, Charles. Thompson, Sandra A.

Format: Computer File

Published: Philadelphia, Pennsylvania: Linguistic Data Consortium

Year: 2000

*Discovery system*

# Thank you.



john.huck@ualberta.ca

## References

<https://bit.ly/2UprJdy>

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

---