

META-SHARE metadata: Overview of the schema & Interoperability with other schemas

Penny Labropoulou
& Maria Gavrilidou
(ILSP/RC Athena)

**CMDI Interoperability Workshop
Utrecht, Netherlands
4-5 June 2013**

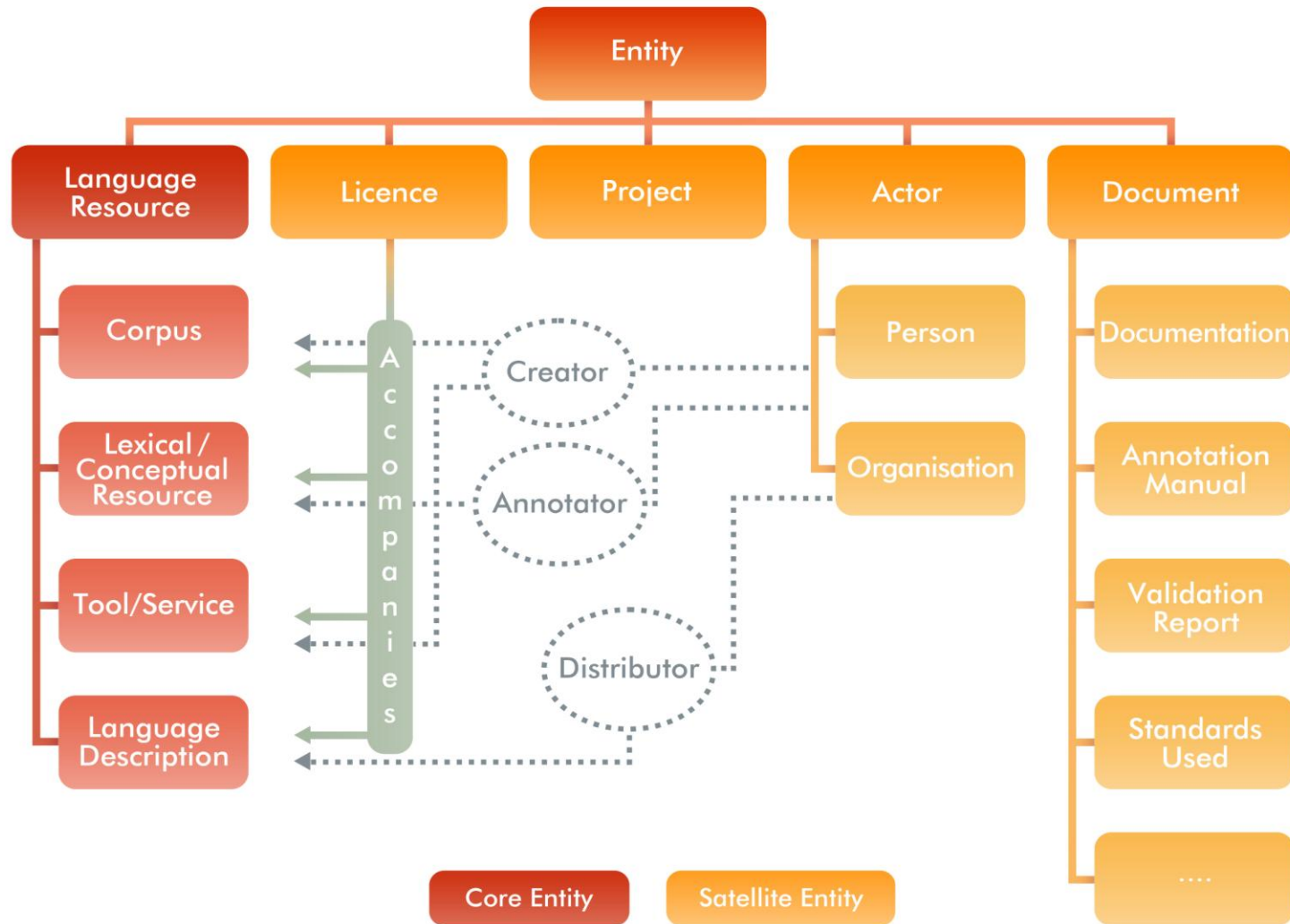
- ❑ META-SHARE is an **open, integrated, secure, and interoperable** exchange infrastructure for language data and tools for the **Human Language Technologies** domain
- ❑ A **marketplace** where language data and tools are documented, uploaded and stored in repositories, catalogued and announced, downloaded, exchanged, discussed, aiming **to support a data economy** (free and for-a-fee LRs/LTs and services)

- ❑ variety of metadata schemas and sets of descriptive elements from LR catalogues in the wider area of language-related activities
- ❑ these come from various backgrounds and focus on the needs of the specific communities that have devised them
- ❑ interoperability problems exist between them
- ❑ **ISO Data Category Registry** caters for semantic interoperability through the registration of *elements*
- ❑ the **Component-based Metadata Infrastructure (CMDI)** complements the ISOcat DCR by introducing the notion of shared *components* and *profiles*

The META-SHARE approach & principles (1)

- ❑ Builds on the CMDI approach
- ❑ Takes into account previous metadata schemas & relevant activities
- ❑ Proposes a schema covering the desiderata of the META-SHARE infrastructure and its users (i.e. LRs providers & consumers) for all facilities provided (incl. search, browsing & retrieval, editing metadata records etc.)
- ❑ Aims to describe (cf. [ontology](#))
 - **LRs**, incl. data (textual, multimodal etc.) resources and tools/services used for their processing
 - their **related entities** (e.g. licences, documentation, actors etc.)

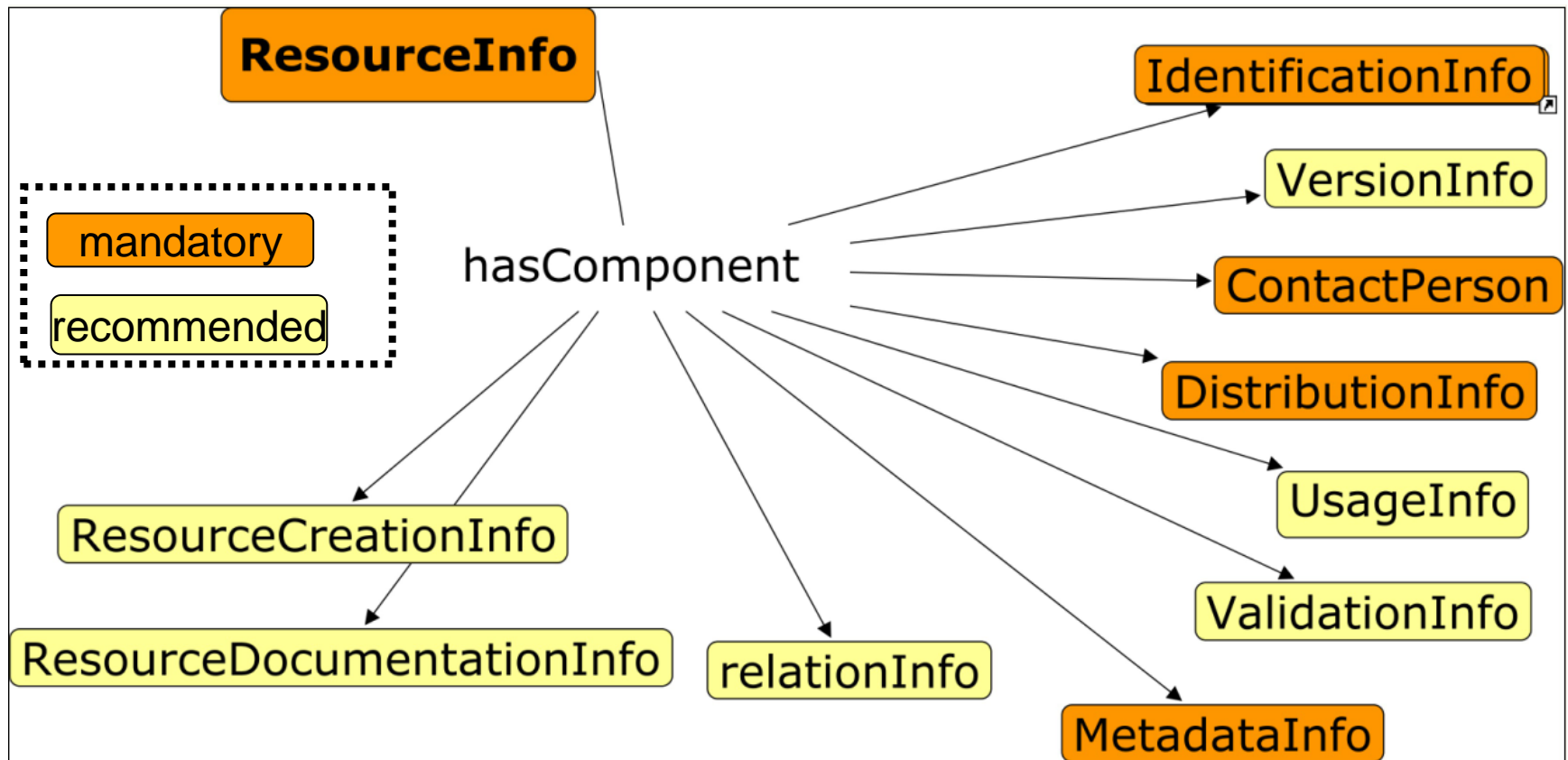
META-SHARE Ontology



The META-SHARE approach & principles (2)

- ❑ Unit of description: *resource* rather than *individual item*, i.e. whole sets of text/audio/etc. files, whole sets of lexical units etc.
- ❑ Aims to cover full lifecycle of the LR production and usage,
 - minimum of mandatory components (**minimal schema**) required for effective LR search, identification and retrieval
 - further recommended/optional (**maximal schema**) components that improve LR use
- ❑ Metadata element values: free text vs. **open** and **closed controlled vocabularies**
- ❑ Highlight **common** elements in LRs → e.g. one profile for all LRs
- ❑ but provide **distinct** elements for differences → e.g. media-type specific components

The core description component



- Two main classification axes:

- **resourceType**

- *corpus*, incl. written/text, oral/spoken, multimodal/multimedia corpora,
- *lexical/conceptual resource*, incl. terminological resources, word lists, semantic lexica, ontologies, etc.,
- *language description*, incl. grammars, typological databases, courseware, etc.,
- *tool/service*, incl. processing tools, applications, web services, etc.

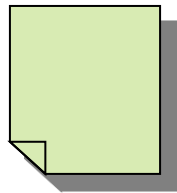
and

- **mediaType** (i.e. the medium on which the LR is implemented)

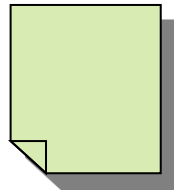
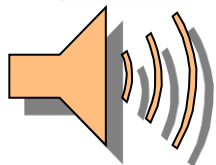
- *text (+textNumerical and textNgram), audio, image, video*

- each LR receives only one *resourceType* value, but may take more than one *mediaType* values (LRs can consist of parts belonging to different types of media)

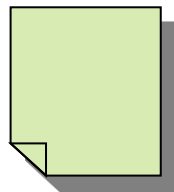
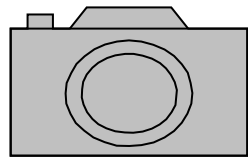
mediaType combinations



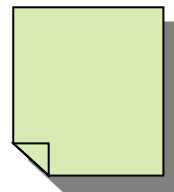
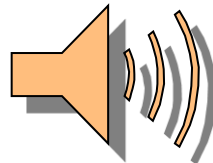
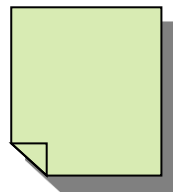
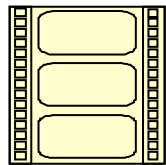
written corpora



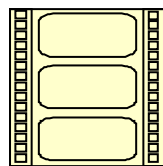
spoken corpora



images (multimedia)

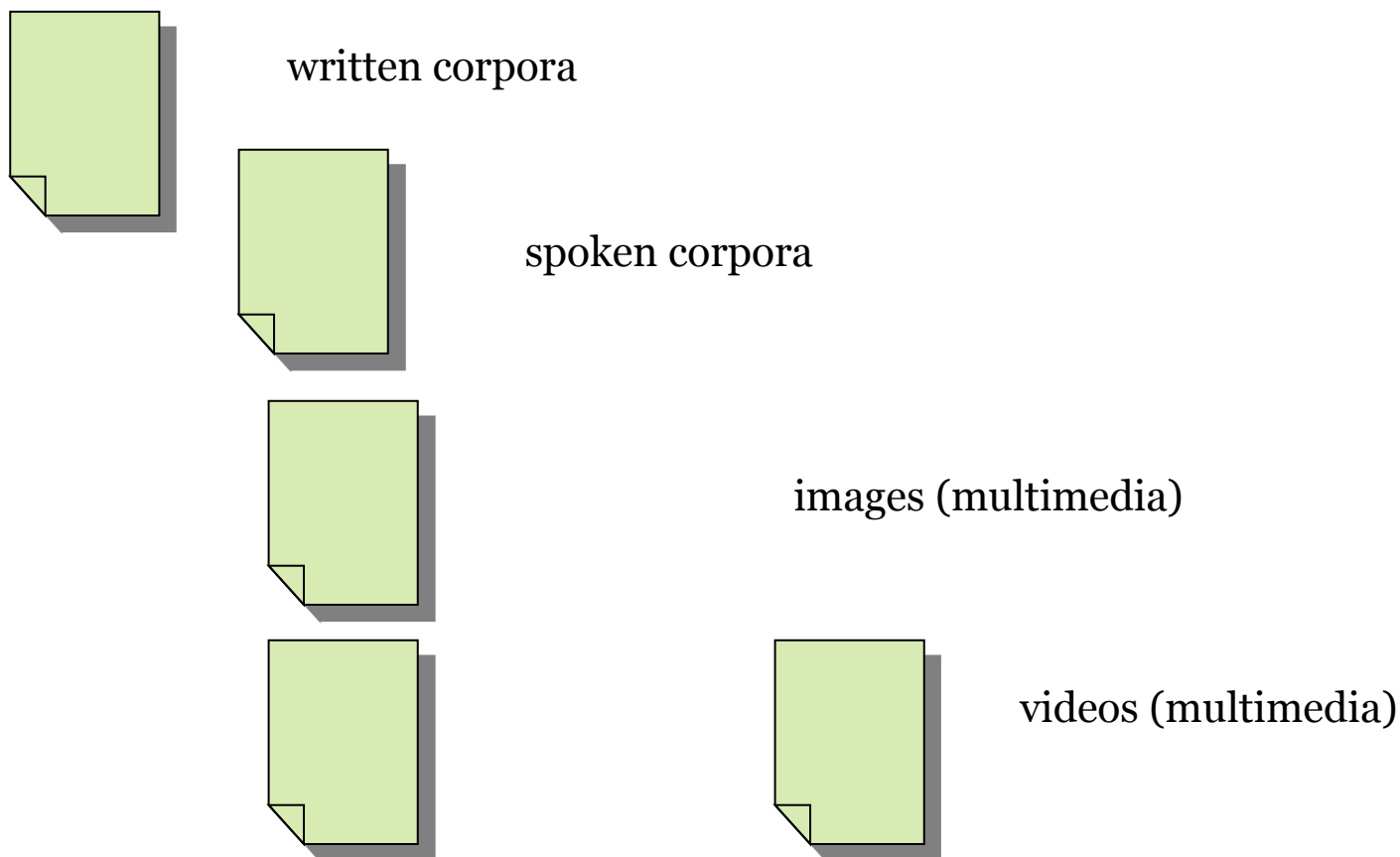


videos (multimedia)

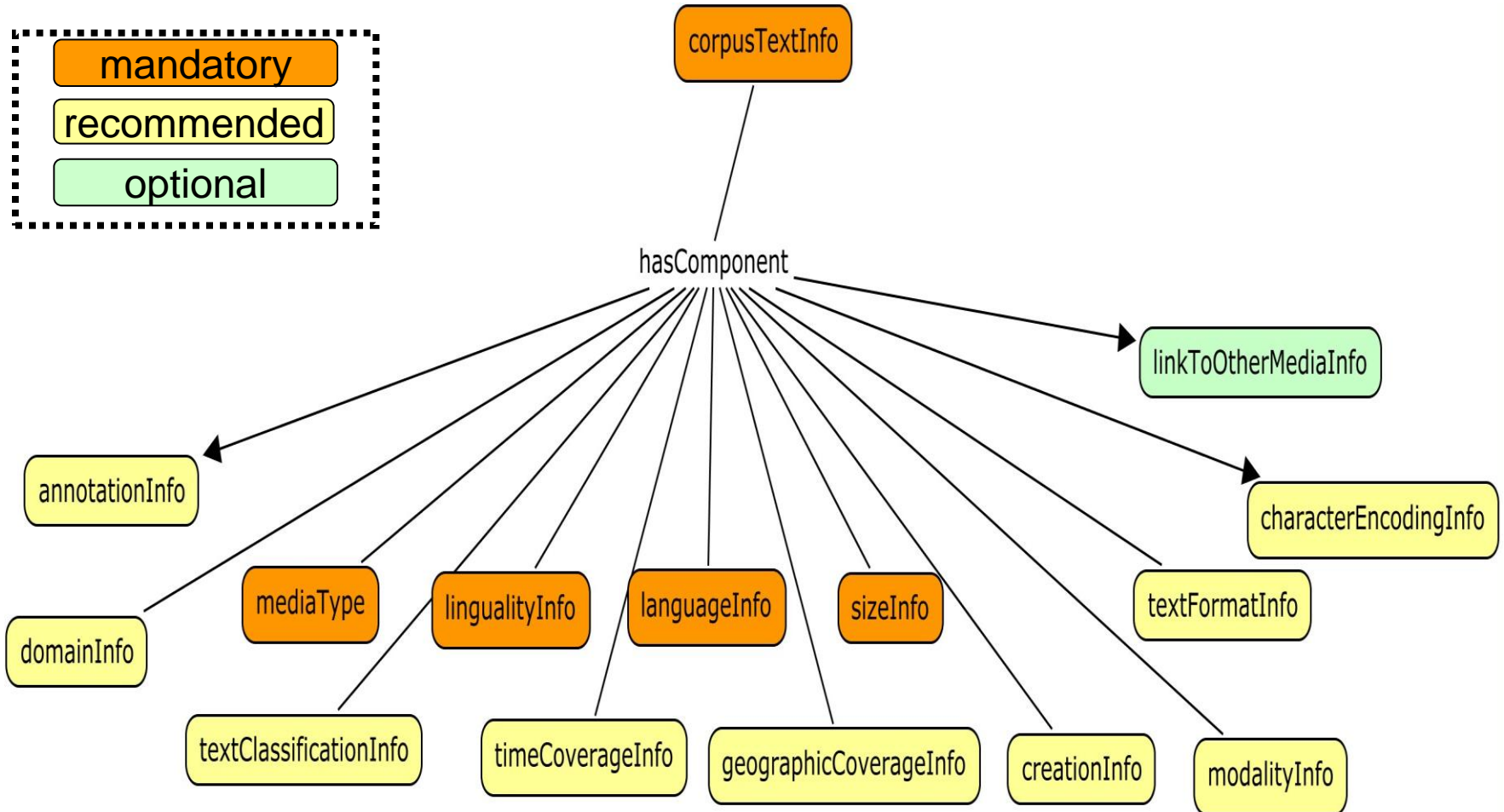


biometrical data (textNumerical)

Identity intact



corpusTextInfo



Implementation of the model

- ❑ the model has been implemented as an XML schema
- ❑ supporting documentation
 - documentation & user manual (with definitions, examples and guidelines)
<http://www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf>
 - knowledge base <http://metashare.ilsp.gr/portal/knowledgebase>
 - user forum
<http://metashare.ilsp.gr/portal/forum/questions/show/all/newest/all/>
- ❑ supporting s/w - META-SHARE platform, incl.
 - [editor](#) and [uploader](#) of XML records
 - metadata [browse](#), search and retrieval

META-SHARE editor

Change Corpus text

- Information
- Required
- Recommended
- Optional

Required information: Linguality, Language, Size

Linguality

Linguality: Indicates whether the resource includes one, two or more languages

Multilinguality: Indicates whether the corpus is parallel, comparable or mixed

Multilinguality details: Provides further information on multilinguality of a resource in free text

Language: Latvian Delete

Language id: The identifier of the language that is included in the resource or supported by the tool/service according to the IETF BCP47 standard

Language name: A human understandable name of the language that is used in the resource or supported by the tool/service according to the IETF BCP47 standard

Language script: Specifies the writing system used to represent the language in form of a four letter code as it is defined in ISO-15924

Size per language: Provides information on the size per language component

Language variety: Groups information on language varieties occurred in the resource (e.g. dialects) Hold down "Control", or "Command" on a Mac, to select more than one.

Available Language variety

- Castilian (dialect)
- Flemish (dialect)
- Chile (other)
- Scottish Gaelic (dialect)
- Belgium (other)
- Argentina (other)
- Colombia (other)
- Valencian (dialect)
- Netherlands (other)
- Mexico (other)
- Modern (1485-) (dialect)
- Costa Rica (other)
- United States (other)

Choose all

Chosen Language variety

Select your choice(s) and click

Clear all

Language: English Delete

Language id: The identifier of the language that is included in the resource or supported by the tool/service according to the IETF BCP47 standard

Language name: A human understandable name of the language that is used in the resource or supported by the tool/service according to the IETF BCP47 standard

Language script: Specifies the writing system used to represent the language in form of a four letter code as it is defined in ISO-15924

Size per language: Provides information on the size per language component

Language variety: Groups information on language varieties occurred in the resource (e.g. dialects) Hold down "Control", or "Command" on a Mac, to select more than one.

Available Language variety

- Castilian (dialect)
- Flemish (dialect)
- Chile (other)
- Scottish Gaelic (dialect)
- Belgium (other)
- Argentina (other)
- Colombia (other)
- Valencian (dialect)
- Netherlands (other)
- Mexico (other)
- Modern (1485-) (dialect)
- Costa Rica (other)
- United States (other)

Choose all

Chosen Language variety

Select your choice(s) and click

Clear all

Language: #3 Delete

Language id: The identifier of the language that is included in the resource or supported by the tool/service according to the IETF BCP47 standard

Language name: A human understandable name of the language that is used in the resource or supported by the tool/service according to the IETF BCP47 standard

Language script: Specifies the writing system used to represent the language in form of a four letter code as it is defined in ISO-15924

Size per language: Provides information on the size per language component

Language variety: Groups information on language varieties occurred in the resource (e.g. dialects) Hold down "Control", or "Command" on a Mac, to select more than one.

Available Language variety

- Castilian (dialect)
- Flemish (dialect)
- Chile (other)
- Scottish Gaelic (dialect)
- Belgium (other)
- Argentina (other)
- Colombia (other)
- Valencian (dialect)
- Netherlands (other)
- Mexico (other)
- Modern (1485-) (dialect)
- Costa Rica (other)
- United States (other)

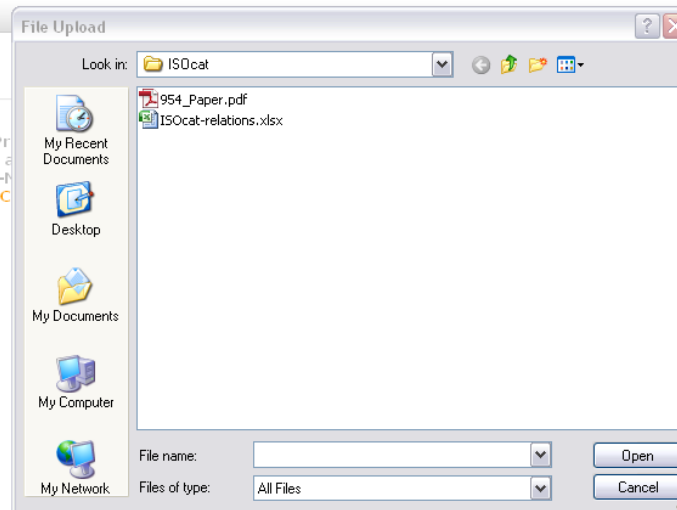
Choose all

Chosen Language variety

Select your choice(s) and click

META-SHARE uploader of metadata records in XML

The screenshot shows the META-SHARE web interface. At the top, there is a navigation bar with a home icon, the text 'META=SHARE', and a 'Logout' button. Below this are several menu items: 'Browse Resources', 'Manage Resources', 'Administration', 'Community', 'Statistics', 'Help', 'About', and 'Your Profile, Penny'. The main content area is titled 'Upload new resource description(s)' and has a sub-header 'Upload'. It contains a form with a 'Resource Description(s):' label, a text input field, and a 'Browse...' button. Below this is a paragraph of instructions: 'You can upload a new resource description in XML format, or many resource descriptions in a ZIP file containing XML files. Please make sure the XML files are Schema-valid before proceeding.' There is also an 'Upload Terms:' section with an unchecked checkbox and a confirmation message: 'By clicking this checkbox, you confirm that you have cleared permissions for the description(s) you intend to upload.' At the bottom right of the form is an 'Upload' button.



Co-funded by the 7th Framework Programme through the contracts T4ME (grant agreement no.: 270893) and META-NET (grant agreement no.: 270893) under Creative Commons Attribution-NonCommercial license.

on (grant

- ❑ v3.0 - minimal schema already in the Component Registry by the Prague University; incl. a profile combining META-SHARE with DC
- ❑ Uploading of the **full schema v3.0** into the [Component Registry](#)
- ❑ Due to technical reasons, modifications were necessary, e.g.
 - four profiles for each *resourceType*
 - some components split into two components, e.g. *actorInfo* into *person* and *organization*
 - ordering of components and elements
 - etc.
- ❑ Converters between the two implementations have been built, catering for these modifications where possible
- ❑ Validation required before uploading to the META-SHARE repo

Interoperability through ISOcat & DC

- ❑ Link to ISOcat elements in the [documentation](#)
- ❑ Link to DC and OLAC elements in the [documentation](#)
- ❑ Link to ISOcat elements (incl. containers) in the Component Registry implementation (*conceptLink*)

Interoperability with other schemas

- ❑ Converters for the ELRA schema, cf. [Gavrilidou M., P. Labropoulou, E. Desipri, I. Giannopoulou, O. Hamon, V. Arranz \(2012\) "The META-SHARE Metadata Schema: Principles, Features, Implementation and Conversion from other Schemas", LREC 2012 – Workshop on Describing Language Resources with Metadata, Istanbul, Turkey.](#)

- ❑ Converters for OLAC and DC – work almost finished, but issues
 - more vs. less detailed schema
 - free text vs. list of values
 - semantic problems: e.g. *publisher* in the META-SHARE context

Interoperability issues at the component/profile level

- ❑ Interoperability - re-usability and better understanding of components
- ❑ Grouping of similar components & comparison/contrast between them
- ❑ Statistics on usage of components/profiles (metadata records, metadata schemas)
- ❑ More view options: e.g. usage of components by other components/profiles
- ❑ Versioning of components

Interoperability issues at the element level

- ❑ How do you decide the most appropriate element for linking?
 - multiple similar elements in the isoCat, e.g. *region* to *region* & *locationRegion*, *resourceCreationDate* to *creationDate* and *startYear*
 - elements with the same name but not exactly the same, e.g. *licence* with *license* (broader than *licence*)
 - elements with free text or different values, e.g. *mediaType* vs. *mediaType*
 - element similar to set of elements, e.g. *contactPerson/givenName & surname* to *contactFullName*
 - same conceptLink inside a component: e.g. *metaShareId* & *identifier* to *identifier*
- ❑ usage of elements – by which components? by which schemas? in which metadata records?
- ❑ grouping and relations between elements in ISOcat: where do we stand?

- ❑ Thank you all!
- ❑ Questions/Discussion