A Metadata Schema for the Description of Language Resources (LRs)

M. Gavrilidou, P. Labropoulou, S. Piperidis
ILSP / R.C. 'Athena'- Greece
{maria, penny, spip}@ilsp.gr

M. Monachini, F. Frontini
CNR – ILC - Italy
{monica.monachini, francesca.frontini}@ilc.cnr.it

G. Francopoulo
TAGMATICA - France
gil.francopoulo@tagmatica.com

V. Arranz, V. Mapelli
ELDA - France
{arranz, mapelli}@elda.org

Abstract

This paper presents the metadata schema for describing language resources (LRs) currently under development for the needs of META-SHARE, an open distributed facility for the exchange and sharing of LRs. An essential ingredient in its setup is the existence of formal and standardized LR descriptions, cornerstone of the interoperability layer of any such initiative. The description of LRs is granular and abstractive, combining the taxonomy of LRs with an inventory of a structured set of descriptive elements, of which only a minimal subset is obligatory; the schema additionally proposes recommended and optional elements. Moreover, the schema includes a set of relations catering for the appropriate inter-linking of resources. The current paper presents the main principles and features of the metadata schema, focusing on the description of text corpora and lexical / conceptual resources.

1 Credits

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

2 Introduction

The very diverse and heterogeneous landscape of huge amounts of digital and digitized resources collections (publications, datasets, multimedia files, processing tools, services and applications) has drastically transformed the requirements for their publication, archiving, discovery and long-term maintenance. Digital repositories provide the infrastructure for describing and documenting, storing, preserving, and making this information publicly available in an open, user-friendly and trusted way. Repositories represent an evolution of the digital libraries paradigm towards open access, advanced search capabilities and large-scale distributed architectures.

META-SHARE (www.meta-share.eu) is 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.

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

3 Design principles for the metadata model

The metadata descriptions constitute the means by which LR users identify the resources they seek. Thus, the META-SHARE metadata model (Gavrilidou et al., 2010) forms an integral part of the search and retrieval mechanism, with a subset of its elements serving as the access points to the LRs catalogue. The model must therefore be as informative and flexible as possible, allowing for multi-faceted search and viewing of the catalogue, as well as dynamic re-structuring thereof, offering LR consumers the chance to easily and quickly spot the resources they are looking for among

a large bulk of resources. Although META-SHARE aims at an informed community (HLT specialists), this is by no means interpreted as a permission to create a complex schema; user-friendliness of the search interface should be supported by a well motivated, easy-to-understand schema.

In this effort, we have built upon three main building blocks:

(a) study of *previous initiatives* (the most widespread in the LT area metadata models & LR catalogue descriptions[1]). The study has focused on the following issues: LR typologies, metadata elements currently in use and/or recommended, value types and obligatoriness thereof.

(b) *user requirements*, as collected through a survey conducted in the framework of the project (Federmann et al., 2011).

(c) *the recommendations of the e-IRG report of ESFRI* (e-IRG, 2009), in what concerns its purpose of usage, its aims and its features.

The basic design principles of the META-SHARE model are:

- semantic clarity: clear articulation of a term's meaning and its relations to other terms
- expressiveness: successful description of any type of resource
- flexibility: provision of complete descriptions of resources but also of minimal but informative descriptions
- customisability: adequate description of all types of resources (from the provider's perspective) and identification of the appropriate resource (user's perspective).
- interoperability (for exchange and harvesting purposes): mappings to at least the Dublin Core metadata & other widely used schemas and link of all elements to the ISOcat Data Categories
- user friendliness: provision of an editor to aid LR description
- extensibility: allow for future extensions, as regards both the model itself and the coverage of more resource types as they become available.
- harvestability: allow harvesting of the metadata (OAI-compatible).

4 The metadata model essentials

As a general framework, the mechanism we have decided to adopt is the *component*-based mechanism proposed by the ISO DCR model grouping together semantically coherent elements which form components and providing relations between them (Broeder et al., 2008). More specifically, ***elements*** are used to encode specific descriptive features of the LRs, while ***relations*** are used to link together resources that are included in the META-SHARE repository (e.g. original and derived, raw and annotated resources, a language resource and the tool that has been used to create it etc.), but also peripheral resources such as projects that created the LRs, standards used, related documentation etc.

The set of all the components and elements describing specific LR types and subtypes represent the ***profile*** of this type. Obviously, certain components include information common to all types of resources (e.g. identification, contact, licensing information etc.) and are, thus, used for all LRs, while others (e.g. components including information on the contents, annotation etc. of a resource) differ across types. The LR provider will be presented with proposed Profiles for each type, which can be used as templates or guidelines for the completion of the metadata description of the resource. Experience has proved that LR providers need guidelines and help in the process of metadata addition to their resources, and the Profiles are to be interpreted in this way and not as rigid structures to be adhered to.

In order to accommodate flexibility, the elements belong to two basic levels of description:

- an initial level providing the basic elements for the description of a resource (***minimal schema***), and

[1] The schemas taken into account include: Corpus Encoding Initiative (CES & XCES - www.xces.org/), Text Encoding Initiative (TEI - www.tei-c.org/index.xml), Open Language Archives Community (OLAC - www.language-archives.org/), ISLE Meta Data Initiative (IMDI - www.mpi.nl/IMDI/), European National Activities for Basic Language Resources (ENABLER - www.ilc.cnr.it/enabler-network/index.htm), Basic Metadata Description (BAMDES - www.theharvestingday.eu/docs/TheBAMDESIn2Pages-June2010.pdf), Dublin Core Metadata Initiative (DCMI - dublincore.org/), ELRA Catalogue (www.elra.info/Catalogue.html), ELRA Universal Catalogue (www.elra.info/Universal-Catalogue.html), LRE map (www.resourcebook.eu), LDC catalogue (www.ldc.upenn.edu/Catalog/), CLARIN metadata activities (www.clarin.eu) and the ISO 12620 – DCR (www.isocat.org/).

• a second level with a higher degree of granularity (***maximal schema***), providing more detailed information on each resource and covering all stages of LR production and use.

This has advantages for addition of metadata descriptions from scratch in two steps, first implementing the minimal schema, and subsequently, but not necessarily, the maximal schema. Harvesting is also served better by distinguishing between the two levels. Finally, LRs consumers can initially identify the resources best suited for their needs through the first level, and by accessing the second level, inspect the exact features of the resource.

The minimal schema contains those elements considered indispensable for LR description and identification. It takes into account the views expressed in the user survey concerning which features are considered sufficient to give a sound "identity" to a resource. It is considered as the "guarantee level" for interoperability as regards LR identification and metadata harvesting.

These two levels contain four classes of elements:

• the first level contains Mandatory (M) and Condition-dependent Mandatory (MC) elements (i.e. they have to be filled in when specific conditions are met), while

• the second level includes Recommended (R, i.e. LRs producers are advised to include information on these elements) and Optional (O) elements.

For each element, the appropriate field type has been chosen among the following options: free text, closed list of values, open list of values (recommended values are provided but users can add their own), numeric fields and special fields (e.g. urls, dates, phone numbers etc.). Special attention has been given to the choice of the field type, taking into consideration user requirements and metadata providers' practices; the intention has been to balance appropriately user-added with system-driven values in order to make the most of each approach. Consistency checking of user-added values will enhance the final results in the course of the META-SHARE operation.

Currently, the schema has been implemented as an XML schema (XSD), while implementation in RDF is also under consideration.[2]

To cater for semantic interoperability with other metadata schemas, all elements will be linked to existing ISOcat DCR data categories (ISO 12620, 2009) and, if they have no counterpart, they will be added to the DCR with appropriate definitions.

5 The META-SHARE ontology

META-SHARE takes a more global view on resources, which aims to provide users not only with a catalogue of LRs (data and tools) but also with information that can be used to enhance their exploitation. For instance, research papers that document the production of a resource as well as standards and best practice guidelines can play an informative role for LR users and an advisory role for prospective LR producers; similarly, information on the usage of a certain resource, as pointed out in the user interviews, is considered valuable for LR users wishing to find whether a certain resource is appropriate for their own application and the steps that they should take to get the best results.

Thus, the metadata model and its associated taxonomy should cover all types of resources (in the broad sense) to be included in META-SHARE.

In the proposed META-SHARE ontology, a distinction is made between LR per se and all other related resources/entities, such as:

• reference documents related to the resource (e.g. papers, reports, manuals etc.)

• persons and organizations involved in their creation and use (e.g. creators, funders, distributors etc.)

• related projects and activities (e.g. funding projects, activities of usage etc.)

• licenses (for the distribution of the LRs).

In the META-SHARE ontology, some of the entities will correspond to digital objects: for instance, all LRs descriptions will have a pointer to the resource itself, licenses and reference documents will point to document files (included in META-SHARE) etc. Entities such as persons and organizations, of course, can optionally be linked to external links (e.g. URL pointers for personal webpages). All these entities will be included in META-SHARE only so far as they are related to a LR.

The metadata model focuses on LRs per se (data and tools). For all other entities of the ontology, we take into account metadata schemas and relevant formats that have been de-

[2] In the current version, all relations are represented in the form of elements.

vised specifically for them, e.g. CERIF for research entities (projects, actors etc.), BibTex for bibliographical references etc.

6 Proposed LR taxonomy

Central to the model is the LR taxonomy, which allows us to organize the resources in a more structured way, taking into consideration the specificities of each type.

The study of the existing LR taxonomies has revealed their diversity, which hampers the request for interoperability.[3]

The proposed LR taxonomy constitutes an integral part of the metadata model, whereby the types of LRs (attributes and values) belong to the element set. The *resourceType* is the basic element according to which the LR types and subsequently the specific profiles are defined and may take one of the following values:

- **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, language models, typological databases, courseware etc.)
- **technology / tool / service** (including basic processing tools, applications, web services etc. required for processing data resources)
- **evaluation package** (for packages of datasets, tools and metrics used for evaluation purposes).

It should be noted here that, according to the practice of the HLT community, the term "language resource" is reserved for a collection/compilation of items (text, audio files etc.), mainly of considerable size or (in the case of tools) able to perform a well-defined task. Parts of LRs clearly identifiable can also be considered as LRs on their own: for instance, monolingual components of multilingual corpora can (and should) be regarded as monolingual corpora themselves. But the focus is on the set rather than the unit (e.g. single text / audio file, in the case of corpora, or word / entry, in the case of lexica).

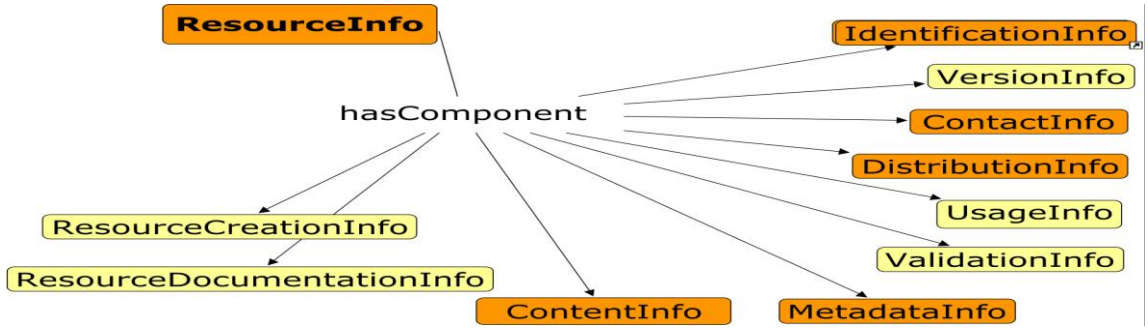

Figure 1 - ResourceInfo - the common components for all LRs

Further sub-classification is dependent upon sets of type-dependent features, which allow the viewing of the same resource along multiple dimensions. Thus, for instance *language* as an organizing feature can be used to bring together monolingual corpora / lexica and monolingual parts of multilingual corpora / lexica. Similarly, *domain*, *format*, *annotation* features etc. can be used as different dimensions according to which the catalogue of LRs can be accessed.

7 Contents of the model

The core of the model is the *ResourceInfo* component (Figure 1), which contains all the information relevant for the description of a LR. It subsumes components and elements that combine together to provide this description. A broad distinction can be made between the "administrative" components, which are common to all LRs, and the components that are idiosyncratic to a specific LR type (e.g. *CorpusInfo*, *LexicalConceptualResourceInfo* etc., as explained further below). For instance, elements needed for the description of video resources are only used for the specific *mediaType*.

The set of components that are common to all LRs are the following:

- the *IdentificationInfo* component includes all elements required to identify the resource, such as the resource full and short name, the persistent identifier (PID, to be as-

[3] For a more detailed discussion on the LR taxonomy discrepancies, cf. Gavrilidou et al. (2011).

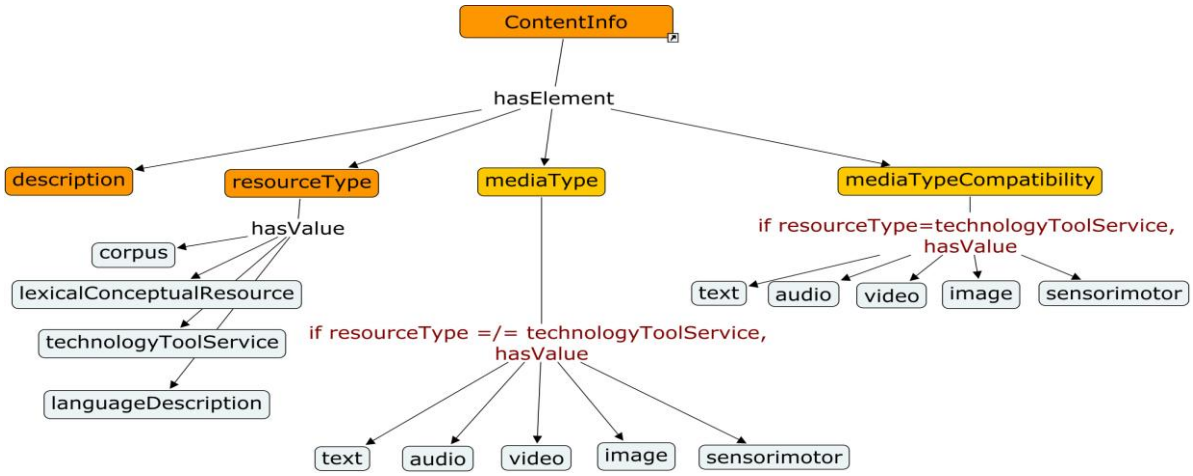

Figure 2 - The ContentInfo component and its elements

signed automatically by the system), identifiers attributed by the source organization or other entities (e.g. ELRA, LDC identifiers) etc.

- the *PersonInfo* component provides information about the person that can be contacted for further information or access to the resource
- all information relative to versioning and revisions of the resource is included in the *VersionInfo* component
- 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 *LicenseInfo* component, which gives a description of the licensing conditions under which the resource can be used; linking to the license documents themselves is also possible through the relevant relation.
- 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 *ResourceCreationInfo* and its dependent components group together information regarding the creation of a resource (creation dates, funding information such as funder(s), relevant project name etc.)
- the *UsageInfo* component aims at providing information on the intended 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 *MetadataInfo* is responsible for all information relative to the metadata record creation, such as the catalog from which the harvesting was made and the date of harvesting (in the case of harvested records) or the creation date and metadata creator (in case of records created from scratch using the metadata editor) etc.
- the *ResourceDocumentationInfo* provides information on publications and documents describing the resource; basic documents (e.g. manuals, tagset documents) can (and should be) included in the META-SHARE repository; the possibility to introduce links to published web documents and/or import bibliographic references in standard formats will be catered for
- finally, the *ContentInfo* component describes the essence of the resource, specifying the *resourceType* and the *mediaType* elements, which give rise to specific components, distinct for each LR type, as presented below.

A further set of four components enjoy a "special" status in the sense that they can be attached to various components, namely *PersonInfo*, *OrganizationInfo*, *CommunicationInfo* and *SizeInfo*. For instance, *PersonInfo* and *OrganizationInfo* can be used for all persons/organizations acting as resource creators, distributors etc. Similarly, *sizeInfo* can be used either for the size of a whole resource or, in combination with another component, to describe the size of parts of the resource (e.g. per domain, per language etc.).

The *ContentInfo* component (Figure 2) is meant to group together descriptive information as regards the contents of the resource. The elements included are:

- *description*: free text of the resource

• *resourceType* with the values corpus, lexical/conceptual resource, language description, technology/tool/service, evaluation package

• *mediaType* (used for data resources) & *mediaTypeCompatibility* (used for tools): the notion of medium constitutes an important descriptive and classificatory element for corpora but also for tools; it is preferred over the written/spoken/multimodal distinction, as it has clearer semantics and allows us to view resources as a set of modules, each of which can be described through a distinctive set of features. The following media type values are foreseen:

• text: used for resources with only written medium (and modules of spoken and multimodal corpora),

• 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 will be described by the relevant Text feature set

• image (+ text): the Image feature set is used for photographs, drawings etc., while the Text set will be reserved for its captions

• video: moving image (+ text) (+ audio (+ text): used for multi-media corpora, with Video for the moving image part, Audio for the dialogues, and Text referring to the transcripts of the dialogues and/or subtitles

• sensorimotor: used for sensorimotor resources which contain data collected through the use of relevant equipment (gloves, helmets, body suits, laryngographs, etc.) and used to measure the activity of non-verbal modalities (such as gestures, facial expressions, body movements, gaze, articulatory activity, etc.) and their interaction with objects, be it common objects or control sequences of human-machine interaction (keyboard, mouse, touch screen).

A resource may consist of parts belonging to different types of media: for instance, a multimodal corpus includes a video part (moving image), an audio part (e.g. dialogues) and a text part (subtitles and/or transcription of the dialogues); a multimedia lexicon includes the text part, but also a video and/or an audio part; a sign language resource is also a good example for a resource with various media types. Similarly, tools can be applied to resources of particular types of medium: e.g. a tool can be used both for video and for audio files.

Each of the values of the *resourceType* and *mediaType* gives rise to a new component, respectively:

• *CorpusInfo*, *LexicalConceptualResourceInfo*, *LanguageDescriptionInfo*, *TechnologyToolServiceInfo* and *EvaluationPackageInfo* which include information specific to each LR type (e.g. subtypes of corpora and lexical/conceptual resources, tasks performed for tools etc.)

• *TextInfo*, *AudioInfo*, *VideoInfo*, *ImageInfo* and *SensorimotorInfo* which provide information depending on the media type of a resource; this information can be broadly described as belonging to one of the following categories (all represented in the form of components and elements):

• content: it mainly refers to languages covered in the resource and classificatory information (e.g. domains, geographic coverage, time coverage, setting, type of content etc.)

• format: file format, size, duration, character encoding etc.; obviously, this information is more media-type-driven (e.g. we have different file formats for text, audio and video files)

• creation: this is to be distinguished from the *ResourceCreationInfo* which is attached to the resource level; at the resource level, it is mainly used to give information on funding but also on anything that concerns the creation of the resource as a whole; at the media-type level, it refers to the creation of the specific files, e.g. the original source, the capture method (e.g. scanning and web crawling for texts, vs. recording methods for audio files)

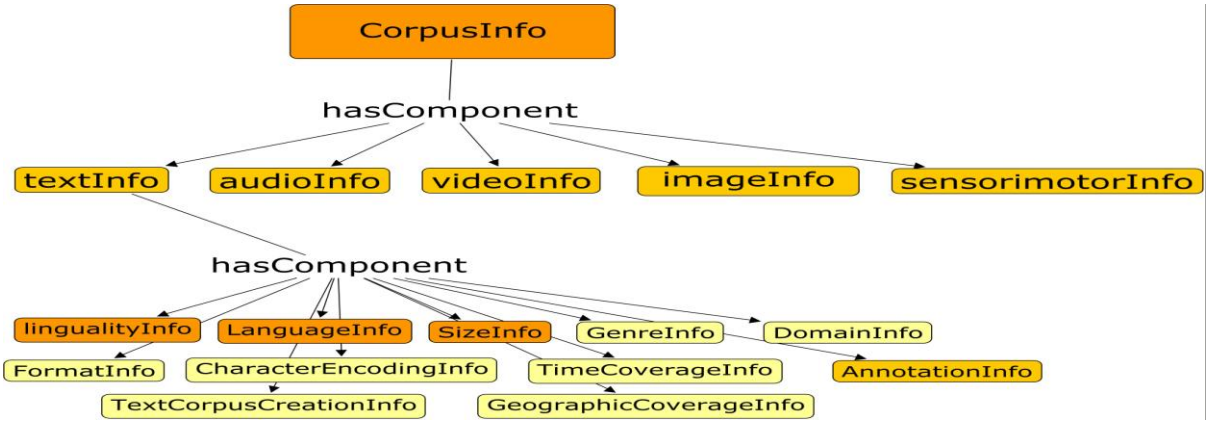

Figure 3 - Excerpt of the CorpusInfo component focusing on text corpora

• linguistic information encoding: the relevant components include information on the types, theoretic models, methods, tools etc. used for adding linguistic information to the resource, which takes the form of encoding for lexica and annotation for corpora and tools; it is both resource-type- and media-type-driven (e.g. morpho-syntactic tagging, parsing, semantic annotation is used for text files, while transcription, prosody annotation etc. for audio parts/corpora etc.).

The mandatory generic components and elements thereof for the description of a resource (for the **minimal schema**) are:

• *IdentificationInfo*, incl. name of the resource and persistent identifier

• *ContentInfo*: all elements (*description*, *resourceType* & *mediaType*) are mandatory

• *DistributionInfo*: *availability* must be filled in and depending on the type of availability, further elements are mandatory (e.g. license, distributor and distribution/access medium for all available resources, types of restrictions for resources available under restrictions etc.)

• *MetadataInfo*: depending on the way the metadata record has been created (harvesting vs. manual creation), a different set of elements must be filled in, some of which are automatically provided (e.g. *metadataCreationDate* vs. *harvestingDate*, *metadataCreator* vs. *source* etc.)

• *PersonInfo*: at least an *email* must be provided for the contact person.

Depending on the resource type, a further set of components are mandatory.

In the next sections, we provide a more detailed view of text corpora and lexical / conceptual resources as exemplary cases of the model.

8 Text corpora

Text corpora are marked as such by the element *resourceType=corpus* & *mediaType=text* and their description must include a *CorpusInfo* component and a *TextInfo* one (Figure 3). As aforementioned, here we include, alongside the traditional text corpora, also the textual parts of audio corpora (transcriptions) and video ones (e.g. subtitles).

Besides the generic components, the type dependent information for text corpora is represented in the following components:

• *LingualityInfo*: it provides information on the linguality type (mono-/bi-/multilingual corpora) and multilinguality type of text resources (parallel vs. comparable corpora)

• *LanguageInfo*: it comprises information on the language(s) of a resource and can be repeated for all languages of the resource; a *LanguageVarietyInfo* component is foreseen to supply further information if the resource includes data in regional language varieties, dialects, slang etc.

• *SizeInfo*: it provides information on the size of the whole resource but it can also be attached to every other component that needs a specification of size (e.g. size per language, per format etc.);

• *AnnotationInfo*: it groups information on the annotation of text corpora, such as specification of the types of annotation level (e.g. segmentation, alignment, structural annotation, lemmatization, semantic annotation etc.), annotation methods and tools etc.

The above four components are obligatory for all text corpora. A further set of components are recommended:

• *FormatInfo*: it gives information on the format (in the form of mime-type) of the corpus

• *CharacterEncodingInfo*: it includes information on character encoding of the resource

• *TextCorpusCreationInfo*: it is used to provide specific information on the creation of the text files, as aforementioned;

• finally, four components are used to give information on the classification of the corpus, namely: *TimeCoverageInfo* (for the time period of the texts), *GeographicCoverageInfo* (for the geographic region from which the texts are collected), *DomainInfo* (presenting the domains covered by the corpus) and *TextGenreInfo* (for the text genre / text type of the texts).

9 Lexical / Conceptual resources

The type dependent subschema for lexical / conceptual resources (LCRs) is activated if the *resourceType* element of the *ContentInfo* component has the value *lexicalConceptualResource* (Figure 4). If this condition is verified, the *LexicalConceptualResourceInfo* component becomes mandatory. In this component a first mandatory element is *lexicalConceptualResourceType*, where the provider is asked to define the type of LRC under description. There is still an open debate on what should be the values to be given in this part and as to which should be the labels thereof. An open list is currently proposed, its suggested values being: *wordList; computationalLexicon; ontology; wordnet; thesaurus; framenet; terminologicalResource; machineReadableDictionary*. Providers can choose to add other values if they consider these not appropriate.

Two optional components are foreseen:

• *LexicalConceptualResourceCreationInfo*, where information on the *originalSource*, a string field where the main sources (dictionaries, grammars, lexica, corpora,) for the creation of the LCR are listed; *creationMode*, with a closed list of values (automatic, semi-automatic, manual, mixed interactive); *creationModeDetails*, which allows to further specify the theoretical and practical principles that guided the creation of the resource; *creationTool*, a repeatable element where either a string, a url or a hyperlink can be entered, the latter enabling the provider to create a connection between the resource and the tool(s) used for its development.

• *LexicalConceptualResourceEncodingInfo* (which is recommended) groups all information regarding the contents of the LCR; it includes the following elements: the mandatory element *encodingLevel* with an open list of values (e.g. phonetics; phonology; semantics), the optional but more detailed *linguisticInformation* with a complex set of suggested values of a varying degree of granularity (e.g. partOfSpeech, syntax-SubcatFrame, semantics-relations, semantics-Relations-Synonyms, semantics-Relations-Antonyms etc.) and the optional *extratextualInformation* (with values images, videos, soundRecordings); this last element can be used for multimedia lexica; if a more detailed account is considered appropriate, the *AudioInfo*, *VideoInfo*, *ImageInfo* components can also be used.

The *TextInfo* and its subsumed components are also to be used for the description of LCRs; the only exceptions are the *TextGenre* and *Annotation* components, which are specific to text corpora.

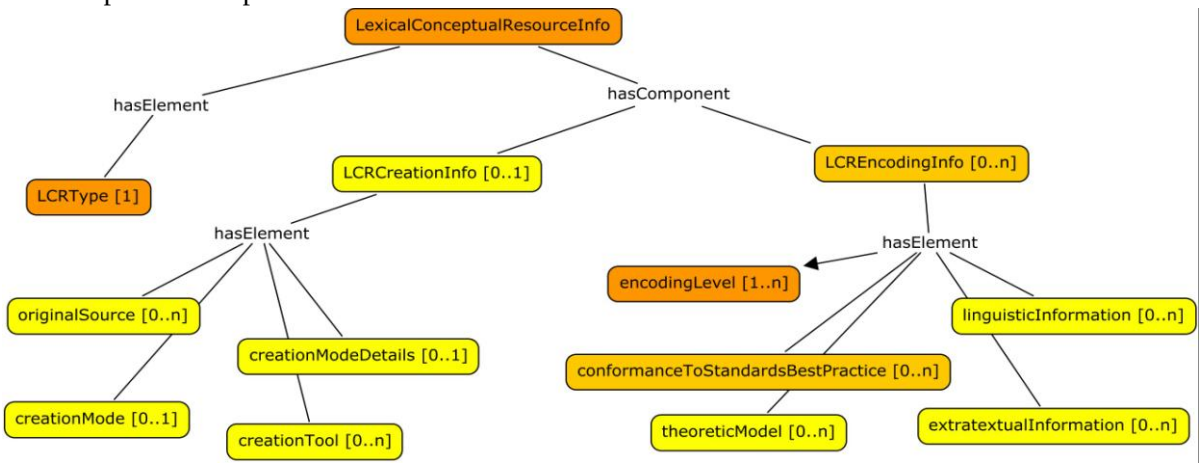

Figure 4 - The components specific to lexical/conceptual resources

10 Conclusions and future work

The current version contains, besides the general presentation of the model, the application of the model to text corpora & to LCRs as presented above. The next steps include:

- extension to other media and LR types: the application of the model to the remaining media types (*audio, video, image, sensorimotor*) and LR types (*languageDescription; technologyToolService; evaluationPackages*) is ongoing. In this process, the expressive power of the model is being tested and it is expected that new components and elements will arise.
- exemplary instantiations: a set of resources selected to represent all LR and media types is being described according to the model, in order to test its functionality; these resources with their descriptions will be uploaded in the prototype infrastructure for testing and exemplification purposes.
- discussion with experts group: this version of the model will be communicated to the metadata experts group that has been set up within WP7, with the purpose of getting feedback for its improvement.
- implementation of the schema for the description of LRs produced or collected by three collaborating projects, namely META-NET4U, CESAR and META-NORD.

References

Broeder, D., T. Declerck, E. Hinrichs, S. Piperidis, L. Romary, N. Calzolari and P. Wittenburg (2008). Foundation of a Component-based Flexible Registry for Language Resources and Technology. In Proceedings of the 6th International Conference of Language Resources and Evaluation (LREC 2008).

e-IRG (2009). eIRG Report on Data Mangement. http://www.e-irg.eu/images/stories/publ/task_force_reports/dmtfjointreport.pdf

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

Gavrilidou M., P. Labropoulou, E. Desipri, S. Piperidis (2010). Preliminary Proposal for a Metadata Schema for the Description of Language Resources (LRs) in Proceedings of the Workshop 'Language Resource and Language Technology Standards – state of the art, emerging needs, and future developments', LREC 2010, Malta 2010.

Gavrilidou, M., P. Labropoulou, S. Piperidis, M. Speranza, M. Monachini, V. Arranz, G. Francopoulo (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