ELRA Activities for the Distribution of Language Resources

Valérie Mapelli

ELRA/ELDA
55-57 Rue Brillat-Savarin, F-75013 Paris, France
Tel. +33 1 43 13 33 33 -- Fax. +33 1 43 13 33 30
Email: mapelli@elda.org
http://www.elra.info/ or http://www.elda.org/

With the support of:

FLaReNet

Overview

- Before ELRA was established
  - … once upon a time
  - rationale behind its foundation and its mission(s)
- ELRA Distribution Activities:
  - Legal issues / Licensing
  - Identification and Distribution
- Other activities related to LRs:
  - Production of LRs
  - Evaluation of Human Language Technologies
  - Dissemination

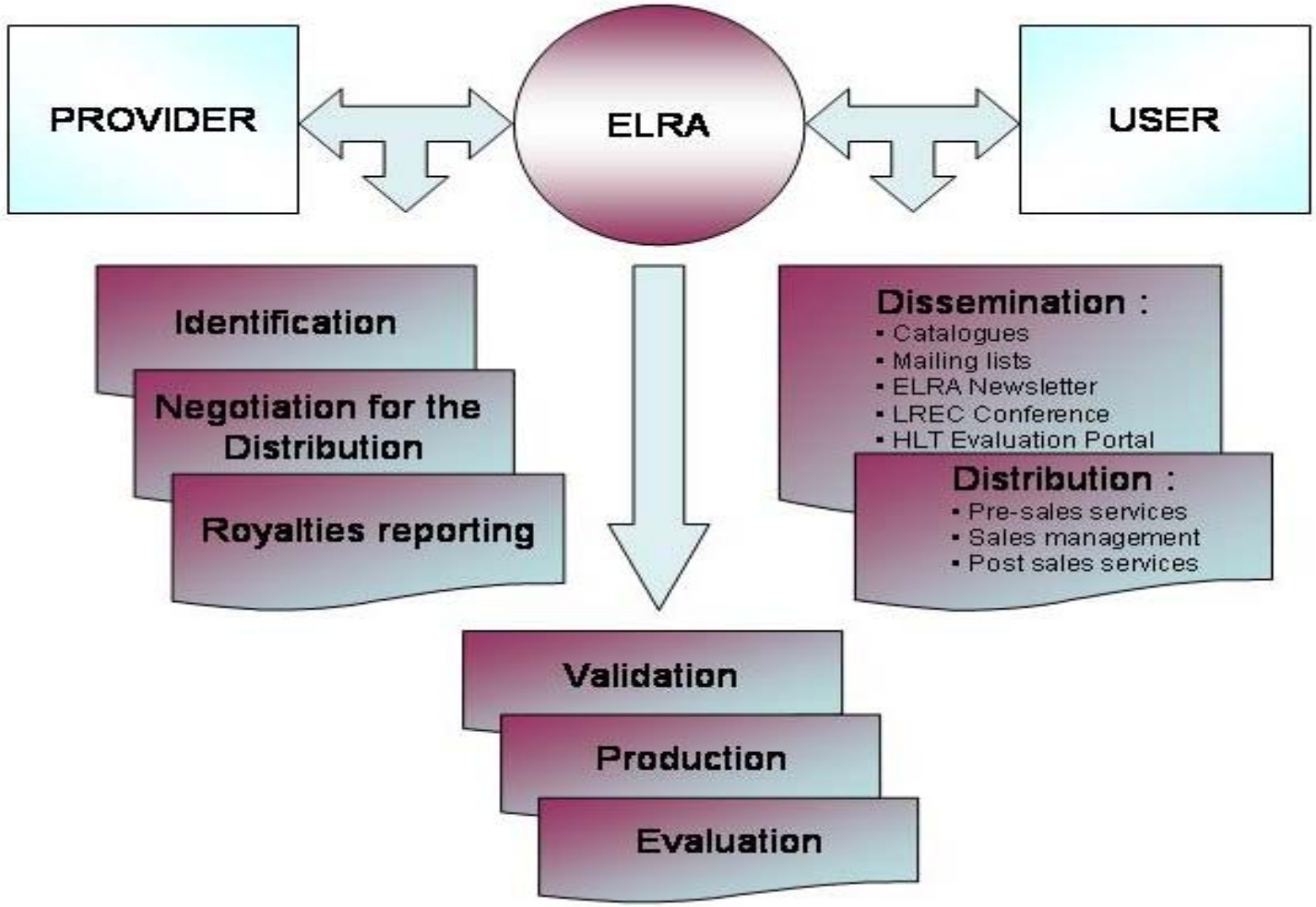
ELRA's Foundation & Mission

- Created in February 1995
- Funding from the European Commission: 3 years
- 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:
  - **Technical & Logistic issues**
  - **Commercial issues (prices, fees, royalties)**
  - **Legal issues (Licensing, IPR)**
  - **Information Dissemination**
- Infrastructure for the evaluation of Human Language Technologies providing resources, tools, methodologies, logistics, Exit strategies / Capitalization on evaluation packages
- Operational body: ELDA

The Association

- **Steering Committee (Board) from profit & non-profit/academic sectors**
- **Self-sustainable (longetivity)**
- **Membership Drive:**
  - **ELRA is Open to European & Non-European Institutions**
  - Resources are available to Members & Non-Members
- **Pay per Resource…. Many are free for R&D**
- **Some of the benefits of becoming a member:**
  - **Substantial discounts on LR prices (over 70%),**
  - **Substantial discounts on LREC registration fees**
  - **Legal and contractual assistance with respect to LR matters**
  - **Access to Validation and production manuals (Quality assessment)**
  - **Figures and facts about the Market (results of ELRA surveys)**
  - **Newsletter and other publications inc. JLREC ←**
  - **Since 2005 Fidelity program … earn miles and get more benefits**

Activities overview

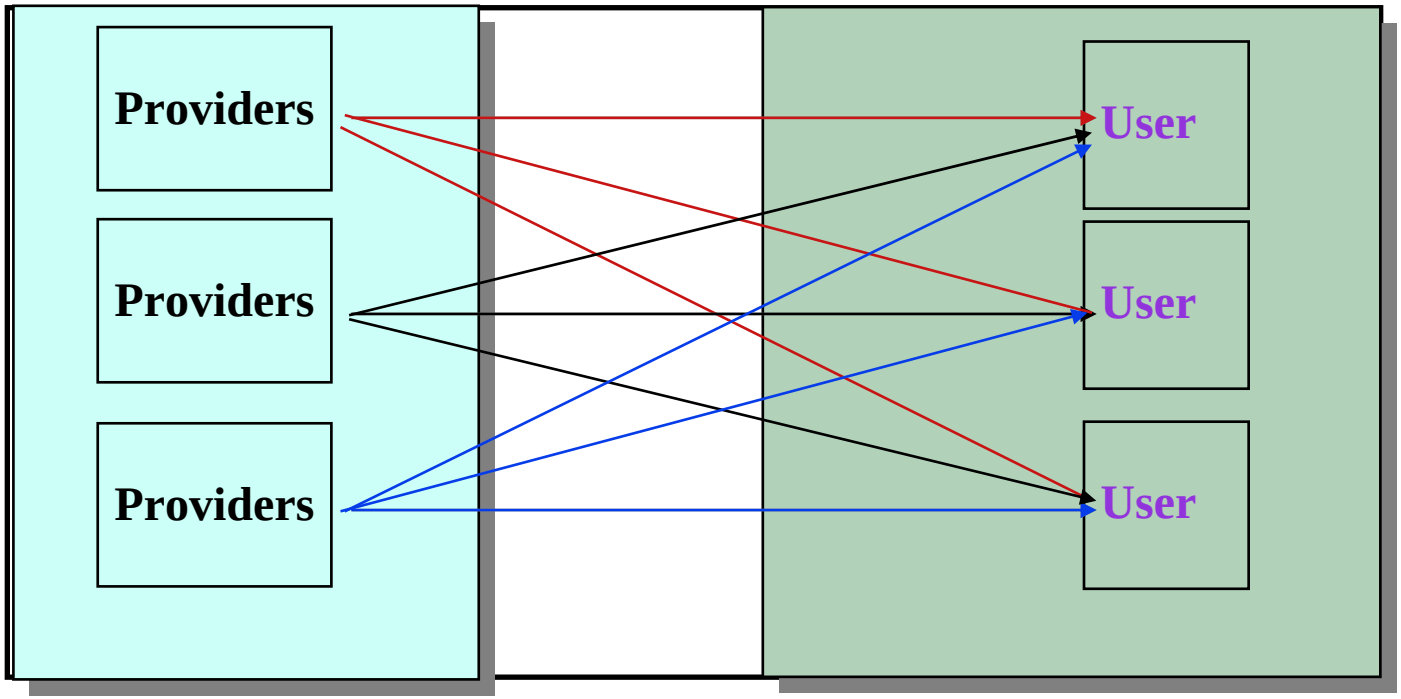


Legal Issues – LR Features

- Legal issues :
  - IPR issues: owner, copyright holder, author…
  - Institutional users: academic organisations and commercial organisations
  - Usage: R&D, development of technology, (re)distribution, distribution of derived products, evaluation of technology
- LR distribution policy:
  - "Open source" (vs APIs)
  - Pricing (free, at cost, market price...)

Legal Issues - Licensing

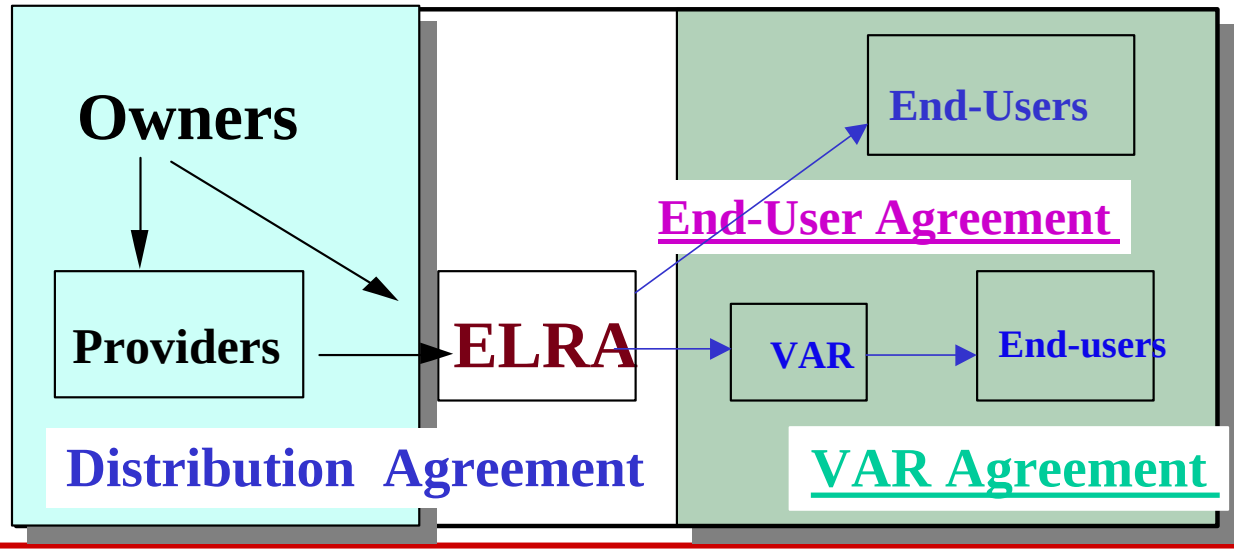
Simplify relationship between providers and users -> draft of generic contracts



Provider-User Agreements

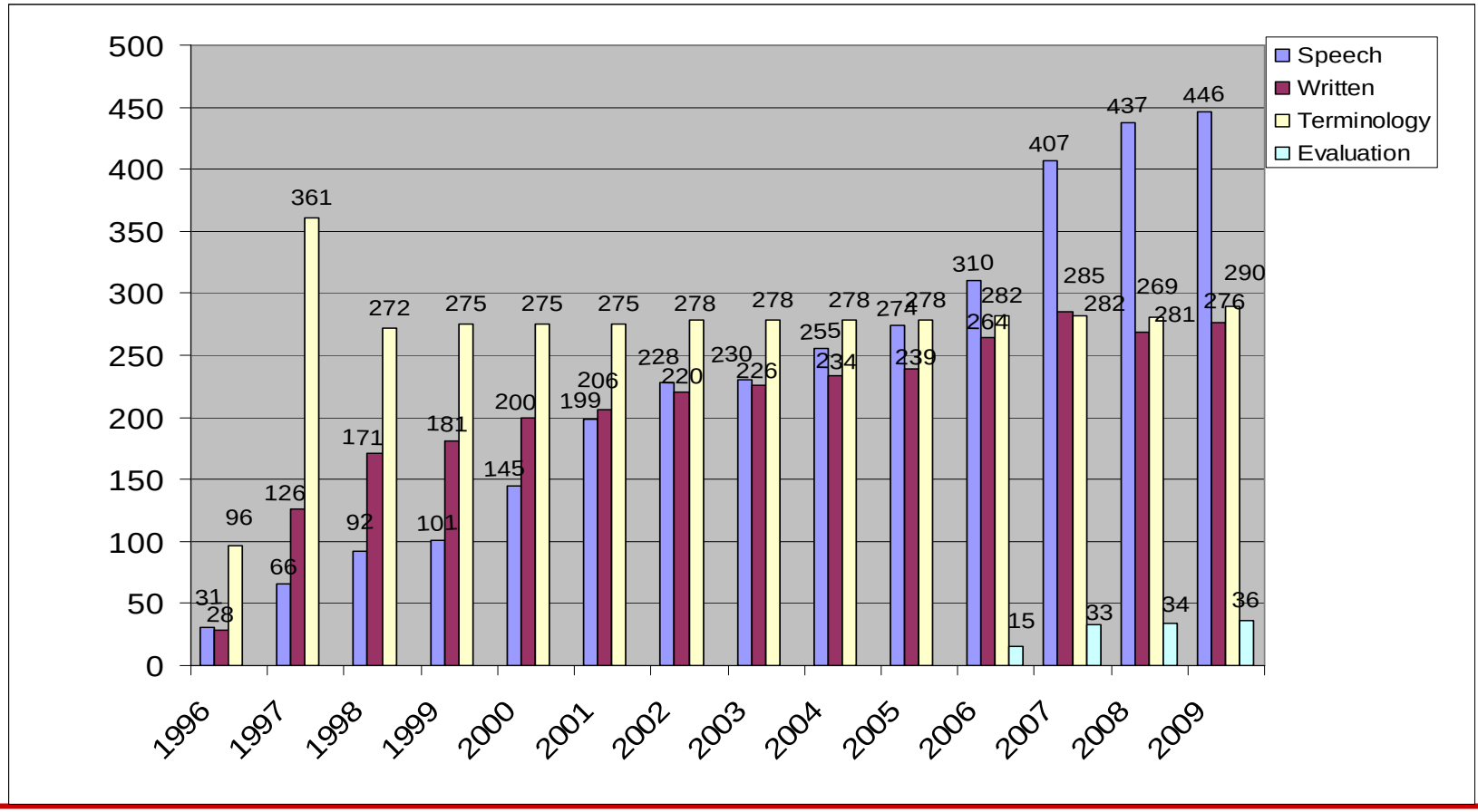
Legal Issues - Licensing

- Drafted Contracts :
  - support of lawyers for basic principles on LRs licensing
  - establish usage: research / technology development
  - protect data owners and their LRs
  - available on www.elda.org/article1.html
  - designed before CC licenses: future mergings or joint designing?
  - ELRA licenses (© ELRA) used more than 5000 times inc. customized ones
- Contract Model:

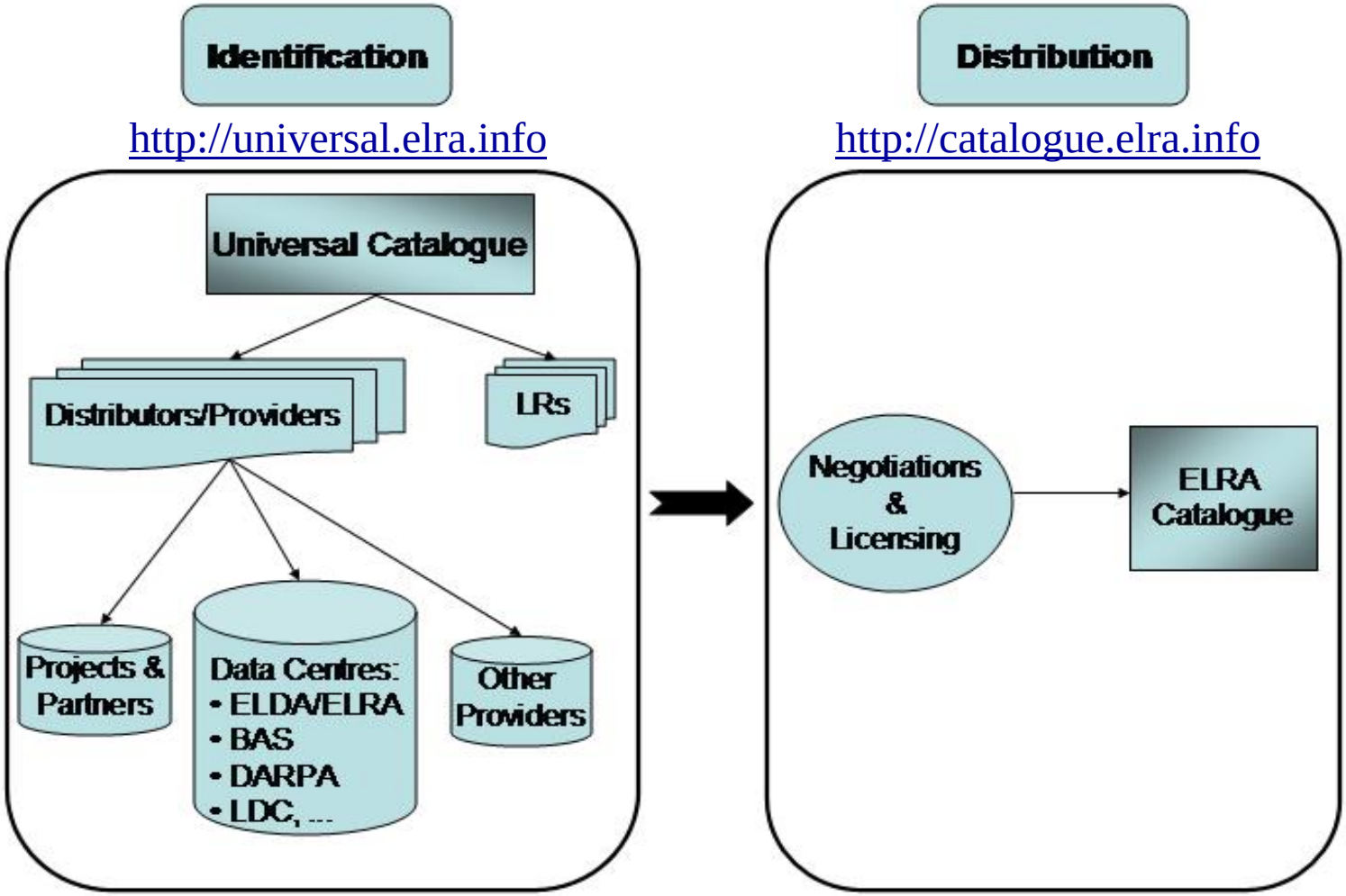


Identification and Cataloguing

- More than 1,000 LRs catalogued and available: ELRA Catalogue of Language Resources: http://catalog.elra.info



The Universal Catalogue & the ELRA Catalogue



The Universal Catalogue & the ELRA Catalogue

- Over 1,700 LRs compiled in the Universal Catalogue: http://universal.elra.info
- Antechamber of the ELRA Catalogue
- Window-shopping nature: allows users to realise about existence of LRs for future availability? ELDA team helps to clear out legal situation
- New feature: simplified collaboration form (following users' feedback)

The Universal Catalogue & the ELRA Catalogue

- Also related: LRE Map initiative:

See: *Calzolari, N., Soria, C., Del Gratta, R., Goggi, S., Quochi, V., Russo, I., Choukri, K., Mariani, J. and Piperidis S.: The LREC 2010 Resource Map. LREC 2010.*

– Cooperation ELRA & FlaReNet
(Fostering LRs Network - http://www.flarenet.eu)

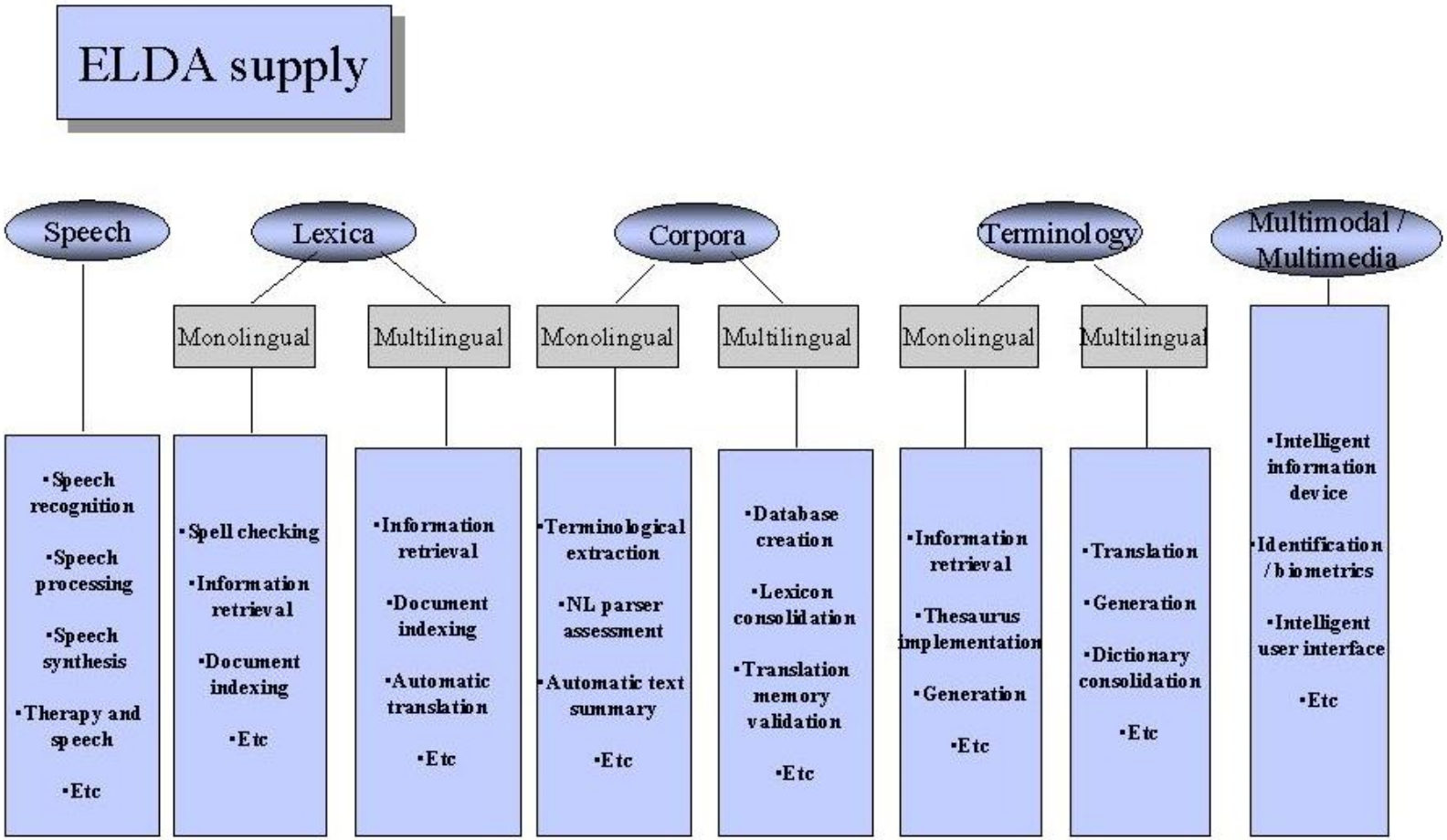
– LR identification tool during LREC 2010 submission time (ELRA & FlaReNet)

– LR description form incl.: LR type, name, production status, use, languages, modality, availability, URL, description, size, license, documentation

– About 2000 LRs: 785 corpora, 289 lexica, 181 taggers/parsers, 134 annotation tools, 73 ontologies, 40 evaluation data, ..., 42 different languages

LRs & Human Language Technologies

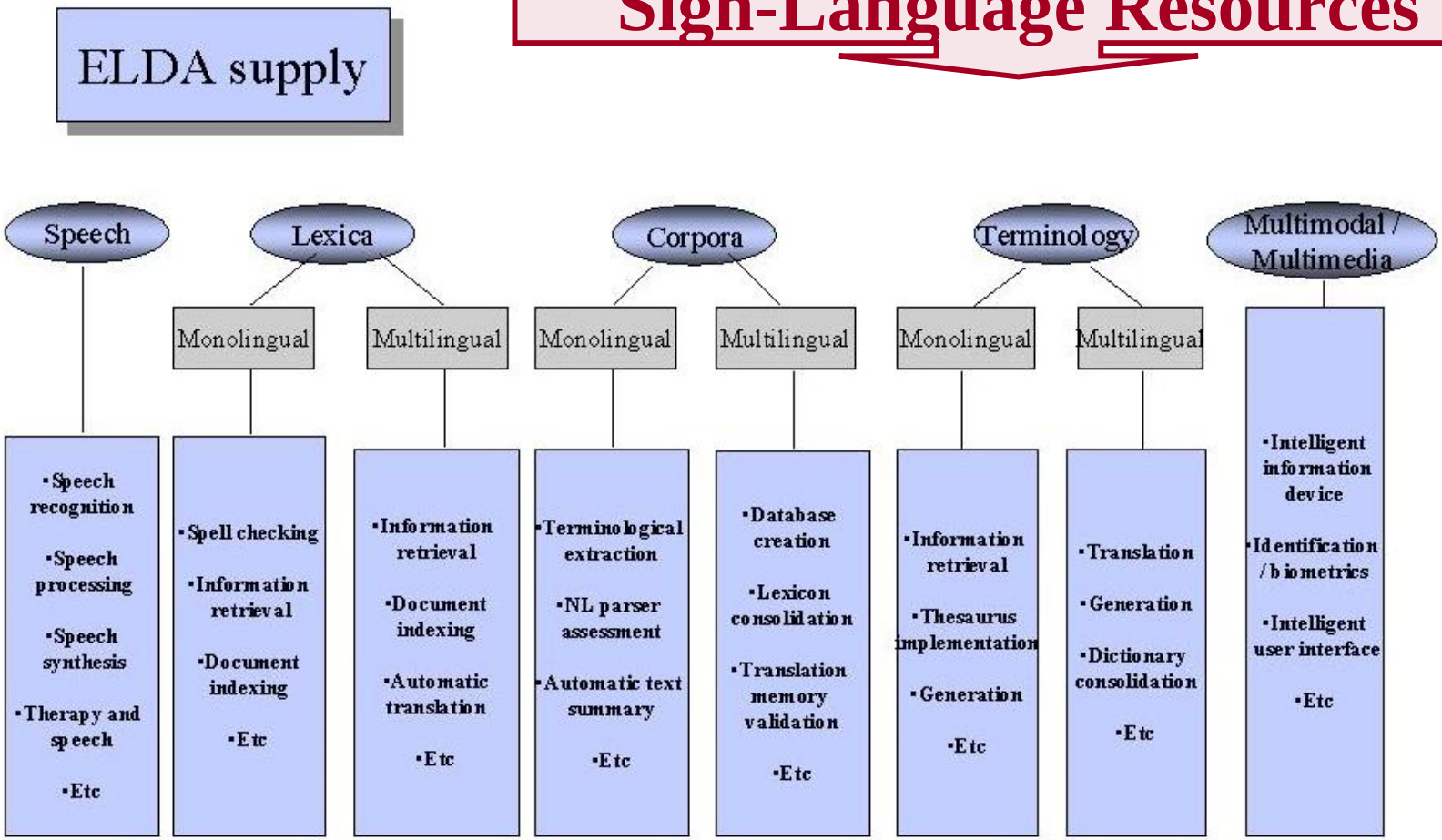
➢➢➢ Current situation



LRs & Human Language Technologies

➢➢➢ New trend

Sign-Language Resources



LRs for sign language in the ELRA catalogue

S0285 Czech Sign Language Corpus for Recognition – Amateur Signer (UWB-06-SLR-A)

(Available since 05/11/2008.)

This is an amateur sign-language database comprising 25 signs from Czech sign language. 15 signers (4 women and 11 men) carried out 5 repetitions of each sign and were recorded from 3 different views. The first is a frontal view of the upper part of the body. The data contain 5685 avi files (one per sign performance), using up 7 GB of disk space, and are stored on DVDs. **More details...**

Language(s) :

Membres	Academic org.	Commercial org.
Research Use	125.00 EUR	125.00 EUR
Commercial Use	275.00 EUR	275.00 EUR

Non Membres	Academic org.	Commercial org.
Research Use	175.00 EUR	175.00 EUR
Commercial Use	425.00 EUR	425.00 EUR

S0286 Czech Sign Language Corpus for Recognition – Professional Signer (UWB-07-SLR-P)

(Available since 05/11/2008.)

This database comprises 378 signs from Czech sign language as performed by 4 everyday sign-language users (4 women, 2 of them deaf). 5 repetitions of each sign were recorded from 3 different views. The data contain 21000 avi files (one per sign performance), using up 20 GB of disk space, and are stored on DVDs. **More details...**

Language(s) :

Membres	Academic org.	Commercial org.
Research Use	225.00 EUR	225.00 EUR
Commercial Use	525.00 EUR	525.00 EUR

Non Membres	Academic org.	Commercial org.
Research Use	325.00 EUR	325.00 EUR
Commercial Use	825.00 EUR	825.00 EUR

S0300 SIGNUM Database

(Available since 20/05/2009.)

The SIGNUM Database contains both isolated and continuous utterances of various signers. The corpus was recorded on video. For quick random access to individual frames, each video clip is stored as a sequence of images. The vocabulary comprises 450 basic signs in German Sign Language (DGS) representing different word types. Based on this vocabulary, overall 780 sentences were constructed. Each sentence ranges from two to eleven signs in length. The entire corpus was performed once by 25 native signers of different sexes and ages. One of them was chosen to be the so-called reference signer. His performances were recorded three times. **More details...**

Language(s) :

Membres	Academic org.	Commercial org.
Research Use	600.00 EUR	600.00 EUR
Commercial Use	600.00 EUR	600.00 EUR

Non Membres	Academic org.	Commercial org.
Research Use	1000.00 EUR	1000.00 EUR
Commercial Use	1000.00 EUR	1000.00 EUR

Distribution of Resources vs Usage

- ELRA has distributed over 3,500 LRs:
  - 48% research in academia
  - 37% research and technology development in industry
  - 16% evaluation
- Further 1,500 copies distributed within evaluation campaigns

Distribution of LRs & Pricing Policy

- Pricing Policy: crucial issue
- ELRA:
  - established a new market: trade of LRs
  - Had to take into account requirements & restrictions imposed by provider
- Market knowledge and contacts with potential providers: reliable & useful info on *demands* + *needs*
- ELRA's approach:
  - Simplify price-setting
  - Clarify possible uses of LRs
  - Reduce restrictions imposed by producer

Distribution of LRs & Pricing Policy

- ELRA:
  - not owner of LRs
  - sets fair price with owner, based on:
    - Production costs
    - Expected revenues
    - ELRA's distribution policy: try to offer discounted price for members
    - ELRA's effort to obtain low cost / free LRs for R&D

Production of LRs …… Over hundred different resources

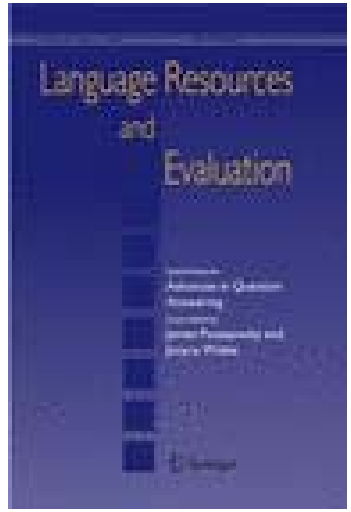
- Production or commissioning the production of LRs
- Within the framework of European and international projects
- In support of companies or institutions
- Production Services:
  - it has already compiled LRs in more than 25 languages
  - high quality LRs + strict validation
  - involved in every stage of production
  - 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, etc. etc.

HLT Evaluation

- Infrastructure for technology Evaluation
- Production of LRs for Evaluation
- Conducting Evaluation Campaigns & Capitalization & ROI (Exit strategies)
- Some covered technologies:
  - Text processing: Information retrieval, Question Answering, Machine Translation, Automatic Summarization, Parsing, Multilingual Text Alignment, Terminology Extraction,
  - Speech processing: Automatic Speech Recognition, Speech Synthesis, Speech Translation, Broadcast News Transcription, Acoustic Person Tracking, Acoustic Speaker Identification, Speech Activity Detection,
  - Multi-modal interfaces: Multimodal Person Tracking, Audiovisual Speech Recognition, Multimodal Person Identification.
- Portal for HLT Evaluation: http://www.hlt-evaluation.org
  - all sorts of information related to evaluation
  - quick and easy reference about protocols, metrics, tasks, resources, projects, campaigns, etc.

Dissemination

- ELRA has increased its activities for the dissemination of information on LRs:
  - "Speaker's Corner" for the researchers and developers of the area
  - Events:
    - Language Resources and Evaluation Conference (7th edition): http://www.lrec-conf.org
    - LangTech
    - European LR and Technologies Forum (within FlaReNet)
    - MEDAR Conferences
    - Worshops of less-resourced languages (within LTC'09)
  - Language Resources and Evaluation Journal (Springer): http://www.springerlink.com/
  - Newsletter
  - Members' News
  - BLARK web site: http://www.blark.org/ + other web sites



Concluding Remarks

- Overview on latest developments in ELRA's services around:
  - Identification & Distribution
  - Evaluation
  - Production
  - Dissemination
- From early archiving and distribution to LR identification, collection, validation and distribution platform....
- ...with clear and well-established legal frameworks...
- ...enhancing work on evaluation (with new techniques, covering more technologies and languages, providing more evaluation packages, setting up the HLT Eval portal....)
- ... increasing work on LR production and its coverage
- ...big push to dissemination...

Concluding Remarks

- ...encouraging international cooperation...
- ...after years of consolidation...ELRA is looking forward to the new challenges emerging from new trends...
- ...a new LR open exchange infrastructure is being launched (META-SHARE project).

Thank you for your attention