DEMOCRITUS UNIVERSITY of THRACE

selected papers of the 10th International Conference of Greek Linguistics

Edited by

Zoe Gavriilidou
Angeliki Efthymiou
Evangelia Thomadaki
Penelope Kambakis-Vougiouklis

Komotini 2012

Οργανωτική Επιτροπή Συνεδρίου
Organizing Committee

Zoe Gavriilidou
Angeliki Efthymiou
Evangelia Thomadaki
Penelope Kambakis-Vougiouklis

Γραμματειακή Υποστήριξη
Secretarial Support

Ioannis Anagnostopoulos
Maria Georganta
Polyxeni Intze
Nikos Mathioudakis
Lidija Mitits
Eleni Papadopoulou
Anna Sarafianou
Elina Chadjipapa

ISBN 978-960-99486-7-8

Τυπογραφική επιμέλεια

Νίκος Μαθιουδάκης
Ελένη Παπαδοπούλου
Ελίνα Χατζηπαπά

Σχεδιασμός εξώφυλλου

Νίκος Μαθιουδάκης

Εργαστήριο Σύνταξης, Μορφολογίας, Φωνητικής, Σημασιολογίας, *+ΜόρΦωΣη* ΔΠΘ
Laboratory of Syntax, Morphology, Phonetics, Semantics, *+MorPhoSE* DUTH

Διεθνές Συνέδριο Ελληνικής Γλωσσολογίας
International Conference of Greek Linguistics
www.icgl.gr

A PROPOSAL FOR A METADATA MODEL FOR LANGUAGE RESOURCES

Maria Gavrilidou
R.C. Athena/ILSP,
Greece
maria@ilsp.athena-innovation.gr

Penny Labropoulou
R.C. Athena/ILSP,
Greece
penny@ilsp.athena-innovation.gr

Stelios Piperidis
R.C. Athena/ILSP,
Greece
spip@ilsp.athena-innovation.gr

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 LR type 'corpora' as an exemplification case and concludes with work to be done in the future for the improvement of the model.

1. Introduction[1]

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 proved that a critical mass of LRs can make advancement in language research possible and quicker (Calzolari, Quochi, and Soria 2011).

Language data collection has started in the 50's with a shift of focus from the native speaker's intuition to the actual use of language. Technological advancements and the advent of the web have moved the attention of researchers to the quick and efficient analysis of huge bulks of data. Digital repositories constitute a valuable tool in the effort of publishing, archiving, discovery and long-term maintenance 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.

META-SHARE (www.meta-share.eu) is an open, integrated, secure and interoperable exchange infrastructure dedicated to LRs; it serves as a marketplace 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,
- by facilitating the LRs search and retrieval processes,
- by facilitating the evaluation of LRs through comparison between similar LRs,
- by encouraging (re-)use and new use of LRs through the monitoring of actual LRs use.

[1] 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.

In Z. Gavriilidou, A. Efthymiou, E. Thomadaki & P. Kambakis-Vougiouklis (eds), 2012,
Selected papers of the 10th ICGL, pp. 284-293. Komotini/Greece: Democritus University of Thrace.

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 an integral part of the search and retrieval mechanism, with a subset of its elements serving as the access points to the LRs catalogue. In this effort, we have built upon three main building blocks:

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

(b) the recommendations of the e-IRG report of ESFRI (e-IRG 2009, http://www.e-irg.eu), in what concerns purpose of usage, aims and features[2],

(c) a study of widespread metadata models in HLT and LR catalogue descriptions. The schemas and catalogues taken into account include:

- the Corpus Encoding Standard (CES, http://www.cs.vassar.edu/CES/) & its XML version (XCES, http://www.xces.org/), which instantiates the EAGLES CES DTDs for linguistic corpora;
- the Text Encoding Initiative (TEI, http://www.tei-c.org/index.xml), which develops and maintains a standard for the representation of digital texts. TEI has most notably provided guidelines for the encoding of machine-readable texts, mainly in the fields of humanities, social sciences and linguistics;
- the Open Language Archives Community (OLAC, http://www.language-archives.org/), which aims at developing best practices for the digital archiving of language resources, and at implementing a network of interoperating repositories and services for hosting and accessing such resources;
- the Dublin Core Metadata Initiative (DCMI, http://dublincore.org/), which, as part of its mission, develops and maintains specifications in support of resource description;
- the ISLE MetaData Initiative (IMDI, http://www.mpi.nl/IMDI/), which is a metadata standard for the description of multi-media and multi-modal language resources;
- the metadata model proposed by the European National Activities for Basic Language Resources project (ENABLER, http://www.ilsp.gr/en/infoprojects/meta?view=project&task =show&id=121);
- the metadata-related activities of the CLARIN project (Common Language Resources and Technology Infrastructure, http://www.clarin.eu/external/), aiming to offer persistent services and provide easy access to language processing resources;
- the Basic Metadata Description (BAMDES), a minimal metadata set used for harvesting purposes by the Harvesting Day initiative (http://theharvestingday.eu/), a routine in which a robot collects metadata descriptions of resources and tools, as published at their websites;
- the European Language Resources Association (ELRA, http://www.elra.info/) resources, namely the ELRA Catalogue (resources distributed by ELRA), the ELRA Universal Catalogue (which comprises information regarding LRs identified all over the world) and the LRE map (a mechanism intended to monitor the use and creation of LRs by collecting information on both existing and newly-created resources);
- the Linguistic Data Consortium (LDC, http://www.ldc.upenn.edu/) catalogue of available resources. The LDC supports language-related education, research and technology development by creating and sharing linguistic resources;
- and last but not least, 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.

[2] For a detailed presentation, cf. (Gavrilidou et al. 2011).

The study of these initiatives revealed that, although general trends can be spotted, there is no consensus as regards LRs typology. The various typologies present different views on LRs categorisation, and two tendencies have been attested in practice: on the one hand there are well-structured typologies for the classification of resources, and on the other hand there is the trend for free categorisation, whereby the provider declares the type of the resource. The first solution lacks flexibility (some resources might not fit into the predefined types), while the latter lacks uniformity and consistency. Furthermore, diverging uses of terminology hinder interoperability between metadata schemas.

The concept of *resource type* seems to be crucial to all metadata schemas and cataloguing practices, given that it constitutes the basic concept for the organization of language resources and determines a critical subset of elements related to their description. As regards the set of descriptive elements selected by each schema, consensus up to a certain degree is attested. The naming of the elements may vary but fundamental properties of LRs (e.g. identification details, resource name, free-text description) are in general present in all schemas.

From the standards and models surveyed, the DCMI standard is the most widespread metadata initiative, going back to the 90's with the advent of the internet, originating in works of library and archive cataloguing. The DC metadata element set refers to a basic set of 15 elements; refinements to this set have already been made and are documented in the DC Metadata Terms. As for LR typology, DC obviously is not restricted to LRs, and, since it was not built for this specific purpose, its terms are not sufficient for the description of language resources.

Inspired by the advantages and disadvantages of the surveyed standards, the basic principles of the METASHARE model were formulated.

- The semantic discrepancies between the standards dictated the need for **semantic clarity**, i.e. clear articulation of a term's meaning and its relations to other terms.
- The fact that certain standards focus on specific language resource types but do not cover all, led to the formulation of the principle of **expressiveness**, i.e. the ability of the model for successful description of any type of resource.
- The differing tendencies attested as regards granularity led to the principle of **flexibility**, i.e. the possibility for exhaustive but also for minimal descriptions.
- The constant emergence of new types of resources which were not covered by existing standards dictated the principle of **extensibility**, i.e. catering for future extensions, as regards the coverage of more resource types as they become available.
- Given that the metadata descriptions should be usable by other initiatives, the principle of **interoperability** is adhered to, which foresees mappings to widely used schemas (mainly DC, OLAC and ISO-DCR).
- Finally, the need for open metadata available to other initiatives led to the adoption of the principle of **harvestability**, allowing the harvesting of the metadata.

Based on these principles, the META-SHARE metadata model was designed and implemented as described in the following sections.

3. The metadata model essentials

As a general framework, the mechanism adopted for the META-SHARE metadata model is the *component-based* mechanism proposed by the ISO DCR (ISO 12620/2009), according to which semantically coherent elements are grouped together to form components (Broeder et al. 2008). *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. raw and annotated resources, a language resource and the tool that has been used to create it etc.), but also satellite resources such as standards used, related documentation etc.

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 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 resource types. The user is presented with proposed profiles for each LR type, which can be used as templates or guidelines for the completion of the metadata description of the resource.

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). 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. LRs producers are asked to fill in at least the minimal schema and, thereafter, enrich the descriptions of their LRs with recommended and optional elements, should they wish to do so.

In addition, the schema specifies the type allowed for all elements[3] (e.g. if the values are of type *string, number, closed set of values* etc.).

4. The META-SHARE ontology

META-SHARE takes a global view on resources, aiming at providing users not only with a catalogue of LRs 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 guidelines are informative for LR users and advisory for prospective LR producers.

In the proposed META-SHARE ontology (Figure 1), a distinction is made between LRs per se and all other related entities, such as reference documents related to the resource (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.) and licences (for the distribution of the LRs).

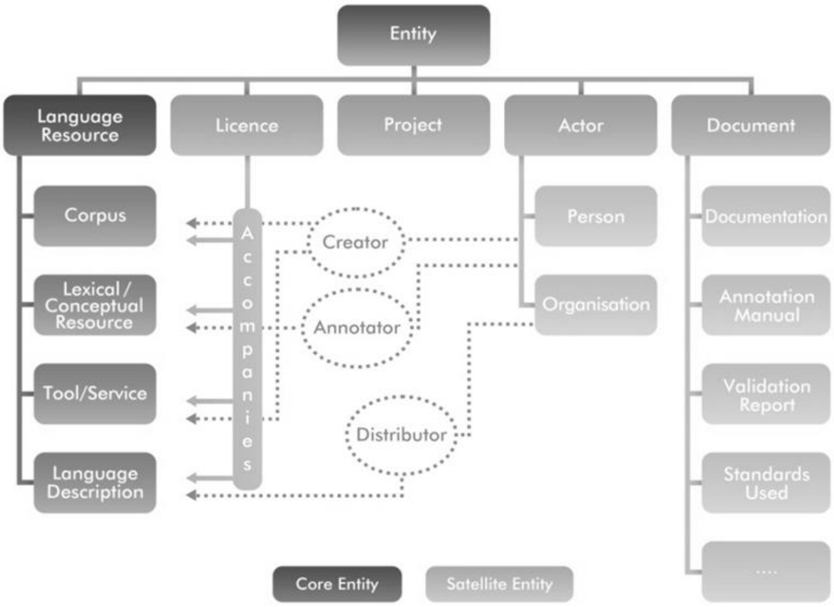

Figure 1 META-SHARE ontology

Thus, the META-SHARE model recognizes the following distinct entities:

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

[3] Relations are also implemented as elements.

- the *document,* and
- the *licence*.

From these, however, the main interest of META-SHARE lies in resources, which constitute the central entity to be described; actors, projects, documents etc. are described when the case arises, i.e. when they are linked to a specific resource. Therefore, it is not expected from META-SHARE, for instance, to provide a bibliographical list of all documents that are relevant to the HLT domain, but only those documents that are related to specific resources (e.g. articles describing the creation and/or use of a resource, documentation manuals, annotation guidelines etc.).

Consequently, the META-SHARE metadata model aims at covering only LRs per se. For all other entities of the ontology, metadata schemas and formats that have been devised specifically for them (e.g. BibTex for bibliographical references) have been taken into account.

5. Proposed LRs taxonomy

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-NET to provide a simple albeit descriptive schema for LRs.

To encompass this issue, the META-SHARE LRs taxonomy is based on intrinsic criteria, i.e. makes use of elements included in the schema. The proposed classification forms an integral part of the metadata model, whereby the types of LRs (attributes and values) belong to the element set itself. A two-level hierarchy, with a coarse "main type" classification and further subclassifying features dependent on each type, is proposed. For the first level, 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.)
- *tool / service* (including processing tools, applications, web services etc. required for processing data resources)
- *language description* (including grammars, typological databases, courseware etc.).

Depending on the *resourceType* values, the LR types and subsequently the specific profiles (i.e. aggregations of components and elements) are defined.

The second element considered crucial for the description and classification of the resources is the physical medium (element *mediaType*). It is preferred over the written/spoken/multimodal distinction, as it has clearer semantics. Moreover, each medium type enforces for the description of the resources a particular set of features which differs across media.

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 (dialogues) and a text part (subtitles and/or transcription of the dialogues); a multimedia lexicon includes the text part, but may also include 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, comparable or parallel
- *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 probability measures for the latter. This categorization allows us to better treat them in the metadata model.

Finally, LR users can devise their own LR taxonomy, by browsing through the META-SHARE inventory using any of the metadata elements (and combinations thereof) as classification criteria. Thus, for instance *lingualityType* as an organizing feature can be used to bring together monolingual data resources and monolingual parts of multilingual ones. Similarly, *languageName*, *domain*, *format*, annotation features etc. can be used as different dimensions according to which the catalogue of LRs can be accessed.

6. Contents of the model

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

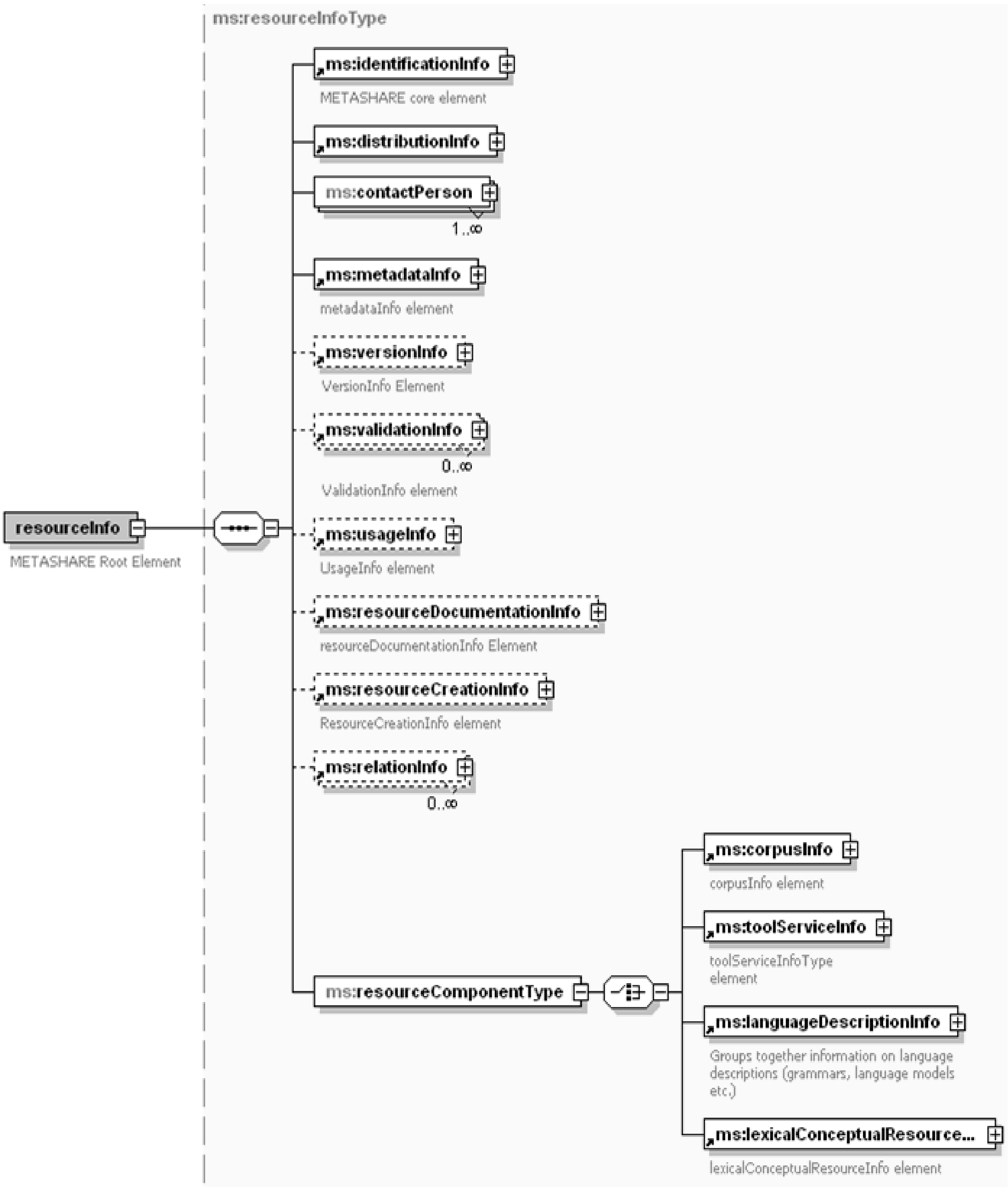

Figure 2 Common components for all LRs and resourceType components

A broad distinction can be made between the "administrative" components, which are common to all LRs, and the resource type- and media type-specific components.

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 resource full and short names, the META-SHARE id (to be assigned automatically by the system) etc.; the *description* element is obligatorily used for the free text description of the resource contents.

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 catalogue 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 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 codification 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.

The 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 recorded information for them 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). The *resourceType* and *mediaType* elements encode the two classification axes of the schema, while each of the values of these two elements is associated with the appropriate component. The set of *resourceType* and *mediaType* components includes:

- *corpusInfo*, *lexicalConceptualResourceInfo*, *languageDescriptionInfo*, *toolServiceInfo* encode information specific to each LR type; the values *corpus*, *lexical/conceptualResource*, *languageDescription* and *toolService* are used for the element *resourceType* respectively
- *corpusTextInfo*, *corpusAudioInfo*, *corpusVideoInfo*, *lexicalConceptualResourceTextInfo*, *lexicalConceptualResourceVideoInfo* 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.

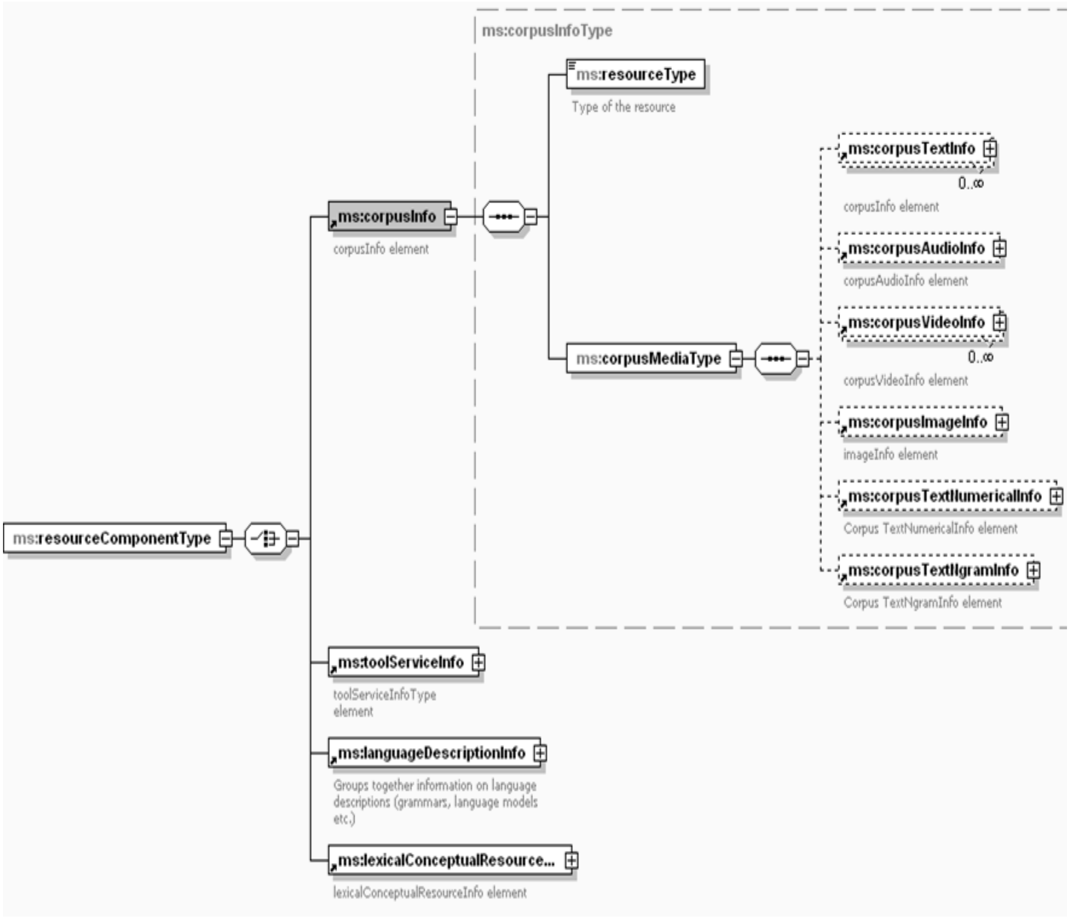

Figure 3 Components for corpora

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

- contents: it mainly refers 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: it includes 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: 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 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)
- 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 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, for multimedia resources, a special component, *linkToOtherMediaInfo*, is provided for linking between the various modules of the resource.

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-share id* 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 license 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.)
- *resourceComponentType*: as aforesaid, this groups together the various LR-type-dependent components; thus, depending on the type of LR described, one of the following components is obligatory: *corpusInfo* for corpora, *lexicalConceptualResourceInfo* for lexical / conceptual resources 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. Corpora

This section presents the metadata schema, using corpora as an exemplary case.

Depending on the ***mediaType***, at least one of the following components must appear: *corpusTextInfo*, *corpusAudioInfo*, *corpusVideoInfo*, *corpusImageInfo*, *corpusTextNumericalInfo* and/or *corpusTextNgramInfo* (cf. figure 3). Thus, for text corpora, the user will select to encode a *corpusTextInfo* component. A multimedia/multimodal corpus that includes videos, the transcribed dialogues thereof, the scenario used for the videos and the motion data captured by sensors (e.g. gloves, full-body equipment) and represented in the form of numerical text will require the encoding of a *corpusVideoInfo*, two *corpusTextInfo* and a *corpusTextNumericalInfo* components; the *linkToOtherMediaInfo* will provide the necessary information as to the linking and synchronization between them.

All of these components subsume further obligatory, recommended and optional components. As aforesaid, two types of information are obligatory for all of these components: *language(s*) and *size*. More specifically, as regards languages, *lingualityInfo* groups information regarding the number of languages included in the resource and the relation between them (e.g. monolingual, multilingual parallel corpus etc.), while *languageInfo* groups information respective to the specific languages and, if applicable, language varieties covered by the resource; in the case of languages, the two relevant elements (*languageId* and *languageName*) must conform to the IETF BCP47 standard (http://www.rfc-editor.org/rfc/bcp/bcp47.txt). Size information is recorded in the *sizeInfo* component; an effort to standardize the recorded information as far as possible has been made: thus, the *sizeInfo* component includes two elements, namely *sizeUnit* with values taken from an open controlled vocabulary[4] and

[4] "Open controlled vocabularies" are an important tool in the META-SHARE model as they bring together the advantages of two competing tendencies in metadata editing: controlled vocabularies allow for the standardization of information by providing a closed set of values from which users can choose but which cannot be easily updated and/or extended; user-added values, on the other hand, give more freedom but quite often the result is a list of similarly expressed values (e.g. *txt*, *TXT*, *text*, *texts*, *textual* etc. as alternatives for *text*). META-SHARE proposes the use of an intermediary tool, where users are provided with a set of predefined values for a given element, but

size, which specifies the size of the resource with regard to the *sizeUnit* measurement in the form of a number.

The *timeCoverageInfo*, *geographicCoverageInfo* and *domainInfo* can be used to provide information on the time, geographic and domain classification of the resource and/or resource parts. Further classification is also provided dependent on the *mediaType*: for the text corpora, *textGenre*, *textType*, *register* etc. are the suggested elements, for the video, *videoGenre* etc.

Creation information is also medium-type dependent: for instance, for audio and video resources (or modules), users can provide information on the recording equipment, the setting, the group of participants, the capturing equipment etc.

Further information can be provided for the *modalities* included in each resource module (in the *modalityInfo* component): so, for instance, for text resources, written vs. spoken language can be specified, while for audiovisual resources one can specify whether gestures, body movements etc. are contained.

The *formatting* information, which is important for the interoperability with tools and services, can also be encoded for each medium-specific part of the resource (e.g. the video part of the resource consists of WAV files, the text part consists of TXT and XML files etc.); character encoding is also included but obviously pertains only to text resources.

Finally, the *annotationInfo* component groups information on the *annotation* of the resource: annotation type, tool, method, process used, standards/best practices adopted etc. For each different type of annotation (e.g. lemmatization, semantic annotation, modality annotation etc.) the component is repeated, thus providing the correct linking between the various annotation details (e.g. which annotation tool has been used for which annotation type, in case of multiple annotation tools used).

9. Conclusions and future work

The current schema has been adopted and utilized for the description of 1,277 resources (datasets and tools), covering a broad variety of languages, resource and media types, available through META-SHARE. The model has been implemented as an XML schema, documented also in the form of a manual with detailed information, including definitions, examples and guidelines for the usage of the whole schema and each element (Desipri et al. 2012). Future work focuses of the completion of the schema as regards both breadth (i.e. coverage of more types) as well as depth (i.e. improvements on the schema based on LR providers' feedback).

References

Broeder, Dan, Thierry Declercq, Erhard Hinrichs, Stelios Piperidis, Roland Romary, Nicoletta Calzolari, and Peter 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*.

Calzolari, Nicoletta, Valeria Quochi, and Claudia Soria. 2011. "The Strategic Language Resource Agenda". FLaReNet. http://www.flarenet.eu/sites/default/files/FLaReNet_Strategic_Language_Resource_Agenda.pdf.

Desipri, Elina, Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis, Francesca Frontini, Monica Monachini, Victoria Arranz, Valerie Mapelli, Gil Francopoulo, and Thierry Declercq. 2012. "META-NET Deliverable D7.2.4 – Documentation and User Manual of the META-SHARE Metadata Model (final)." Ed. Penny Labropoulou and Elina Desipri.

e-IRG. 2009. "e-IRG Report on Data Management." http://www.eirg.eu/images/stories/publ/task_force_reports/dmtfjointreport.pdf.

Federmann, Christian, Byron Georgantopoulos, Ricardo del Gratta, Bernardo Magnini, Dimitris Mavroeidis, Stelios Piperidis, and Manuela Speranza. 2011. "META-NET Deliverable D7.1.1 – METASHARE Functional and Technical Specifications."

Gavrilidou, Maria, Penny Labropoulou, Stelios Piperidis, Manuela Speranza, Monica Monachini, Victoria Arranz, and Gil 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.

they are also allowed to add their own values (by choosing the "other" value and inputting a new value); a regular checking of the new values will allow the better monitoring of the controlled vocabularies.