“On the way to Language Resources sharing: principles, challenges, solutions”

Stelios Piperidis

ILSP, RC Athena, Greece
spip@ilsp.gr

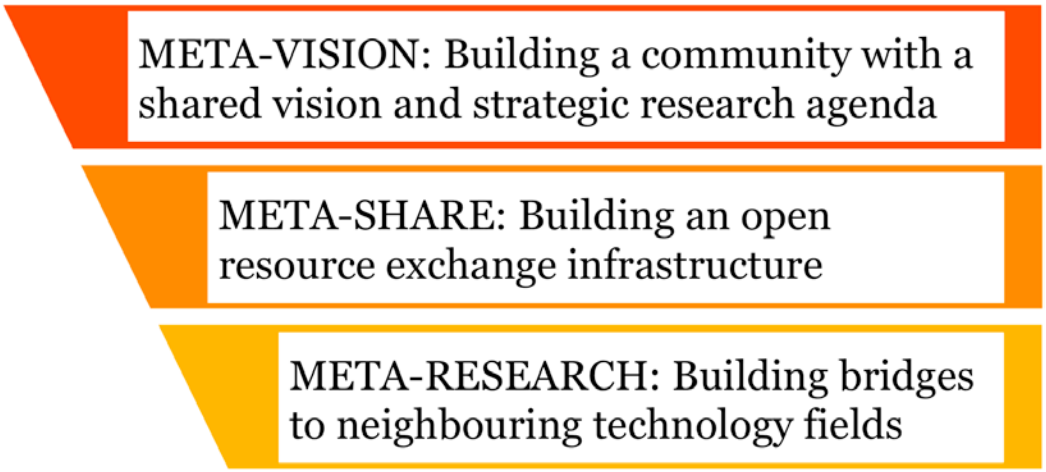
„Content on the Multilingual Web“, 4-5 April, Pisa, 2011

Co-funded by the 7th Framework Programme of the European Commission through the contract T4ME, grant agreement no.: 249119.

Outline

- META-NET
- META-SHARE : Intro & Rationale
- Architecture
- META-SHARE vO and next steps

META-NET: Objectives



- META-NET is a Network of Excellence dedicated to fostering the technological foundations of the European multilingual information society:
  - Build META, a strategic alliance that includes multiple stakeholders to prepare the ground for a large-scale concerted effort.
  - Strengthen the European research community.
  - Approach open problems in MT in collaboration with other fields.

Introduction
Rationale & Objectives

Building META-SHARE, an open resource exchange

- Data has become a key factor in LT R&D. A few indicators:
  - Increasing size and importance of the LREC conference, corpora mailing list etc.
  - Citation ranks of publications on language resources
  - High-ranking demand in all three META-NET Vision Groups

- No matter what technology or application one intends to build, a substantial, bulky data set together with the associated basic processing tools/services is indispensable
  - (Statistical) machine translation, speech recognition/synthesis, ...
  - Information extraction and higher level text and media analysis and annotation (e.g. sentiment, persuasion, etc)
  - ...

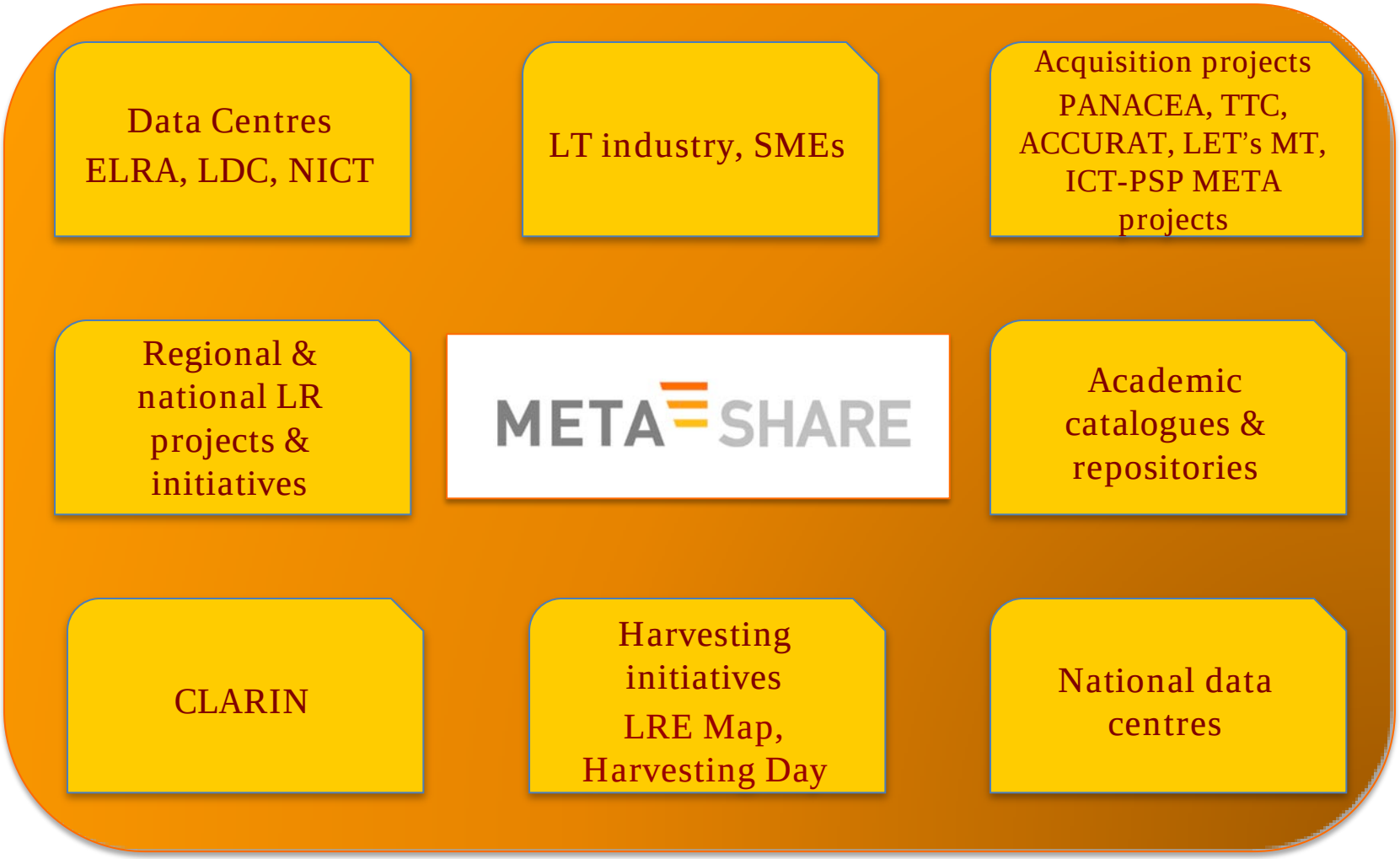
A few observations

- Data collection, cleaning, annotation, curation, maintenance, etc is a very costly business
- Data become considerably valuable through sharing.
- Commissioner Neelie Kroes, Vice-President of the EC (responsible for the Digital Agenda): *"Scientific data has the power to transform our lives for the better – it is too valuable to be locked away."*
- High-Level Group on Scientific Data report *: "A fundamental characteristic of our age is the rising tide of data – global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge."*
- The long demanded and well-contemplated instruments for managing and sharing this data are *still missing*.

META-SHARE: Key Features

- 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)
- Standards-compliant, overcoming format, terminological and semantic differences.

META-SHARE


Data Centres
ELRA, LDC, NICT
LT industry, SMEs
Acquisition projects
PANACEA, TTC, ACCURAT, LET's MT, ICT-PSP META projects
Regional & national LR projects & initiatives
META SHARE
Academic catalogues & repositories
CLARIN
Harvesting initiatives
LRE Map, Harvesting Day
National data centres

Architecture

META-SHARE architecture

- META-SHARE is implemented as a network of distributed repositories
  - Local (organisation-based), and
  - Non-local (central) repositories
- Local repos store and maintain the organisation's LRs (data sets and tools)
- Non-local repos act as storage and documentation facilities for LRs of organisations not wishing to set up their own repository, or donated or orphan LRs, etc.
- LRs are described according to a metadata schema, including their rights of use

META-SHARE architecture (2)

- Actual LRs and their metadata (MD) reside in the local repositories.
- Each repository
  - maintains an inventory (a local inventory) with all MD of their LRs
  - exports MD
  - allows their harvesting.
- Harvested MD are stored in the META-SHARE central servers, which . share MD in a p2p fashion
- Central servers create, host and maintain a central inventory with all MD descriptions of all LRs available in the distributed network.

META-SHARE architecture (3)

- Users (language resources seekers/consumers) will be able to
  - **log-in once** www.meta-share.eu or www.meta-share.org
  - **search** the central inventory using multifaceted search facilities, and
  - access the actual resources by visiting the **local** (or **non-local**) repositories for **browsing and downloading** them.
- To access LRs (data, tools, language processing services) users need to agree with the terms and conditions of use spelt out in the licence of the respective LR
- Rights of use and related restrictions under the control and responsibility of LR owners and the repository where the LR resides
- META-SHARE favours and aligns with open data and open source movements
- Does not exclude LRs for a fee, fosters commercial use of LRs

Priorities

- **Type of resources and technologies**:
  - language data description, collection and cataloguing,
  - language processing tools description, collection and cataloguing,
  - evaluation data and evaluation tools and services description and cataloguing,
  - language data processing services through tools and technologies (starting from basic ones),
  - workflows by integrating simple services

Metadata schema – basic principles (1)

- Descriptions of
  - **LRs**, encompassing both **data** (*textual, multimodal/multimedia and lexical*) and **tools/technologies** used for their processing
  - **related objects** (*reference documents, actors, activities* etc.)
- External metadata only (referring to LR description and related processes)
- Aim: to support META-SHARE users (incl. LRs providers and consumers) in all services provided (LR description, search and retrieval, metadata harvesting/updating, monitoring of LRs and related objects, etc.)
- We're not reinventing the wheel: **harmonize** existing schemas and related initiatives and **adapt** them to the requirements of the HLT community

Metadata schema – basic principles (2)

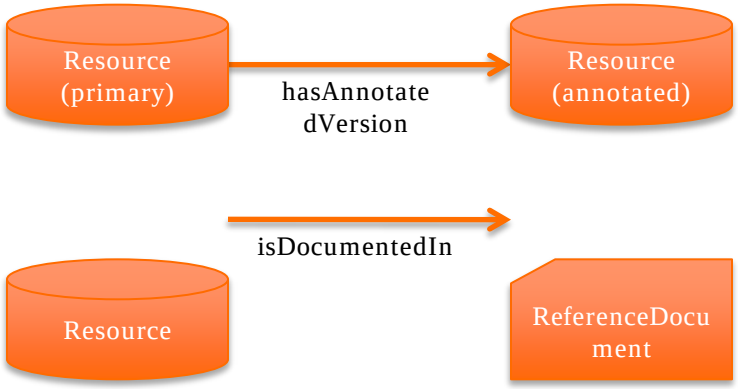
- main desiderata:
  - clarity of semantics - expressiveness
  - flexibility - customisability
  - interoperability - user friendliness
  - extensibility - harvestability
- methodology
  - survey of existing schemas & relevant initiatives
    - ISOcat DCR (CLARIN), IMDI, ENABLER, BAMDES, TEI, XCES, DC, OLAC, etc.
    - catalogues: ELRA, LDC, Universal Catalogue, NLSR etc.
  - user requirements surveys and usage scenarios (ongoing in project)

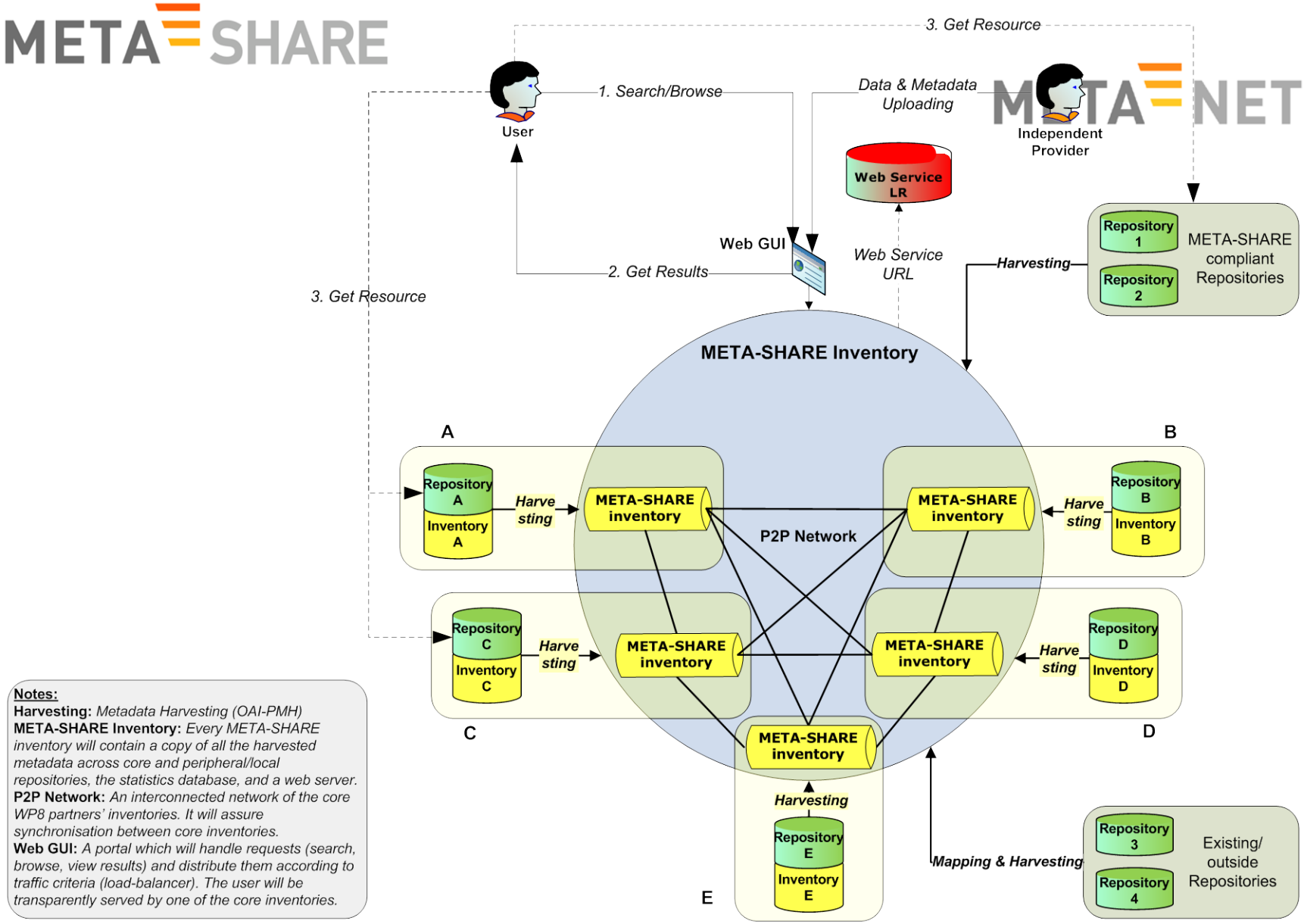
Metadata schema - main features (1)

- ISOcat-compatible
- includes:
  - ***elements*** (linked to ISOcat Data Categories): used to describe specific features of the resources (e.g. title, description, format, languages etc.

  - ***relations*** (extension of ISOcat): used to link together resources included in the META-SHARE (e.g. original and derived corpus, raw and annotated corpus, a corpus and the tool that has been used to create it, a corpus and its documentation etc.)

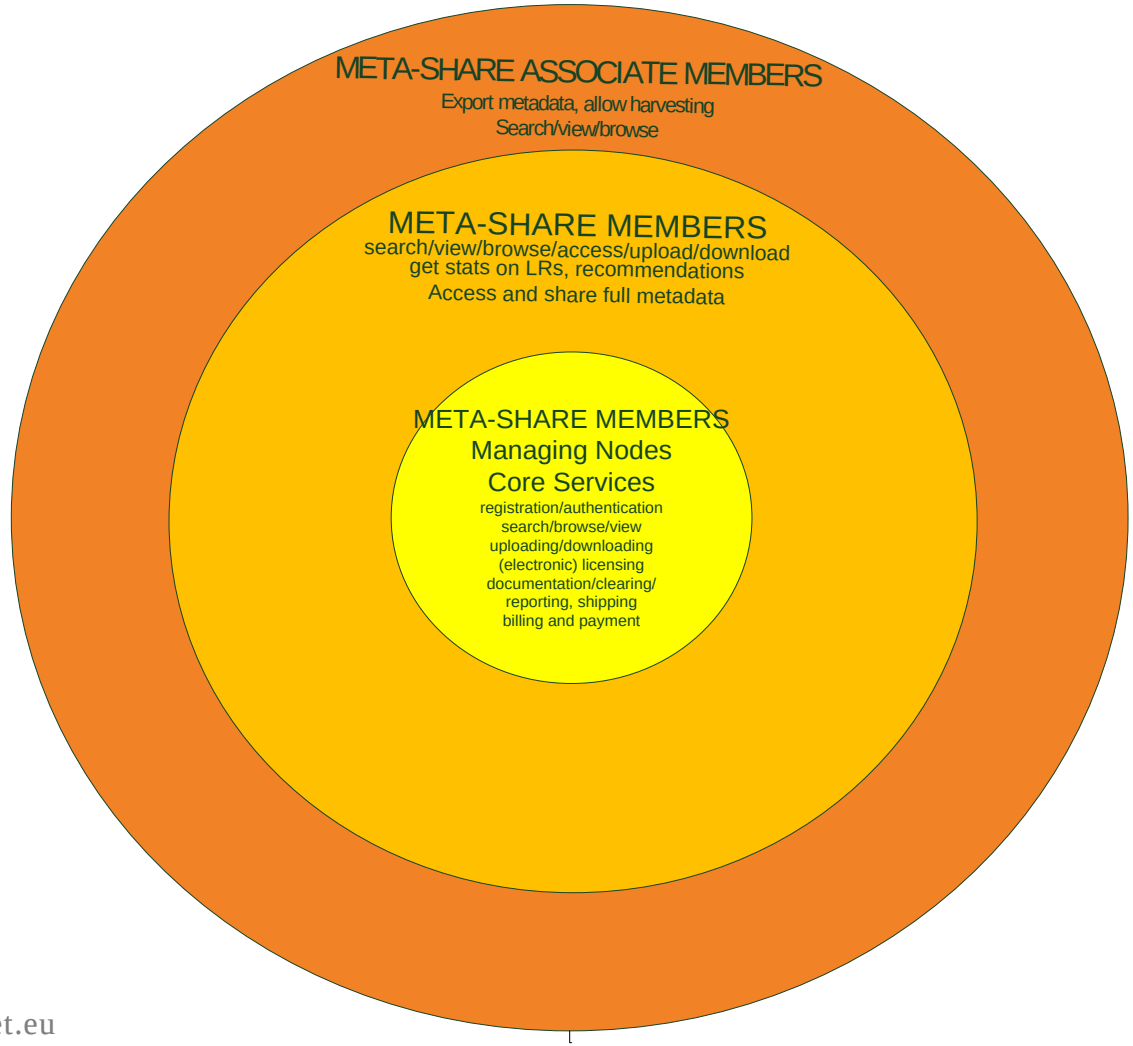
ResourceTitle: String
Description: String
NumberOfLanguages: Integer
LanguageName: Enumerated
...



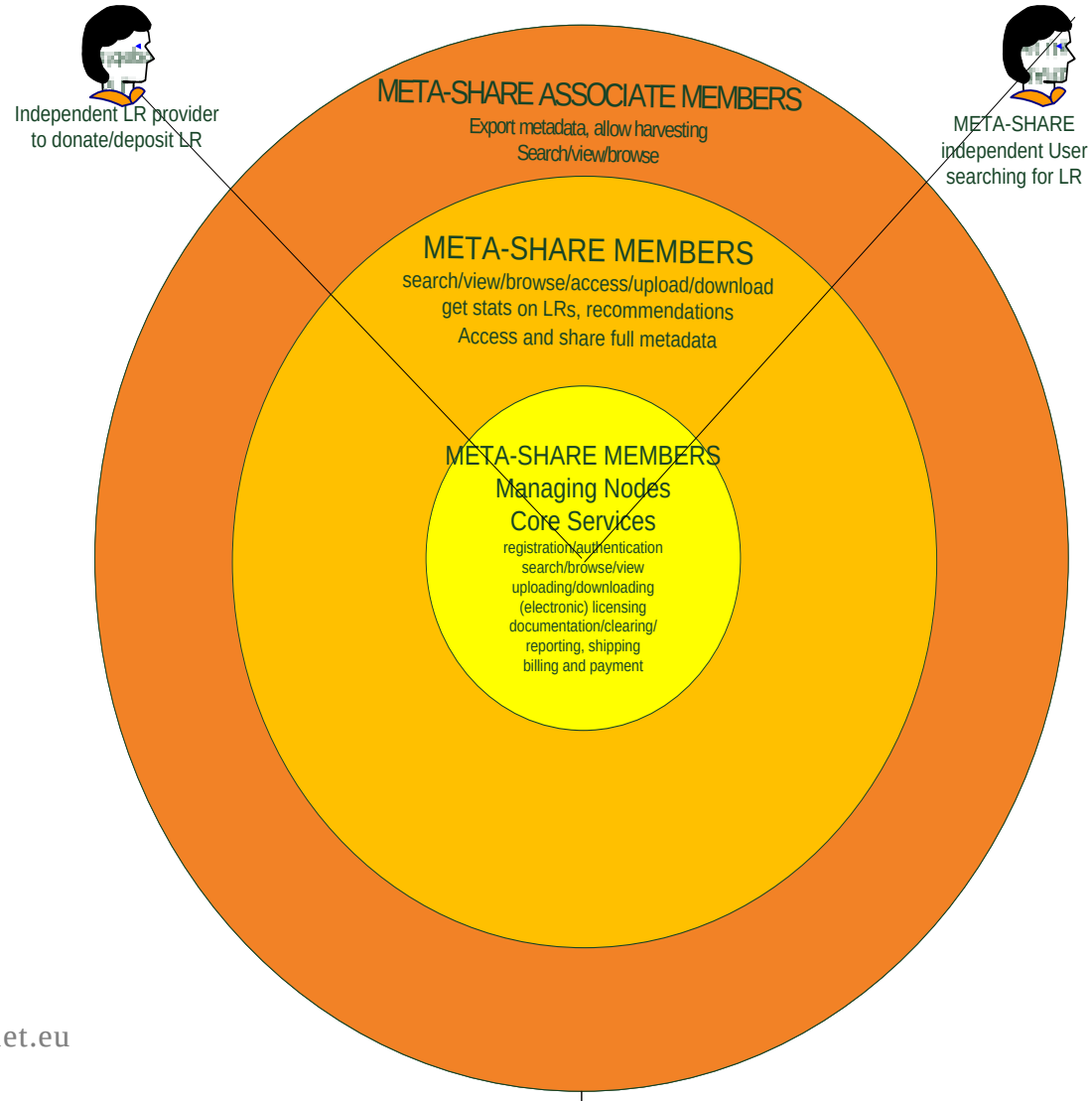


Notes:
Harvesting: *Metadata Harvesting (OAI-PMH)*
META-SHARE Inventory: *Every META-SHARE inventory will contain a copy of all the harvested metadata across core and peripheral/local repositories, the statistics database, and a web server.*
P2P Network: *An interconnected network of the core WP8 partners' 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.*

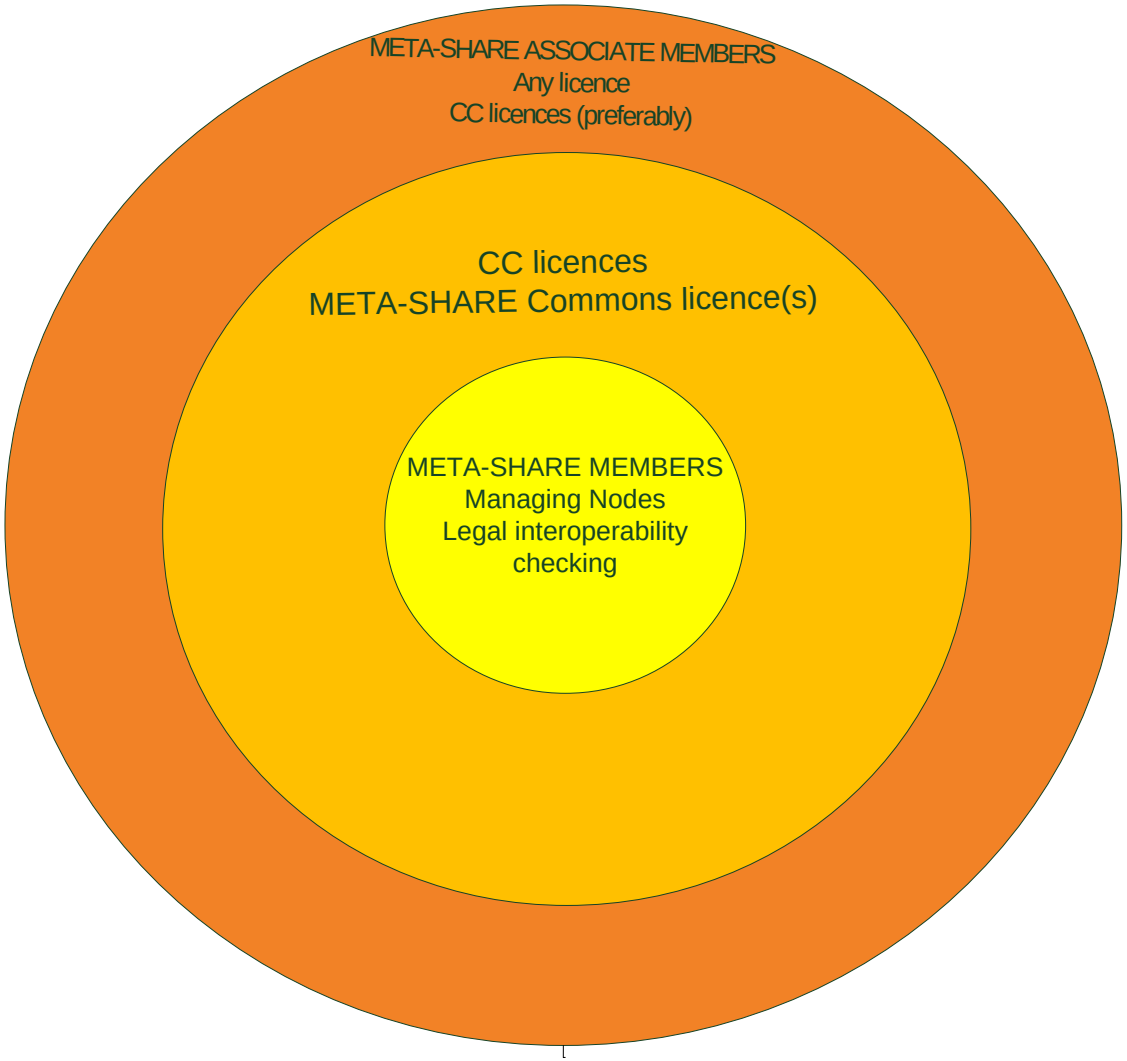
Governance


META-SHARE ASSOCIATE MEMBERS
Export metadata, allow harvesting
Search/view/browse
META-SHARE MEMBERS
search/view/browse/access/upload/download
get stats on LRs, recommendations
Access and share full metadata
META-SHARE MEMBERS
Managing Nodes
Core Services
registration/authentication
search/browse/view
uploading/downloading
(electronic) licensing
documentation/clearing/
reporting, shipping
billing and payment

META-SHARE third parties



META-SHARE legal domain



Features

- Open Source
- Distributed
- Metadata Harvesting
- Replication/Backup
- Easy Administration
- Single Sign-On
- Intuitive Search
- Persistent LR Identification (PIDs)
- Easy licensing
- Reporting & Statistics

META-SHARE

Version O

Welcome to META-SHARE!

META-SHARE is developed within the META-NET Network of Excellence

About the project

META-NET is designing and implementing META-SHARE, a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources. Data and tools can be both open and with restricted access rights, free and for-a-fee. META-SHARE targets existing but also new and emerging language data, tools and systems required for building and evaluating new technologies, products and services.

About the partners

META-SHARE will start by integrating nodes and centres represented by the partners of the META-NET consortium. It will gradually be extended to encompass additional nodes/centres and provide more functionality with the goal of turning into an as largely distributed infrastructure as possible.

Select network node

Please select one of the following META-SHARE network nodes to proceed:

CNR — National Research Council of Italy

DFKI — Deutsches Forschungszentrum für künstliche Intelligenz

ELDA — Evaluations and Language resources Distribution Agency

FBK — Fondazione Bruno Kessler

ILSP — Institute for Language and Speech Processing

Co-funded by the 7th Framework Programme of the European Commission through the grant agreement no. 249119.

META SHARE Browse metadata catalogue

Contains all known metadata information

	Title	Date	Provider
1.	Corpus	11/09/2010 at 00:00	CNR Fedora (OAI Dublin Core)
2.	Cultural Thesaurus of the Greek Language POTHEG	11/15/2010 at 14:01	ILSP Fedora (OAI Dublin Core)
3.	Dictionary	11/09/2010 at 00:00	CNR Fedora (OAI Dublin Core)
4.	Excerpt of BootStrep DB	11/07/2010 at 11:46	CNR Fedora (OAI Dublin Core)
5.	Excerpt of Simple DB: Semantic	11/08/2010 at 16:04	CNR Fedora (OAI Dublin Core)
6.	Excerpt of Simple DB: morpho-phono	11/08/2010 at 16:09	CNR Fedora (OAI Dublin Core)
7.	Greek Dependency Treebank GDT	11/15/2010 at 11:45	ILSP Fedora (OAI Dublin Core)
8.	Heart of Gold HoG	11/15/2010 at 14:45	DFKI Fedora (OAI Dublin Core)
9.	Hellenic National Corpus HNC	11/10/2010 at 13:55	ILSP Fedora (OAI Dublin Core)
10.	INTERA corpus	11/15/2010 at 13:57	ILSP Fedora (OAI Dublin Core)
11.	Kyoto - Corpus Estuaries	11/12/2010 at 15:25	CNR Fedora (OAI Dublin Core)
12.	Kyoto Ontotagger	11/12/2010 at 15:34	CNR Fedora (OAI Dublin Core)
13.	MMorph	11/15/2010 at 14:51	DFKI Fedora (OAI Dublin Core)
14.	MT Server Land	11/15/2010 at 14:45	DFKI Fedora (OAI Dublin Core)
15.	Mary Text To Speech	11/15/2010 at 14:31	DFKI Fedora (OAI Dublin Core)
16.	PET	11/15/2010 at 14:20	DFKI Fedora (OAI Dublin Core)

META SHARE Metadata search interface

Search within all known metadata information

Keywords: greek corpus

Search Catalogue | Advanced Search

Search results

Metadata object matching your query

	Title	Date	Provider
1.	Greek Dependency Treebank GDT	11/15/2010 at 11:45	ILSP Fedora (OAI Dublin Core)
2.	Cultural Thesaurus of the Greek Language POTHEG	11/15/2010 at 14:01	ILSP Fedora (OAI Dublin Core)
3.	POETICON Multisensory and Multimedia Recordings of Everyday Interaction POETICON recordings	11/15/2010 at 11:45	ILSP Fedora (OAI Dublin Core)
4.	INTERA corpus	11/15/2010 at 13:57	ILSP Fedora (OAI Dublin Core)
5.	Hellenic National Corpus HNC	11/10/2010 at 13:55	ILSP Fedora (OAI Dublin Core)
6.	Corpus	11/09/2010 at 00:00	CNR Fedora (OAI Dublin Core)
7.	Kyoto - Corpus Estuaries	11/12/2010 at 15:25	CNR Fedora (OAI Dublin Core)

META SHARE Browse metadata catalogue

Contains all known metadata information

Header fields

Title	Greek Dependency Treebank GDT
Date	11/15/2010 at 14:05
Provider	ILSP Fedora (OAI Dublin Core)

Dublin Core fields

Title	Greek Dependency Treebank GDT
Creator	Institute for Language and Speech Processing/R.C. "Athena"
Subject	European Parliament sessions Politics Health Travel
Description	70K words, Non-validated sentence segmentation. Non-validated POS tagging, Manual annotation of syntactic dependencies and dependency labels, Manual annotation of semantic roles, Manual annotation of events based on a shallow domain specific ontology (only for a 31K words subset of GDT).
Publisher	Institute for Language and Speech Processing/R.C. "Athena"
Date	2005-2010
Type	Collection Text
Format	text/plain
Identifier	ilsp.langres:2
Source	ilsp.langres:2/datastreams/Datastream/content
Language	ell
Relation	ILSP Dependency Parser
Rights	CC-BY-SA-NC

META SHARE

Browse metadata catalogue

Download resource: Greek Dependency Treebank (1/1)

License agreement

CC-BY-SA-NC
You are free:

to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Under the following conditions:

Attribution — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Noncommercial — You may not use this work for commercial purposes.
Share Alike — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.
With the understanding that:

Waiver — Any of the above conditions can be waived if you get permission from the copyright holder.
Public Domain — Where the work or any of its elements is in the public domain under applicable law, that status is in no way affected by the license.
Other Rights — In no way are any of the following rights affected by the license:
Your fair dealing or fair use rights, or other applicable copyright exceptions and limitations;
The author's moral rights;
Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.
Notice — For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this web page.

☑ I **agree** to these license terms and want to download the resource.

Download Resource

META SHARE Metadata advanced search

Search within all known metadata information

Title:	multimedia
Language:	eng
Type:	
Rights:	
Format:	video
Subject:	

Search Catalogue | Simple Search

Search results

Metadata object matching your query

	Title	Date	Provider
1.	POETICON Multisensory and Multimedia Recordings of Everyday Interaction POETICON recordings	11/15/2010 at 17:34	ILSP (OAI Dublin Core)

META-SHARE Browse metadata catalogue

Contains all known metadata information

Header fields

Title POETICON Multisensory and Multimedia Recordings of Everyday Interaction
POETICON recordings

Date 11/10/2010 at 17:28

Provider ILSP (OAI Dublin Core)

Dublin Core fields

Title POETICON Multisensory and Multimedia Recordings of Everyday Interaction
POETICON recordings

Creator Institute for Language and Speech Processing/R.C. "Athena"

Subject scenes of everyday life (cleaning, table setting, etc.)

Description The corpus comprises of six everyday human:human interaction scenes, each one performed 3 times by 4 different English-speaking couples (interaction between a male and a female actor), each couple acting each scene in two settings: a fully naturalistic setting in whi+E21ch 5-camera multi-view video recordings take place, and a high-tech setting, with full body motion capture for both individuals, a 2-camera multiview video recording, and 3D tracking of focus objects. All recordings include full language-based interaction (dialogue) which though pre-scripted for providing a guide to the actors, it is natural and spontaneous due to the actors left free to improvise based on the general script lines. Each scene lasts approximately 2-7 minutes depending on the scene and the actors, while the duration of the whole corpus is approximately 12 hours. The scenes are related to activities one may perform in a dinning room/kitchen, such as changing the pot of a plant, cleaning the room, setting the table, preparing a Greek salad, preparing Sangria, and making a parcel.

Publisher Institute for Language and Speech Processing/R.C. "Athena"

Date 2008-2010

Type Collection
MovingImage
Sound
Image
Text

Format text/xml
audio/x-wav
image/jpeg
video/x-ms-wmv
video/x-msvideo
video/mp4

Identifier ilsp.langres:3

Language eng

Rights available for viewing / educational purposes/ academic research / NonCommercial Research

META SHARE

Browse metadata catalogue

Contains all known metadata information

Header fields

Title	POETICON Multisensory and Multimedia POETICON recordings
Date	11/15/2010 at 18:33
Provider	ILSP (OAI Dublin Core)

Dublin Core fields

Title	POETICON Multisensory and Multime POETICON recordings
Creator	Institute for Language and Speech Pr
Subject	scenes of everyday life (cleaning, table setting, etc.)
Description	The corpus comprises of six everyday human:human interaction scenes, each one performed 3 times by 4 different English-speaking couples (interaction between a male and a female actor), each couple acting each scene in two settings: a fully naturalistic setting in whi+E21ch 5-camera multi-view video recordings take place, and a high-tech setting, with full body motion capture for both individuals, a 2-camera multiview video recording, and 3D tracking of focus objects. All recordings include full language-based interaction (dialogue) which though pre-scripted for providing a guide to the actors, it is natural and spontaneous due to the actors left free to improvise based on the general script lines. Each scene lasts approximately 2-7 minutes depending on the scene and the actors, while the duration of the whole corpus is approximately 12 hours. The scenes are related to activities one may perform in a dinning room/kitchen, such as changing the pot of a plant, cleaning the room, setting the table, preparing a Greek salad, preparing Sangria, and making a parcel.
Publisher	Institute for Language and Speech Processing/R.C. "Athena"
Date	2008-2010
Type	Collection MovingImage Sound Image Text
Format	text/xml audio/x-wav image/jpeg video/x-ms-wmv video/x-msvideo video/mp4

META-SHARE: Next Steps

Implementation Level

- **META-SHARE Version 1: July 2011**
  - Stable, working version of META-SHARE to be rolled out within the META-NET network.
- **META-SHARE Version 2: February 2012**
  - Stable version, ready for production use.

Increase your share in
META-SHARE!

It’s simple! It’s free! It’s yours!

META NET

Thank you!