

Integrating ELRA/LDC Metadata into OLAC Repository

Andrew W. Cole **Khalid Choukri**

andrew.cole@ldc.upenn.edu choukri@elda.fr

Linguistic Data Consortium
University of Pennsylvania
www.ldc.upenn.edu

ELRA/ELDA
55 Rue Brillat-Savarin
F-75013 Paris, France
http://www.elda.fr

■ OLAC 2002 IRCS UPenn 1

Net-DC

- Net-DC : Networking Data Centers, an initiative funded by NSF and EC to coordinate activities of data providers – specifically LDC and ELRA but in ways that should encourage other centers to join
- Included a task for joint LDC/ELRA dissemination of information on resources being distributed
- LDC/ELRA concluded having NetDC fund the integration of their catalog into OLAC was the best solution.
- In the division of tasking, LDC agreed to write the converters for both LDC and ELRA/ELDA catalogs.

■ OLAC 2002 IRCS UPenn 2

ELRA Architecture

• System Overview

```

graph LR
    A[ELRA Catalog MS/Access Table] --> B[Visual Basic Conversion Module in MS/Access]
    B --> C[ELRA Catalog XML file]
  
```

• Access Table Match

<u>OLAC</u>	<u>OLAC/ELDA</u>
Subject.language	Subject.language
Type	Type
Type.linguistic	Not in ELDA Catalog
Coverage	Not in ELDA Catalog
Date	Not Applicable

■ OLAC 2002 IRCS UPenn 3

LDC/ELRA Code

- Coding/Knowledge Problems
 - Error in ELDA OLAC program, link to online description is between an <identifier> tag, <description> would be better.
- Nonetheless Access System is Simple and Robust
 - MS/Access Visual Basic Module of 280 lines.
 - Single MS/Access Table Converted to Single XML file.

■ OLAC 2002 IRCS UPenn 4

LDC Architecture

• System Overview

```

graph LR
    A[LDC Catalog Oracle Table] --> B[PERL Conversion Script (cron)]
    B --> C[LDC Catalog XML file]
  
```

• Oracle Table Problems

ldc_catalog_id:	LDC94817
name:	OGI Multilanguage Corpus
language:	English, Farsi, French, German, Hindi, Japanese, Korean, Chinese, Spanish, Tamil, Vietnamese,

■ OLAC 2002 IRCS UPenn 5

LDC PERL Coding

- Coding/Knowledge Problems

```

<record spec="lexicon">
<header>
  <recordId>olac:ldc:LDC94L2</recordId><datestamp>2002-10-16</datestamp>
</header>
<metadata>
<olac>
<identifier>LDC94L2</identifier>
<title>*COMLEX English Syntax Lexicon</title>
<type>lexicon</type>
...
  
```

- Nonetheless System is Simple and Robust
 - PERL Script of 150 lines (lots of comments).
 - Single Oracle Table converted to Single XML file.
 - Repairs taking less than a day, done by non-experts.

■ OLAC 2002 IRCS UPenn 6

• **OLAC Issues.**

- Existing OLAC vocabulary assumes that linguistic data is for traditional linguistic research (ie. linguistic field) and that language technology developers are only interested in software not data.
- Difficult to determine/find the correct or applicable type and vocabulary from OLAC web site with unknowledgeable staff (eg., me, Andy).
- Need OLAC vocabularies encode information about pricing.
- Providers ramp-up to full meta-data compliance.

• **Web Links**

- **ELDA ECI** <http://www.elda.fr/cata/text/W0004.html>
- **LDC ECI** <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94T5>
- **OLAC LDC/ECI** <http://saussure.linguistlist.org/cfdocs/new-website/LL-WorkingDirs/olac/olac-search3.cfm?id=112715>
» (Bulgarian)
- **OLAC ELDA/ECI** <http://saussure.linguistlist.org/cfdocs/new-website/LL-WorkingDirs/olac/olac-search3.cfm?id=58000>
» (Turkish)