

Language Resources

By the Other Data Center ... over 15 years for partnership

Khalid Choukri

ELRA

55 Rue Brillat-Savarin 75013 Paris, France

Tel. +33 1 43 13 33 30 Fax. +33 1 43 13 33 30

choukri@elda.org



www.elda.org/ or <http://www.elra.info/>

Past ... Present ... future frameworks for LRS



META  SHARE
META  NET

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)



EUROPEAN
ASSOCIATION
EL
RA
LANGUAGE
RESOURCES



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 LR
- An (not for profit) Association of Users of Language Resources
- A Repository Center that takes care of all issues related to LR from Discovery to sharing and distribution, for R&D and Commercial purposes
- LR 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>

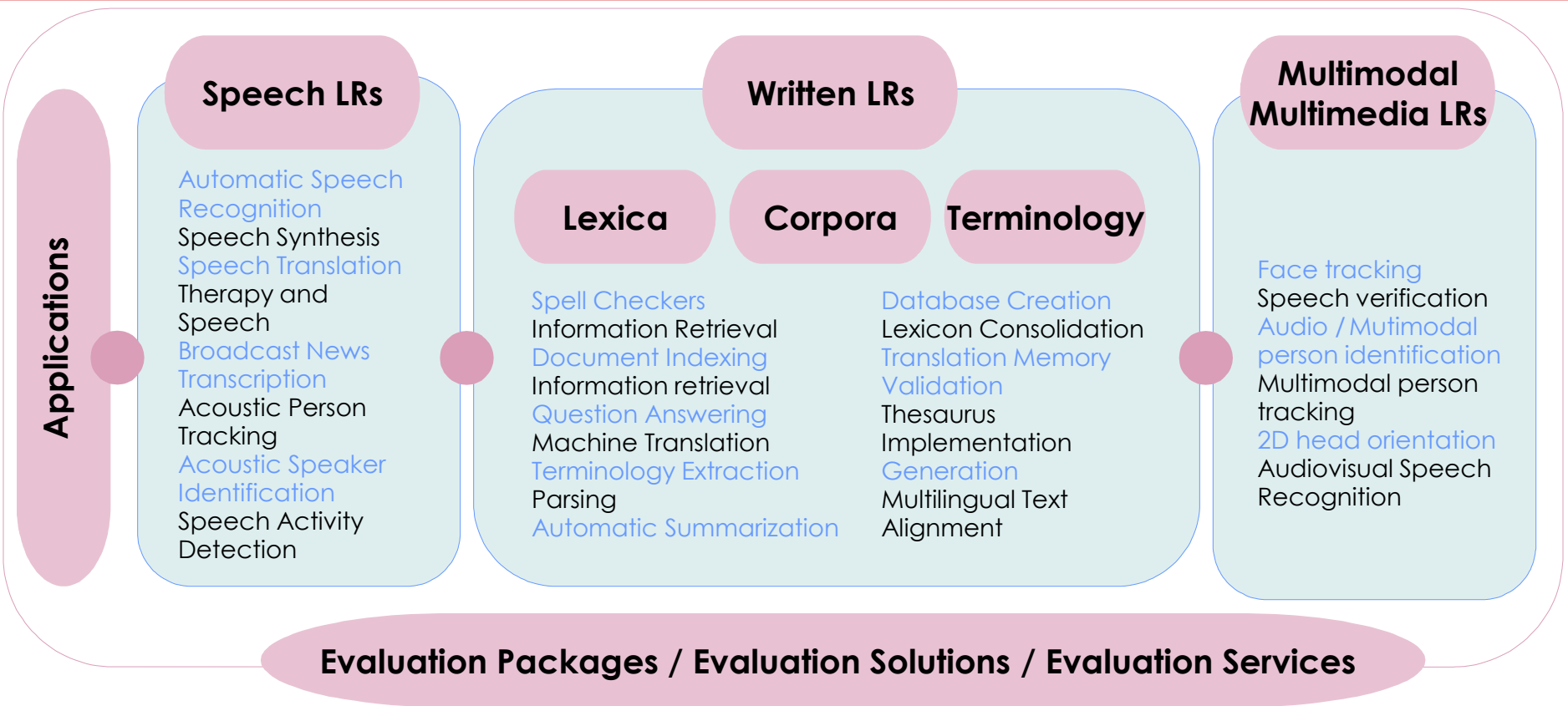
ELRA's Foundation & Mission

An Improved infrastructure for Data sharing & HLT evaluation

➤ Infrastructure for the evaluation of Human Language Technologies

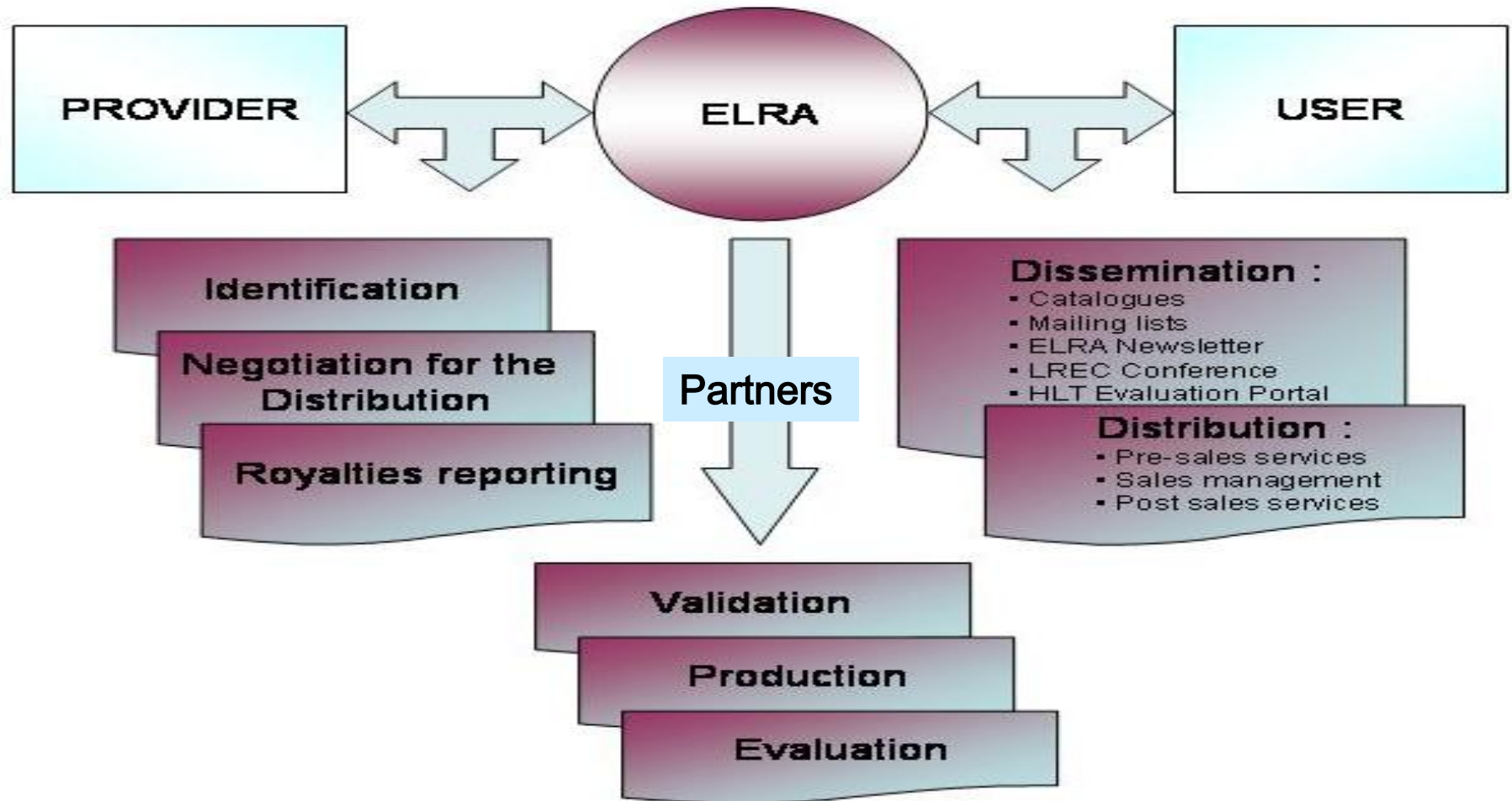
- providing resources, tools, methodologies, logistics,
- Online evaluation platforms
- Exit strategies / Capitalization on evaluation packages

ELRA offer



➤➤➤ **LRs & Evaluation Packages**

ELRA Activities.... Identification, Distribution, Production of LR's and HLT Evaluation & Dissemination



Identification and distribution of LRs

- ELRA Catalogue (>1000 LRs): <http://catalogue.elra.info/>
 - About 30% are/will be free of charge
- Universal Catalogue: (>2000 LRs) <http://universal.elra.info/>
- **LRE-MAP (4000+ LRs & LT)** <http://www.resourcebook.eu/>
- Types of LRs : All modalities associated (or not) with language (images, videos, sign, OCR, ...)

- 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)

Written Data (Ranked)	Bulgarian	Czech	Danish	Dutch	English	Estonian	Finnish	French	German	Greek	Hungarian	Irish	Italian	Latvian	Lithuanian	Maltese	Polish	Portuguese	Romanian	Slovak	Slovene	Spanish	Swedish	Other Europe	Regional Europe	Multilingual	L.I.	N.A.	Total
Corpus	7	12	6	17	206	3	3	44	43	10	8	1	32	9	4	1	7	19	12	2	5	29	19	19	18	5	9	2	552
Lexicon	6	7	2	8	77	1	2	24	15	3	4	0	16	0	0	0	2	6	7	0	1	19	4	11	8	3	3	0	229
Ontology	1	2	0	2	18	0	0	3	4	2	0	0	4	0	2	0	1	1	1	0	0	4	0	3	0	1	16	2	67
Grammar/Language Model	1	1	2	1	11	0	1	4	2	0	1	0	2	0	0	1	2	1	1	1	0	5	1	3	1	0	2	1	45
Terminology	1	1	0	2	10	1	0	5	3	0	1	0	0	1	1	0	1	0	0	0	0	2	0	2	3	1	1	0	36
A syntactic judgments database	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	3
Resource: morphology	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2
Thesaurus	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
A Knowledge Base with Lexical-Semantic Relations between words	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
A list of categories with examples of language use	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Controlled Legal Language	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Corpus-Based Online Dictionary	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Database	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Encyclopedic knowledge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Event Semantics	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Lexicon/corpus	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Online Encyclopedia	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Repository of bilingual lexicons	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
Resources integration	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Text Book	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Virtual Game World	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Yahoo!'s local listings in Chicago	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Encyclopedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Part-of-Speech Tagset	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Psycholinguistic Database	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	16	23	11	30	327	5	6	80	70	15	14	1	56	10	7	2	13	27	21	3	6	61	25	39	30	11	36	5	950

E.3.a. Written Language Data (Ranked order)

- 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

EUROPEAN
ASSOCIATION
ELRA
LANGUAGE
RESOURCES



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-ata), 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 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 LR_s and related objects and processes and foster their use

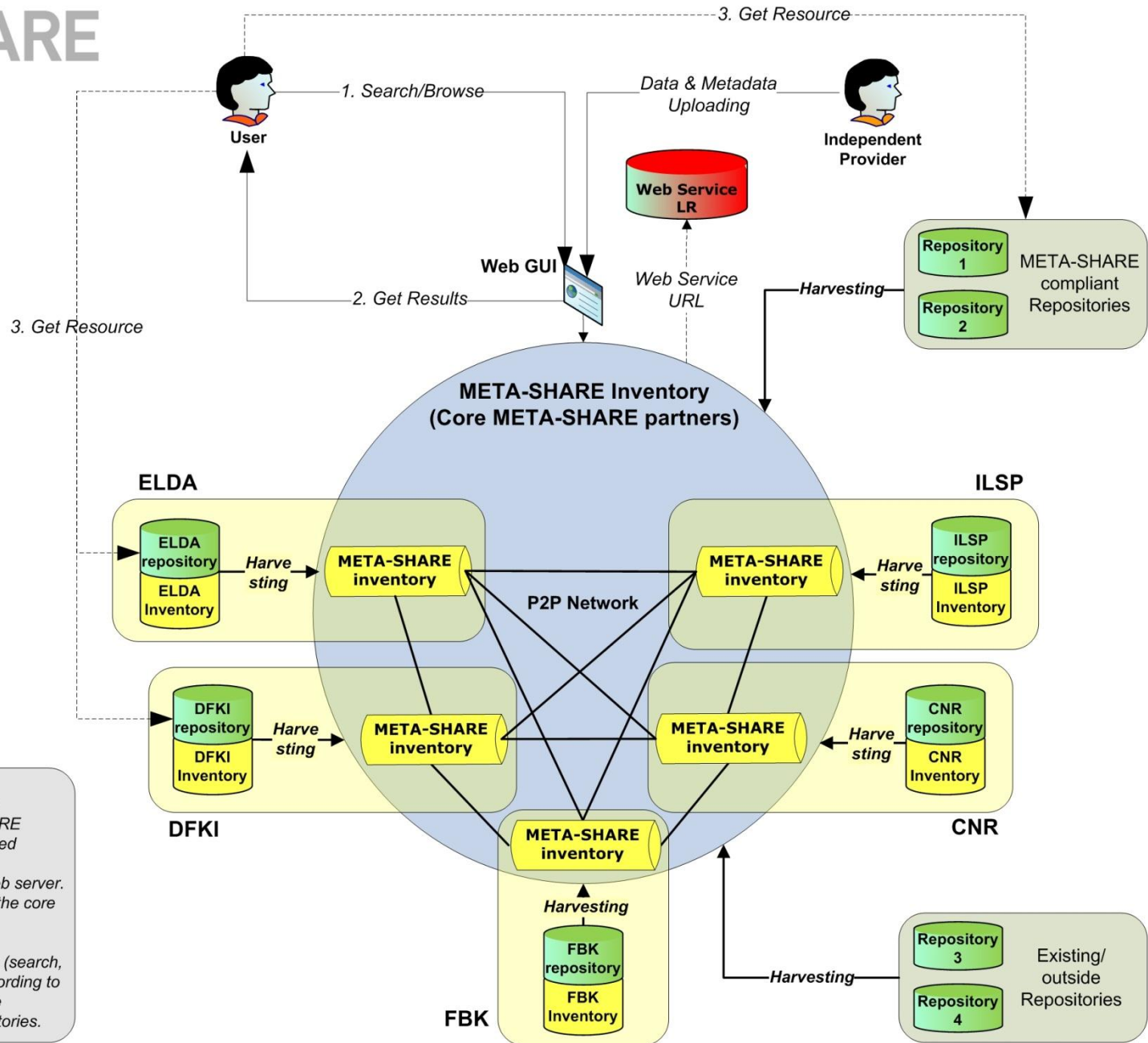
❑ **How?**

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

- ➔ Adoption of a distributed network of Repositories
- ➔ Adoption of a *uniform* metadata schema
- ➔ Adoption of a common and trustable legal framework

META-SHARE

Version 0



<http://www.meta-net.org>

Notes:

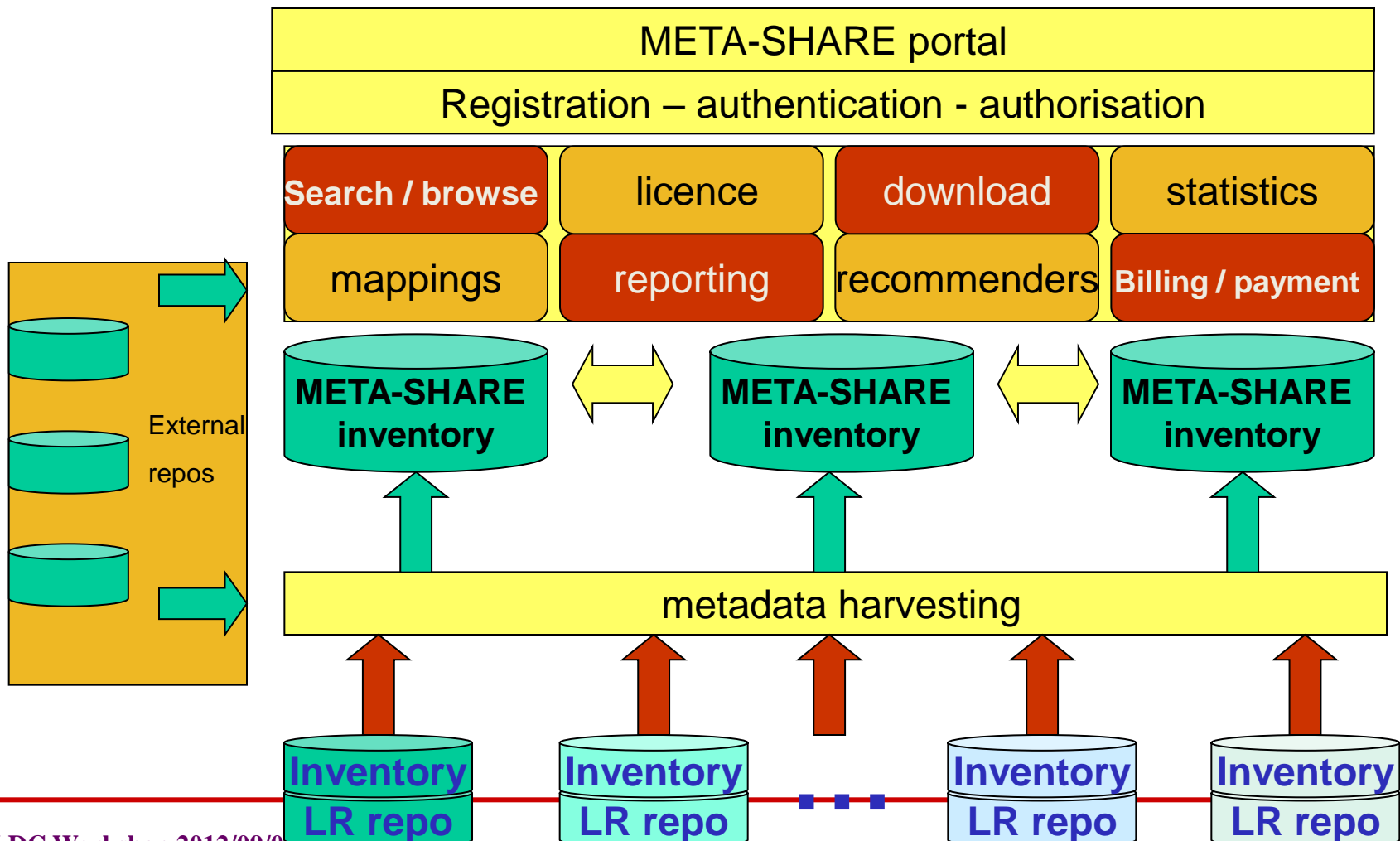
Harvesting: Metadata Harvesting (OAI-PMH)

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 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)

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

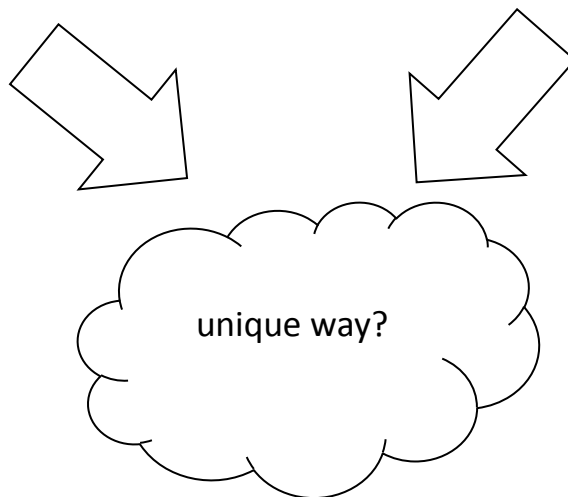
Translanguage English Database (TED):



ELRA-S0031



LDC2002S04

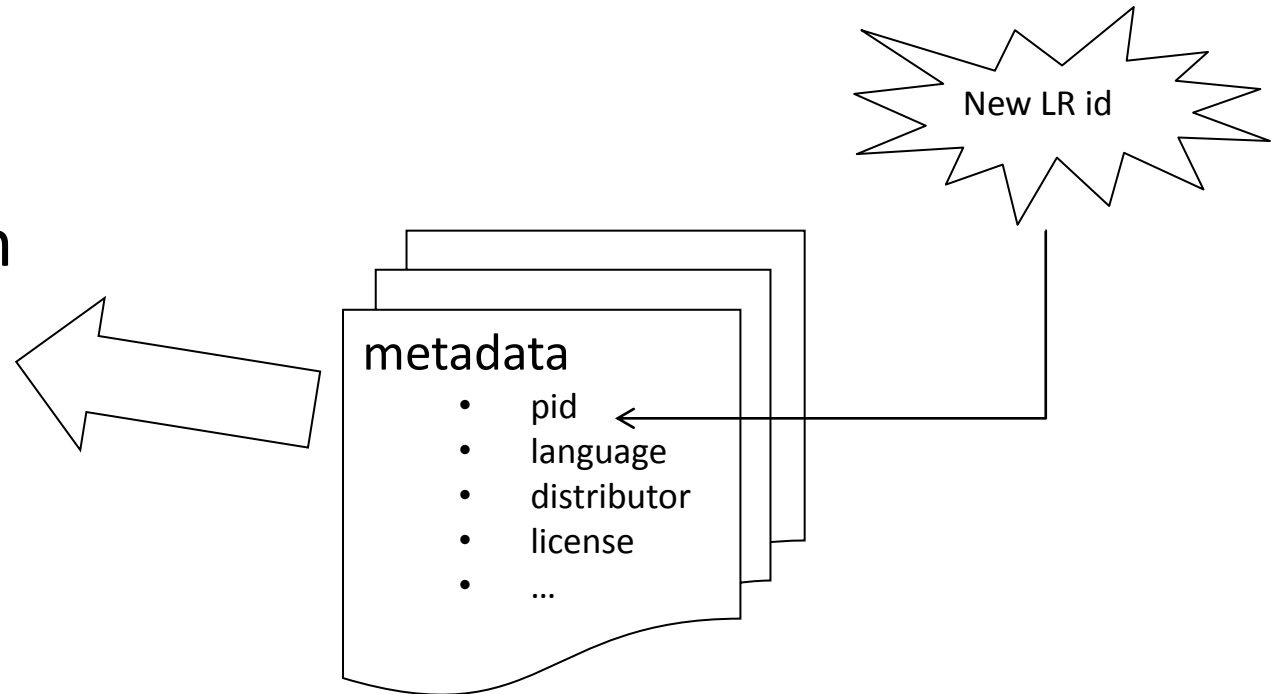


Usage of some major data centers

	ELRA	LDC	NICT	GSK	BAS	Chinese LDC	HLT-Centrale
Publisher	X	X		X		X	
Category	X	X	X			X	
Year		X		X		X	
Digit ID	X (4)	X (2)	X (6)			X (3)	
Letter ID				X			
Free ID					X		X
Software	X	X		X	X		X
Example	ELRA-S0035	LDC2004L01	G-00035	GSK2010-C	SC10	CLDC-SPC-2007-002	CORN

- 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.

- Persistence
- Uniqueness
- Identification
- Discovery



Highlight of the ISLRN

ISLRN: XXX-XXX-XXX-XXX-X

Random +Check-sum 13 digits

e.g. **ISLRN: 193-187-955-056-0**

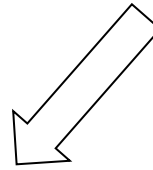
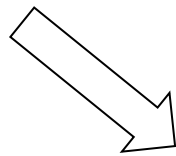
Translanguage English Database (TED):



ELRA-S0031



LDC2002S04



ISLRN: 001-920-233-002-2

- 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

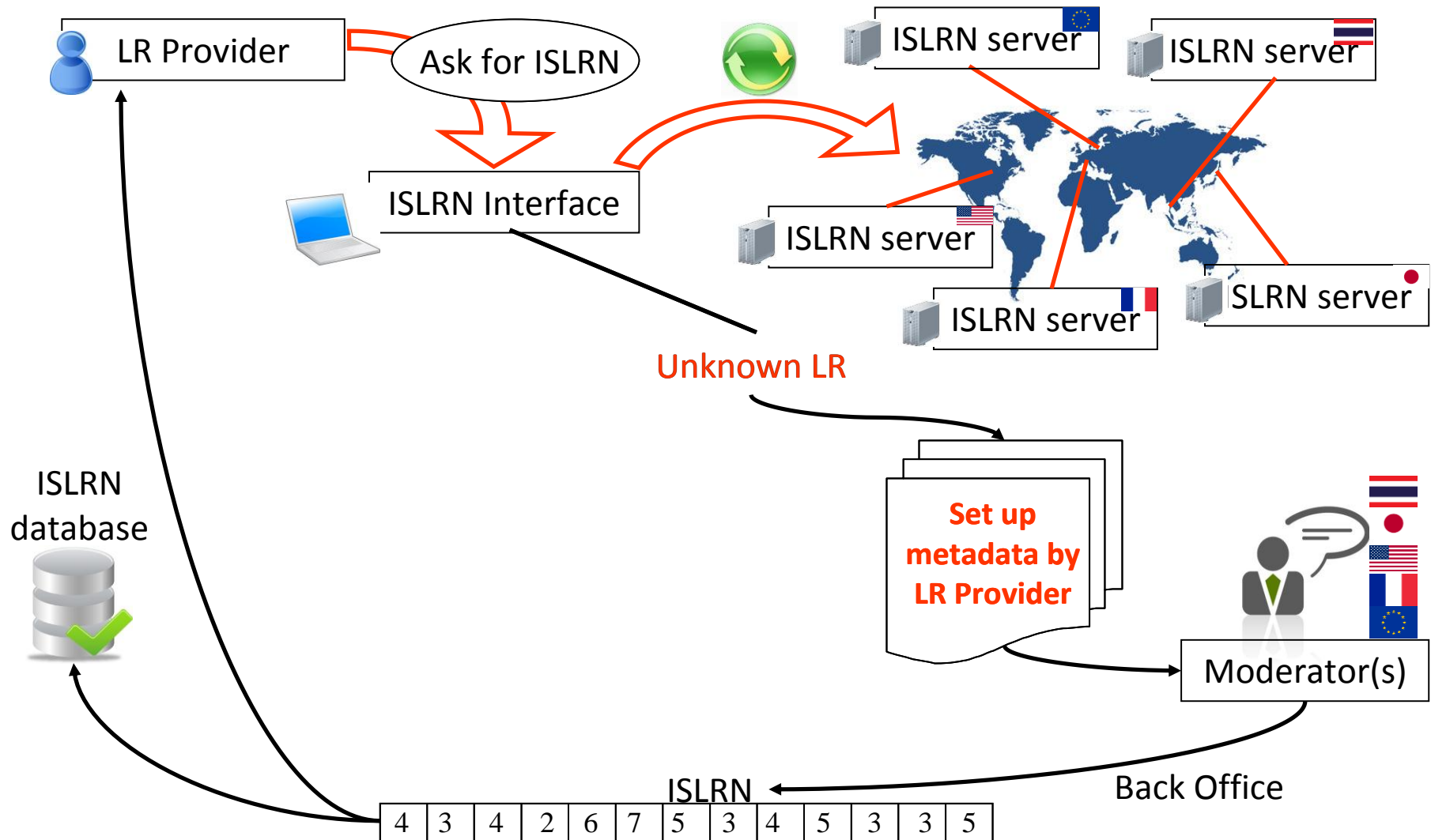
• ...

- 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

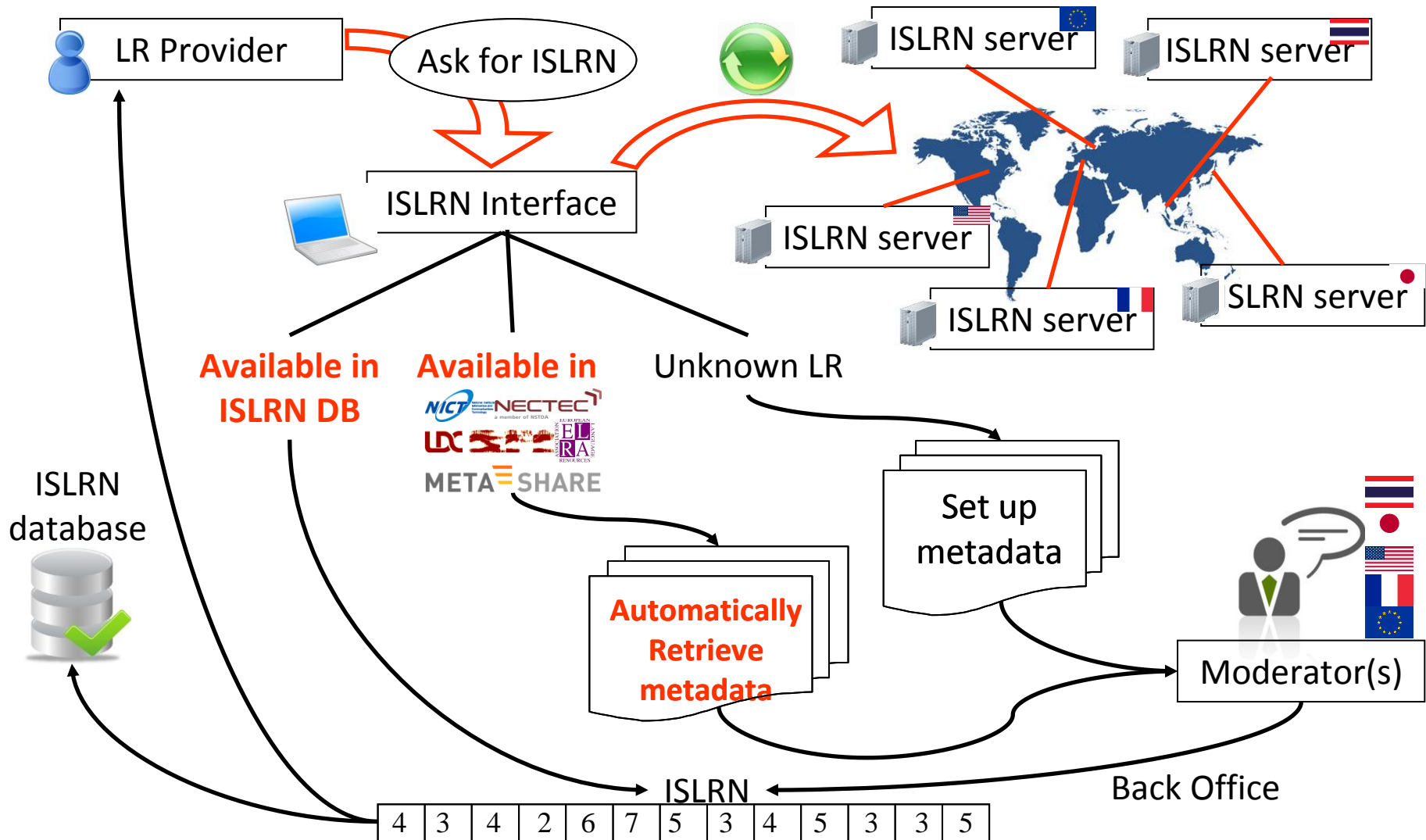
a Proposal the ISLRN management structure



ISLRN attribution approach during a first phase



ISLRN attribution approach during a 2d phase



- 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



EUROPEAN
ASSOCIATION
EL
RA
LANGUAGE
RESOURCES



Thank you for your attention

