This document is part of the Network of Excellence "Multilingual Europe Technology Alliance (META-NET)",
co- funded by the 7th Framework Programme of the European Commission through the T4ME grant agreement no.: 249119.

A Network of Excellence forging the
Multilingual Europe Technology Alliance

Specification of metadata-based descriptions for language resources and technologies

Author(s): Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis, Manuela Speranza, Monica Monachini, Victoria Arranz, Gil Francopoulo

Dissemination Level: Restricted

Date: January 20, 2011

Grant agreement no.	249119
Project acronym	T4ME Net (META-NET)
Project full title	Technologies for the Multilingual European Information Society
Funding scheme	Network of Excellence
Coordinator	Prof. Hans Uszkoreit (DFKI)
Start date, duration	1 February 2010, 36 months
Distribution	Restricted
Contractual date of delivery	January 31, 2011
Actual date of delivery	January 20, 2011
Deliverable number	D7.2
Deliverable title	Specification of metadata-based description for language resources and technologies
Type	Report
Status and version	Pre-final version
Number of pages	44
Contributing partners	DFKI, CNR, **ILSP**, CNRS, UU, ELDA, FBK
WP leader	ILSP
Task leader	ILSP
Authors	Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis, Manuela Speranza, Monica Monachini, Victoria Arranz, Gil Francopoulo
EC project officer	Hanna Klimek
The partners in META-NET are:	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
	Barcelona Media (BM), Spain
	Consiglio Nazionale Ricerche – Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR), Italy
	Institute for Language and Speech Processing, R.C. "Athena" (ILSP), Greece
	Charles University in Prague (CUP), Czech Republic
	Centre National de la Recherche Scientifique – Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (CNRS), France
	Universiteit Utrecht (UU), Netherlands
	Aalto University (AALTO), Finland
	Fondazione Bruno Kessler (FBK), Italy
	Dublin City University (DCU), Ireland
	Rheinisch-Westfälische Technische Hochschule Aachen (RWTH), Germany
	Jozef Stefan Institute (JSI), Slovenia
	Evaluations and Language Resources Distribution Agency (ELDA), France

For copies of reports, updates on project activities and other META-NET-related information, contact:

DFKI GmbH
META-NET
Dr. Georg Rehm
Alt-Moabit 91c
10559 Berlin, Germany

office@meta-net.eu
Phone: +49 (30) 3949-1833
Fax: +49 (30) 3949-1810

Copies of reports and other material can also be accessed via http://www.meta-net.eu

Table of Contents

1 Acknowledgements

The present report documents the work that has been done in the framework of WP7.2. Many people (members of the META-SHARE metadata working group) have contributed and/or provided feedback on the model. The group has worked together in two meetings:

- one meeting in Malta during the LREC 2010 Conference and included the metadata experts as well as representatives of the ICT projects that will implement the META-SHARE metadata model, and
- two Technical Meetings of the project in Athens (Kick-off /1st Technical Meeting and mainly 3rd Technical Meeting).

Besides the meetings, several consultation rounds facilitated the provision of valuable feedback and comments by the contributing partners.
Many thanks are due to all these people for their contribution.

Malta Metadata experts meeting (list of participants)
Project members: Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis (ILSP), Pavel Straňák, Jan Hajič (Charles Univ.), Brigitte Jörg, Georg Rehm, Hans Uszkoreit (DFKI), Victoria Arranz, Djamel Mostefa, Khalid Choukri (ELDA), Bernardo Magnini, Manuela Speranza (FBK), Monica Monachini, Claudia Soria (CNR-ILC), Marta Villegas, Maite Melero (Barcelona Media).
ICT projects representatives: Voula Giouli (*ACCURAT*, ILSP), Ahmet Aker (*ACCURAT*, Univ. Sheffield), Ulrich Heid (*TTC*, Univ. Stuttgart), Carla Parra (*PANACEA*, UPF).
Experts: Laurent Romary (LORIA), Marc Kemps-Snijders, Peter Wittenburg, Dieter van Vytvanck, Daan Broeder (MPI), Christopher Cieri (UPenn), Roberto Cencioni (EU).

Project members' contributors
Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis (ILSP), Victoria Arranz, Khalid Choukri, Valerie Mapelli, Priscille Schneller (ELDA), Nicoletta Calzolari, Monica Monachini (CNR-ILC), Gil Francopoulo, Gilles Adda, Joseph Mariani (LIMSI), Georg Rehm (DFKI), Pavel Straňák (Charles Univ.).

2 Executive Summary

This document reports on the metadata model that will be used for the description of Language Resources (LRs) made available through META-SHARE, the open distributed facility for the sharing and exchange of resources META-NET is building – hereafter, referred to as "META-SHARE metadata model". First of all, it puts the model in the context

of its application (LR sharing), delineating the intended goals of use and factors to be taken into account for its design. It moves on to the presentation and discussion of the two main building blocks that have influenced the design of the model, namely (a) the user requirements, as collected through a survey conducted in the framework of the project and (b) an overview of the most widespread metadata models and catalogue descriptions of LRs. Finally, it presents the model itself, i.e. its basic concepts, the META-SHARE ontology and LR taxonomy, as well as a synopsis of the main descriptive mechanisms and elements included. Particular emphasis is put to the presentation of the minimal schema, i.e. the subset of the META-SHARE model which, consisting of elements considered indispensable for the description of LRs, will be the minimum required level of description for resources to be uploaded in META-SHARE. The report concludes with future work considerations.

3 Introduction

The current deliverable presents the metadata model proposed for the description of language resources (LRs) made available through META-SHARE, the ***open distributed facility for the sharing and exchange of resources*** META-NET is building. As foreseen in the DoW, Deliverable D 7.2 has two versions:

- the present one (version 1, due M12) which focuses on the META-SHARE metadata model basic principles and concepts, and its implementation as regards one LR type, namely **written corpora**, for validation and exemplification purposes, and
- the final one (version 2, due M24), which extends the model to cover all other LR types (spoken, multimodal, lexical and tools/technologies).

In the context of META-SHARE, the term **metadata** refers to descriptions of **Language Resources**, encompassing both ***data resources*** (textual, multimodal/multimedia and lexical data, grammars, language models etc.) and ***tools/technologies/services used for their processing***. These are also found in the literature as Language Resources and Technologies (LRTs).

4 Purpose and goals

The META-SHARE metadata descriptions will constitute the means by which LR users will identify the resources they seek in the META-SHARE context. Thus, the META-SHARE 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. 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 should by no means be 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 intend to build upon previous initiatives so that the model is *easily adopted* by the target community. The aim is not to create yet another competing metadata model but rather to adapt existing resource description models to a unified proposal catering for the specific requirements of the community.

The proposed model takes into account the results of the user requirements survey, aiming to cover the needs expressed by the interviewees (cf. section 4), but also aims to follow the recommendations of the e-IRG report of ESFRI (e-IRG Report on Data Management, Nov. 2009), in what concerns its purpose of usage, its aims and its features. Thus, it aims to

- be useful to LRs providers, service providers and users alike
- provide clear, semantically transparent (as far as possible) terminology, supported by definitions, recommended values and examples
- contain elements suitable for the description of all stages of a resource's life-cycle (provenance, creation, annotation, distribution etc.) and critical facets of its identity (IPR, licensing issues, administrative data etc.)
- adhere to standards for the adoption of methodologies, elements' names etc. in order to be of high quality, persistent and interoperable with other schemas and tools (for instance, harvesting tools)
- if that is not possible, follow best practices and *de facto* standards.

It is in this way that recommendation R10 (availability) of the e-IRG report "It is a MUST for all resource and service providers to create and provide quality metadata descriptions" can be met.
Based on these recommendations, the basic design principles of the META-SHARE model are:

- semantic clarity: the meaning of each term and its relations to the other terms should be clearly articulated
- expressiveness: it should be able to successfully describe any type of resource

- flexibility: it should be able to constitute a tool for exhaustive and complete descriptions of resources (should the resource provider wish so) but also allow for minimal but informative descriptions of the resource
- customisability: the model should be able to adequately describe all types of resources (from the provider's perspective) and to aid the user to identify the resource appropriate to his/her needs (user's perspective). Therefore, it aims to be
    - adaptable to each LR type
    - adaptable to each LR producer description requirements
    - adaptable to each LR user search requirements
- interoperability (for exchange and harvesting purposes)
- the model should lend itself to the development of mappers to at least the Dublin Core metadata standard (cf. section 5.1.8) & other widely used schemas (being persistent, compatible with standards and best practices)
- user friendliness
- it should guide the LR producer through an editor to choose the appropriate metadata set for the description of the LR
- extensibility: it should allow for future extensions, as regards both the model itself and as regards the coverage of more resource types as they become available.
- harvestability: it should allow harvesting of the metadata (OAI-compatible) but also metadata production from scratch for LR providers who have not as yet added any kind of metadata to their resources.

To accommodate the above requirements, WP7.2 relies on the following:

- a user requirements survey and usage scenarios discussed in the framework of the same pillar (presented in section 5); and
- a comparative study of the most widespread in the area metadata standards and relevant practices (presented in section 6).

5 User requirements survey

The user requirements survey was envisaged as the first step by which we would be able to analyse the needs of the various HLT players in the endeavour to define the best model for

the infrastructure. The collection of users' requirements and its analysis will lead to the specification of the major features and the desirable principles of the infrastructure.
As specified in the DoW, the following classes of target users have been identified:

- **Providers**: this class of users includes any individual or organisation who accesses META-SHARE to make available a LR. Significant functionalities for this class of users include: 1) uploading, linking etc. facilities; 2) metadata tools such as mappers for the conversion of the original description to the META-SHARE model, as well as editors for the addition of metadata from scratch in the case of LRs with no previous metadata descriptions; 3) licensing templates to serve as models for the providers to specify the terms of use of their LR; 4) information on similar resources (e.g. the descriptions other providers have used, statistics on the demand for a similar resource, its users etc.).
- **Consumers**: this class of users includes any individual or organisation accessing META-SHARE to search for LRs. Relevant functionalities for consumers include: 1) registration, authentication, authorisation; 2) search, browsing and filtering etc. of the LR catalogue; 3) retrieval and download of LR descriptions; 4) browsing, viewing etc. of the LRs themselves; 5) licensing, that is, any kind of procedure the consumer goes through to be allowed to use a specific LR, complying with restrictions imposed by the provider; 6) exploitation of the functionalities related to the actual usage of a resource: downloading, providing feedback (if needed/appropriate), but also possibly exploiting computational services directly provided by the infrastructure (e.g. pos tagging or syntactic annotation of a corpus).
- **Repositories**: this is a special class of Providers, and includes aggregators. These are individuals or organisations making available aggregated LRs, by linking them to the infrastructure. They implement their own policies for resource aggregation, (re)use, validation, harvesting etc. The functionalities are: 1) links to/provision of repositories and inventories i.e. catalogues/ directories); 2) LR inventories (registering, updating etc.); 3) metadata harvesting (from announced trusted sites); 4) archiving and preservation services; 5) IPR, licensing etc. (help desk); 6) distribution services.

For all three types of users, various "reporting" services will be made available, such as the number of times a specific LR has been accessed/downloaded, different counts (e.g. top downloads), statistics according to different metadata elements (languages, applications etc.), recommendation services (e.g. resources commonly used with other specific resources etc.)

5.1 The user survey methodology

In May 2010, FBK conducted interviews at LREC 2010 (MALTA, 18-21 May 2010) with the goal of producing functionality specifications based on existing experiences and also through the direct involvement of potential users. The basic user requirements for the META-SHARE platform, which are described in detail in Deliverable 7.1, are the result of a four-step process: (i) collection of the interviews during LREC 2010 in Malta, (ii) transcription on the content of those interviews, (iii) synthesis of the content of the interviews, and (iv) extraction of user requirements.

For the collection of input from users, we have opted for interviews over more structured methodologies (e.g. questionnaires) due to the flexibility of this methodology, which allowed for the collection of very specific requirements that would hardly have emerged from the latter. In addition, this choice was further motivated by the fact that the LREC 2010 conference would take place exactly in the period in which the collection had also been scheduled.

With the help of the T4ME consortium, we first defined a list of candidates to be interviewed; they were in part members of the institutions involved in T4ME and in part not, which helped us to build a mixed sample of people, e.g. to include also people involved in other projects and initiatives in the field of LR infrastructures.

Candidates for interviews were asked to prepare a small usage scenario based on a specific (type of) LR (i.e. data or tool) they intended to provide to / get from META-SHARE, so that the interviews were anchored in a concrete situation. As META-SHARE users are divided into the three classes we mentioned above (i.e. consumers, providers and repositories), each with specific perspectives, we formulated different questions for each group of users.

In total, we collected 23 interviews (10 for the consumer typology, 11 for the provider typology and 2 representatives of a repository), each lasting on average about 20-30 minutes. As regards the first two classes (consumers and providers), we interviewed a sample of people composed of 3 PhD students, 7 Post Doc and 11 Tenured Researchers/Professors having spent most of their activity in the last three years in 8 different countries: Czech Republic, Italy, Germany, Spain, Netherlands, Greece, Latvia and France.

5.2 User survey results

Both consumers and providers think that information about a LR is very important. All consumers, in fact, expect to find some information to help them understand if a certain LR is relevant for them and all providers are willing to provide information about the resource they are making available.

The pieces of information that consumers and providers mention more frequently (in order of importance) are the **language/s** (mentioned by 9 in total), the **size** (mentioned by 8), the **author/s** (7), and the **source and acquisition modalities** (6). Other important pieces of information perceived as very important are **annotations** available (5), **licensing information** (5), **type** of resource (4), **documentation** or link to documentation (4), **date** of **creation** (4), **uses** of the resource (3) and **contact address** (3). As for contact information, in particular, it is suggested that, upon provider's acceptance, consumers might also view additional pieces of information about the provider.
Two interviewees mention the name of the resource, its operating system, its domain, its format, a sample of it, its webpage, why it has been developed and, if annotations are available, whether they are manual or automatic. Other pieces of information mentioned are: tagging scheme, project webpage if created in a project, content, links to (automatic) annotations provided by other people, known bugs, download link, versions available, for annotations the quality (e.g. inter-annotator agreement), software requirements, relevant publications, programming language, published in (if different from the documentation), stage to define quality/completeness (e.g. alpha, beta, production), memory, if web service or executable, technical data and requirements, text encoding and format of the input.
More specifically about consumers, five of them are interested in having information about the use of the resource by others, mentioning mainly publications about the resource and other people who are working on it and which kind of work they are doing, but also download counts. When asked directly, also the other consumers say that they consider such information useful, mainly to get a feeling of the quality, but also to increase comparability.

As for the format in which information about a resource should be provided, all providers consider it fine to provide information by filling a form with metadata, but most of them would like to also have the possibility to add some free text.

The requirements recorded by the user needs survey were carefully studied in the process of defining the metadata model to be used for LR description in META-SHARE.

6 Overview of LR description schemas and activities

A survey has been conducted among the most widely used *metadata schemas* and *standards* in the LR area, as well as the sets of elements used for the description of resources in *catalogues* of LRs and related software, focusing on the following issues: LR typologies, metadata elements currently in use and/or recommended, value types and obligatoriness thereof. Hereafter, a brief presentation of each schema/catalogue is given followed by a short discussion of the findings. Correspondences with their elements are provided for the META-SHARE model elements (Appendix A); these will be utilized in the implementation of the respective schema mappers/converters to the META-SHARE schema.

6.1 Summary of metadata schemas and related efforts

6.1.1 ISO 12620 - Data Category Registry (ISO DCR)

The ISO Data Category Registry (ISOcat DCR, http://www.isocat.org/,) is "an attempt to achieve interoperability among the various metadata schemas". "ISO 12620:2009 provides guidelines concerning constraints related to the implementation of a Data Category Registry (DCR) applicable to all types of language resources, for example, terminological, lexicographical, corpus-based, machine translation etc. It specifies mechanisms for creating, selecting and maintaining data categories, as well as an interchange format for representing them." Thus, interoperability is achieved through the registration of elements ("data categories"), which refer to widely used concepts in the linguistics domain; users can subsequently link their own elements to them (or add new ones according to the ISO 12620 framework requirements), thus achieving common terminology.

Data categories are categorized in thematic areas (metadata, morphosyntax, syntax, language resource ontology etc.). The thematic area of Metadata, which is relevant to our work, is responsible for the "set of data categories that can be used to describe language data resources, web services and applications with keyword type of metadata. It supports all linguistic data types that are required in the emerging eScience scenario, such as speech and multimedia recordings, written resources, annotations, lexicons, data category registries and ontologies, schemas, tools of different sorts and many others."
Until now, 274 data categories have been registered in the metadata area, most of them by the CLARIN metadata working group; this set of data categories are the outcome of a critical review of the most widespread metadata schemas (also taken into account in our survey).

In contrast to the usual hierarchical systems employed until now by most metadata schemas, data categories are listed in a flat structure, promoting a more flexible and user-driven approach, where the same element can be re-used at various levels of the resource description.
As with other generic schemas, users can create "profiles", using the elements they consider appropriate for the description of particular types of resources and/or specific projects (e.g. "written corpus profile", "lexicon profile", or "project-specific profile"). "Components" are an important ingredient in this mechanism, as they group together elements (or even other components) – for instance, the "Organization" component can group together details necessary for the description of organization, including name and communication data for an organization.

Regarding LR typology, in accordance to this philosophy, ISOcat does not propose any specific hierarchical system and, in addition, remains open as to the values of possible LR types. Two data categories included so far are specific to LR typology, both implemented as complex/open elements (i.e. with an open set of values):

- *lexicon type*, with example values: *monolingual dictionary, bilingual dictionary, word list, thesaurus, glossary term base*, and
- *corpus type*, with example values: *comparable corpus, parallel corpus, monitor corpus, general corpus, specialised corpus, learner corpus, reference corpus etc.*

6.1.2 Corpus Encoding Initiative (CES & XCES)

The Corpus Encoding Initiative (CES) and its XML version (XCES, http://www.xces.org/) adopts the TEI philosophy (see below) catering particularly for the needs of linguistic corpora. It has been influential in the design and construction of major corpora in the HLT area and is still currently being used. XCES recommendations take the form of the "corpus header", grouping all information relevant to the construction, identification and description of a corpus, and the "file headers", which are pertinent to the description of each text file; moreover, the CES includes recommendations for the structural annotation of texts (e.g. recognition of textual segments, typesetting conventions etc.) as well as for the representation of linguistic annotation.

In our survey we have focused on the corpus header, as this is considered more relevant to our purposes, especially with respect to monolingual raw and annotated corpora. Obligatoriness of elements has been considered for the drafting of the minimal core subset.

6.1.3 Text Encoding Initiative (TEI)

The Text Encoding Initiative (TEI, http://www.tei-c.org/index.xml) consortium has developed a "standard for the representation of texts in digital form", currently the most widely used one in the area of humanities. The TEI P5 guidelines (the most recent version of TEI) include recommendations both for the bibliographic-style description of texts and text collections (external metadata in the form of "headers" – a notion that has influenced other metadata standards also) as well as for the representation of the internal structure of the texts themselves.

Obviously, in our survey, we have restricted the study to the external metadata. It should be noted that our evaluation has taken into consideration the fact that both TEI and OLAC (see below) are more humanities-oriented.

6.1.4 Open Language Archives Community (OLAC)

The Open Language Archives Community (OLAC, http://www.language-archives.org/) is "an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources."
The OLAC metadata standard, which has been used for the description of a huge bulk of LRs[1], is DC-compliant (cf. section 5.1.8), i.e. it uses the complete set of DC metadata terms with OLAC-specific extensions (and the possibility to define further extensions, in the form of community-specific qualifiers, where required) for the description of LRs. OLAC archives are harvestable using the OAI (Open Archives Initiative) protocol.

With regard to the taxonomy of LRs, OLAC recommends the use of one of the values of the DCMItype vocabulary as well as the use of an extension, namely *linguistic type* with values *primary text, lexicon* and *language description*. The repository also includes resources classified as *other resources about the language* and *other resources in the language*. It should be noted that the use of another OLAC classificatory element, *discourse type*, as a means to classify LRs is too refined and is not used in OLAC either for the broad classification.
As far as the elements are concerned, OLAC includes deliberately only a small set thereof, giving priority to interoperability over wealth of information. Moreover, important for our

[1] OLAC archives in January 2011 contain approximately 35,000 records, covering resources in half of the world's living languages.

evaluation is the fact that it includes also paper resources, which are of no value to META-SHARE.

6.1.5 ISLE Meta Data Initiative (IMDI)

The ISLE Meta Data Initiative (IMDI, http://www.mpi.nl/IMDI/) is a metadata standard proposed for the description of multi-media and multi-modal LRs, accompanied with tools that support both corpus and resource-specific browsing and search.

Together with other standards, it has been a primary source for the ISO DCR, and an important standard for our own work; one of its assets is the multi-media/multi-modal view that it brings in to the survey, as other metadata activities mainly have the written medium as a starting point.
In terms of LR taxonomy, as envisaged in this overview, it seems that IMDI recognizes the following types of resources: *MediaFile*, *WrittenResource* and *LexiconResource*. *MediaFile* groups information for all media resources, including text, audio and video resources. For each of these, a *type* element exists, but only the *MediaFile.type* is a closed vocabulary set.

6.1.6 European National Activities for Basic Language Resources (ENABLER)

The European National Activities for Basic Language Resources (ENABLER, http://www.ilc.cnr.it/enabler-network/index.htm) project was a Network aimed at coordinating national activities established by European states concerning LRs.

Within the framework of the ENABLER network and, more specifically, for the construction of the Catalogue of Language Resources, a metadata schema has been elaborated for the description of LRs and related tools. The ENABLER model is supported by tools both for adding LRTs descriptions and for searching the catalogue and includes a set of metadata elements generic to all resources (e.g. identification and creation information) and a set specific to each resource type (e.g. annotation information for corpora etc.), which have also been considered for inclusion in the ISO-DCR through the CLARIN metadata working group.

ENABLER includes a two-level hierarchical LR taxonomy. The first level distinguishes between *text resources*, *speech resources*, *multimodal resources* and *lexical resources* and *tools*. The second level of taxonomy is type-dependent; we should note here that in the questionnaires used for the survey, the second level is not always as apparent. A number of features can be used for the subclassification of resources. In the presentation of the survey results, one or more of these features have become more prominent:

- for written language resources, there is no explicit subtype; various dimensions can be used for subclassification (e.g. *multilinguality type*, *language coverage* etc.)
- for spoken language resources, development type has the values: read, spontaneous, monologue, dialogue, text and other
- for multimodal resources, *format type* has the values: *speech, gestures, text, sound* and *other*
- lexical resources are distinguished (lexicon type) into computational lexicon, MRD, printed, multimedia dictionary and other
- for tools, the types used are tagger, parser, tokenizer, lemmatizer, NE recognizer and other.

The values are predefined but users can add their own values – when selecting the option *other* they are prompted to further specify the type.

6.1.7 Basic Metadata Description (BAMDES)

The Basic Metadata Description (BAMDES, http://www.theharvestingday.eu/docs/TheBAMDESIn2Pages-June2010.pdf) is a proposal for a minimal set of metadata elements to be used for the description of LRs in the framework of the Harvesting Day Initiative (http://www.theharvestingday.eu). It follows the ENABLER model (with two subsets of elements, generic and resource-type-specific) but limits the metadata to only a very basic set, in an effort to persuade LRs producers to engage in the activity and make their resources visible to interested communities.

In our context, the set of BAMDES elements has been considered mainly for the specification of the META-SHARE minimal subset.
As for LR taxonomy, five basic types are acknowledged: *lexical resource, written corpus, multimodal corpus, oral corpus* and *tool*. Each type can be further subclassified by predefined values, specific to the type. Thus:

- for lexical resources, two criteria are mixed together, namely that of number of languages and a more classical *lexicon type* approach;
- for all types of corpora, the *corpus type* values are the same, although some values are appropriate only for specific types (e.g. *transcribed* should apply only to oral or multimodal corpora); the criterion of number of languages

included in the corpus is again present, showing the importance attributed to it;

- for tools, there is no specific *type* feature, although *operations* can be used as a classificatory element.

6.1.8 Dublin Core Metadata Initiative (DCMI)

The Dublin Core Metadata Initiative (DCMI, http://dublincore.org/) is "an open organization engaged in the development of interoperable metadata standards that support a broad range of purposes and business models." 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.

The importance of DC lies in its widespread use as the basis for exchange of metadata descriptions between various schemas. Thus, the minimum requirement for the META-SHARE minimal core is to be DC-compliant in order to achieve interoperability.
As for LR typology, DC obviously is not restricted to LRs. It includes an element *type* used to document "the nature or genre of the resource". The recommended values included in the DCMIterms vocabulary, which are of relevance to our work, are: *collection, dataset, image, movingImage, service, software, sound, stillImage* and *text*. Moreover, these values can be used accumulatively: e.g. a written language corpus could be coded as *collection* and *text*, a video instance could be an aggregation of *movingImage*, *sound* and *text*.

6.1.9 ELRA Catalogue

The European Language Resources Association's (ELRA) missions are "to promote language resources for the Human Language Technology (HLT) sector, and to evaluate language engineering technologies"; one of its major services in this capacity is the identification and distribution of LRs. The ELRA catalogue (http://www.elra.info/Catalogue.html) includes all resources available through ELDA, ELRA's distribution agency, described according to a resource-type-specific set of elements.

The ELRA LR typological system is a two-level one: the first level distinguishes between *spoken, written, terminological* and *multimedia/multimodal resources* while the second level is type-specific and, in fact, is used only for the *spoken* and *written resources*. Lexical

resources are subsumed under the *written* or *spoken* resources depending on the type of information included in them. Tools/technologies are not included in the catalogue.
It is noteworthy, however, that in the section requesting information from interested resource providers (http://catalog.elra.info/distribution_procedure.php), description templates are structured according to a different typology: *corpus, lexicon, speech* and *multimodal*. Although not explicitly stated, terminological resources seem to be catered for under *lexicon* in this schema. Again, a second level of classification is introduced for the different LR types.
The ELRA set of elements and recommendations for LR descriptions have been a valuable source in setting up the META-SHARE metadata model given that they are more tuned to the HLT community and, more importantly, they incorporate the experience of LRs distribution practices and user preferences.

6.1.10 ELRA Universal Catalogue

The ELRA Universal Catalogue (http://www.elra.info/Universal-Catalogue.html) is a repository established by ELRA comprising information regarding LRs identified all over the world. Unlike the ELRA catalogue which is limited to LRs distributed through the agency, it includes information on all LRs, available or not through any distribution channel. Addition of LRs to this Catalogue can be made both by the ELRA team as well as interested LRs producers.

The description of LRs is made according to a minimal set of elements, which has been considered for the minimal META-SHARE model.
In the Universal Catalogue, LRs can be searched via the same *typological* system used in the ELRA catalogue, notably with the addition of *tools*.

6.1.11 LRE map

The LRE Map (http://www.resourcebook.eu) is an initiative undertaken for the first time in the framework of the LREC2010 conference and extended to other LR-related conferences (COLING & EMNLP until now). During the paper submission procedure, authors were asked to submit a form providing information on the resource(s) and tools related to their paper.

The form was kept deliberately simple in order not to overburden authors and, thus, consisted of a set of 12 elements. Values have also been oversimplified: for each element a list of suggested values (the most common/popular ones) is supplied with the addition of *other*, which prompts users to add their own values. This approach of mixing pre-set values with user-added free text has yielded interesting conclusions as to preferences of metadata

providers, some of which are discussed in (Calzolari et al. 2010, "The LREC 2010 Resource Map", *LREC 2010 Proceedings*). Moreover, the LRE map is an important source for information relative to the use of resources.

The descriptor *Resource type* has been used to elicit classificatory information from the users. The main values used as a response (i.e. those that appeared more than 20 times) are: *corpus, lexicon, tagger/parser, annotation tool, ontology, evaluation data, representation-annotation formalism/guidelines, grammar/language model, evaluation tool, terminology, named entity recognizer, representation-annotation standard/best practice*. Two comments should be made here:

- the LRE map takes a very broad view on the resources documented, including not only data and tools but also metadata, guidelines etc. - as evident by the values *representation-annotation formalism/guidelines* and *representation-annotation standard/best practice*
- given the fact that the values are user-driven, we note a mixture of values recognized in other schemas/descriptions as higher-order values (e.g. *corpus, lexicon*) together with lower-order ones (e.g. *evaluation tool, named entity recognizer* instead of the more generic *tool*).

6.1.12 LDC catalogue

The Linguistic Data Consortium (LDC) is "an open consortium of universities, companies and government research laboratories. It creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes." The LDC catalogue (http://www.ldc.upenn.edu/Catalog/) includes "corpus resources" (incl. text, speech, video and lexicon resources) distributed through the LDC. The catalogue is browsable and searchable while the resource description is provided by the LRs producer in accordance to the set of elements and recommendations provided by LDC for that purpose.

Search of the catalogue can be made through ready-made menus as well as a simple search form. One of the criteria used for the predefined search facilities is that of *type*. LDC considers all resources as "corpora" which are further "divided into major categories according to the type of data they contain": *lexicon, speech, text, transcripts* and *video*; combinations thereof are also possible (e.g. *lexicon & speech & text*). As in the ELRA catalogues, a two-level hierarchy is used; the second level is based on the uniform criterion of *data source*, in contrast to the different criteria used for the subclassification of major types in the ELRA catalogues (e.g. mainly data source for spoken resources, number of languages for lexica etc.).

The *data type* and *data source* questions included in the form to be filled in by resource providers (http://www.ldc.upenn.edu/Providing/) are both free text, although the exemplary values listed guide the users to the expected answers. The degree to which the answers are subsequently harmonized is unknown.

6.1.13 CLARIN activities

The CLARIN (Common Language Resources and Technology Infrastructure, www.clarin.eu) project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable. LR typology constitutes an issue that has been viewed from various perspectives in a number of its activities, which are presented in the following subsections.

6.1.13.1 CLARIN inventory of LRs and LTs

The CLARIN consortium (WP5 – Language Resources and Technologies Overview) has carried out over the internet a questionnaire-based survey of LRTs. The results can be found at: http://www.clarin.eu/view_resources (LRs) and http://www.clarin.eu/view_tools (LTs).

A two-level hierarchy is adopted in this inventory again. The main distinction is between *Language Resources* and *Tools*. Additionally, each of these categories is further broken down into minor categories, different for each type: both questionnaires include a *type* question, where a pre-defined set of values is provided, from which users can choose multiple values (*other* among them); explanations are provided for tool types in a help page. The suggested type values have been selected from previous relevant initiatives but make up together a mixture of values based on different criteria: e.g. *aligned corpus*, *treebank* and *written corpus* are all possible values at the same level, but the first two values refer to the annotation dimension while the latter to the medium dimension. The multiple choice option serves as a remedy to this drawback. Currently[2] 27 out of 887 LRs (3%) and 38 out of 228 tools (16%) have been assigned *other* for "type".

As for the questions, they cover a rather broad set of descriptive features, different for resources and tools. Attentions should be drawn to the fact that metadata descriptions are rather poor, while the metadata fields that the respondents to the survey have decided to fill in for the various resources differ a lot between them. This lack of consensus as to which features are considered important by LR providers should be taken into account for the specification of the META-SHARE minimal schema.

[2] The LRT catalogue facility remains open and is constantly updated; the current deliverable takes into account findings of the version of January 2011.

6.1.13.2 CLARIN metadata Working Group

The CLARIN metadata Working Group, formed by experts representing the various domains in the wider LR area, is largely responsible for the population of the ISOcat DCR metadata thematic group (see section 5.1.1). The work has been based on a detailed comparison and contrast of metadata elements used for LR description in major metadata schemas and related activities (e.g. IMDI, DC, OLAC, ENABLER etc.).

At the stage of discussion, as regards typology, a distinction was made between *annotations, lexica, lists, media, texts* and *tools*: metadata elements are characterized as being appropriate for the description of one or more of these categories (http://www.clarin.eu/view_datcats). However, this distinction has not been transferred to the ISOcat DCR. Moreover, from the three elements that can be considered as classificatory elements found in the original list, namely *LexiconType*, *WrittenResource-Type* and *WrittenResource-Subtype*, only the former has made it to the ISOcat DCR.

6.1.13.3 Virtual Language Observatory

The Virtual Language Observatory (VLO, http://www.clarin.eu/vlo/) is a CLARIN-initiated activity aiming to present LRs from various perspectives, ranging from more traditional approaches, such as the original menu-driven viewing of the CLARIN inventory of LRs and tools and the browsing of CLARIN harvested metadata using the IMDI browser, to more advanced ones, as the geographical browsing of LRs (exploiting Google Earth) and the faceted browsing of harvested metadata.
In the two latter approaches, the notion of *LR* taxonomy is more loosely defined:

- in the geographical browsing approach, LRs are organized along the geographical dimension, i.e. resources are found attached to countries of origin;
- in the faceted search and browsing facility, the distinction is made only between *language resources* and *tools*. For LRs, there is no mention of LR type while browsing is organized along the following dimensions: origin, continent, country, language, organization, genre and subject. For tools, the *type of tool* is one of the suggested modes of browsing, along with the following elements: contributor, language, platform, organization and license. The values for *type of tool* come from the harvested metadata, e.g. inherit the values from the CLARIN inventory and the NLSR catalogue.

6.1.14 Natural Language Software Registry

The Natural Language Software Registry (NLSR, http://registry.dfki.de/) mainly lists language tools and technologies. As stated in the site, it is a "concise summary of the capabilities and sources of a large amount of natural language processing (NLP) software available to the NLP community." Although its focus is on tools, data resources are not excluded from the catalogue.

The NLSR taxonomy is structured at two levels, but there is no distinction between tools and resources; in fact *language resource* is one of the values alongside all other values that could be considered as subtypes of *tool*. As for the tool type distinction, this mixes two criteria:

- the type of content to be covered by the tool: *written language*, *spoken language*, *multimedia* and *multimodality*,
- the task performed: *annotation tool, evaluation tool, NLP development aid*.

Multiple choice is possible when submitting an entry in order to cater for appropriate encoding. For most of these values, type-dependent subtypes are also provided (predefined values) to allow for a more detailed classification. For the *Language Resource* type, values follow the classical distinctions: *written language corpora*, *spoken language corpora*, *multimodal corpora*, *lexica*, *grammar resources* and *terminology tools*[3]. Tool subtypes mainly refer to specific tasks.

6.1.15 LT World

The Language Technology World (LT World, http://www.lt-world.org/) is a portal intended to provide constantly updated information on LTs. The service is provided by the German Language Technology Competence Center at DFKI. Information is provided not only for LRs and LTs but also on players, projects, conferences etc.
The two sections "R & D – Systems" and "Products" list LRs and tools/technologies, with information taken mainly from the NLSR. It is not clear how entries from the original NLSR catalogue have been allocated to either of or both of these two new lists.

Both lists have inherited the NLSR *taxonomy* with the addition of a *misc* type and a few changes/additions/deletions at the *subtype* level (e.g. replacement of *written language corpora system* instead of *written language corpora*, addition of *evaluation system* under

[3] The value "terminology tools" comes as a surprise rather than "terminological resources". In fact, the resources for which this value has been selected either include a terminological/domain component (e.g. lexicon, thesaurus) or are indeed tools catering for some special domain.

evaluation tool etc.), and a transfer under another type (*theory developments and resources* has been moved from *NLP development aids* to *misc tools*).

6.2 Discussion of the survey findings

The notion of ***resource type*** is crucial to all metadata schemas and cataloguing practices: it determines a critical subset of elements related to the technical description of resources and, most importantly, it constitutes the basic feature according to which they are organized. However, there is as yet no consensus as to the values of LR type in the community, although general trends can be spotted (e.g. distinction of corpora and lexica). The following table summarizes values of resource types for the major metadata schemas and LRT catalogues.

Repository	Typology used
ELRA	- spoken resources - terminological resources - multimodal / multimedia resources - written resources written corpora monolingual lexicon multilingual lexicon
ELRA UC	same LRs typology + tools
ENABLER	- written resources - lexical resources - spoken resources - multimodal resources
CLARIN metadata	- media - texts - lexica - lists - annotations - tools
CLARIN catalogue	- multimodal corpus - spoken corpus - written corpus - aligned corpus - tree bank - grammar - n-gram model - terminological resource - lexicon/knowledge source
CLARIN report on Metadata	- Textual / Written Resources - Speech Resources - Multimedia/Multimodal Resources - Images - Annotations - Lexica

Repository	Typology used
	- Conceptual lists / Terminologies - Ontologies - Tools/Services - Typological databases - Grammars - Treebanks - Wordlists - Transcripts - Training data sets
LDC catalogue	corpus types "according to the type of data they contain" - lexicon - lexicon, speech, text - speech - speech + text - speech + transcripts - text - transcripts - video
DC	- collection - dataset - event - image - InteractiveResource - MovingImage PhysicalObject -Service - Software - Sound - StillImage - Text
OLAC	type from DC olac extension: linguistic type - lexicon - language description - primary text olac extension: discourse type - drama - formulaic discourse - oratory - ...
BAMDES	written corpus multimodal corpus oral corpus

Table 1: LR typology in the most popular metadata schemas and LRT catalogues

In accordance with the overall philosophy of the META-SHARE design, we have purposefully decided not to create a new LR taxonomy but rather to build upon previous initiatives, by harmonizing existing proposals and practices, and adapting them to the requirements of the

HLT community. Still, as can be seen from the table above, the harmonization of currently existing typologies is not an easy task, a problem that stems from a number of sources:

- the various typologies present different views on LRs categorisation, they do not share the same perspectives on the relation between types of resources (raw data related to annotation, subtitles considered as annotation of a multimedia resource or as an independent text file, transcribed spoken corpora can be viewed as a different modality of the original resource aligned to it or as annotation of the raw data etc.). There are two tendencies attested in the actual practice: on the one hand there are well-structured typologies according to which a resource should be classified, 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.
- most typologies (and elements thereof) are affected by diverging views on terminology which hinder interoperability between metadata schemas:
    - use of ambiguous terms: for instance, the term "written" refers to the text medium as well as to the communicative situation whereby a message is produced to be read rather, which makes the term "written corpus" ambiguous: if the term refers to the medium, then transcripts of audio files should be included under "written corpora", but if it refers to communicative situation, then they should be included in "spoken corpora";
    - use of semantically close terms (such as "spoken"/"oral"/"speech" or "written"/"text") without a clear indication of whether they are used with the same meaning;
- the approach adopted as regards the selection of values for LR types: most of the typologies are based on a pre-defined set of values while others take a bottom-up approach, where the values are entered by the users, and a few combine the two approaches together; as a consequence, the set of values rarely coincides between the various typologies.

Another issue with existing typologies is the fact that they do not reveal the "hidden" complexity of LRs. In fact, parts of LRs can be viewed as LRs on their own: for instance, monolingual parts of parallel corpora can be viewed as monolingual corpora, transcripts can be viewed as written corpora etc.

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 covered. There are discrepancies, however, as to two important points:

- the type of values used for each element, i.e. whether they will be free-text or constrained to a set of pre-defined values;
- obligatoriness of each element, i.e. whether it is considered obligatory, recommended or optional; the set of obligatory elements of the various schemas has been considered for the minimal schema.

Finally, we should note tendencies attested in the metadata records creation: it seems that free text is preferred by LR providers, while LR consumers prefer to search using pre-defined values. At the same time, LR providers are reluctant to providing detailed metadata records, while LR consumers ask for more informative descriptions.

7 The META-SHARE metadata model

7.1 Basic concepts

Based on the findings of the user requirements survey and of the overview of metadata schemas and catalogues, the principles of our proposal stem from the following observations on the needs of the HLT domain:

- the need for a taxonomy of LRs, which would define the various types of LRs (corpora, collections, annotations, speech corpora, multimodal corpora…) and the relations between them
- the need for a common shared terminology
- the need for minimal sets of metadata that would facilitate and not hamper LRs description and harvesting
- the need for a clear and non-complex structure of elements
- the need for clear semantics of the elements (definitions, relations)
- the need for the interoperability of metadata between repositories and between resources and tools/services.

As aforesaid, the META-SHARE metadata model builds upon previous initiatives. 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*** and ***relations*** as well as other components. More specifically, elements are used to encode specific descriptive features of the LRs, while relations are used to link together resources that are both 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.).

Central to the model is the ***LR taxonomy*** as outlined in section 6.3, 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), being modality dependent, differ across types. The user 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 users indeed 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. Moreover, exemplary instantiations (e.g. for wordnet-type resources, for parallel corpora, for treebanks etc.) as guiding assistance to LRs metadata providers will be available.

The model comprises ***all elements and relations required for the description of LRs***, i.e. any kind of information, such as identification parameters, administration information, technical information required for their manipulation, information as to the production and usage (intended and actual) processes etc. These constitute the maximal / fully fledged META-SHARE model. However, a ***minimal core subset of metadata*** is also foreseen, detailed in section 6.5.

Acceptance of the model by resource providers is a challenge, as they usually prefer to present information about their resources in an unstructured way, often limiting the description to the minimal set of elements; thus, the META-SHARE metadata description templates must be appealing to them in a number of ways, e.g. they should be easy- and fast-to-complete, stressing common points and differences with other LRs, promoting particular features of each LR etc. In order to meet requirements of both users and providers, the design of the model must allow for both coarse and fine-grained descriptions in a scalable

approach, where the minimal set of core elements is mandatory, supplemented by a varying degree of other elements, to be filled in at the discretion of the LR provider. The maximal set of elements must be able to give the full amount of information required by the LR consumers in order to select and make the utmost use of the most appropriate resource for their needs. In doing so, it conceptualizes the whole lifecycle of the LR production and deployment in a structured form, describing and associating resources (e.g. original and derived ones), processes and documentation (e.g. manuals) linked to the core resource.

In order to be useful, the META-SHARE catalogue should bring together the maximum number of resources available, together with their accompanying documentation, at least in the form of metadata descriptions. ***Harvesting*** from existing collaborating initiatives and projects will be an important source; to this end, the META-SHARE model needs to cater for the minimum loss of information in the process of converting existing resource descriptions to the ones adhering to the proposed model[4]. At the same time, the mechanism of ***uploading new resources*** directly to the META-SHARE mechanism will incorporate a module for their description to be filled in by the resource provider, in the form of guided templates, via a metadata editor.

7.2 The META-SHARE ontology

META-SHARE takes a more global view on ***resources***, aiming at providing 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; manuals documenting the use of a tool are important both for prospective users as well as for developers that wish to integrate this tool in an application.

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

We should note here that a similar approach but with different focus and objectives is also inherent in two of the initiatives we have presented in the overview, namely:

- the LRE map, where "language resource" is meant to encompass all types of relevant resources, and

[4] Semantic interoperability of meta-data is a major issue as the META-SHARE repository should allow harvesting both ways: harvesting from other resources and be harvestable on its own right.

- the LT World, which includes information on other entities besides LRs per se, such as information sources, actors, events etc.; these are kept at separate sections of the portal.

In the proposed META-SHARE ontology (Figure 1), a distinction is made between LR per se and all other related resources/entities[5], such as:

- reference documents (e.g. papers describing the resource, associated reports, tagset manuals, guidelines for LR production etc.)
- persons and organizations involved in their creation and use (e.g. creators of resources, funders, distributors etc.)
- related projects and activities (e.g. projects that have funded the creation of an LR, where an LR has been exploited etc.)
- licenses (for the distribution of the LRs).

In the META-SHARE ontology, some of the entities will have physical counterparts: 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) but they cannot have a similar physical counterpart. It is highly desirable that all these entities will be included in META-SHARE *only so far as they have some interlinking between them*: for instance, we will not have a list of people involved in LT (as in the LT World) but only of those people that are explicitly mentioned in the metadata records of the LRs included in the META-SHARE catalogue (e.g. authors, contact persons etc.).

The META-SHARE metadata model aims at covering only *LRs per se* (data and tools). For all other entities of the ontology, we will take into account metadata schemas and relevant formats that have been devised specifically for them, e.g. CERIF for research entities (projects, actors etc.), BibTex for bibliographical references etc.

[5] In this respect, it differs from the LRE map approach in two ways:
- the LRE map lists only resources excluding entities such as projects, persons etc.
- META-SHARE makes a distinction between LRs per se and other resources (documents, licenses etc.)

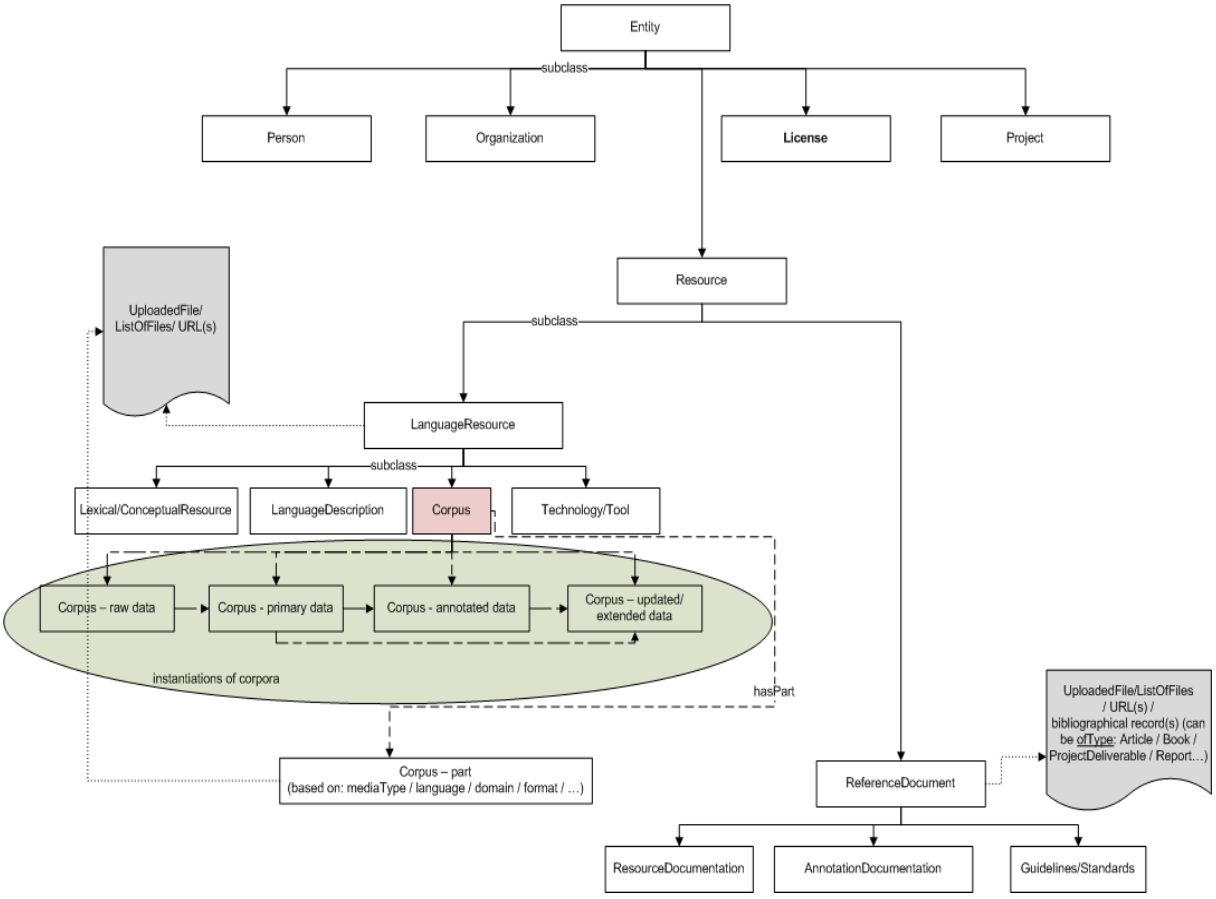

Figure 1: META-SHARE ontology excerpt

7.3 Proposed LR taxonomy

The study of the existing LR typologies 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.

The LR taxonomy proposed forms an integral part of the metadata model, whereby the types of LRs (attributes and values) belong to the element set (cf. section 6.4). The *resourceType* is the basic element according to which the LR types and subsequently the specific profiles (i.e. aggregations of components and elements) are defined.

A **two-level hierarchy**, with a coarse "main type" classification and further subclassifying features dependent on each type, is suggested. More specifically, the following four values are suggested for the **first level**:

- ***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*** (including basic processing tools, applications, web services etc. required for processing data resources).

It should be noted here that, given the practices 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***.[6] Parts of LRs clearly identifiable and serving the needs of the community 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 / image file, in the case of corpora, or word, in the case of lexica).

For **the second level of the taxonomy**, we suggest a ***type-dependent subclassification***, where the same resource can be viewed 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.
The notion of ***medium*** (element *mediaType*) constitutes the most important element employed in the classification of corpora; it is preferred over the written/spoken/multimodal distinction, as it has clearer semantics and allows us to view corpora as a set of modules, each of which can be described through a distinctive set of features. Thus, the following media type values are foreseen:

- ***text***: used for corpora 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

[6] This constitutes a crucial difference with the CLARIN inventory of LRTs and VLO approach, where language resource refers to the lowest identifiable unit (e.g. text file in the case of written corpora) while corpora are considered aggregate resources. This approach befits CLARIN purposes, which targets the wider Social Sciences and Humanities community and propagates the concept of "virtual collection": users are encouraged to select resources fitting their criteria in order to create their own collection; thus, for instance, corpora are seen only as sources from which users can draw segments for their own collection of texts. On the other hand, HLT users prefer to search for and use whole corpora rather than single texts.

- ***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.

For each of these values, a component is created including the appropriate set of descriptive elements, which are medium-dependent. A subset of these elements can be used to further classify the various LRs discussed below[7].

For text corpora, the most used feature for written corpora distinction is the *number of languages* (monolingual vs bi/multilingual corpora); the *language name* itself is a feature also used, but more so in reporting practices and menu-driven searches. The *annotation* feature is also used as a distinguishing feature, mainly as regards aligned corpora, but also for tagged corpora, treebanks etc. *Transcribed corpora* are identified in various taxonomies as a particular subtype of written (but also spoken) corpora (irrespective of whether the related audio component is available or not). The LDC includes *data source* as a distinguishing feature in its taxonomy; in fact, this feature combines two types of information: the *setting* for transcripts (e.g. microphone, broadcast etc.) which is deemed an important feature for audio files (see below) and a *text type/genre* classification (e.g. articles, speech, news, journal, government documents, email etc.). Consequently, the metadata elements recommended for the second hierarchical level of text resources are:

- languages dimension: number of languages, multilinguality type (with values: parallel, comparable), language name;
- annotation dimension: annotation type (with values: *morphosyntactic tagging, shallow parsing, treebanking, alignment etc.*); absence of an annotation component is a value on its own.

7.4 Contents of the model

As aforesaid, the full/maximal META-SHARE metadata model comprises all elements and relations required for the description of LRs put together in components. Elements will be linked to existing ISOCcat DCR data categories and, if they have no counterpart, these will be added with appropriate definitions. Specific profiles will be built for distinct LR types (and subtypes) using the various components, providing also exemplary instantiations (e.g. for wordnet-type resources, for parallel corpora, for treebanks etc.) as guiding assistance to LRs metadata providers.

[7] Other metadata elements can also be used by users to search the META-SHARE inventory (e.g. format, size etc.). Here we only discuss the ones we consider important for the classification of LRs.

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
- a second level with a higher degree of granularity (***maximal schema***), providing more detailed information on each resource (cf. section 6.4).

This has the advantage that LRs producers can enter information in an easy- and quick-to-fill-in form (implementing the minimal schema); this information can then be elaborated through appropriate forms (implementing the maxima schema) whenever they choose. Harvesting is also served better by distinguishing between the two levels, as well as LRs consumers: they 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.
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:

- string / free text
- closed list of values (enumeration-closed)
- open list of values (enumeration-open): recommended values are provided but users can add their own
- integer / numeric field
- 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.
In addition, ***XML attributes*** are introduced in order to clarify the meaning/usage of a generic element in specific contents. For instance, a person can act as the resource creator or

distributor, as a contact person or as the metadata creator; in this case, the "role" attribute is used to distinguish between his/her various capacities accordingly.

The generic XML attribute "lang" caters for multilinguality. It is used for all free text elements, with the default value set to English; metadata creators will have the chance to input text in other languages as well (e.g. resource titles, person and organization names etc. in their original language, if different from English) by repeating the relevant element and specifying the "lang" value.

Figure 2 gives a graphic representation (UML diagram) of the components included so far in the schema, while the full set of components and elements together with their mappings to elements of the ISOcat DCR (where possible) and of the most popular metadata schemas is presented in Appendix A.

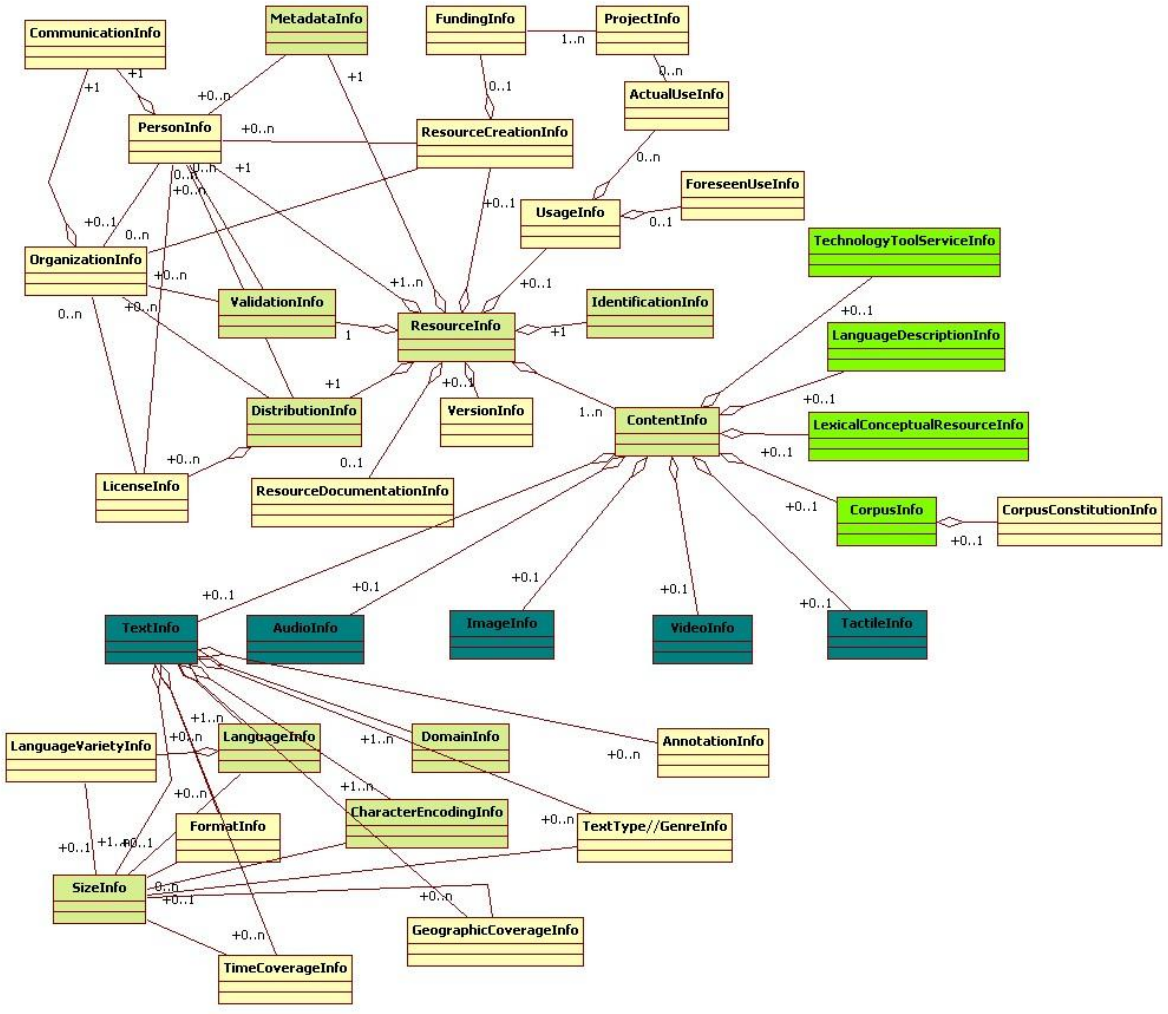

Figure 2: Graphic representation of the META-SHARE model (excerpt)

The current proposal focuses on ***text corpora***, while work on the other LR types has already commenced; the general principles of the schema as well as components common to all LRs have been laid down and agreed upon.

The core of the model is the ***ResourceInfo*** component**,** 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 and are, thus, located only in one place in the schema. Thus, elements needed for the description of *tactile* resources are only used for the specific *mediaType*.

The set of components that are common to all LRs are: *IdentificationInfo*, *PersonInfo*, *VersionInfo*, *DistributionInfo*, *ValidationInfo*, *CreationInfo*, *UsageInfo*, *MetadataInfo,* *ResourceDocumentationInfo* and *ContentInfo*. More specifically:

- the *IdentificationInfo* component includes all elements required to identify the resource, such as the resource title and acronym, the PID (to be assigned automatically by the system), internal 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
- 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), project name etc.)

- the *UsageInfo* component aims at providing information 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 *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 META-SHARE 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 input links to published over the internet documents and/or import bibliographic references in standard formats should 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 for the further description of the resource, distinct for each LR type, presented below.

A further set of three 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 giving 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 is meant to group together descriptive information as regards the contents of the resource. The elements included are:

- *description*: free text description of the resource
- resourceType with the values suggested in the LR taxonomy section (corpus, lexical/conceptual resource, language description, technology/tool/service)
- *mediaType*: values include *text*, *audio*, *image*, *video* and *tactile*. A resource may consist of parts attributed 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, and for some words a video and/or audio part may be included; a sign language resource is also a good example for a resource with various media types; a tool can be used both for video and for audio files. Thus, this element can take multiple values.

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

- *CorpusInfo*, *LexicalConceptualResourceInfo*, *LanguageDescriptionInfo* and *TechnologyToolServiceInfo* 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 *TactileInfo* which provide information depending on the media type of a resource; this can be broadly described as belonging to one of the following categories (all represented in the form of components and elements – cf. Appendix A):
    - *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)
    - *annotation*: information relative to the various annotation levels (tiers) of a resource applies only to corpora, and is media type-driven in the sense that we can distinguish between types of annotation performed on text parts/corpora (e.g. morpho-syntactic tagging, parsing, semantic annotation), audio parts/corpora (e.g. transcription, prosody annotation, speaker annotation), video parts/corpora (e.g. shot categorization, gesture annotation, facial expression annotation) etc.

7.5 The minimal schema

The proposed model in its totality is very large and detailed, as it aims to cover all types of LRs (written, spoken, multimodal, lexical/conceptual data, language descriptions and tools). Completeness and descriptive power are requisites for the model, but so are interoperability, user-friendliness and efficiency. To achieve this we have adopted the notion of "minimal schema".

The minimal schema contains those elements considered indispensable for LR description (from the provider's perspective) and identification (from the consumer's perspective), covering all stages of LR production. It takes into account the views expressed by the interviewees (see section 4) concerning which features are considered sufficient to give a sound "identity" to a resource.
The minimal schema contains only the first level components and elements (cf. section 6.1), i.e those which should be included in the description (whether created form scratch, harvested and/or converted from an existing schema to the META-SHARE schema) in order for a language resource to be included in the infrastructure. The minimal schema is considered as the "guarantee level" for interoperability as regards LR identification and metadata harvesting. The minimal schema with the mandatory elements will be the *sine qua non* condition for interoperability between the META-SHARE model and the other models; mappers / converters will cater for migration from one model to the other based on the set of mandatory elements.

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

- *IdentificationInfo*: groups together information needed to identify the resource and comprises the elements
    - *resourceTitle:* the complete title of the resource without any abbreviations
    - *pid*: a persistent identifier that refers to the resource or tool/service this metadata information describes
    - *identifier*: unique identifier; the attribute *type* is obligatorily used for further specification
- *ContentInfo*: groups together information on the contents of the resource, and comprises the elements *description, resourceType* (element which entails the use of type-specific elements and components) and *mediaType.*
        - *description:* free text description of the resource in prose

- *resourceType:* specifies the type of the resource (list of possible values: *corpus; lexicalConceptualResource; languageDescription; technologyToolService*)
- *mediaType:* specification of the media type of the resource; can be multiple if the resource is a multimodal set (values: *text; audio; video; image; tactile*)

- *DistributionInfo*: groups information on the distribution of the resource and comprises the elements *availability* and *distributionMedium* and the component *licenseInfo*
  - *availability*: declaration of the terms of availability of the resource in simple words
  - *licenseInfo:* description of the licensing conditions under which the resource can be used (recommended values are: *GNU; CC; own; ELRA_END_USER; ELRA_VAR; ELRA_EVALUATION*)
  - *distributionMedium:* specifies the format used for the delivery of the resource (recommended values are: *internetBrowsing; download; CD-ROM; DVD-R; bluRay; hardDisk; paperCopy; other*)
- *ValidationInfo*: Indication of the validation status of the resource, contains only one element (boolean)
  - *validated:* values *yes/no*
- *MetadataInfo*: groups information on the metadata record itself
  - *metadataCreationDate:* for creation of metadata from scratch, the date of creation of the specific metadata description
  - *source:* for harvested metadata, the catalogue from which the harvesting was made (CLARIN, OLAC, META,...)
  - *harvestingDate:* for harvested metadata, date of harvesting of the metadata
  - *originalMetadataLink:* for harvested metadata, link to the metadata of the original source.
- *FundingInfo*: information on all projects that have funded the resource; repeated for each project, includes the component *ProjectInfo* with elements
  - *projectTitle:* the full title of the project that led to the creation of the resource or tool/service
  - *fundingType:* type of funding (e.g. *EU, national funds, private organisation funds, own funds* etc.)
- *PersonInfo*: groups information on the contact person
  - *surname*
  - *givenName*

- *CommunicationInfo:* information on communication details (address etc.)

- *OrganizationInfo*: groups information the organization
  - *organizationName*: name of an organization
  - *CommunicationInfo*: information on communication details (address etc.)
- *CommunicationInfo*: groups information on communication details (address, email etc.) and can be attached to either *PersonInfo* or *OrganizationInfo*

In the case where the *resourceType* is specified as *mediaType=text*, the type dependent components and elements are the following:

- *LanguageInfo*: information on the language(s) of a resource; repeated for each language, contains the elements
  - *languageCoding:* designation of the standard used to code the name of the languages (ISO-639-3)
  - *languageId:* identifier of the language
  - *languageName*: a human understandable name of the language that is used in the resource or supported by the tool/service
- *SizeInfo:* as mentioned above, this component can be attached to every component that needs a specification of size; it includes two elements, namely
  - *size:* the size of the resource with regard to the SizeUnit measurement in form of a number.
  - *sizeUnit:* Specification of the unit of size that is used when specifying the size (exemplary values: *words; tokens; bytes; sentences; texts*).
- *FormatInfo*: the mime-type of the resource which is a formalized specifier for the format included. Takes values from the Internet Assigned Numbers Authority (IANA http://www.iana.org/assignments/media-types/).
- *CharacterEncodingInfo:* Groups together information on character encoding of the resource; repeated if parts of the resource have different character encodings. Includes
  - *characterEncoding:* name of the character encoding used in the resource or accepted by the tool/service. Recommended values: *ISO 8859-1;UTF-8;ISO 2022;* etc.
  - *SizeInfo*

- *DomainInfo:* Groups together information on domains of a resource; can be repeated for parts of the resource with distinct domain and includes

- *domain:* indicates the application domain of the resource or the tool/service.
- *SizeInfo*

- *AnnotationInfo*
  - *annontationType:* specification of the types of annotation levels provided by the resource. Values: *segmentation; alignment; structural annotation; lemmatization; stemming; PosTagging; bPosTagging…*

8 Conclusions and future work

This report documents the design of the META-SHARE metadata model. It analysed its purpose in the framework of the META-SHARE infrastructure and the goals set as regards the features that should characterise the model; it discussed the stages followed for the definition of the model, namely: (1) the user requirements survey that resulted in the recording of the user needs in what concerns description and identification of LRs to be catered for by the infrastructure, and (2) the overview of similar initiatives, which revealed the approaches adopted by others, but also served as the initial step for the mapping of the elements of the META-SHARE model to those of the most widely used schemas); it proceeded to present and explain the principles and basic concepts of the model, the ontology and the proposed taxonomy and concluded with the detailed presentation of the model itself, focusing on the notions of maximal and minimal schema.

As specified in the DoW, the current version contains, besides the general presentation of the model, the application of the model to the *text* mediaType. The next steps in WP7.2 include:

- extension to other media and LR types: the model will be applied to the rest media types (*audio, video, image, tactile*) and LR types (*lexicalConceptualResource, languageDescription; technologyToolService)*. In this process, the expressive power and the coverage of the model will be tested and it is expected that new components and elements will arise. Work on some of these types has already started (for instance for the lexical resources and the multimedia/multimodal types), and the first application of the model to them is very promising. The full model containing the components and elements for all the media and resource types will be documented in the final report.
- exemplary instantiations: we plan to implement the metadata model to describe a set of resources selected to represent all LR and media types, in order to test its

functionality. These resources with their descriptions will be uploaded in the prototype infrastructure for testing 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.
- development of the model as a schema: the model will be implemented most probably as an XML schema.

Appendix A: The META-SHARE model & mappings

The file META-SHARE metadata model v1_0.xlsx contains the fully fledged version of the model. More specifically, it contains 3 worksheets:

- **components only**: it includes only the components (i.e. not the elements) of the proposed schema. For each component in the 1st column, the components included in it are given in the 2nd column. For an explanation of all columns, see below.
- **components & elements gradual**: it includes all components and elements (i.e. not relations) of the proposed schema. For each component in the leftmost column, the next column includes all components and elements included in it; if it includes components, the next column includes its own components and/or elements and so on. For an explanation of all columns, see below.
- **relations for written corpora**: it contains a preliminary set of the relations identified until now for text resources (they most probably serve other resource types as well).

Explanations of columns

- **Component**: Name of the component; all components are marked in orange fonts, start with a capital letter and their name ends with "Info" (e.g. *DistributionInfo*, *ResourceDocumentationInfo*). A component groups together a specific type of information (in the form of elements and/or components), e.g. information on distribution, documentation, format of a resource etc. The elements and/or components included in a component are given in the next column.
- **Component/Element**: Component or element included in the component of the previous column.
- **Element**: Element included in the component of the previous column. Element names start with a small letter and, if consisting of more than one words, a capital letter is used for the first letter of each following word (e.g. *license*, *givenName*).

- **Optionality**: Optionality of the component/element; the following symbols are used:
    - M: *Mandatory*; must always appear; in the UML diagram noted as 1 / 1..n
    - MC: *mandatory under conditions*; must appear if certain conditions are met; in the UML diagram noted as 1 / 1..n
    - R: *recommended*; information that metadata creators are encouraged (not obliged) to fill in because it is considered useful for the LR description by prospective LR users
    - O: *optional*; metadata creators are free to fill in for a full description of hte resource.
- **Conditions for MC**: if a component/element is marked as MC, we give here the conditions for the mandatory state
- **Repeatable** (Y/N): whether a certain element/component can be repeated (Y) or no (N); in the UML diagram noted as "..n".

N.B.: Repeatability here does not take into account the repeatability of the "lang" attribute: all fields of type "string" (free text) can be repeated if the "lang" attribute is used with a different value (e.g. title of a resource in English, Greek, Chinese etc.); so *ResourceTitle* in this table is marked as *repeatable=N*.

- **Field type**: The following symbols are used:
    - *cmp* (component)
    - *date* (for normalised format, check http://www.w3.org/TR/NOTE-datetime)
    - email
    - *enumeration-closed* (users select from a closed list of values)
    - *enumeration-open* (recommended values are given but users can add their own values)
    - integer
    - *string* (free text)
    - *tel* (telephone)
    - url

- **XML attributes**: attributes and respective values to be added in the XML version of the schema; note that we have not inserted the "lang" attribute anywhere as it is a general attribute to be used for all elements of type "string" (free text).
- **Recommended values**: if an element is of type "enumeration", a list of the values is given; if "enumeration-closed" the list must include all the values a user can choose from; if "enumeration-open", the list includes values that are recommended but users can add their own; note also that if the element is not repeatable, then users must choose only one value, otherwise multiple values are allowed.
- **Definition / Description**: a short definition/explanation of the component/element. When the element is mappable to the ISOcat DCR, the definition is taken from there; italics are used for deviations from the ISOcat.
- **Examples**: examples for the element.
- **ISO DCR – identifier**: the corresponding data category from the Metadata thematic group.
- **ISO DCR – PID**: the PID of the data category.
- All other columns refer to popular **metadata schemas** and **catalogue descriptions**.