Language Resources
By the Other Data Center … over 15 years fruitful partnership

Khalid Choukri
ELRA

55 Rue Brillat-Savarin, 75013 Paris, France
Tel. +33 1 43 13 33 33 Fax. +33 1 43 13 33 30

choukri@elda.org
www.elda.org/ or http://www.elra.info/
Outline … Past – Present – Future of LRs

- The ELRA foundation, Mission, and activities
- ELRA Catalogues

THE META-SHARE INITIATIVE
- New distributed networked LR repository
- Identification, cataloguing, licensing, sharing, information dissemination

- Unique Identifier for each LR
  - International Standard LR Number

- Future … and what's next …
Past ….. Inefficient Chaos
Some knew where to shop … Many did not
Highly centralized (Data Centers)
... Past – Present – Future of LRs

199x ...1995  2000  2010  2012  2020

Past .............................. Present .............................. future(s)

Inefficient Chaos ...................... Centralized data centers ....... ????
PAST .... Few words about ELRA
ELRA’s Foundation & Mission
An Improved infrastructure for Data sharing & HLT evaluation

• Created in February 1995, initial funding from the European Commission
• Main rationale: bring into focus the need for a mutual exchange and use of LRs
• An (not for profit) Association of Users of Language Resources
• A Repository Center that takes care of all issues related to LRs from Discovery to sharing and distribution, for R&D and Commercial purposes
• LRs Broker /middle man
• Infrastructure for the evaluation of Human Language Technologies
• Information dissemination, i.e. LRECs,
• Operational body: ELDA
ELRA’s Foundation & Mission (as a Repository Center)
An Improved infrastructure for Data sharing & HLT evaluation

- Discovery and Identification of LRs (LR repositories: ELRA Catalogue, Universal Catalogue, LRE Map, etc.)
- Distribution and sharing of LRs
- Identifying of gaps in the LR landscape (Surveys, Requests, Scientific publications analysis, etc.)
- Filling in of gaps through LR Production, Packaging, Repurposing, Production or commissioning the production of LRs
- Assistance for the analysis of sustainability (see our new analytical model, within FlareNet)
- Assistance with legal issues (IPR clearing, licensing, help desk for members, etc.)
- Information dissemination on LR & Evaluation, i.e. LRECs, Portal for HLT Evaluation: http://www.hlt-evaluation.org
ELRA’s Foundation & Mission
An Improved infrastructure for Data sharing & HLT evaluation

- Infrastructure for the evaluation of Human Language Technologies
  - providing resources, tools, methodologies, logistics,
  - Online evaluation platforms
  - Exit strategies / Capitalization on evaluation packages
ELRA offer

Speech LRs
- Automatic Speech Recognition
- Speech Translation
- Therapy and Speech
- Broadcast News Transcription
- Acoustic Person Tracking
- Acoustic Speaker Identification
- Speech Activity Detection

Written LRs
- Lexica
  - Spell Checkers
  - Information Retrieval
  - Document Indexing
  - Information retrieval
  - Question Answering
  - Machine Translation
  - Terminology Extraction
  - Parsing
  - Automatic Summarization
- Corpora
- Terminology
  - Database Creation
  - Lexicon Consolidation
  - Translation Memory Validation
  - Thesaurus
  - Implementation
  - Generation
  - Multilingual Text Alignment

Multimodal Multimedia LRs
- Face tracking
- Speech verification
- Audio / Multimodal person identification
- Multimodal person tracking
- 2D head orientation
- Audiovisual Speech Recognition

Evaluation Packages / Evaluation Solutions / Evaluation Services

>>>> LRs & Evaluation Packages
ELRA Activities.... Identification, Distribution, Production of LRs and HLT Evaluation & Dissemination

Diagram:
- **Provider**
- **ELRA**
- **User**

**Partners**
- **Identification**
- **Negotiation for the Distribution**
- **Royalties reporting**

**Dissemination**:
- Catalogues
- Mailing lists
- ELRA Newsletter
- LREC Conference
- HLT Evaluation Portal

**Distribution**:
- Pre-sales services
- Sales management
- Post sales services

**Validation**
**Production**
**Evaluation**
Identification and distribution of LRs

  - About 30% are/will be free of charge


• LRE-MAP (4000+ LRs & LT) [http://www.resourcebook.eu/](http://www.resourcebook.eu/)

• Types of LRs: All modalities associated (or not) with language (images, videos, sign, OCR, …)
LRE-MAP as of mid-2012

- The concept: collect and compile data from all submissions to LREC,
- Extend it to other conferences dealing with LRs and LTs
- Associate LRs and Publications
- Key figures: over 4000+ identified item

- Gaps identified through the LRE-Map Matrices
- (Retrieve data from past publications and associate LR & publications)
<table>
<thead>
<tr>
<th>Written Data (Ranked)</th>
<th>Bulgarian</th>
<th>Czech</th>
<th>Danish</th>
<th>Dutch</th>
<th>English</th>
<th>Estonian</th>
<th>Finnish</th>
<th>French</th>
<th>German</th>
<th>Hungarian</th>
<th>Irish</th>
<th>Italian</th>
<th>Lithuanian</th>
<th>Latvian</th>
<th>Malay</th>
<th>Polish</th>
<th>Portuguese</th>
<th>Romanian</th>
<th>Slovak</th>
<th>Slovene</th>
<th>Spanish</th>
<th>Swedish</th>
<th>Other European Languages</th>
<th>Multilingual</th>
<th>L1</th>
<th>N/A</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>7</td>
<td>12</td>
<td>6</td>
<td>17</td>
<td>206</td>
<td>3</td>
<td>3</td>
<td>43</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>32</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>12</td>
<td>2</td>
<td>5</td>
<td>29</td>
<td>19</td>
<td>18</td>
<td>15</td>
<td>39</td>
<td>36</td>
<td>9</td>
</tr>
<tr>
<td>Lexicon</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>8</td>
<td>77</td>
<td>1</td>
<td>2</td>
<td>24</td>
<td>15</td>
<td>3</td>
<td>4</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>1</td>
<td>19</td>
<td>4</td>
<td>11</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Ontology</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Grammar/Language Model</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Terminology</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>A syntactic judgments database</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Resource: morphology</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Thesaurus</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A Knowledge Base with Lexical-Semantic Relations between words</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A list of categories with examples of language use</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Controlled Legal Language</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Corpus-Based Online Dictionary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Database</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Encyclopedic knowledge</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Event Semantics</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lexicon/corpus</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Online Encyclopedia</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Repository of bilingual lexicons</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Resources integration</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Text Book</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Virtual Game World</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yahoo!'s local listings in Chicago</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Encyclopedia</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Part-of-Speech Tagset</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Psycholinguistic Database</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>23</td>
<td>11</td>
<td>30</td>
<td>327</td>
<td>5</td>
<td>6</td>
<td>80</td>
<td>70</td>
<td>15</td>
<td>14</td>
<td>1</td>
<td>56</td>
<td>10</td>
<td>7</td>
<td>23</td>
<td>17</td>
<td>21</td>
<td>3</td>
<td>61</td>
<td>25</td>
<td>39</td>
<td>30</td>
<td>11</td>
<td>36</td>
<td>5</td>
<td>9</td>
</tr>
</tbody>
</table>

E.3.a. Written Language Data (Ranked order)
Production of LRs ……… Over 100 different resources

• Production or commissioning the production of LRs
  – Within the framework of European and international projects:
  – In support of companies or institutions (may be with confidentiality statements)

• Production Services (in addition to work for HLT evaluations):
  ▪ ELRA has already compiled LRs in more than 25 languages
  ▪ high quality LRs + strict validation
  ▪ involved in every stage of production (from Specs to Distribution)
  ▪ covering different types of LRs and for different technologies
  ▪ some recent achievements: LILA Hindi and Korean databases, Turkish Corpus, Kazak, OrienTel Arabic(s), Broadcast News Speech Corpus for Arabic, French, Spanish, Telephony databases (several thousands of speakers), Aligned textual corpora for SMT (several languages), video annotations with audio transcriptions, SMS; OCR Docs, etc.
Production of LRs .......... cost-effectiveness ...

• Cost of Production vs Cost of Repurposing /Repackaging
  – About -70% compared to a decade ago
  – Producing vs Buying
  – Collaborative work (see LREC language Library, MT-abstract, etc.)
  – New efficient methods .... e.g. Crowdsourcing

• Crowdsourcing platforms (which ones suitable?)
  ➢ See and contribute to the ELRA Crowdsourcing survey
  ➢ What are the issues :
    ➢ ethics, direct access to turkers, fair compensations, extend beyond USD, Roupies
Present – Future of LRs

199x ...1995  2000  2010  2012  2020

Past ........................................ Present .............................. future(s)
ELRA in the open data era, new strategy for next (decade)

- More Open and Free resources to share
- More accurate and detailed Descriptions to represent data (Meta-data), common & interoperable data descriptions
- More on line, easy to get resources
- More Connected, distributed, networked set of Repositories of LR & T

► More Sharing and openness of LRs, more META-SHARE
THE NEW EUROPEAN HLT INITIATIVE FOR LANGUAGE RESOURCES

A distributed network of Repositories & Data centers
META-SHARE: what it is

- META-SHARE is an **open, integrated, secure, and interoperable** exchange infrastructure for language data and tools for the Human Language Technologies domain.

- A **marketplace** where language data and tools are documented, uploaded and stored in repositories, catalogued and announced, downloaded, exchanged, discussed, aiming to **support a data economy** (free and for-a-fee LRs/LTs and services).

- It brings together several organisations and initiatives.
META-SHARE aim:

- bring together knowledge about LRs and related objects and processes and foster their use

How?

- by providing easy, uniform, one-step access to LRs through the aggregation of LR sources into one catalogue
- by facilitating the LRs search and retrieval processes
- by facilitating the evaluation of LRs through comparison between similar LRs
- by encouraging (re-)use and new use of LRs through the monitoring of actual LRs use

- Adoption of a distributed network of Repositories
- Adoption of a uniform metadata schema
- Adoption of a common and trustable legal framework
Notes:
Harvesting: Metadata Harvesting (OAI-PMH)
META-SHARE Inventory: Every META-SHARE inventory will contain a copy of all the harvested metadata across core and peripheral/local repositories, the statistics database, and a web server.
P2P Network: An interconnected network of the core WP8 partners’ inventories. It will assure synchronisation between core inventories.
Web GUI: A portal which will handle requests (search, browse, view results) and distribute them according to traffic criteria (load-balancer). The user will be transparently served by one of the core inventories.
Architecture

META-SHARE portal
Registration – authentication - authorisation

Search / browse  licence  download  statistics
mappings  reporting  recommenders  Billing / payment

META-SHARE inventory
META-SHARE inventory
META-SHARE inventory

metadata harvesting

External repos

Inventory LR repo
Inventory LR repo
Inventory LR repo
Inventory LR repo
Achievements

- META-SHARE Networked Repositories but ONE catalogue
  - and associated software management (open source software)
    - http://www.meta-share.org/

- Local versus non-local Repositories

- META-SHARE Common Licenses

- Common rich Meta-data for all types of LRs with conversion/import tools
  - (e.g. ELRA, OLAC, Other sets)

- Help Desk (Legal, Technical, meta-data issues)

- download free/for-a-fee resources (soon to be open to all)
Identification of LRs

Now that my LRs are described in several catalogues and places:

• How can I identify a LR?

• How can I refer to a LR?
Actual LR Identifiers

Translanguage English Database (TED):

ELRA-S0031

LDC2002S04

unique way?
### Usage of some major data centers

<table>
<thead>
<tr>
<th></th>
<th>ELRA</th>
<th>LDC</th>
<th>NICT</th>
<th>GSK</th>
<th>BAS</th>
<th>Chinese LDC</th>
<th>HLT-Centrale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publisher</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Category</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Digit ID</td>
<td>X (4)</td>
<td>X (2)</td>
<td>X (6)</td>
<td></td>
<td>X</td>
<td>X (3)</td>
<td></td>
</tr>
<tr>
<td>Letter ID</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free ID</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Software</td>
<td>ELRA-S0035</td>
<td>LDC2004L01</td>
<td>G-00035</td>
<td>GSK2010-C</td>
<td>SC10</td>
<td>CLDC-SPC-2007-002</td>
<td>CORN</td>
</tr>
<tr>
<td>Example</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unique Identifiers ...Purpose

• Every object in the world requires a kind of identification to be correctly recognized.
• For example, traditional printed materials like a *book*:
  • International Standard Book Number (ISBN),
  • Library of Congress Control Number (LCCN),
  • Digital Object Identifier (DOI),
  • Amazon Standard Identification Number (ASIN) and
  • several other numeric identifiers as unique identification scheme.
• The main goal is to get a unique way for naming a resource through the several LR distribution institutions.
  – Even for LRs which may not have the referable web site.
PID

- Persistence
- Uniqueness
- Identification
- Discovery

metadata
- pid
- language
- distributor
- license
- ...

New LR id
Highlight of the ISLRN

ISLRN: XXX-XXX-XXX-XXX-X

Random +Check-sum 13 digits

e.g.       ISLRN: 193-187-955-056-0
Translanguage English Database (TED):

ELRA-S0031

LDC2002S04

ISLRN: 001-920-233-002-2
Scholarly societies behind ISLRN

- All international key players (LR production, distribution, sharing, archiving, etc.) should be involved

- Should have: Legitimacy & Trustability

- Involvement in NLP & HLT essential

- Example of key players
  - ELRA, LDC
  - ACL
  - EAMT/IAMT
  - ISCA
  - AFNLP, ALAGIN, GSK
  - Oriental-Cocosda
Infrastructure for the ISLRN Assignment

- Agree on a naming protocol and assignment procedures.
  - Define an efficient and adequate Infrastructure (Robust Server, mirrored, etc.).
  - Agree on a validation process and moderating (accept/reject a query for ISLRN).
- Interoperability with key Catalogue Portals
- Implementation to start by .... 2012
a Proposal the ISLRN management structure

Steering Committee

Executive Committee

ISLRN Service
ISLRN attribution approach during a first phase

LR Provider

Ask for ISLRN

ISLRN Interface

ISLRN server

ISLRN server

ISLRN server

Unknown LR

Set up metadata by LR Provider

Moderator(s)

ISLRN database

Back Office
ISLRN attribution approach during a 2d phase
Remarks / Conclusions

• ISLRN:
  – a Unique Identifier that allows to name and discover LRs
  – ISLRN is not about access, not about rights, not about archiving.

• META-SHARE framework
  – Networked repositories
  – Software to set-up a node within the network or independent
  – Meta-data
  – Licensing schema and licenses
Past ..... Inefficient Chaos
Some knew where to shop … Many did not
Highly centralized (Data Centers)
Efficient … distributed …. Networked Centers /Chaos
... Past – Present – Future of LRs

199x ...1995 2000 2010 2012 2020

Past ......................... Present .................... future(s)

Inefficient Chaos ......................... Centralized data centers ......... Distributed/
(Efficient) Chaos (lack of coordination)
Thank you for your attention