The META-SHARE Metadata Schema for the Description of Language Resources

Maria Gavrilidou*, Penny Labropoulou*, Elina Desipri*, Stelios Piperidis*, Haris Papageorgiou*, Monica Monachini†, Francesca Frontini†, Thierry Declerck^, Gil Francopoulo‡, Victoria Arranz& and Valerie Mapelli&

*Athena R.C./ILSP, †ILC/CNR, ^DFKI, ‡CNRS-LIMSI+IMMI, &ELDA

*Athens, Greece, †Pisa, Italy, ^Saarbrücken, Germany, ‡Paris, France, &Paris, France

E-mail: {maria, penny, elina, spip, xaris}@ilsp.athena-innovation.gr, {monica.monachini, francesca.frontini@ilc.cnr.it}, Thierry.Declerck@dfki.de, gil.francopoulo@limsi.fr, {arranz, mapelli}@elda.org

Abstract

This paper presents a metadata model for the description of language resources proposed in the framework of the META-SHARE infrastructure, aiming to cover both datasets and tools/technologies used for their processing. It places the model in the overall framework of metadata models, describes the basic principles and features of the model, elaborates on the distinction between minimal and maximal versions thereof, briefly presents the integrated environment supporting the LRs description and search and retrieval processes and concludes with work to be done in the future for the improvement of the model.

Keywords: metadata, META-SHARE, LRs description

1. Introduction

The importance of Language Resources (LRs) for language-related and language-based research and applications is undeniable. Language technology applications, in particular, such as multilingual information extraction, machine translation, automatic document indexing etc., include LRs as critical components. Even language technologies that consist of language independent engines rely on the availability of language-dependent knowledge under the form of LRs for their real-life implementation. It has also been shown that a critical mass of LRs can make advancement in language research possible and quicker (Calzolari, Quochi & Soria 2011).

Digital repositories constitute a valuable tool in the effort of publishing, archiving, discovery and long-term maintenance and curation of huge amounts of digital data (publications, datasets, multimedia files, and even processing tools and services), as they provide the infrastructure for describing and documenting, storing, preserving, and making this information publicly available in an open, user-friendly and trusted way. In this framework, interoperability at all levels is a crucial issue. META-SHARE (www.meta-share.eu) is an open, integrated, secure and interoperable exchange infrastructure dedicated to LRs; it serves as a space where LRs are documented, uploaded and stored in repositories, catalogued and announced, downloaded, exchanged and discussed, aiming to support a data economy. META-SHARE brings together knowledge about LRs and related objects and processes and fosters their use by providing easy, uniform, one-step access to LRs through the aggregation of LR sources into one catalogue; it facilitates LRs' search and retrieval processes, and encourages (re-)use and new use of LRs (Piperidis, 2012). The adoption of a uniform metadata schema, i.e. a common terminology for the external description of LRs, is crucial to the success of the endeavour.

In the context of META-SHARE, the term *metadata* refers to descriptions of LRs, encompassing both data (textual, multimodal/multimedia and lexical data, grammars, language models, etc.) and technologies (tools/services) used for their processing.

2. Design principles for the metadata model

The metadata descriptions constitute the means by which LR producers describe their resources and LR users identify the resources they seek. Thus, the META-SHARE metadata model forms the core engine driving the META-SHARE access interfaces to the LRs catalogue.

For the design of the metadata schema we have taken into consideration the user needs (as collected through interviews with a variety of stakeholders and documented in (Federmann et al., 2011) and the advantages but also the shortcomings of previous efforts for the efficient and adequate description of LRs, via an overview of widespread metadata models in HLT as well as LR catalogue descriptions (Gavrilidou et al., 2011).

The overview studied models that put emphasis on the 'minimalist nature' of the schema, such as Dublin Core (DC, http://dublincore.org/), and BAMDES, the Basic Metadata Description, used for harvesting purposes by the Harvesting Day initiative (http://theharvestingday.eu/), but also very granular and elaborated schemas, such as the ISLE MetaData Initiative (IMDI, http://www.mpi.nl/IMDI/), which originally focused on multimedia and multimodal language resources, and the Open Language Archives Community (OLAC, http://www.language-archives.org/), which constitutes an extension of the Dublin Core schema devoted to language resources. It also reviewed older standardization activities, such as the Corpus Encoding Standard (CES, http://www.cs.vassar.edu/CES/) and its XML version (XCES, http://www.xces.org/), which instantiates the EAGLES CES DTDs for linguistic corpora and, obviously, the Text Encoding Initiative (TEI,

http://www.tei-c.org/index.xml), which has developed and maintains a standard for the representation of digital texts, as well as recommendations' initiatives such as the European National Activities for Basic Language Resources project (ENABLER, http://www.ilsp.gr/en/infoprojects/) and the metadata model it proposed, and the most recent activities such as the metadata-related activities of the CLARIN project (Common Language Resources and Technology Infrastructure, http://www.clarin.eu/external/). Finally, the overview studied metadata used by well-known catalogues and LRs agencies, such as the ELRA Catalogue and the ELRA Universal Catalogue of the European Language Resources Association (ELRA, http://www.elra.info/) and the Linguistic Data Consortium catalogue of available resources (LDC, http://www.ldc.upenn.edu/). Last but not least, the overview studied the ISO 12620 – Data Category Registry (ISOcat DCR, (ISO 12620, 2009), http://www.isocat.org/), which defines widely accepted linguistic concepts, including metadata for the description of language resources.

This overview concludes with a set of observations, which led to the formulation of the basic design principles of the META-SHARE model. The needs identified are:

- need for a language resources typology identifying and defining all types of LRs and the relations holding between them,
- need for a common terminology, or at least, for terminology with clear semantics,
- contradicting needs for minimal schemas with simple structures (for ease of use) but also for extensive, detailed schemas (for exhaustive description of LRs),
- need for interoperability between LRs and tools, and between repositories.

In answer to these needs, we came up with the following design principles:

- expressiveness: through the proposed LR typology we aim at covering any type of resource;
- extensibility: the modularity of the schema allows for future extensions, to cover more resource types as they become available; the schema will also cater for combinations of LR types for the creation of complex resources;
- semantic clarity: to achieve clear articulation of a term's meaning and its relations to other terms, each element of the schema is accompanied by a bundle of information constituting its identity, comprising its definition, its type, its domain and range of values, an example, the relations to other components/elements and links to the appropriate DC and ISOcat DCR terms (where applicable);
- flexibility: by the definition of a two-tier schema (minimal and maximal), we cater for the possibility for exhaustive but also for minimal descriptions;
- interoperability: this is guaranteed through the mappings to widely used schemas (mainly DC, and ISOcat DCR).

3. The META-SHARE ontology

The META-SHARE focus lies on the description of LRs; as aforesaid, this covers both data resources and tools/services used for their processing.

META-SHARE remains at the level of *resource* rather than *individual item*, in the sense that it targets to describe whole sets of text/audio/video etc. files (corpora), sets of lexical entries (lexical/conceptual resources), integrated tools/services and so on, rather than individual items. For individual items, the META-SHARE model refers users to the recommended standards and/or best practices reported in (Monachini et al., 2011). However, this does not mean that the schema cannot handle *resource parts* (crucial for all multimedia-type resources). These are detailed in Section 4.

Resource collections are also in the process of being defined and will be available shortly within META-SHARE. These collections comprise both *evaluation packages*, which are composite resources made up of all elements necessary to reproduce an evaluation (e.g., data, tools, metrics, protocols, etc.) and *bundle resources*, which are resources grouped together mainly for administrative reasons (e.g. belonging to the same resource owner, distributed by the same organization etc.).

The central entity of the META-SHARE ontology is, as already discussed, the LR per se. However, in the ontology, LRs are linked to other satellite entities through relations that in the model are represented as basic elements (Figure 1). The interconnection between the LR and these satellite entities pictures the LR's lifecycle from production to use: reference documents related to the LR (papers, reports, manuals etc.), persons/organizations involved in its creation and use (creators, distributors etc.), related projects and activities (funding projects, activities of usage etc.), accompanying licenses, etc. Thus, the META-SHARE model recognizes the following distinct entities:

- the *resource* itself, i.e. the LR being described,
- the *actor*, further distinguished into *person* and *organization*,
- the *project*,
- the *document*, and
- the *licence*.

It should be noted, however, that the satellite entities are described only when the case arises, i.e. when they are linked to a specific resource. For their description, the metadata schema takes into account schemas and guidelines that have been devised specifically for them (e.g. BibTex for bibliographical references).

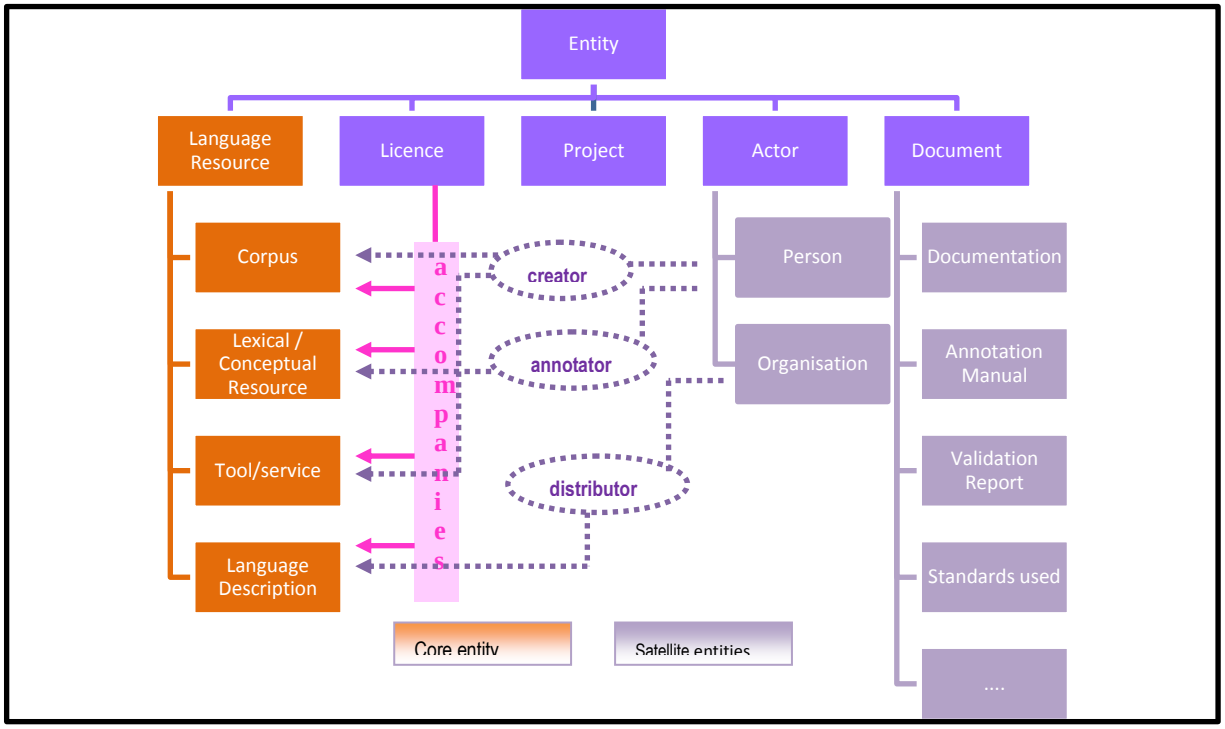

Figure 1: The META-SHARE ontology: the two types of entities & example relations holding between them

4. Proposed LRs typology

The study of existing LR typologies (Gavrilidou et al., 2011) has revealed their diversity, which hampers the request for interoperability and jeopardizes the mandate of META-SHARE to provide a simple albeit descriptive schema for LRs.

The META-SHARE model uses metadata elements as criteria for classifying LRs; the identity of a resource is the outcome of the combination of specific elements and does not originate with a top-down procedure. In this sense, the LR typology and the schema form a coherent universe.

Two are the main classification axes: *resourceType* and *mediaType* (i.e. the medium on which the LR is implemented). This choice has been dictated by the fact that they both bring to the description of the LRs distinct sets of features; for instance, *resourceType*-specific information includes annotation features (for corpora), types of encoding contents (for lexica and grammars), performance (for grammars), while *mediaType*-specific information refers to the actual medium of the LR, and includes features like format (wav/avi etc. for videos, txt/doc/pdf/xml for texts etc.) and size (sentences/words/ bytes for text corpora, duration for audio/video corpora, entries/items for lexica etc.).

More specifically, the following four values are suggested for the element *resourceType*:

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

Each LR receives only one *resourceType* value, but naturally it may take more than one *mediaType* values since LRs can consist of parts belonging to different types of media: for instance, a multimodal corpus includes a video part (moving image), an audio part (dialogues) and a text part (subtitles and/or transcription of the dialogues); a multimedia lexicon, besides the textual part, also includes a video and/or an audio part; a sign language resource is also a resource with various media types (video, image, text). Similarly, tools can be applied to resources of different media types: e.g. a tool can be used both for video and for audio files. Thus, for each part of the resource, the respective feature set (components and elements) should be used: e.g. for a spoken corpus and its transcriptions, the audio feature set will be used for the audio part and the text feature set for the transcribed part. The following media type values and combinations are foreseen:

- *text*: used for data resources with only written medium (and modules of audio and multimodal corpora, see below), whether monolingual or multilingual;
- *audio* (+ text): the audio feature set will be used for a whole resource or part of a resource that is recorded as an audio file; its transcripts are to be described by the relevant *text* feature set;
- *image* (+ text): the *image* feature set is used for photographs, drawings, images of sensorimotor data etc., while the *text* set can be used for the description of its captions

• *video*: moving image (+ text) (+ audio (+ text)): used for multimedia corpora, with *video* for the moving image part, *audio* for the dialogues, and *text* referring to the transcripts of the dialogues and/or subtitles.

Two additional values are introduced in the model, although they are not really distinct media type values: these correspond to numerical text resources (value *textNumerical*) and n-grams (value *ngram*). These are actually subtypes of text resources but they present further descriptive particularities due to their contents: numerical data (e.g. biometrical, geospatial data, etc.) for the former, and items with frequency counts for the latter.

In addition to the two main classification elements described in this section, metadata elements (and combinations thereof) can be treated as classification criteria in the process of unfolding the inventory of LRs: faceted browsing and filtering of the catalogue is also possible on the basis of these features. Thus, for instance *lingualityType* as an organizing feature can be used to distinguish between mono- bi- and multilingual data resources. Similarly, *languageName*, *domain*, *format*, *annotation* features, etc. can be used as different dimensions according to which the catalogue of LRs can be accessed.

5. The essentials of the metadata model

The general framework for the development of the metadata model is inspired by the component-based mechanism proposed by the ISOcat DCR, according to which semantically coherent elements are grouped together to form components (Broeder et al., 2010).

Components are the core building blocks of the metadata model and act as placeholders for well defined categories of information (i.e. information on usage, validation, licensing, etc.). They are organized in terms of the two main axes of the model (resource and media type). Components consist of *elements* (categories) that are used to encode specific descriptive features and are grouped and combined in terms of semantic coherence.

Elements are also used to represent *relations* in the current version of the schema. The relation mechanism represents the encoding of linking features between resources. Relations hold between various forms of a LR (e.g. raw and annotated resource), different LRs included in the META-SHARE repository (e.g. a language resource and a tool that has been used to create it, etc.) but also between LRs and satellite resources such as standards used, related documentation, etc.

Central to the model is the LR taxonomy, which allows the structuring of the components around the two aforementioned main axes of the schema, i.e. the resource and media type, taking into consideration the specificities of LR type (combination of resource and media type).

The set of all the components describing specific LR types and subtypes constitute the profile of each type. Components are distinguished in three classes: (a) components common to all types of resources (e.g. identification, contact, licensing information, etc.), (b) components re-usable for more than one resource / media types but not globally applicable (e.g. capture information for audio, video and image resources) and (c) the ones strictly applied to specific resource and media types (e.g. evaluation for tools, audio content for audio resources).

The user is presented with proposed profiles for each type, which can be used as templates and guidelines for the completion of the metadata description of the resource. Experience has shown that users indeed need guidelines and help in the process of metadata addition to their resources. Moreover, exemplary instantiations (e.g. for wordnet-type resources, for parallel corpora, for multimodal resources, for treebanks, etc.) will be made available as guiding assistance to LRs metadata providers.

In order to accommodate flexibility, the elements belong to two basic levels of description (stepwise approach):

- an initial level providing the basic elements for the description of a resource (*minimal schema*), and
- a second level with a higher degree of granularity (*maximal schema*), providing detailed information on a resource and covering all stages of LR production and use.

The minimal schema contains those elements considered indispensable for LR description (from the provider's perspective) and identification (from the consumer's perspective).

In addition, the schema specifies the type allowed for all elements (e.g. if the values are of type *string, number, closed set of values*, etc.).

6. Contents of the model

The core of the model is the *resourceInfo* component (Figure 2), which contains all information relevant for the description of a resource. It subsumes components that combine together to provide the full description of a resource.

Administrative components are common to all LR types and provide information on the various phases of the resource's life cycle, i.e. creation, validation, usage, distribution, etc. It should be noted that these components encode most of the relations of the LR per se to all other satellite entities, i.e. persons, organizations, licences, etc.

The set of components that are common to all LRs are: *identificationInfo*, *distributionInfo*, *contactPerson*, *metadataInfo*, *versionInfo*, *validationInfo*, *usageInfo*, *resourceDocumentationInfo*, *creationInfo* and *relationInfo*. More specifically:

The *identificationInfo* component includes all elements required to identify the resource, such as the LR's full and short names, the META-SHARE ID (to be automatically assigned by the system)[1] etc.; the *description* element is obligatorily used for the free text description of the resource contents.

[1] The ISLRN (International Standard Language Resource Number) is also foreseen to be assigned in a coming version.

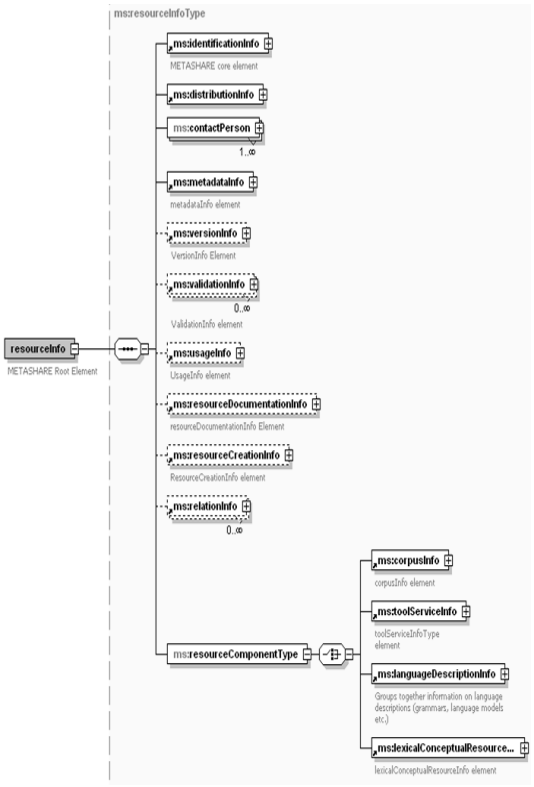

Figure 2: Common components for all LRs and resourceType components

Crucial is the information on the legal issues related to the availability of the resource, specified by the *distributionInfo* component, which provides a description of the terms of availability of the resource and its attached *licenceInfo* component, which gives a description of the licensing conditions under which the resource can be used.

The *contactPerson* component provides information about the person that can be contacted for further information or access to the resource.

The *metadataInfo* is responsible for all information relative to the metadata record creation, such as the source of the metadata record, the creation date and metadata creator (in case of records created from scratch using the META-SHARE metadata editor), etc.

All information relative to versioning and revisions of the resource is included in the *versionInfo* component.

The *validationInfo* component provides at least an indication of the validation status of the resource (with boolean values) and, if the resource has indeed been validated, further details on the validation mode, results, etc.

The *usageInfo* component aims at providing information on the foreseen use of a resource (i.e. the application(s) for which it was originally designed) and its actual use (i.e. applications for which it has already been used, projects in which it has been exploited, products and publications having resulted from its use, etc.).

The *resourceDocumentationInfo* provides information on publications and documents describing the resource; links to documents over the internet enhances this feature.

The *resourceCreationInfo* and its dependent components group together information regarding the creation of a resource (creation dates, funding information such as funder(s), project name, etc.).

Finally, the *relationInfo* component allows the encoding of relations that have not been foreseen by the metadata model; the resource providers have the chance to encode the relation type and the related resource.

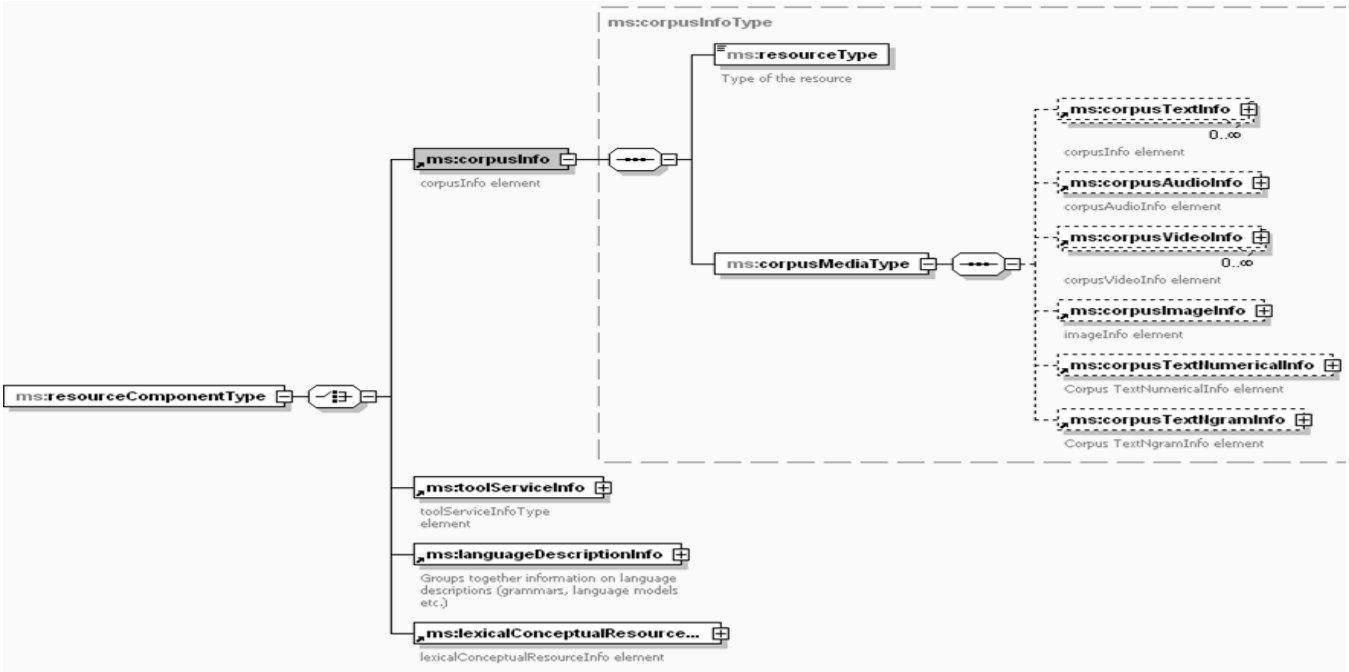

Figure 3: Components for corpora

The LR type- and media- specific components are organized around the elements *resourceType* and *mediaType* that encode the two classification axes of the schema.

LR type-specific components are all located under the *resourceComponentType* component. Similarly, for each LR type, particular medium-dependent components are created to group together sets of features relevant to each LR/media type, given that media types and the relevant information differs across LR types; these are again grouped under an *xMediaType* component, where x stands for each of the LR type values (see Figure 3). *corpusTextInfo*, *corpusAudioInfo*, *corpusVideoInfo*, *lexicalConceptualResourceTextInfo*, *lexicalConceptual-ResourceVideoInfo* etc. provide information depending on the media type of each LR type and include the *mediaType* element with the values *text*, *audio*, *video* etc. accordingly.

Broadly speaking, the resource / media type-specific components cover the following types of information:

- contents: components mainly referring to languages covered in the resource, types of content (e.g. for images: drawings, photos, histograms, animations etc.), modalities included (e.g. written / spoken language, gestures, eye movements, etc.), etc.
- classificatory information: components including resource-type subclassification (e.g. subtypes of lexical/conceptual resources, tools/services etc.) as well as classification of the contents of the resource; this can be cross-media (e.g. domains, geographic coverage, time coverage, etc.) as well as media-dependent (e.g. text type, audio genre, setting, etc.)
- formatting: file format, character encoding etc.; obviously, this information is more media-type-driven (e.g. different file formats for text, audio and video files)
- information on creation: it refers to the creation of the specific resource parts e.g. the original source, the capture and recording methods (e.g. scanning and web crawling for texts vs. recording methods for audio files). These components are to be distinguished from the resourceCreationInfo component attached at the resource level, which is used to give information on anything concerns the creation of all resource and media types (e.g. creation dates)
- performance: information regarding the performance of the resource; it is resource-type driven, given that the measures and criteria differ across resource types
- operation: information relevant to the operation requirements of the resource (e.g. the hardware and software prerequisites for running a tool/service)
- input and output: these components are specific to tools/services; they can be used to provide information on the media type, format, language, etc. that the tool/service can take as input and the resulting output
- finally, a special component, *linkToOtherMediaInfo*, is provided for linking between the various media type parts of the resource. This component is to be applied to multimedia resources.

7. Minimal schema

The obligatory components and elements thereof that constitute the minimal schema are presented here below:

- *identificationInfo*: groups together information needed to identify the resource; the obligatory elements are the *resourceName*, the *meta-shareId* and the *description*
- *distributionInfo*: groups information on the distribution of the resource; the element *availability* serves as a first indication of the terms of availability of the resource (with values *available*, *available-restrictedUse*, *available-unrestrictedUse*, *notAvailableThroughMetaShare*, *underNegotiation*); in case the resource is available, the component *licenceInfo* provides obligatorily further information regarding the licensing conditions under which the resource can be used (at least the licence must be specified)
- *contactPerson*: groups information on the contact person; the only obligatory information is the *surname* and *email* of the person
- *metadataInfo*: groups information on the metadata record itself; the only mandatory element is the *metadataCreationDate*, which encodes the date of creation of the metadata record either from scratch or through harvesting; depending on the way the metadata record has been created (harvesting, editing, uploading, etc.) further information can be optionally provided (e.g. metadata creator, original metadata link, etc.)

Further obligatory components and elements are specified for each LR type. In general, the mandatory information is restricted to basic information so as not to intimidate metadata creators: size and languages for datasets, subtype for all (obviously with value sets depending on the resource type), level of encoding for language descriptions and so on.

The further characterisation of specific components and elements as "recommended" prompts the resource providers to input richer descriptions of their resources.

8. Implementation of the model

The model has been implemented as an XML schema, documented also in the form of a user manual (cf. http://www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf), which contains detailed information, including definitions, examples and guidelines for the usage of the whole schema and each element (Desipri et al., 2012).

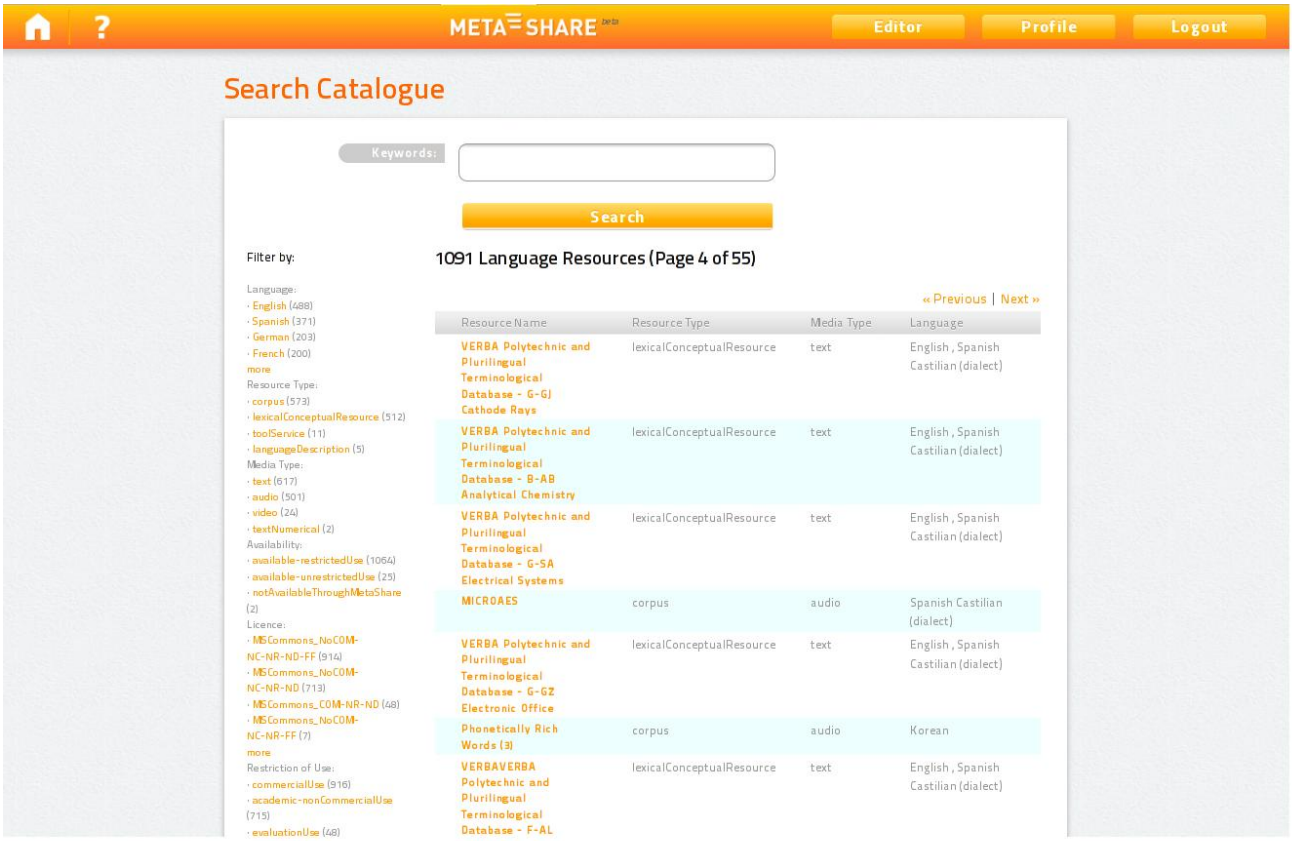

Figure 4: The META-SHARE browser

9. META-SHARE environment

An integrated environment has been developed, which facilitates the description of LRs, either from scratch or through uploading of XML files adhering to the META-SHARE metadata schema, as well as browsing of the LRs (Federmann et al., 2012). Language resources and their metadata reside at the members' repositories, or in case this is not possible or desirable, they are hosted by META-SHARE repositories. Only metadata are exported for harvesting purposes and for populating the network's inventories that include metadata-based descriptions of all LRs in the network. META-SHARE serves both LR providers and users: it offers to the user the possibility to search and browse the catalogue (Figure 4), to view details about a LR, to download a LR, to view general statistics, to have access as a registered user and to describe and upload a LR.

Distinct user profiles have been defined, including related authorisations which enable certain actions and ensure the security of transactions. Users may be registered or non-registered, where the former are divided into end users, providers or administrators of a META-SHARE node. With the exception of non-registered users, every user is given a specific profile containing the information about their rights and obligations.

Consumers of LRs (end users) will be able to: register and create a user profile, log-in to the repository network (single sign-on), browse and search the central inventory using search facilities, access the actual resources by visiting the local (or non-local) repositories for browsing and downloading them, get information about the usage of specific resources, their relation (e.g. compatibility, suitability, etc.) to other resources, as well as recommendations, download resources accompanied by easy-to-use licensing templates, including both free and for-a-fee resources, provide feedback about resources and exploit additional functionalities.

Providers of resources will additionally be able to: create, store and edit resource descriptions by using the metadata editor, get support through mapping services from an existing metadata schema into the META-SHARE metadata model, upload actual resources directly or by contacting support staff for large volume resources, get reports and statistics on number of views, downloads, types of consumers, etc. of LRs, as well as feedback from consumers.

META-SHARE is open-source software, available on github at https://github.com/metashare/META-SHARE.

10. Current situation

The schema has been adopted by the different node repositories within META-SHARE, namely repositories /

catalogues from DFKI, ELDA, FBK, ILC-CNR and ILSP. All of them have converted their data into the latest version of the schema, which allows a common resource search among all the catalogues. These repositories contain 1,277 resources (datasets and tools), covering a broad variety of languages, resource and media types, described according to the META-SHARE schema and available through www.meta-share.eu.

The schema is a living entity and it evolves according to needs and the developments in the field. It is currently being tested by the related projects METANET4U, CESAR and META-NORD. Their data conversion work provides invaluable feedback for the improvement of the schema.

11. Future work

Work in the future naturally includes the evolution of the schema as regards breadth (i.e. coverage of more types as they emerge) and depth (i.e. enrichment and updating of the controlled vocabularies, representation of additional relations, improvements based on future feedback, etc.). Mapping to other schemas is also of priority to support interoperability between LR descriptions. Additionally to the currently existing linking of the elements to the corresponding DC and ISOcat ones, links to OLAC elements is foreseen in the future.

12. Acknowledgements

This paper presents work done in the framework of the project T4ME, funded by DG INFSO of the European Commission through the 7th Framework Program, Grant agreement no.: 249119.

Many thanks are due to all the colleagues of the META-SHARE metadata working group, to the META-SHARE implementation team and to all the colleagues from the projects METANET4U, CESAR and META-NORD for their valuable feedback.

13. References

Broeder, D.; Kemps-Snijders, M.; Van Uytvanck, D.; Windhouwer, M.; Withers, P.; Wittenburg, P. and Zinn, C. (2010). A Data Category Registry- and Component-based Metadata Framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Malta.

Calzolari, N.; Quochi, V. and Soria, C. (2011). *The Strategic Language Resource Agenda*. FLaReNet. http://www.flarenet.eu/sites/default/files/FLaReNet_Strategic_Language_Resource_Agenda.pdf

Desipri, E.; Gavrilidou, M.; Labropoulou, P.; Piperidis, S.; Frontini, F.; Monachini, M.; Arranz, V.; Mapelli, V.; Francopoulo, G. and Declerck, T. (2012). *META-NET Deliverable D7.2.4 – Documentation and User Manual of the META-SHARE Metadata Model (final)*. Available also as a working document at: http://www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf

Federmann, C.; Georgantopoulos, B.; del Gratta, R.; Magnini, B.; Mavroeidis, D.; Piperidis, S. and Speranza, M. (2011). *META-NET Deliverable D7.1.1 – METASHARE functional and technical specifications.*

Federmann, C.; Georgantopoulos, B.; Girardi, C.; Hamon, O.; Mavroeidis, D.; Minutoli, S. and Schröder, M. (2012). META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC2012)*, Turkey.

Gavrilidou, M.; Labropoulou, P.; Piperidis, S.; Speranza, M.; Monachini, M.; Arranz, V. and Francopoulo, G. (2011). *META-NET Deliverable D7.2.1 - Specification of Metadata-Based Descriptions for Language Resources and Technologies.*

ISO 12620. (2009). *Terminology and other language and content resources -- Specification of data categories and management of a Data Category Registry for language resources.* http://www.isocat.org

Monachini, M.; Quochi, V.; Calzolari, N.; Bel, N.; Budin, G.; Caselli, T.; Choukri, K.; et al. (2011). *The Standards' Landscape Towards an Interoperability Framework.* FLaReNet, CLARIN, META-NET. http://www.flarenet.eu/sites/default/files/FLaReNet_Standards_Landscape.pdf

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC2012)*, Turkey.