# "META-SHARE: An Open Resource Exchange Infrastructure for Stimulating Research and Innovation"

Stelios Piperidis

Athena RC, Greece
spip@ilsp.athena-innovation.gr

META FORUM 2011 Solutions for Multilingual Europe
Budapest, Hungary, June 27/28 eu 2011.hu

Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME, CESAR, METANET4U, META-NORD (grant agreements no. 249119, 271022, 270893, 270899).

# Outline

- META-SHARE mission and principles
- Architecture
- META-SHARE metadata model
- Legal framework
- META-SHARE towards version 1

# Introduction

- 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)
  - …
- But…data collection, cleaning, annotation, curation and maintenance is a very costly business
- As evidence from other domains (e.g. biotechnology, geodata) shows data and tools become valuable through opening and sharing.

# META-SHARE: objectives

- META-SHARE tries to treat problems and fill in gaps related to **visibility, documentation, identification, availability, preservation, interoperability**, etc. in the area of language data and (basic language processing) tools

- It launches a rather **long-term endeavour** by which language resources are recognised as important assets that can **boost research, technology and innovation** through **wide availability, pooling, openness and sharing**

# 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

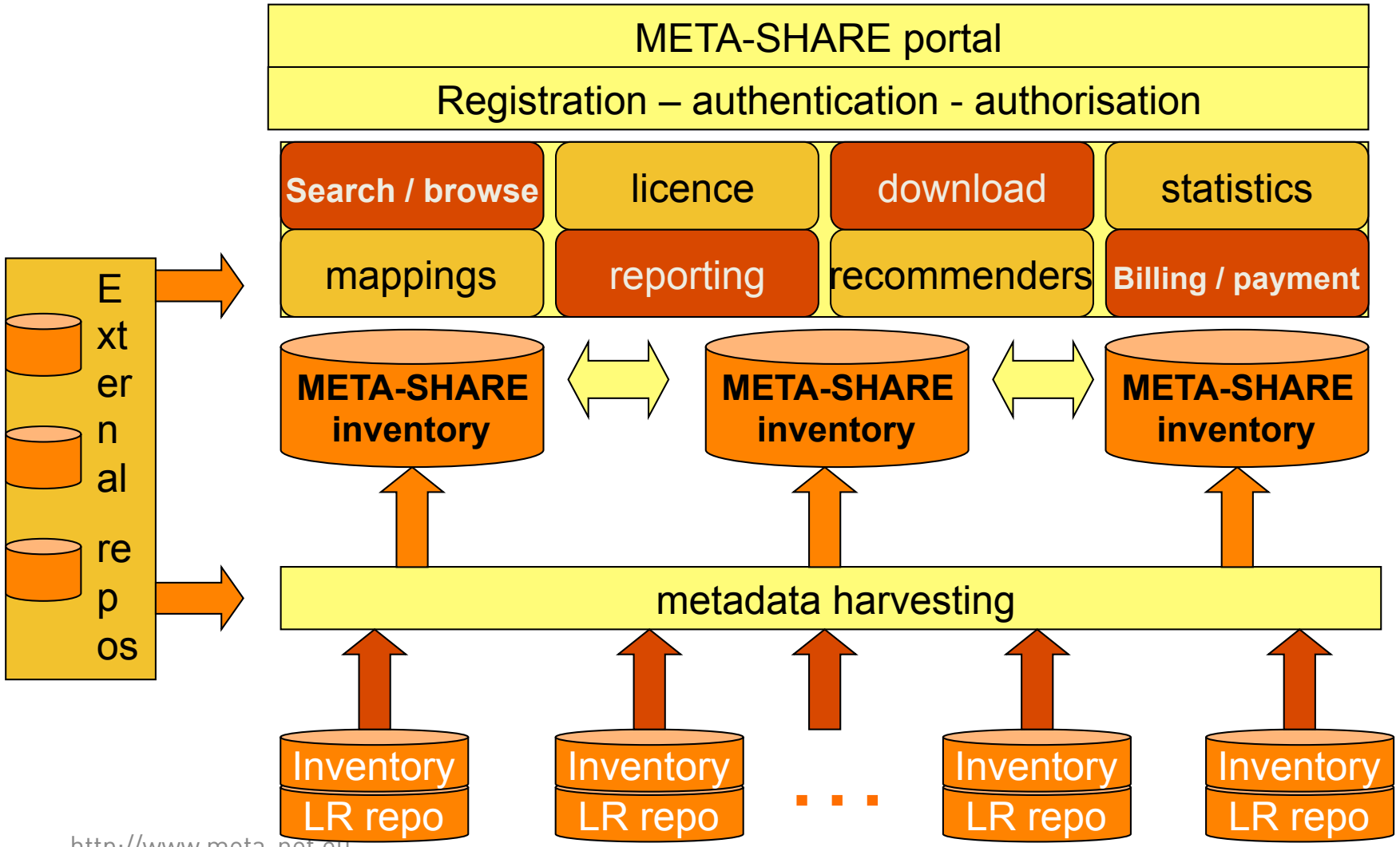
# META-SHARE architecture

- 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.

# META-SHARE architecture (2)

- Actual LRs and their metadata (MD) reside in the local repositories.
  - Rights of use and related restrictions under the control and responsibility of LR owners and the repository where the LR resides
- 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, with synchronised inventories at all times
- Central servers create, host and maintain a central inventory with all MD descriptions of all LRs available in the distributed network.

# Architecture



# META-SHARE Metadata Model

# Main features 1/2

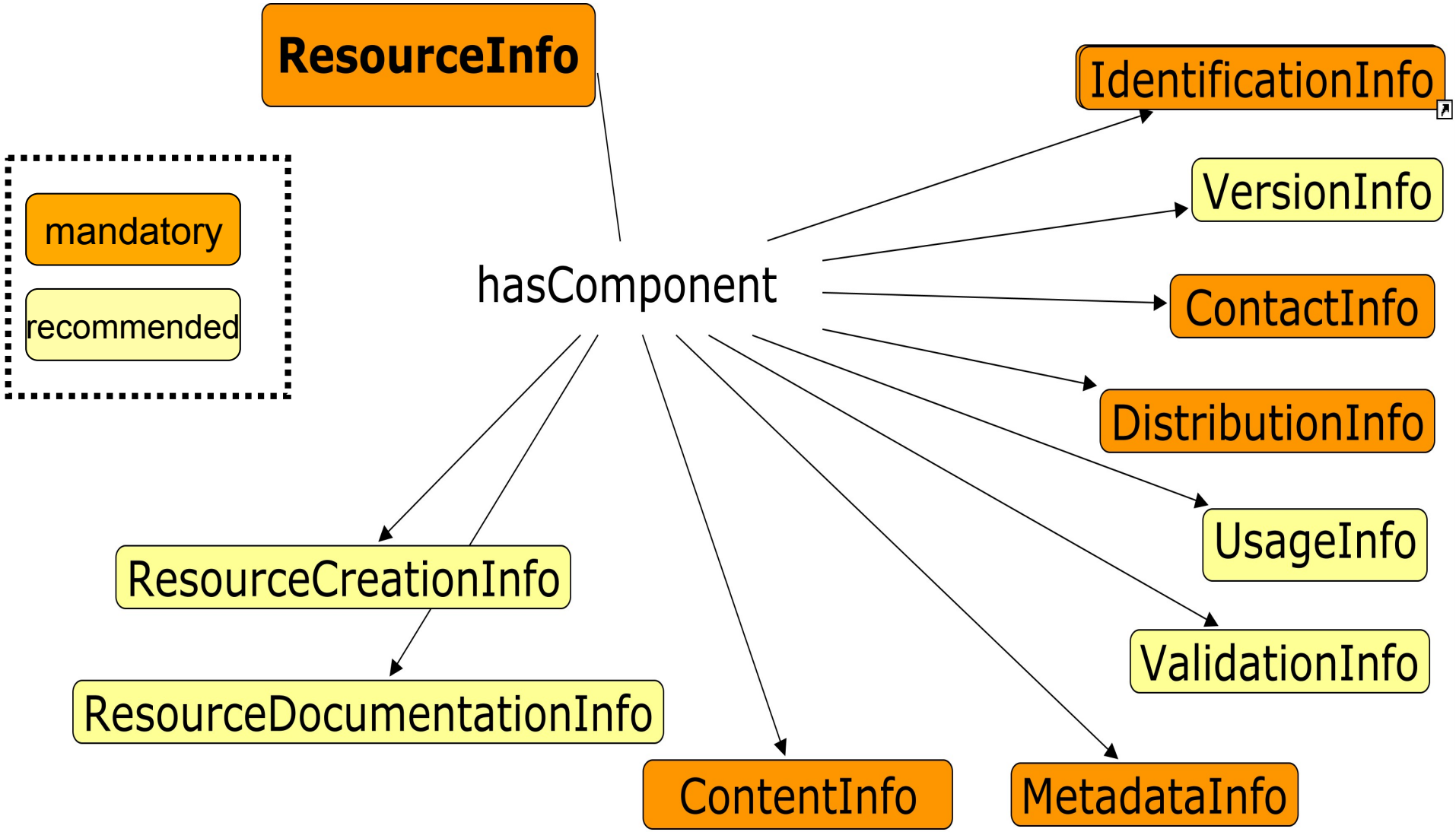
- Descriptions of
  - **LRs**, encompassing both **data** (*textual, multimodal/multimedia and lexical*) and **tools/technologies** used for their processing
  - **related objects** (*reference documents, actors, activities* etc.)

- structure:
  - "***components***" are used to group together elements and relations (and other components)
  - elements linked to ISOcat Data Categories
  - "***profiles***" for each LR type are built upon components, incl.
  - components common to all LR types (e.g. identification, contact, distribution etc.)
  - LR type-specific components (e.g. annotation)

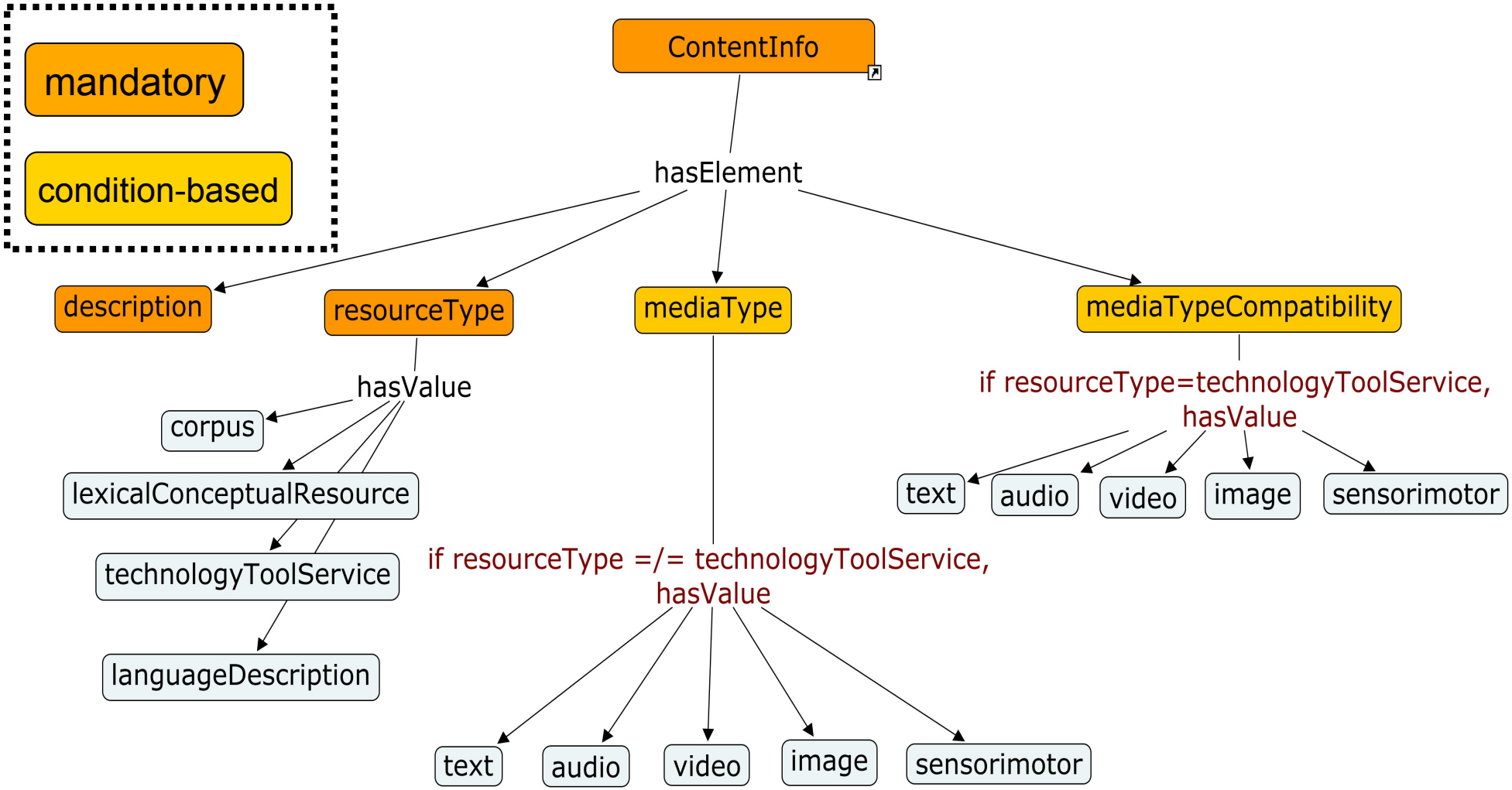
# Main features 2/2

- distinguishes:
  - ***minimal schema***:
    - minimum set of obligatory elements and relations required for effective LR search, identification and retrieval
    - e.g.: identification info (title, unique identifier), contact details, technical info (language, text encoding, size etc.), distribution info
  - ***maximal schema***
    - all elements and relations required for the description of LRs (i.e. additional set of recommended and optional elements and relations for the whole lifecycle of LR production and usage)
    - e.g.: provenance information, creation details, validation/evaluation information, intended and actual use(s) etc.

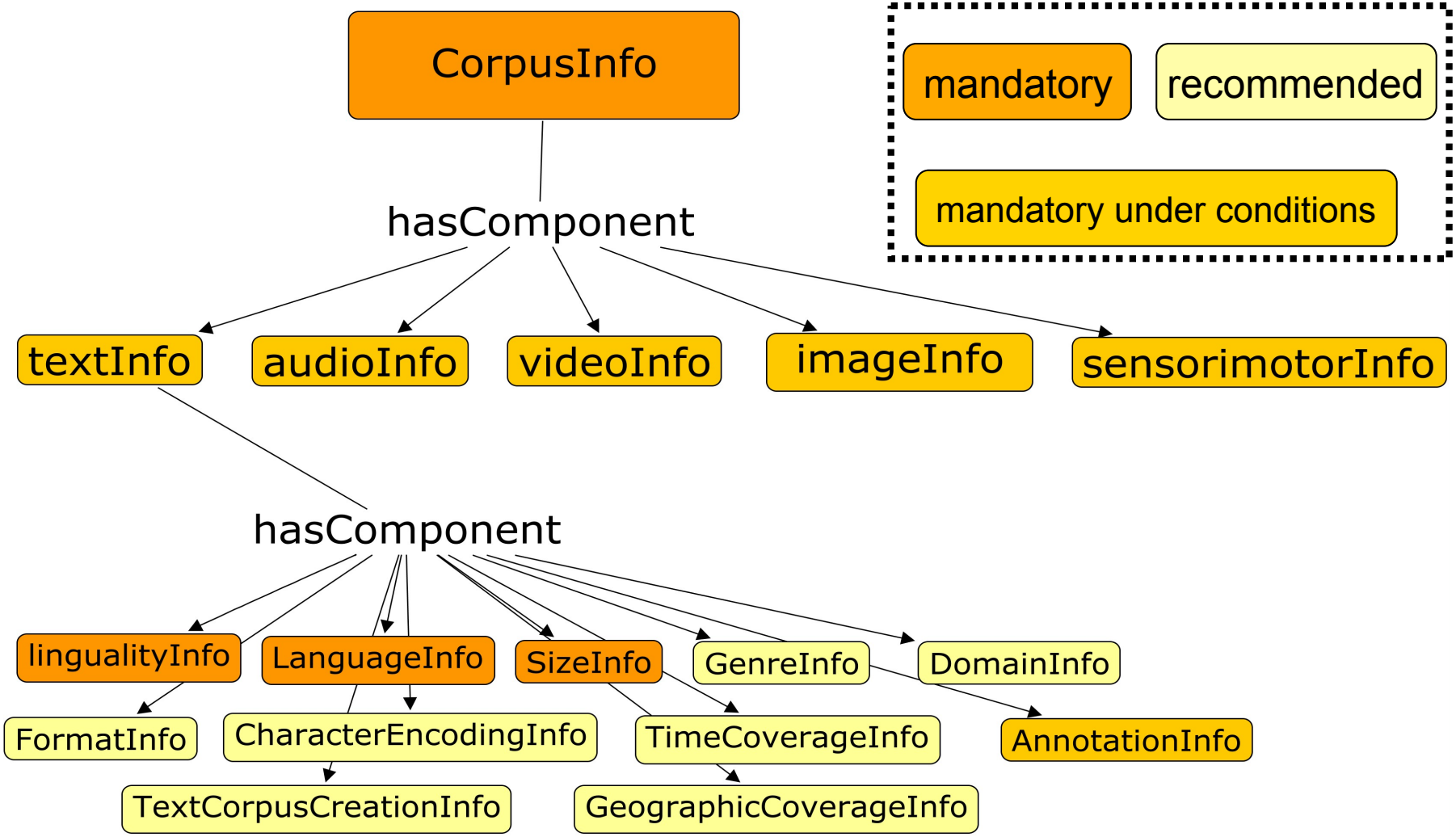
# Metadata upper level



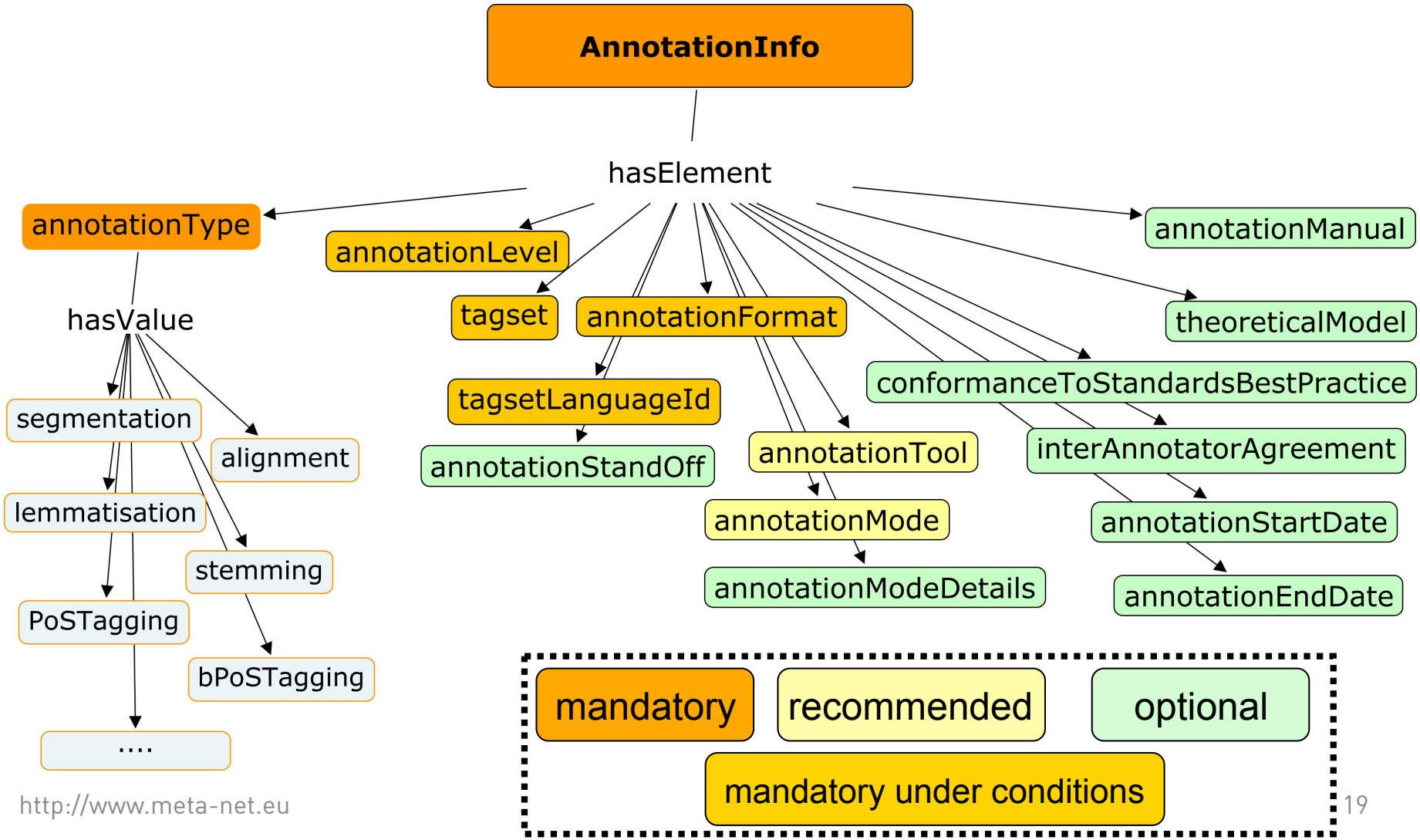
# Content component



# Corpus Component



# Annotation Component



# META-SHARE legal framework

# Language Resources Sharing Charter

- **Aim:** to give a clear signal to
  - Language Resource providers and users,
  - the Language Technology community,
  - market players, policy makers and the public

- that in the digital world LRs should be **shared** and further **re-used** with the **minimum** possible **transaction costs** and **efforts** and under **clear** and easy to understand **rules**

- META-SHARE fully endorses and supports this vision.

# Guidelines 1/2

- **Infrastructure**
  - LRs described with metadata
    - standardized, ideally open to fair harvesting, with web interfaces allowing search and browsing.
  - LRs persistently stored in open and documented format.

- **Standardisation and interoperability**
  - Open Standards and best practices should be used for data sets, tools and metadata, if available.
  - Data formats have to be standardized and ideally open.
  - Software has to use open interfaces and be ideally Open Source.

# Guidelines 2/2

- **Availability / Licensing**
  - All necessary Intellectual Property Rights have to be cleared.
  - If commercial licenses are used, they have to be standardized.
  - Copyright limitations /exceptions need to be harmonised across the EU.
  - Restrictions to use or re-use LRs should be the minimum possible.
  - Public Domain content → no additional restrictions.
  - If LRs are produced entirely with public funding → open or shared at least for research purposes.
  - Any LR infrastructure should support all possible business models without imposing restrictions to market entry.

- **Ethical / privacy issues**
  - Adopt effective privacy policies for anonymisation and consent management.

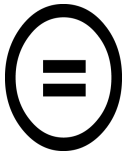
# META-SHARE Legal Domain

- **META-SHARE model licenses** support for
  - Opening
    - Creative Commons –based licences (starting with Creative Commons Zero (CC-0) licences and all possible combinations)
  - Sharing
    - META-SHARE network licences (templates to be used by members when they wish to make available their resources to other network members)
    - Building on CC concepts
    - Re-deposition, as the main driver for e.g. collaborative annotation
  - recommended standardized commercial licences

# CC and MSC, duties/rights of use

- Attribution

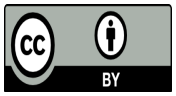
- Non Commercial



- No Derivative

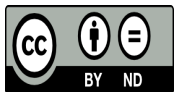
- ShareAlike

# Module Combinations

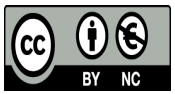



- Attribution


- Attribution – Share Alike




- Attribution – No Derivatives




- Attribution – Non Commercial


- Attribution – Non Commercial – ShareAlike


- Attribution – Non Commercial – No Derivatives

# META-SHARE version 1

www.meta-share.eu,
www.meta-share.org
www.meta-net.eu/meta-share

# 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 w… nodes/centres and provide more functionality with the goal of turning into an as largely distributed infrastructure as p…

Single Sign On enabling hopping from repository to repository

## 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

Collaborating PSP projects — CESAR, META-NORD, METANET4U


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

META SHARE

Logout

Username: cfedermann

# Welcome to Meta-Share!

Type in your keywords, please...

Search | Advanced Search


## What is it? - About the project

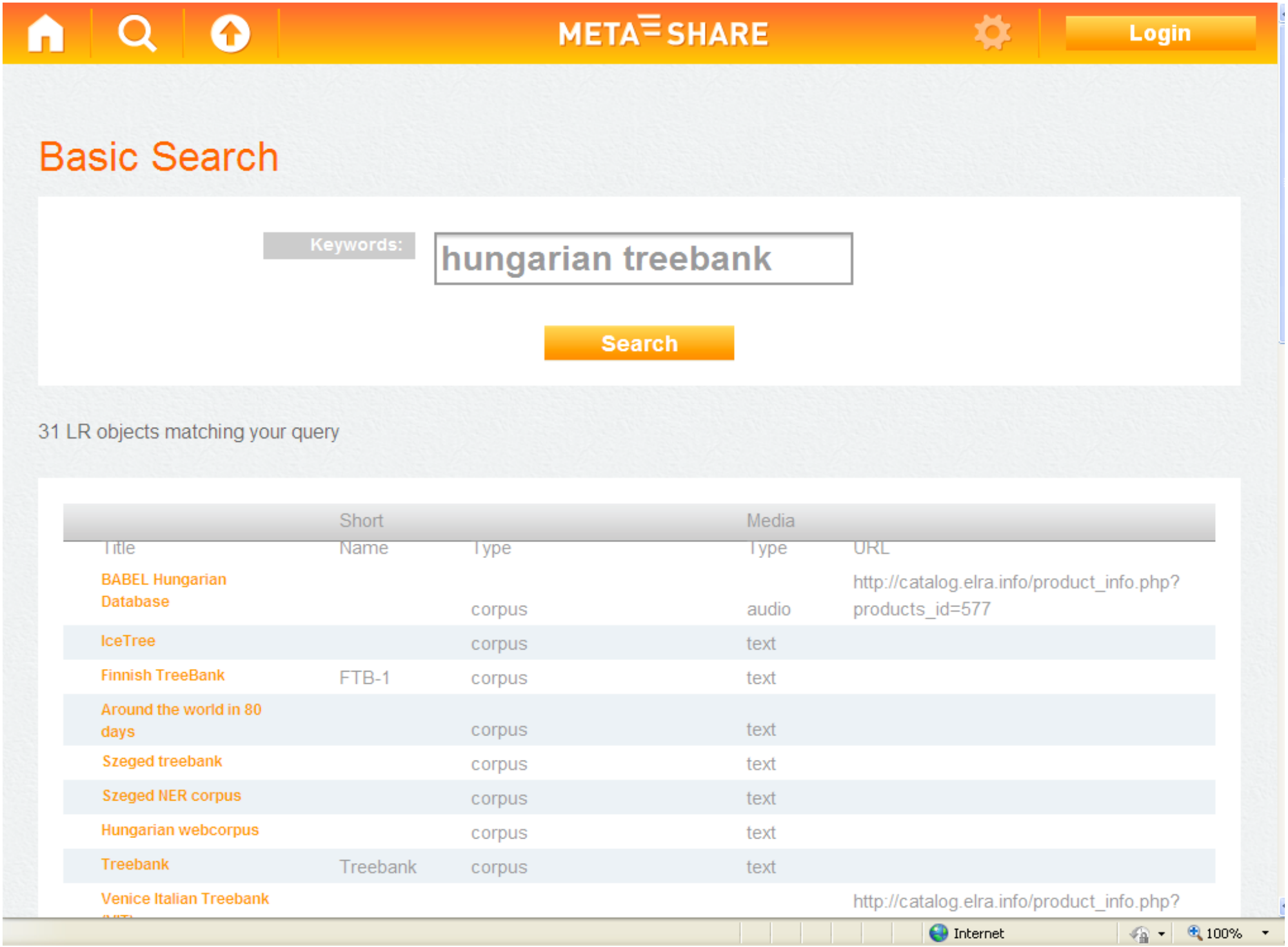
META-NET aims at creating META-SHARE, a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources. Data and tools can be both open and with restricted 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. In this respect, reuse, combination, repurposing and re-engineering of language data and tools play a crucial role.

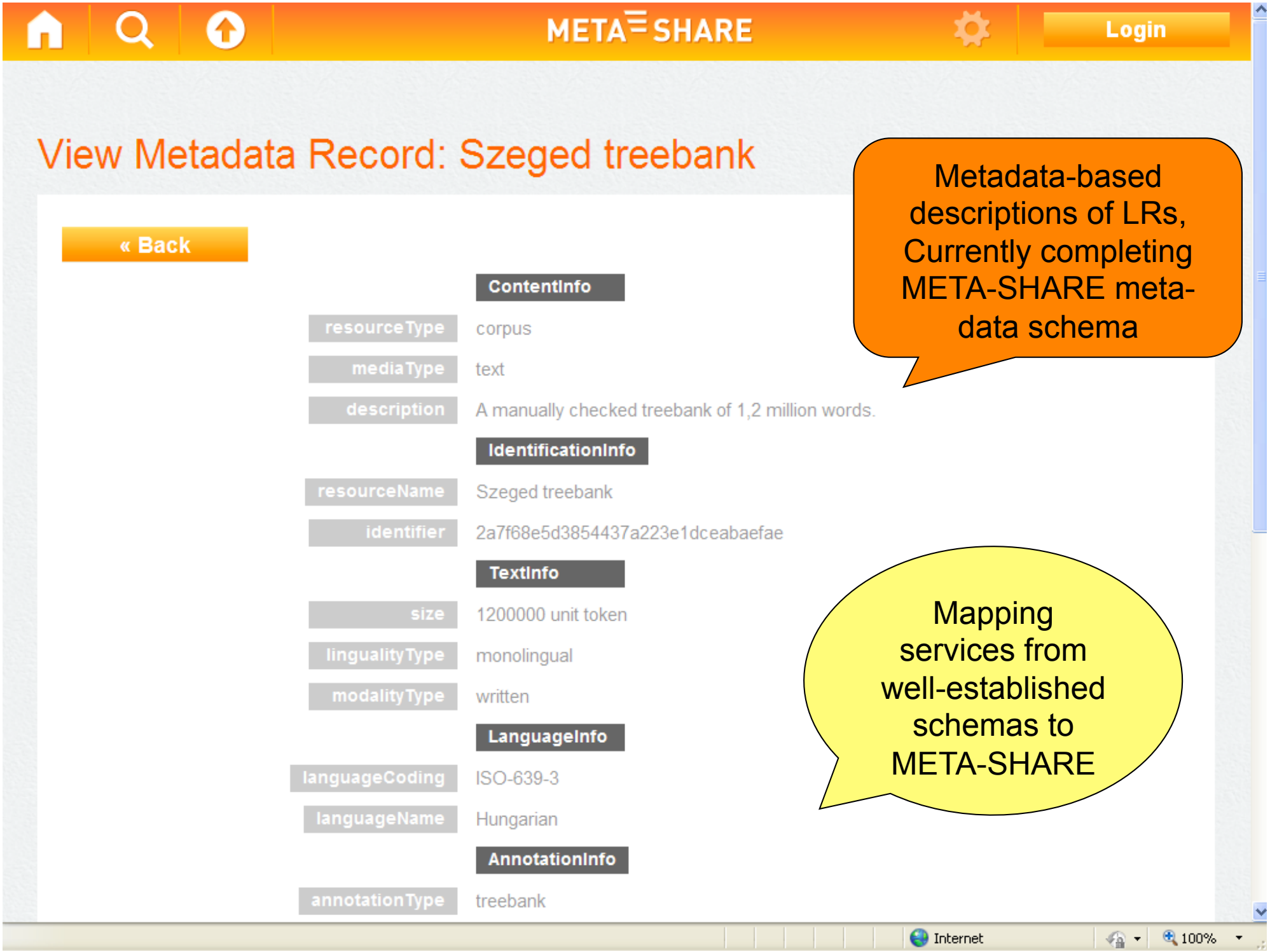
META-SHARE will eventually be an important component of a language technology marketplace for HLT researchers and developers, language professionals (translators, interpreters, content and software localisation experts, etc.), as well as for industrial players, especially SMEs, catering for the full development cycle for HLT, from research through to innovative products and services. Where the work or any of its elements is in public domain under applicable law, that status is in no way affected by the license.

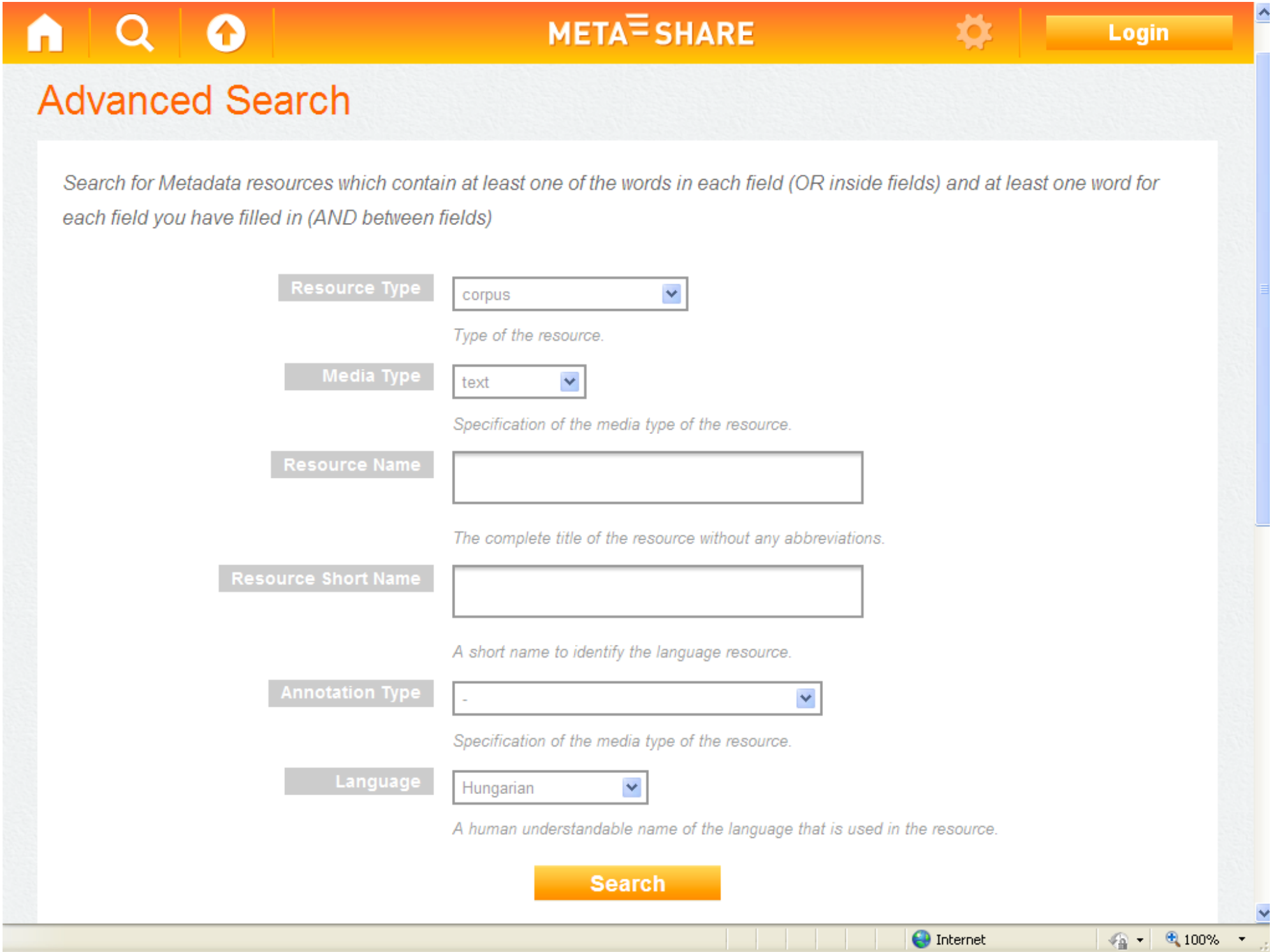
Go to top

## How can I use it? - First steps

More specifically, META-SHARE will be a freely available facility supported by a large user and developer community, based on a distributed networked repositories accessible through common interfaces. Users (consumers, providers


META SHARE
Login
Basic Search
Keywords:
hungarian treebank
Search
31 LR objects matching your query
Title
Short Name
Type
Media Type
URL
BABEL Hungarian Database
corpus
audio
http://catalog.elra.info/product_info.php?products_id=577
IceTree
corpus
text
Finnish TreeBank
FTB-1
corpus
text
Around the world in 80 days
corpus
text
Szeged treebank
corpus
text
Szeged NER corpus
corpus
text
Hungarian webcorpus
corpus
text
Treebank
Treebank
corpus
text
Venice Italian Treebank
http://catalog.elra.info/product_info.php?
Internet
100%


META SHARE
Login
View Metadata Record: Szeged treebank
« Back
ContentInfo
resourceType
corpus
mediaType
text
description
A manually checked treebank of 1,2 million words.
IdentificationInfo
resourceName
Szeged treebank
identifier
2a7f68e5d3854437a223e1dceabaefae
TextInfo
size
1200000 unit token
lingualityType
monolingual
modalityType
written
LanguageInfo
languageCoding
ISO-639-3
languageName
Hungarian
AnnotationInfo
annotationType
treebank
Metadata-based descriptions of LRs, Currently completing META-SHARE meta-data schema
Mapping services from well-established schemas to META-SHARE
Internet
100%


META SHARE
Login
Advanced Search
Search for Metadata resources which contain at least one of the words in each field (OR inside fields) and at least one word for each field you have filled in (AND between fields)
Resource Type
corpus
Type of the resource.
Media Type
text
Specification of the media type of the resource.
Resource Name
The complete title of the resource without any abbreviations.
Resource Short Name
A short name to identify the language resource.
Annotation Type
-
Specification of the media type of the resource.
Language
Hungarian
A human understandable name of the language that is used in the resource.
Search
Internet
100%

Language: Hungarian

*A human understandable name of the language that is used in the resource.*

Search

8 LR objects matching your query

| Title | Short Name | Type | Media Type | URL |
|---|---|---|---|---|
| Hungarian National Corpus | HNC | corpus | text | |
| Hunglish parallel corpus | | corpus | text | |
| Szeged corpus | | corpus | text | |
| Around the world in 80 days | | corpus | text | |
| Szeged treebank | | corpus | text | |
| Szeged NER corpus | | corpus | text | |
| Hungarian webcorpus | | corpus | text | |
| Named entity lexical database | | corpus | text | |

Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME (grant agreement no.: 249119), CESAR (grant agreement no.: 271022), METANET4U (grant agreement no.: 270893) and META-NORD (grant agreement no.: 270899). 

# View Metadata Record: Hunglish parallel corpus

**ContentInfo**

| | |
|---|---|
| resourceType | corpus |
| mediaType | text |

**IdentificationInfo**

| | |
|---|---|
| resourceName | Hunglish parallel corpus |
| identifier | 2c788914e2f84613b8cf3ac17ee5edc5 |

**SizeInfo**

| | |
|---|---|
| size | 1200000 unit token |
| lingualityType | 1 |
| modalityType | written |

**LanguageInfo**

| | |
|---|---|
| languageCoding | ISO-639-3 |

**AnnotationInfo**

| | |
|---|---|
| annotationType | MSD |
| annotationStandoff | False |

**ValidationInfo**

| | |
|---|---|
| validated | False |

**MetadataInfo**

| | |
|---|---|
| source | CESAR |
| metadataCreationDate | June 23, 2011 |

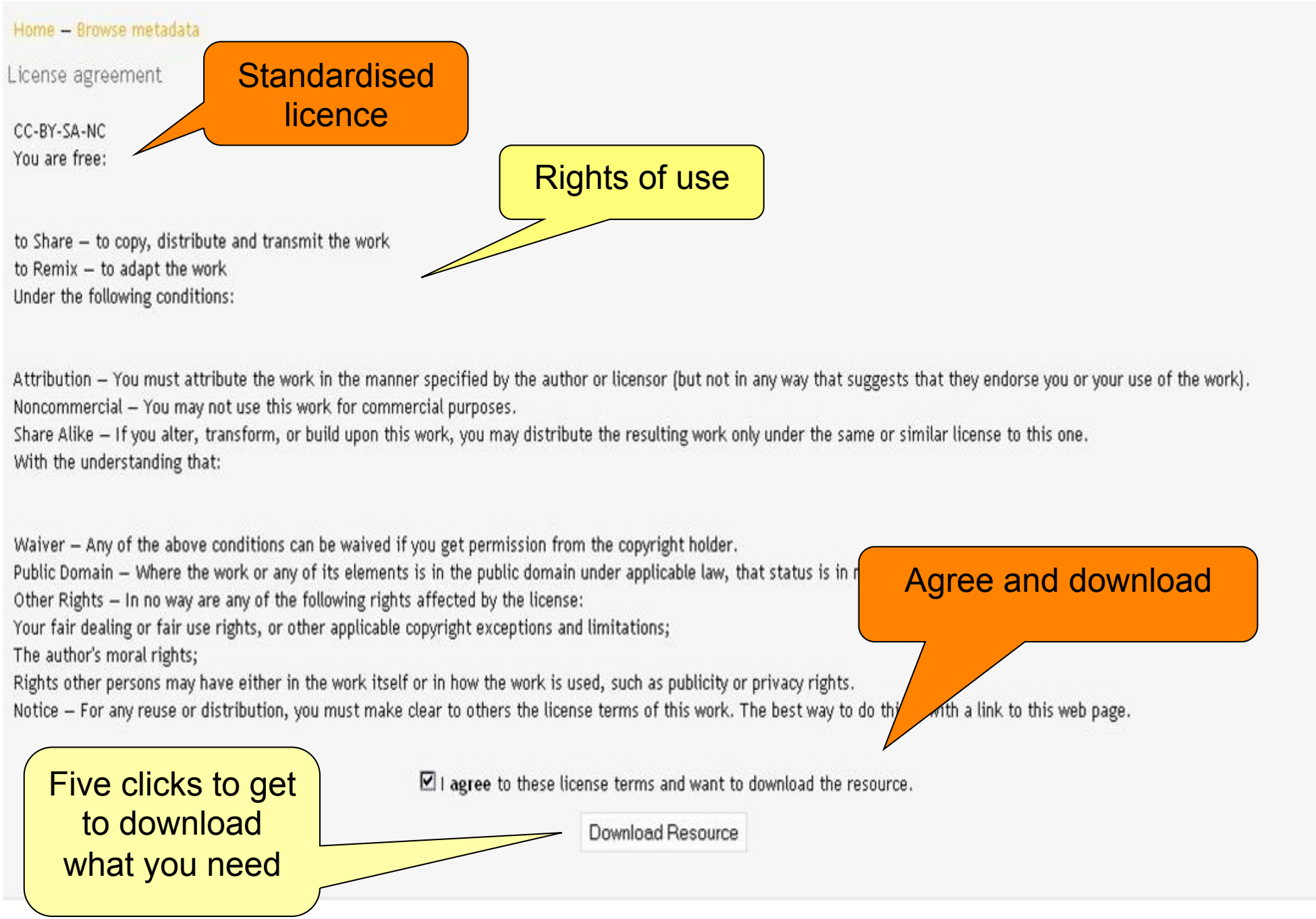
**DistributionInfo**

| | |
|---|---|
| availability | available-unrestrictedUse |

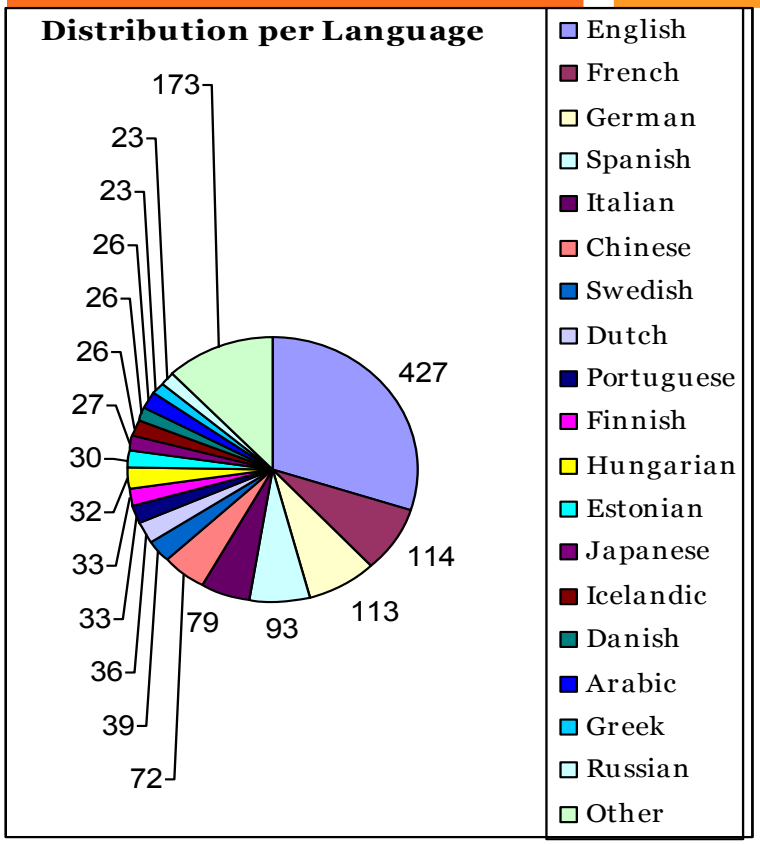
**LicenseInfo**

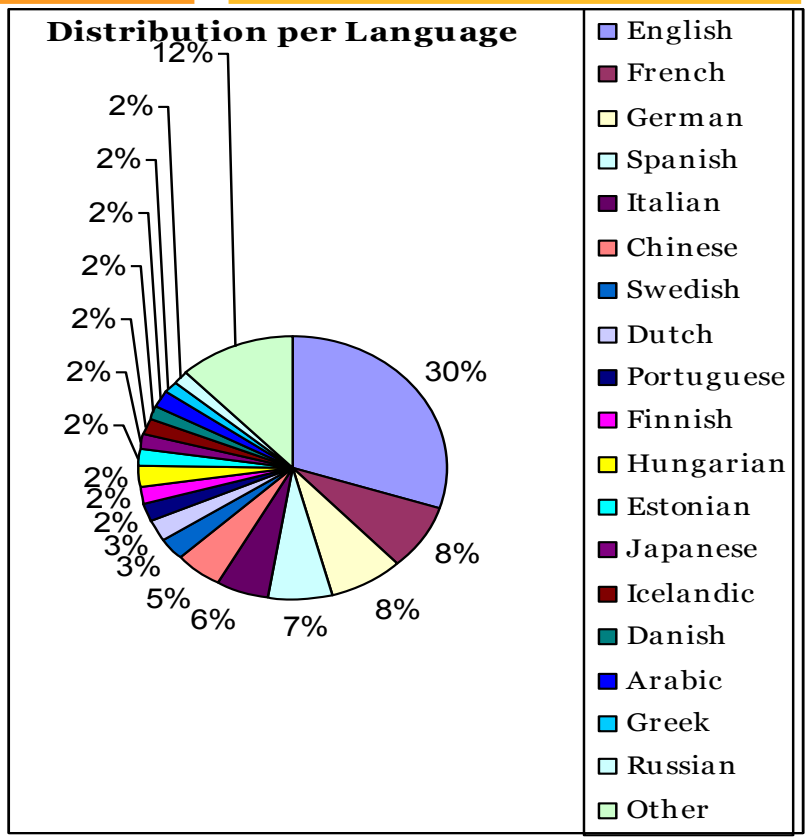
| | |
|---|---|
| license | CC BY |
| restrictionsOfUse | attribution |

Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME (grant agreement no.: 249119), CESAR (grant agreement no.: 271022), METANET4U (grant agreement no.: 270893) and META-NORD (grant agreement no.: 270899). 


Home – Browse metadata
License agreement
Standardised licence
CC-BY-SA-NC
You are free:
Rights of use
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
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 th
ith a link to this web page.
Agree and download
I agree to these license terms and want to download the resource.
Download Resource
Five clicks to get to download what you need

# Population round 0




As of 23/6/2011 : 1425 LR packages catering for 40 Languages

Distribution per language, # of LR packages >=20

# From now on 1/3

Implementation Level

- **META-SHARE Version 1: July 2011**
  - Stable, working version of META-SHARE to be rolled out within the META-NET network.
  - Focus on laying the distributed repository infrastructure repository by making available free sw for repository setup to help organise, document LRs
- **META-SHARE Version 2: February 2012**
  - Stable version, ready for production use.
  - Focus on providing services to LR providers, consumers and aggregators (statistics, reporting, recommendations)

# From now on 2/3

Implementation Level

- Further future versions will provide services to alleviate any manual effort (through scrapers, mapping tools, harvesting, interoperability / conversion programs) Just show us where the relevant information exists and we will do the rest!!

# From now on 3/3

Principles, Operation and Governance Level

- Gathering community feedback on sharing principles
- Testing the applicability and suitability of recommended licencing instruments in an as wide as possible range of resource types, countries and jurisdictions
- Populating repositories with well documented LRs

# Today and tomorrow

Come and visit the META-SHARE booth in

Room Margit, Desk 8


# Thank you!