

Linguistic Data on Demand

John Huck & Anna Bombak
University of Alberta Libraries

Presentation at NEOS Miniconference, June 7, 2019

A project...

...to improve access to
linguistics datasets
in the U of A Libraries' collection

- 1 - Struggling with the status quo
- 2 - Simplifying & updating our access model
- 3 - “Data delivery” services at UAL





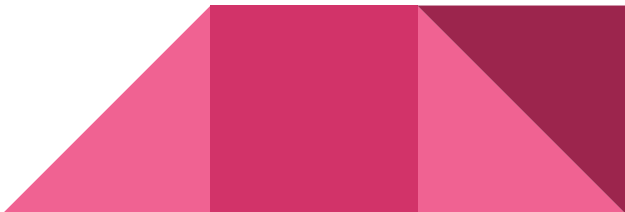
1 - Struggling with the status quo

Linguistic Data Consortium (LDC)

- Based at the University of Pennsylvania
- Publishes linguistic corpora (datasets)
- 40-50 new corpora each year
- 800 releases since 1993
- Data types include text, sound and video
- Datasets created by various researchers and initiatives



Example of an LDC corpora

- **Title:** BOLT Arabic Discussion Forums (LDC2018T10)
 - **Description:** ...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
 - **File types:** HTML and XML files
 - **Size:** 30 GB of zipped files
 - **Released:** March 15, 2018
- 

LDC business model


- All corpora released on DVDs or hard drives
- Most corpora can be downloaded (**LDC Catalog**)
- Annual Membership for organizations
- Members:
 - receive hard copies
 - retain perpetual download access
 - manage download privileges for their users



UAL access in 2017

- 570 corpora in U of A holdings
- Download access through **LDC Catalog**
- Discs catalogued – in-person access through Data Team
- A common arrangement for libraries with LDC subscriptions

But that meant...

- Two places to search
 - No downloads for large corpora
 - Two-year cataloguing backlog
- 

Project goals

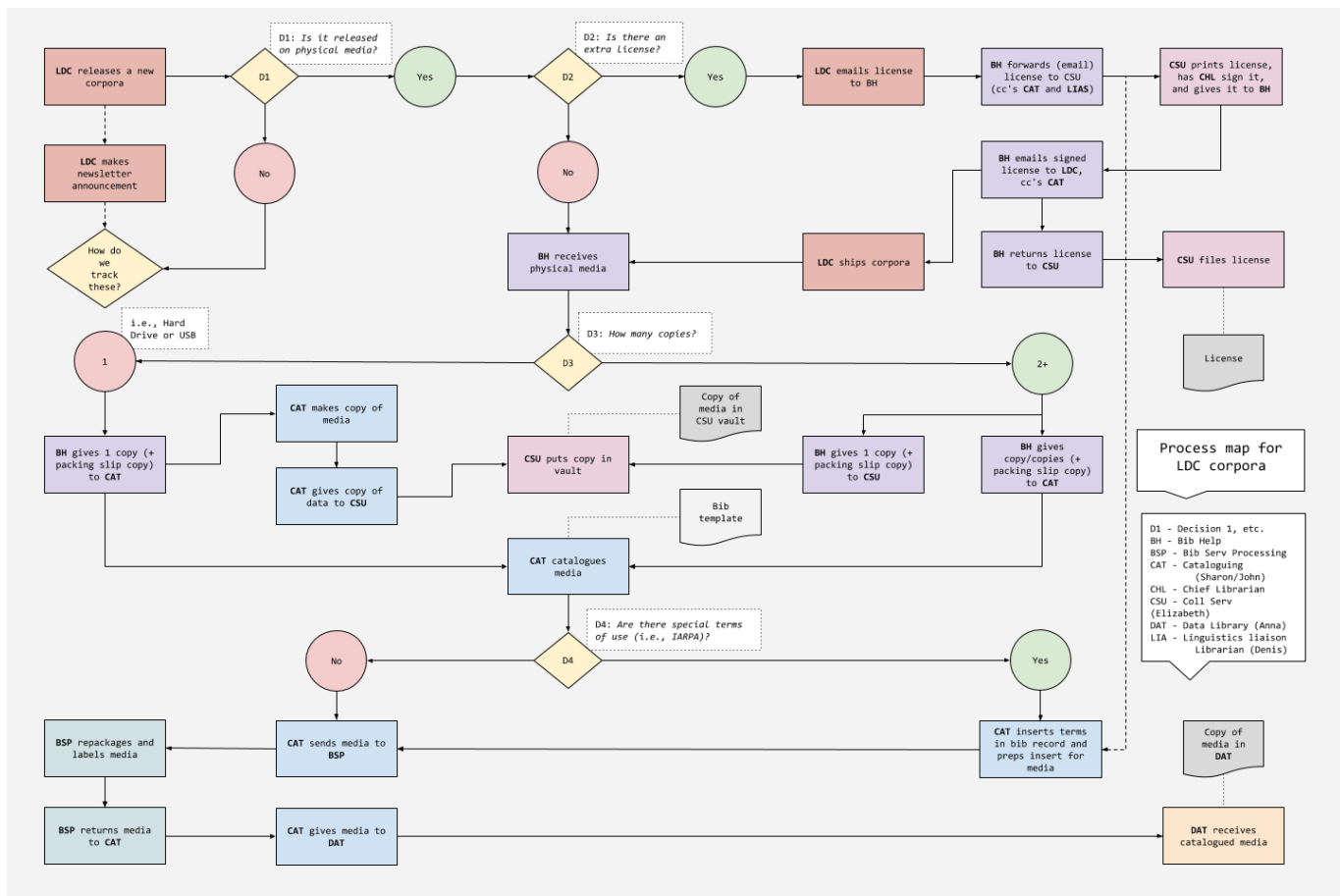
- Restart cataloguing workflow
- Catalogue download access
- Host large corpora locally for download

Obstacles

- Local hosting not possible
- Couldn't easily link to download
- Could we justify continuing to cataloguing discs?



Library processing for discs



Evaluating *LDC Catalog*

Users...

- could order and pay for data
- forced to sign agreements
- could see other agreements and names of other users

Conclusion

- *LDC Catalog* not appropriate for end-users in current form

Linguistic Data Consortium | UNIVERSITY OF PENNSYLVANIA | CONTACT US

LDC Linguistic Data Consortium

Home > Language Resources > Data

2015-2016 CoNLL Shared Task

Item Name: 2015-2016 CoNLL Shared Task

Author(s): Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Atappol T. Rutherford, Bonnie Webber, Chuan Wang, Hong Min Wang, Rashmi Prasad

LDC Catalog No.: LDC2017T13

ISBN: 1-58563-812-9

ISLRN: ID Number

Release Date:

Member Year(s):

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 (LDC2009T12)
- 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 wikinews. Chinese train, dev and test data are news material from Chinese Discourse Treebank 0.5 (LDC2014T21), Chinese blind test data are from wikinews.

Samples

Please view this source sample and annotation sample.

Updates

None at this time.

Copyright

Portions © 1987-1989 Dow Jones & Company, Inc., © 1994-1998, 2006, Xinhua News Agency, © 2008, 2012, The Penn Discourse Treebank Group, © 2016 Atappol Rutherford, © 2001, 2004, 2005, 2007, 2008, 2009, 2010, 2012, 2013, 2014, 2017 Trustees of the University of Pennsylvania

Available Media Fee

Web Download Extra Copy

Request Data

My Account Logout Bin: (Empty)

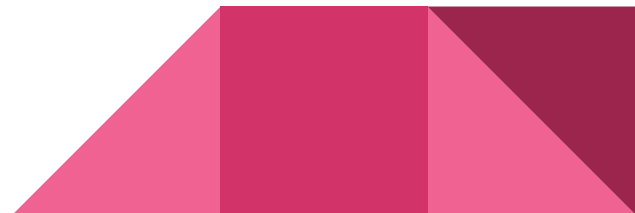
1992-2018 Linguistic Data Consortium, The Trustees of the University of Pennsylvania. All Rights Reserved.

2 - Simplifying the access model

New project goals

Make it simple!

- One place to search
- Digital delivery
- Reduce complexity
- Eliminate the need to use ***LDC Catalog***



Stakeholders

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



Piecing it together

1. LDC metadata in XML
2. Script to generate MARC
3. Data Team's service model
4. Openness to unorthodox records

```
<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
```

LDC Metadata

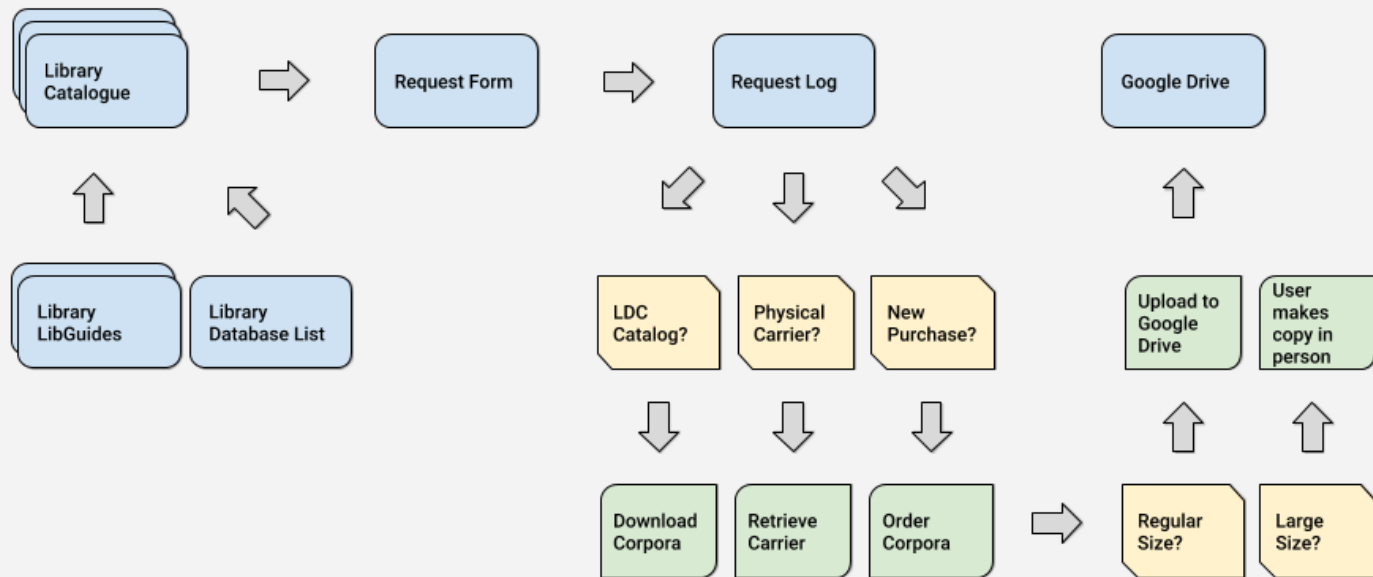
Plan for a new access model

- Generate MARC records from metadata
- Add a link to request form
- Deliver data to users with Google Drive

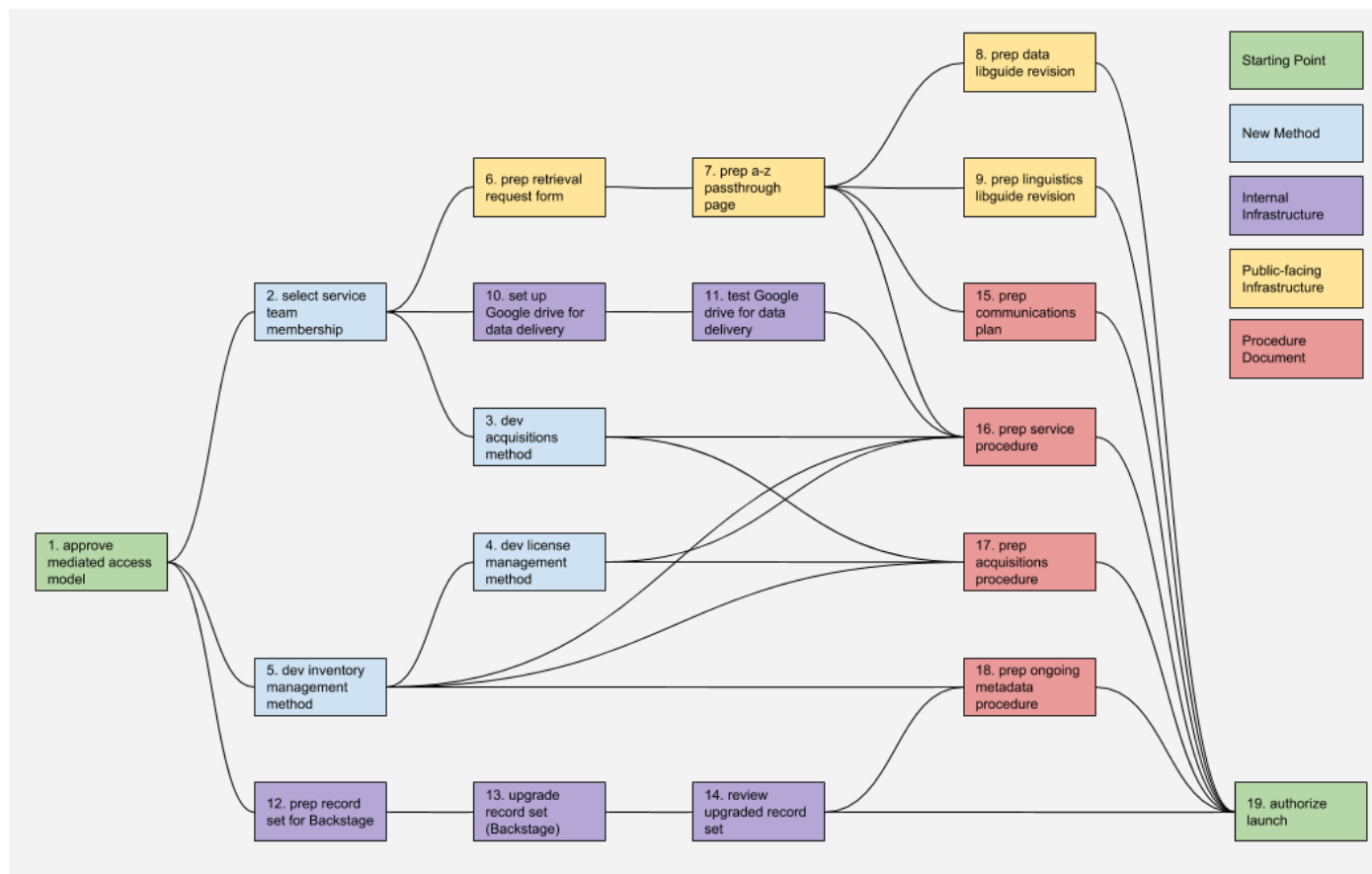
```
LEADER 03337nmm a22005173i 4500
001 8587995
006 m o u
007 cu |||||u|||
008 190313s2000 pau u eng d
020 a| 1585631647
020 a| 9781585631643
024 8 a| LDC2000S85
024 8 a| 4077318196684 q| ISLRN
035 a| on1090038764
039 a| exclude
040 a| AEU b| eng e| rda c| AEU d| AEU
042 a| dc
043 a| n-us---
050 4 a| 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

Mediated access model




Launched in May, 2019



3 - “Data delivery” services at UAL

Sample LDC Catalogue Record

 UNIVERSITY OF ALBERTA
LIBRARIES

[Search & Home](#) [News / Workshops](#) [Services](#) [Subject Guides](#) [Research Support](#) [My Account](#) [ask us](#)

1 of 875 | [Next »](#) [Back to Search Results](#) [Start Over](#)

LDC Spoken Language Sampler

[Click Here for University of Alberta Access \(Request Form\)](#) ←

Additional authors/performers: Castelletto, Anthony.
Format: Computer File
Published: Philadelphia, Pennsylvania: Linguistic Data Consortium
Year: 2008
Physical Details: 1 online resource
ISBN: 1585634956, 9781585634958
General Note: LDC number: LDC2008S08.
Summary: The Linguistic Data Consortium (LDC) at the University of Pennsylvania distributes a wide and growing assortment of resources for researchers, engineers and educators whose work is concerned with human languages. Historically, most linguistic resources were not generally available to interested researchers but were restricted to

Tools

- [Export to Refworks](#)
- [Export to EndNote](#)
- [Librarian View](#)
- [Email Me This Item](#)
- [Text Me A Link To This Item](#)
- [Send Correction](#)
- [Go to Bookmarks](#)
- [View Search History](#)
- [View Citation \(MLA/APA/Chicago\)](#)
- [Bookmark](#)

LDC Request Form

Linguistic Data Consortium (LDC) Corpora Retrieval Request

University of Alberta faculty, students and staff may 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.

* Required

Email address *

Your email

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.

Thank you.



john.huck@ualberta.ca

abombak@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.