META-SHARE
the Open Resource Exchange Facility

Stelios Piperidis

ILSP-Athena RC, Greece
spip@ilsp.gr

META-FORUM 2010: Challenges for Multilingual Europe
Brussels, Belgium, November 17/18, 2010

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

- Language research and language technology belong to the **Data Intensive Sciences**
- Data collection, cleaning, annotation, curation, maintenance, etc is a very costly business
- Data become considerably valuable through sharing.
- However, the long demanded and well-contemplated instruments for managing and sharing this data are *still missing*.

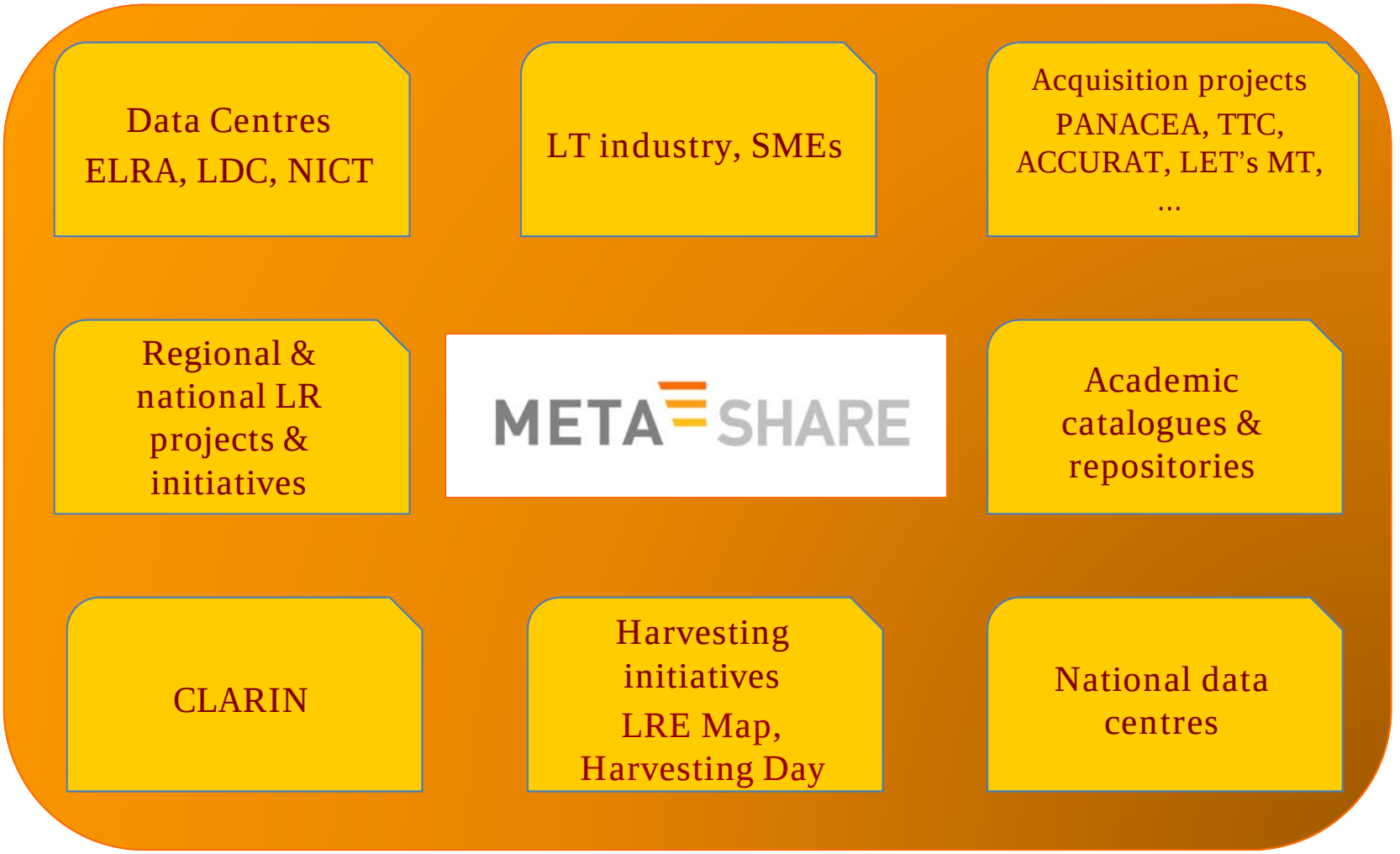
META-SHARE: Key Features

- 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,
...
Regional &
national LR
projects &
initiatives
META-SHARE
Academic
catalogues &
repositories
CLARIN
Harvesting
initiatives
LRE Map,
Harvesting Day
National data
centres

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.

Metadata Schema

- External metadata (description of resources)

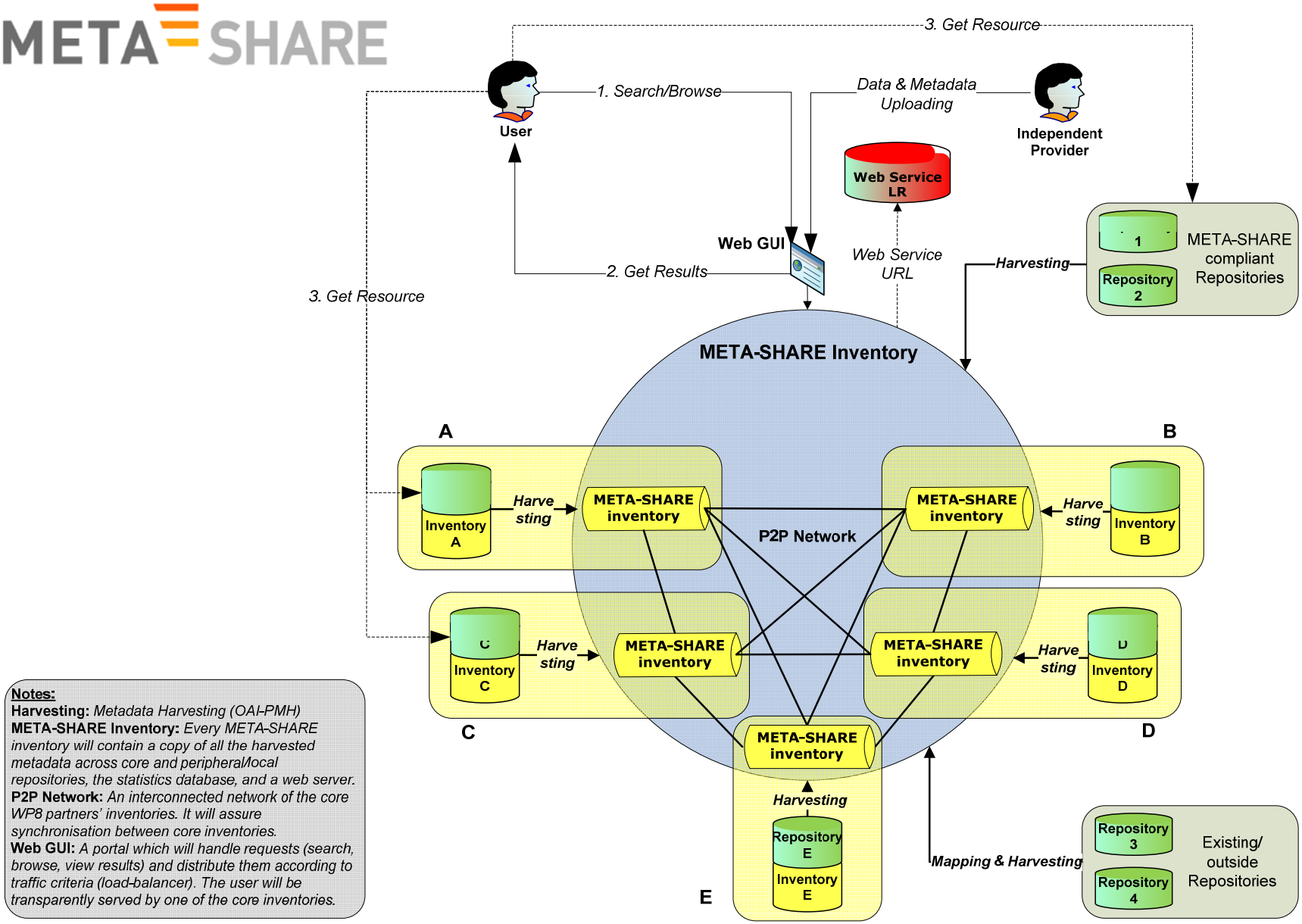
- We're not reinventing the wheel: harmonize existing schemas and adapt them to the requirements of the HLT community
- Mappers for widespread schemas

- Ready-to-be-used profiles depending on the type of a resource
- Metadata are component based

- Main desiderata:
  - clarity of semantics
  - flexibility
  - interoperability
  - extensibility
  - expressiveness
  - customisability
  - user friendliness
  - harvestability

META-SHARE architecture (3)

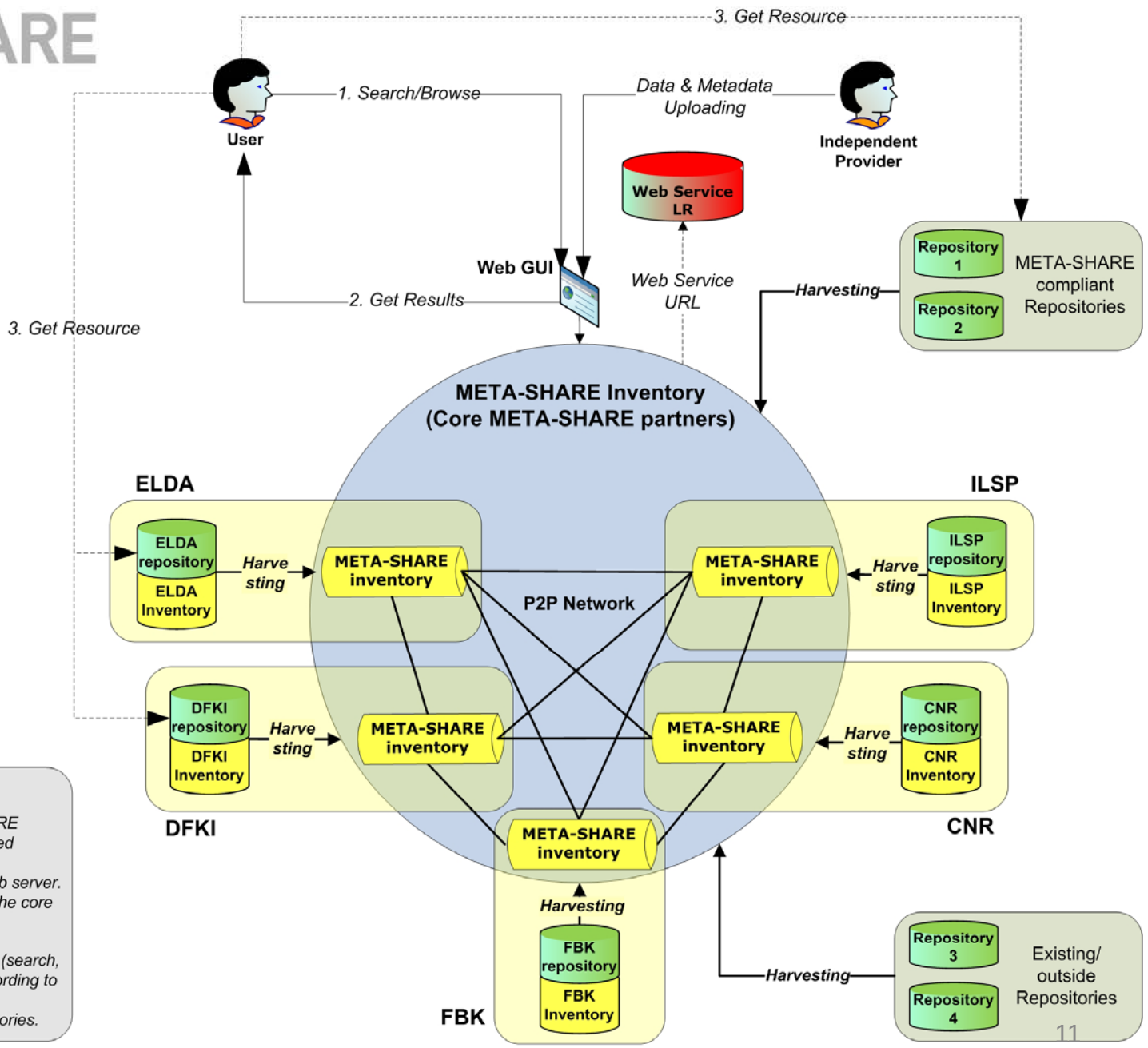
- Users (language resources seekers/consumers) will be able to
  - **log-in once** www.meta-share.eu or www.meta-share.org
  - **search** the central inventory using multifaceted search facilities, and
  - access the actual resources by visiting the **local** (or **non-local**) repositories for **browsing 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



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.*

META SHARE

Version 0



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.*

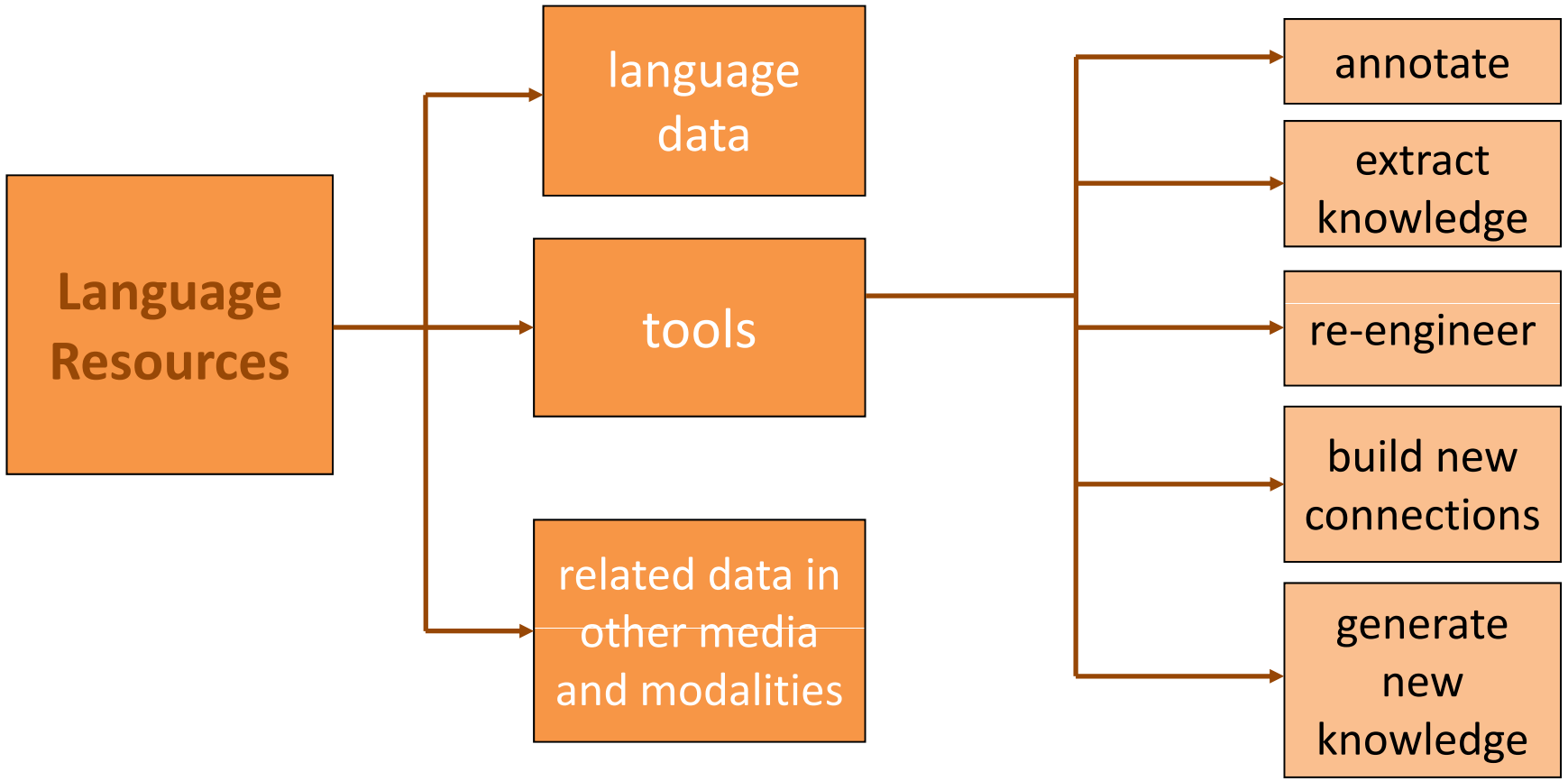
P2P Network: *An interconnected network of the core WP8 partners' inventories. It will assure synchronisation between core inventories.*

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.*

Steps of integration

- Start by integrating relatively few nodes/centres, notably those represented by the partners of the META-NET network
- Gradually extend to encompass more nodes/centres and provide more functionality (richer metadata, recommendation services, collaboration facilities, etc.),
- Turning into an as largely distributed infrastructure as possible as the project progresses.

In the future, within META-SHARE…



In a nutshell :
META-SHARE is now offering

- A channel to share and distribute language data and tools.
- Technical solutions for building your own repositories.
- Protocols and mechanisms for making the descriptions of your resources (and the actual resources) harvestable.
- Guidelines and recommendations on standards used in the LR production and documentation processes.
- Recommendations on data and tools licensing issues.
- **Access to large catalogues of documented, high-quality resources, as well as the actual data and tools.**

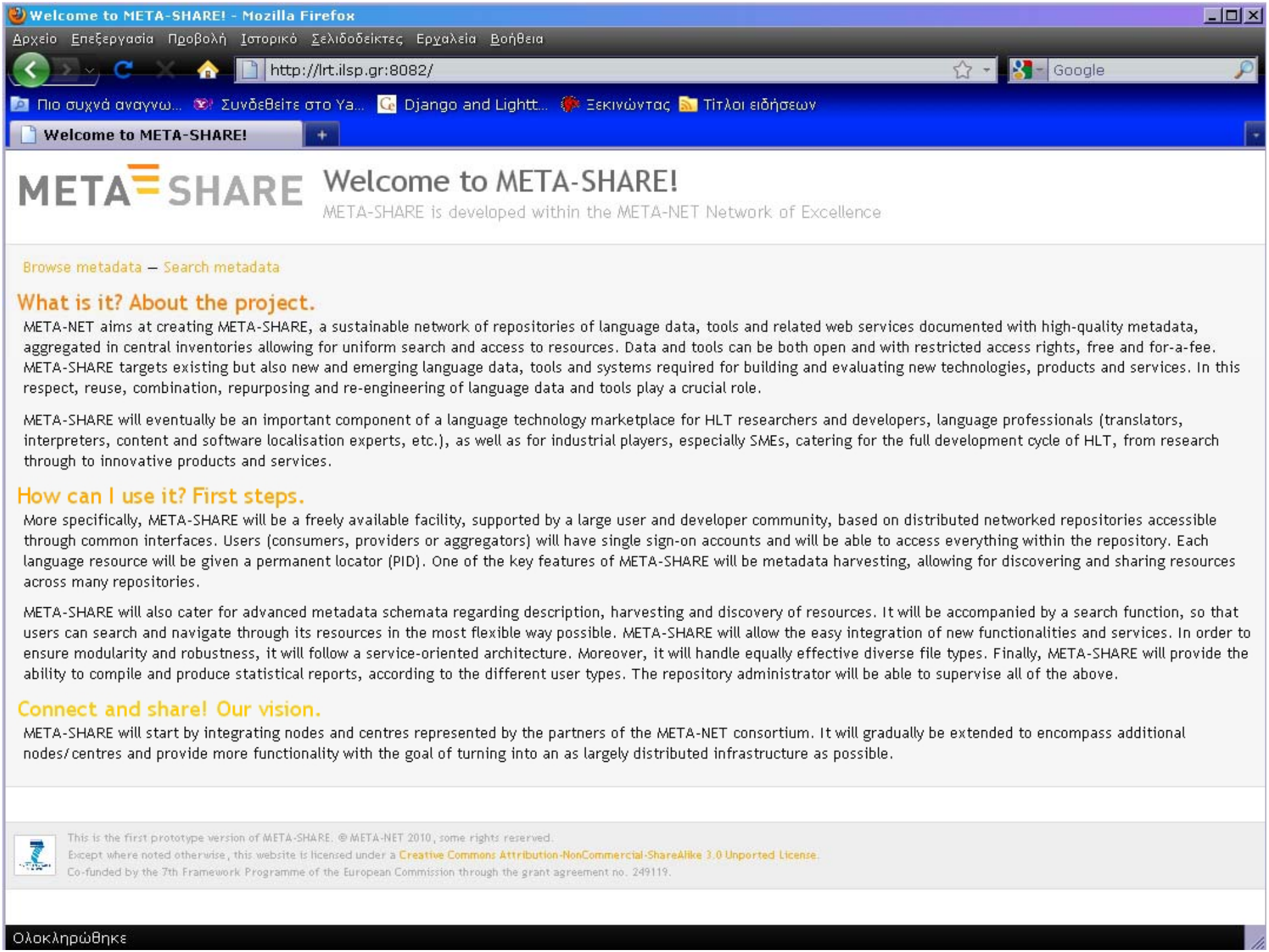
Features

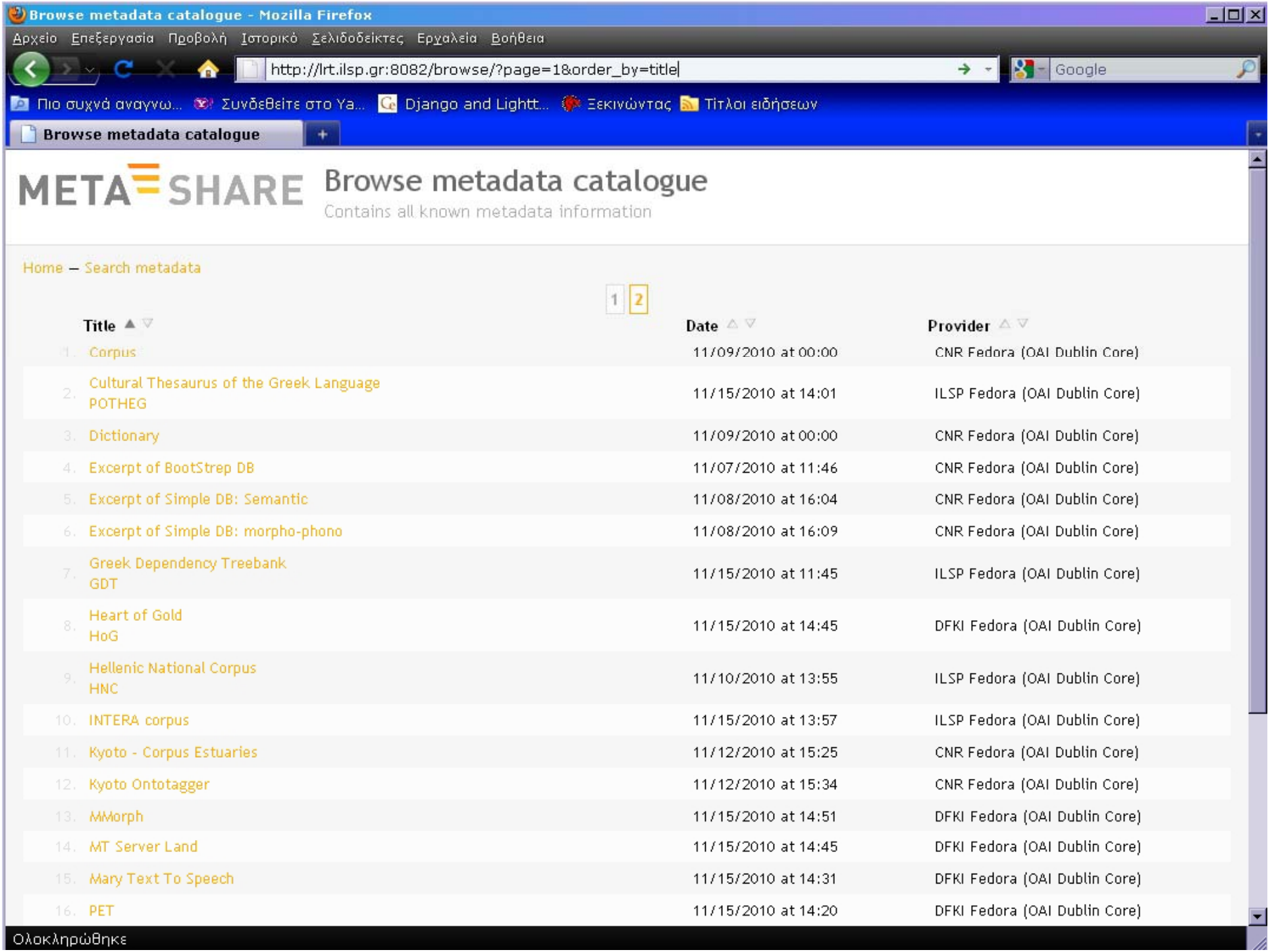
- Open Source
- Service-Oriented
- Distributed
- Replication/Backup
- Reporting & Statistics
- Single Sign-On
- Easy Administration
- Metadata Harvesting
- Persistent Identifiers (PIDs)
- Intuitive Search

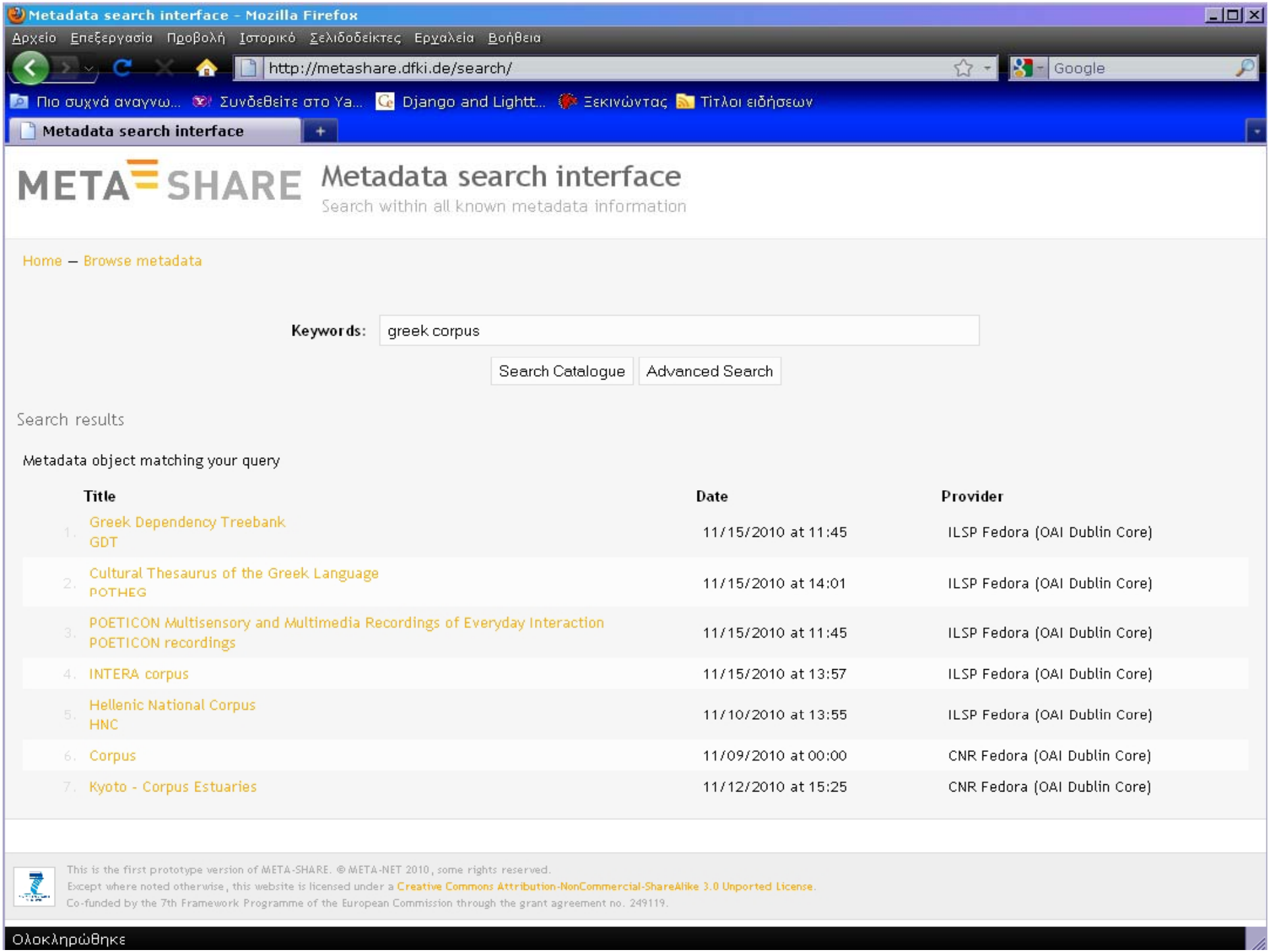
Sneak Peak

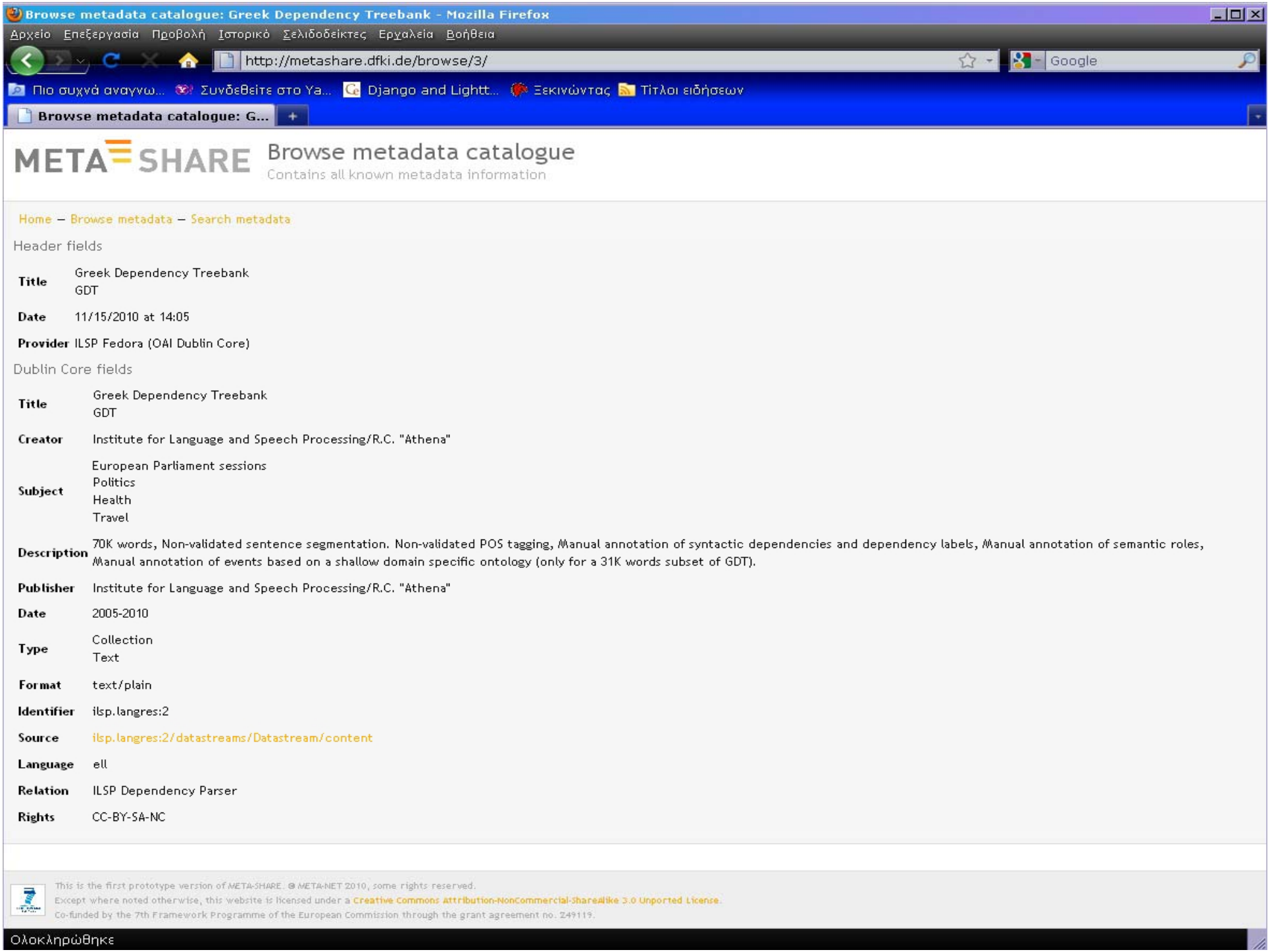
Version 0

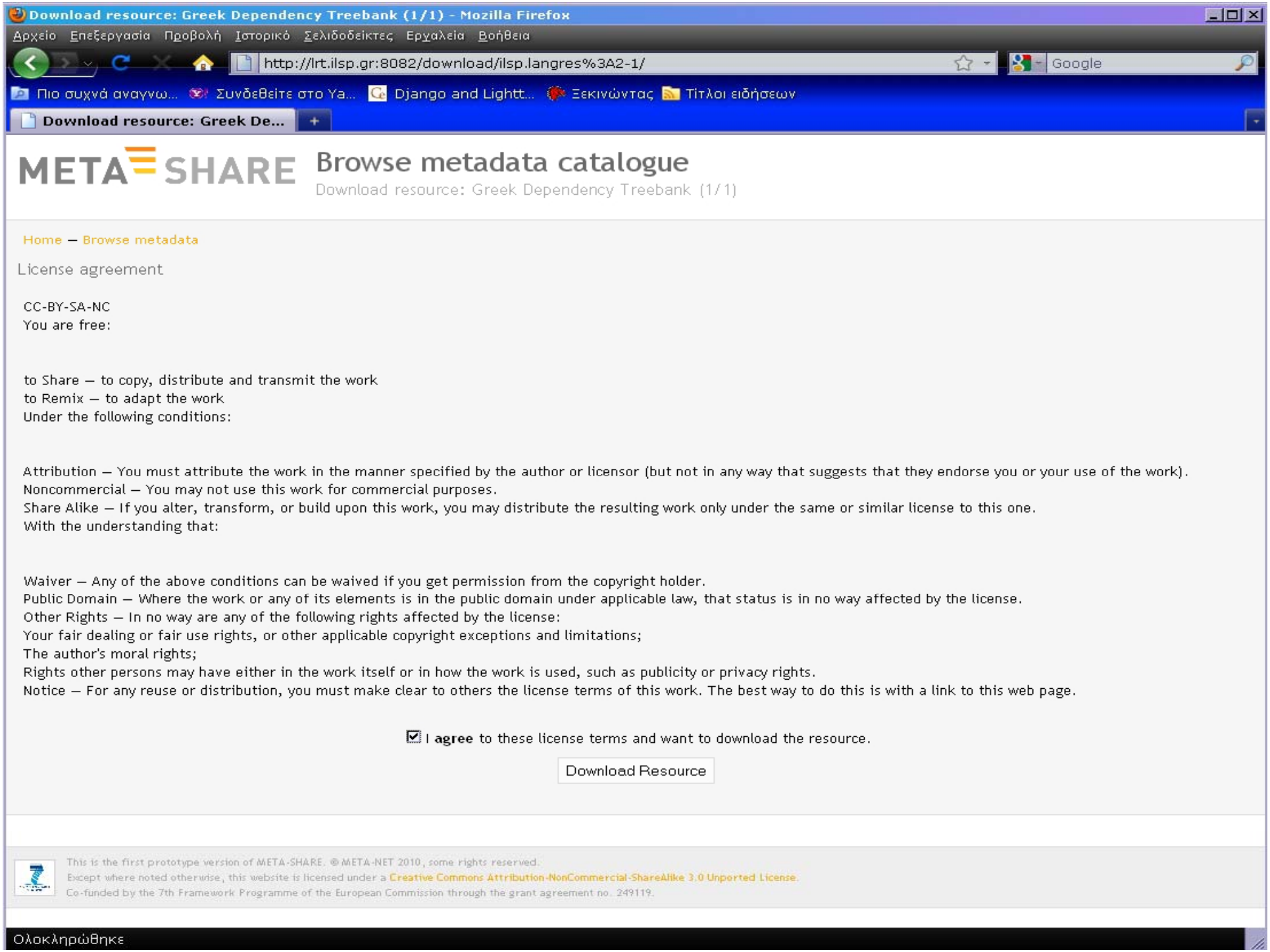

Welcome to META-SHARE! - Mozilla Firefox
Αρχείο Επεξεργασία Προβολή Ιστορικό Σελιδοδείκτες Εργαλεία Βοήθεια
http://www.meta-share.org/
Google
Welcome to META-SHARE!
META-SHARE
Welcome to META-SHARE!
META-SHARE is developed within the META-NET Network of Excellence
About the project
META-NET is designing 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.
META-NET
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
This is the first prototype version of META-SHARE. © META-NET 2010, some rights reserved.
Except where noted otherwise, this website is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
Co-funded by the 7th Framework Programme of the European Commission through the grant agreement no. 249119.
http://www.meta-net.eu/

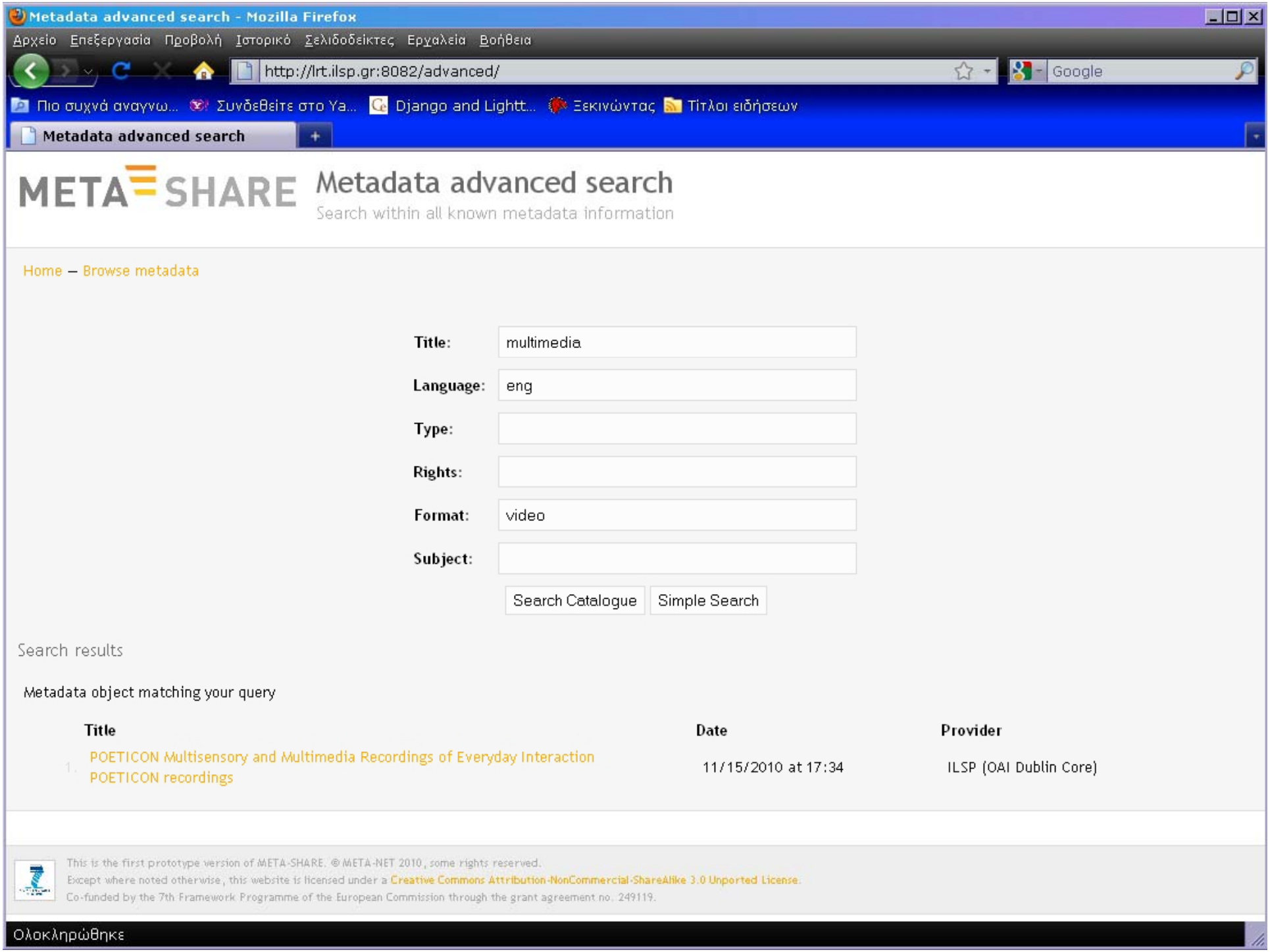

Welcome to META-SHARE! - Mozilla Firefox
Αρχείο Επεξεργασία Προβολή Ιστορικό Σελιδοδείκτες Εργαλεία Βοήθεια
http://lrt.ilsp.gr:8082/
Google
Welcome to META-SHARE!
META SHARE
Welcome to META-SHARE!
META-SHARE is developed within the META-NET Network of Excellence
Browse metadata – Search metadata
What 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 of HLT, from research through to innovative products and services.
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 distributed networked repositories accessible through common interfaces. Users (consumers, providers or aggregators) will have single sign-on accounts and will be able to access everything within the repository. Each language resource will be given a permanent locator (PID). One of the key features of META-SHARE will be metadata harvesting, allowing for discovering and sharing resources across many repositories.
META-SHARE will also cater for advanced metadata schemata regarding description, harvesting and discovery of resources. It will be accompanied by a search function, so that users can search and navigate through its resources in the most flexible way possible. META-SHARE will allow the easy integration of new functionalities and services. In order to ensure modularity and robustness, it will follow a service-oriented architecture. Moreover, it will handle equally effective diverse file types. Finally, META-SHARE will provide the ability to compile and produce statistical reports, according to the different user types. The repository administrator will be able to supervise all of the above.
Connect and share! Our vision.
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.
This is the first prototype version of META-SHARE. © META-NET 2010, some rights reserved.
Except where noted otherwise, this website is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
Co-funded by the 7th Framework Programme of the European Commission through the grant agreement no. 249119.
Ολοκληρώθηκε


Browse metadata catalogue - Mozilla Firefox
Αρχείο Επεξεργασία Προβολή Ιστορικό Σελιδοδείκτες Εργαλεία Βοήθεια
http://lrt.ilsp.gr:8082/browse/?page=1&order_by=title
Google
Πιο συχνά αναγνω... Συνδεθείτε στο Ya... Django and Lightt... Ξεκινώντας Τίτλοι ειδήσεων
Browse metadata catalogue
META SHARE
Browse metadata catalogue
Contains all known metadata information
Home – Search metadata
1 2
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)
Ολοκληρώθηκε


Metadata search interface - Mozilla Firefox
Αρχείο Επεξεργασία Προβολή Ιστορικό Σελιδοδείκτες Εργαλεία Βοήθεια
http://metashare.dfki.de/search/
Google
Πιο συχνά αναγνω... Συνδεθείτε στο Ya... Django and Lightt... Ξεκινώντας Τίτλοι ειδήσεων
Metadata search interface
META SHARE
Metadata search interface
Search within all known metadata information
Home – Browse metadata
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)
This is the first prototype version of META-SHARE. © META-NET 2010, some rights reserved.
Except where noted otherwise, this website is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
Co-funded by the 7th Framework Programme of the European Commission through the grant agreement no. 249119.
Ολοκληρώθηκε


Browse metadata catalogue: Greek Dependency Treebank - Mozilla Firefox
Αρχείο Επεξεργασία Προβολή Ιστορικό Σελιδοδείκτες Εργαλεία Βοήθεια
http://metashare.dfki.de/browse/3/
Google
Πιο συχνά αναγνω... Συνδεθείτε στο Ya... Django and Lightt... Ξεκινώντας Τίτλοι ειδήσεων
Browse metadata catalogue: G...
META SHARE
Browse metadata catalogue
Contains all known metadata information
Home – Browse metadata – Search metadata
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
This is the first prototype version of META-SHARE. © META-NET 2010, some rights reserved.
Except where noted otherwise, this website is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
Co-funded by the 7th Framework Programme of the European Commission through the grant agreement no. 249119.
Ολοκληρώθηκε


Download resource: Greek Dependency Treebank (1/1) - Mozilla Firefox
Αρχείο Επεξεργασία Προβολή Ιστορικό Σελιδοδείκτες Εργαλεία Βοήθεια
http://lrt.ilsp.gr:8082/download/ilsp.langres%3A2-1/
Google
Πιο συχνά αναγνω... Συνδεθείτε στο Ya... Django and Lightt... Ξεκινώντας Τίτλοι ειδήσεων
Download resource: Greek De...
META SHARE
Browse metadata catalogue
Download resource: Greek Dependency Treebank (1/1)
Home — Browse metadata
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
This is the first prototype version of META-SHARE. © META-NET 2010, some rights reserved.
Except where noted otherwise, this website is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
Co-funded by the 7th Framework Programme of the European Commission through the grant agreement no. 249119.
Ολοκληρώθηκε


Metadata advanced search - Mozilla Firefox
Αρχείο Επεξεργασία Προβολή Ιστορικό Σελιδοδείκτες Εργαλεία Βοήθεια
http://lrt.ilsp.gr:8082/advanced/
Google
Πιο συχνά αναγνω... Συνδεθείτε στο Ya... Django and Lightt... Ξεκινώντας Τίτλοι ειδήσεων
Metadata advanced search
META SHARE
Metadata advanced search
Search within all known metadata information
Home – Browse metadata
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)
This is the first prototype version of META-SHARE. © META-NET 2010, some rights reserved.
Except where noted otherwise, this website is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
Co-funded by the 7th Framework Programme of the European Commission through the grant agreement no. 249119.
Ολοκληρώθηκε

META-SHARE

Browse metadata catalogue

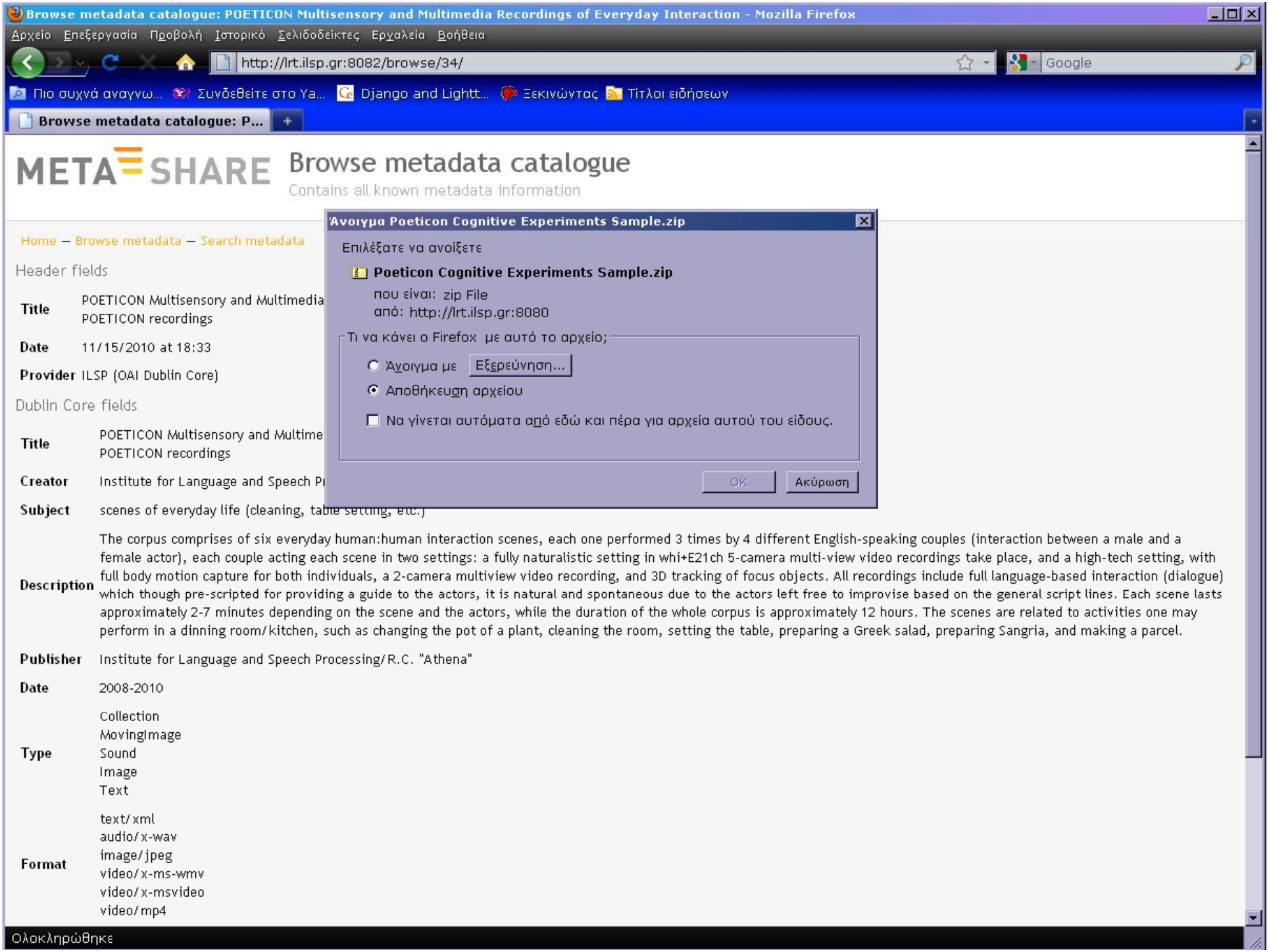
Contains all known metadata information

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

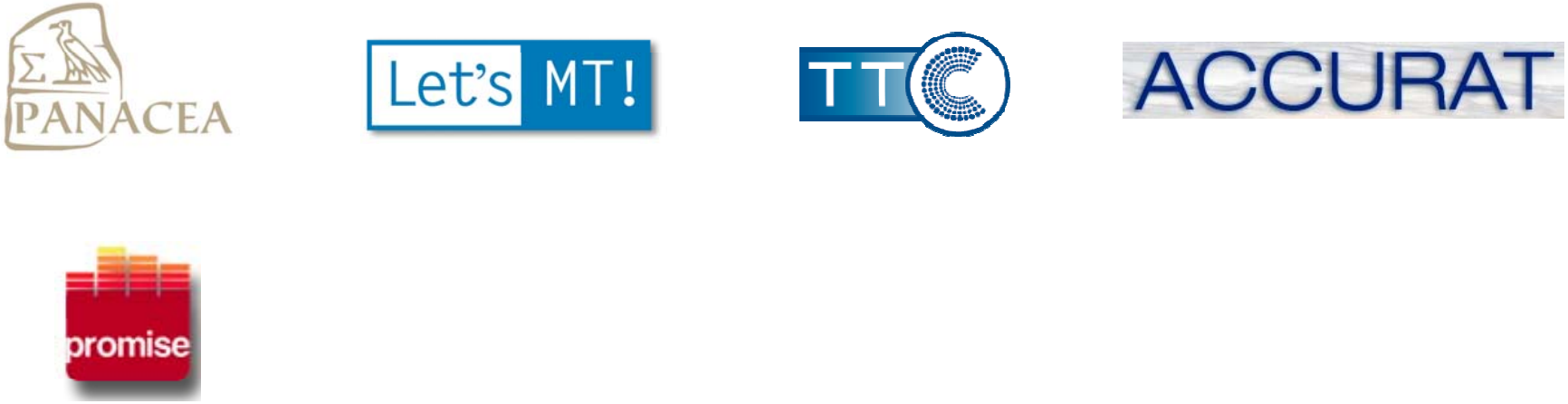

Browse metadata catalogue: POETICON Multisensory and Multimedia Recordings of Everyday Interaction - Mozilla Firefox
Αρχείο Επεξεργασία Προβολή Ιστορικό Σελιδοδείκτες Εργαλεία Βοήθεια
http://lrt.ilsp.gr:8082/browse/34/
Google
Πιο συχνά αναγνω... Συνδεθείτε στο Ya... Django and Lightt... Ξεκινώντας Τίτλοι ειδήσεων
Browse metadata catalogue: P...
META SHARE
Browse metadata catalogue
Contains all known metadata information
Home – Browse metadata – Search metadata
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 P
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
Άνοιγμα Poeticon Cognitive Experiments Sample.zip
Επιλέξατε να ανοίξετε
Poeticon Cognitive Experiments Sample.zip
που είναι: zip File
από: http://lrt.ilsp.gr:8080
Τι να κάνει ο Firefox με αυτό το αρχείο;
Άνοιγμα με
Εξερεύνηση...
Αποθήκευση αρχείου
Να γίνεται αυτόματα από εδώ και πέρα για αρχεία αυτού του είδους.
OK
Ακύρωση
Ολοκληρώθηκε

META-SHARE: Next Steps

- **META-SHARE Version 0: November 2010**
  - First prototype demo'ed at this first META-FORUM.
- **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.

Collaborations

Collaborations



CLARIN and META-NET

- Facilitates research by coordinating and making existing Language Resources and tools available and readily useable for the Social Sciences and Humanities.
- Offers resources and services to allow computer-aided language processing (e.g., querying data and complex processing of data sets).
- Focus on eResearch, eScience.

- Building and offering results for Language Technology at large.
- Clear orientation towards development, innovation and services (including commercial).
- Focus on the distribution of Language Resources (currently).
- End user: the European citizen.
- Goal: to address the problem of multilingualism in Europe.

Join in!

Increase your share

in META-SHARE

http://www.meta-net.eu

Thank you!