CESAR

Central and South-East European Resources

Project no. 271022

Deliverable D4.1

Metadata descriptions and other interoperability standards as agreed with META-NET and partner projects

Version No. 1.3

29/06/2011

Document Information

Deliverable number:	D4.1
Deliverable title:	Metadata descriptions and other interoperability standards as agreed with META-NET and partner projects
Due date of deliverable:	30/04/2011
Actual submission date of deliverable:	19/07/2011
Main Author(s):	György Szaszák, Klára Vicsi
Participants:	
Internal reviewer:	Tamás Váradi - RILHAS
Workpackage:	WP4
Workpackage title:	Cross-national collaboration and pilot service
Workpackage leader:	BME
Dissemination Level:	PP
Version:	1.3
Keywords:	metadata

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
1.3	29/06/ 2011	Final	György Szaszák	BME	Finalization based on CESAR project meeting od 26/06/2011.
1.2	26/06/ 2011	Pre-final draft	György Szaszák	BME	Section 4 on metadata editor tool recommendations added
1.1	06/06/ 2011	draft	György Szaszák	BME	Core version

EXECUTIVE SUMMARY

This report offers an overview of the metadata description model to be adapted within the CESAR project for language resources and tools. The model template provided by META-NET is the base of the work, which needs some extensions to cover all type of language resources and tools, not only the textual ones.
The present document is based on recommendations by META-NET D7.2. and CESAR partners' contribution.

Table of Contents

Abbreviations

Abbreviation	Term/definition
LR	Language Resource
LRT	Language Resources and Tools (either language data or tools)
Component	The metadata model is composed from components The name of the component 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.
Element	Elements (but also components) can be included in the components. 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*).
DCR or ISO-DCR or ISOcat-DCR	**Data Category Registry** ISO 12620 provides a framework for defining data categories compliant with the ISO/IEC 11179 family of standards. According to this model, each data category is assigned a unique administrative identifier, together with information on the status or decision-making process associated with the data category. In addition, data category specifications in the DCR contain linguistic descriptions, such as data category definitions, statements of associated value domains, and examples. Data category specifications can be associated with a variety of data element names and with language-specific versions of definitions, names, value domains and other attributes.
PID	Presistent Identifier in the *ISOcat* Data Category Registry

Table 1. Abbreviations

1. Scope

The Description of Work (DoW) of CESAR defines task 4.1, "Metadata descriptions and other interoperability standards as agreed with META-NET and partner projects" as follows:

"The consortium will agree on standardized top-level resource descriptions (metadata) for all relevant types of resources, based on a recommended set of metadata for documenting resources provided by META-NET (see WP3). It will produce such descriptions for each and every resource contributed to the shared pool. Metadata sets will include mandatory as well as optional elements, together with sets of recommended values whenever possible and appropriate. Metadata will include information for the resource per se, its identification (including a persistent identifier), together with its creation, annotation, provenance, documentation, usage, availability, licensing, distribution data."

The aim of the present document is to specify the metadata description scheme taking into consideration the needs of all CESAR partners. The recommended set of metadata provided by META-NET forms the base of the work. However, the document D7.2 of META-NET focuses on written corpora, although within the CESAR project, a number of spoken and multimodal resources are also included. META-NET is planning to extend its metadata description model to cover all other LRT types (spoken, multimodal, lexical resources and tools and technologies), but this work is still in progress. Therefore the CESAR community should pay a special attention to specify its metadata description model for non-textual LRTs.

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

The META-NET D7.2 document specifies a metadata model template, that is intended to be used for the description of Language Resources and Tools (LRTs) made available through META-SHARE. The proposed model is referred to as "META-SHARE metadata model". The META-NET D7.2 puts the model in the context of its application (LRT sharing), delineating the intended goals of use and factors to be taken into account for its design. The proposed model was elaborated based on two main points: (i) user requirements, as collected through surveys and (ii) an overview of the most widespread metadata models and catalogue descriptions of LRTs (ELRA, LDC, and many others). Particular emphasis is put on the presentation of a minimal schema, which is a subset of the META-SHARE model, consisting of elements considered indispensable and hence compulsory for the description of LRTs.

Metadata descriptions of LRTs offered for META-SHARE will be open distributed. Metadata descriptions should follow a pre-defined, well discribed, machine readable format. As already mentioned, the document D7.2 of META-NET focuses on textual corpora, although within the CESAR project a number of spoken and multimodal corpora are planned to be shared via META-SHARE. The metadata models for lexical and conceptual resources and tools are also being elaborated by META-NET at the time this deliverable is prepared. The description of these resources and tools will be elaborated based on partners' know-how and experiences

and submitted to META-NET, but may need some readjustments after final META-NET guidelines will be known.

In the next section, the META-SHARE metadata description model is presented. Hereafter, the minimal schema is described and extended to other types of resources.

2. The META-SHARE metadata model

2.1 Hierarchy and taxonomy

The META-SHARE model uses a two level taxonomy. The first level (layer) identifies the type of the LRT and suggests using 4 categories (called main types), which are:

- ***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, …)
- ***technology / tool*** (including basic processing tools, applications, web services and so on, required for processing data resources).

The second level of the taxonomy suggests further categorization, introducing subtypes, depending on the main type. For example, the subclassification of main type *corpus* is the best elaborated: the core notion is *medium*, divided into the following classes:

- ***text***: used for corpora with only written medium (or text modules of spoken and multimodal corpora)
- ***audio***: the audio feature set will be used for a whole resource or part of a resource that is recorded as an audio file; its transcripts will be described by the relevant *text* feature set.
- ***image***: the image feature set is used for photographs, drawings etc., while the *text* set will be reserved for its captions
- ***video***: moving image, 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.

As it can be seen from the above example, a text corpus contains only text related description, whilst an audio corpus has an audio feature set to describe audio specific metadata, but shares the same text metadata set, which is used by a pure textual corpus.

For each of these medium types, 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 corpora.

The model scheme is defined in the next section, providing a detailed description of its components. As of 20th June 2011, first level metadata schemes and the text corpora subclass metadata schemes are finalized (which, per se does not exclude updating in the future, but the schemes are ready for use provided by META-NET). Audio and video corpora schemes are foreseen by the end of July 2011. Lexica/conceptual resource main type schemes are also foreseen to come up in June 2011. Tools and technology schemes can be expected for July 2001. Language description main type comes later in 2011.

In practice, META-NET will distribute the finalized schemes across all partner projects, the schemes are likely to be issued as medium dependent metadata models which contain main type specific fields (i.e a text medium metadata model incorporates corpora and lexica main types, but some fields concern only resources of type lexica, for example).

2.2 Model schema

The model scheme is being populated by META-NET. The structure and features behind this template are presented in this section.

2.2.1 Basic structure

The model is composed of *components* and *elements*, linked together and structured less or more into a hierarchy. The model itself is represented by a so-called core component, which has attached components, which can also have attached components or elements and so on. A component or an element can be attached also to more than one other components. Some characteristics of the components and elements are listed below:

- **Optionality**: A component can be globally mandatory, mandatory if certain conditions are fulfilled, recommended or optional:
  - 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. If a component/element is marked as MC, the conditions for the mandatory state should be well defined
  - 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 the resource.
- **Repeatability** (Y/N): whether a certain element/component can be repeated (Y) or no (N); in the UML diagram noted as "..n". Repeatability here does not take into account the repeatability of the language attribute: all fields of type "string" (free text) can be repeated if the language attribute is used with a different value (e.g. title of a resource in English, Greek, Chinese etc.);
- **Field type**: Each component/element is assigned a set of fields containing the descriptive information. 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

Each component or element has its own:

- **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).

- **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.
- **ISO DCR – identifier**: the corresponding data category from the Metadata thematic group.
- **ISO DCR – PID**: the PID of the data category.

2.2.2 Components and elements

The core component of the model is the ***ResourceInfo*****,** which contains all the information necessary for the description of a LRT. It contains further components and elements that provide together the 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.

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

These components include the followings (some components can be "standalone" as attached to the core *ResourceInfo* but can also be attached to other components):

- *IdentificationInfo:* all elements which identify the LRT, such as the resource name and acronym, the PID (to be assigned automatically by the system), internal identifiers, etc.

- *PersonInfo* information about the person who can give further information or access to the resource; this is a special component as it can be attached to many other components in this list (for example, *PersonInfo* can be used for any person acting as resource creator, distributor, etc.

- *VersionInfo:* information relative to versioning and revisions of the LRT

- *DistributionInfo:* legal issues and availability of the LRT

    - *LicenseInfo:* attached to *DistributionInfo,* gives the description of the licensing conditions under which the LRT can be used

- *ValidationInfo:* provides at least an indication of the validation status of the LRT (using boolean values: validated or not) and, if the resource has been validated, further details on the validation mode, results etc.

- *ResourceCreationInfo:* information regarding the creation of a resource (creation dates, funding information such as funder(s), project name etc.) The *ResourceCreationInfo* component has dependent components.

- *UsageInfo:* foreseen use of the LRT (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.)

- *MetadataInfo:* 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.

- *ResourceDocumentationInfo:* publications and documents which describe the resource in details, including basic documents (such as manuals, tagset documents, etc.), which preferably 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

- *ContentInfo:* 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 LRT type, presented below.

- A further set of three other components have a special status as they can be attached to various components seen above such as *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.).

From the above list, the ***ContentInfo*** component needs some further explication, as this is the component through which level two descriptors connect to level one in the taxonomy: the *ContentInfo* component is meant to group together descriptive information as regards the contents of the LRT. This descriptive information includes:

- a free text description of the resource (*description*)

- *resourceType* element with values representing the level 1 in the taxonomy: *corpus, lexical/conceptual resource, language description, tool/ technology*

- *mediaType* element, representing level 2 of the taxonomy, with values: *text*, *audio*, *image*, *video* or *tactile.* A LRT may contain parts attributed to different types of media (e.g. 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) or a tool which can be used both for video and for audio files, for example. For this reason, the *mediaType* element can be assigned multiple values

Each of the values of the *resourceType* and *mediaType* gives rise to a series of other components, namely:

- *CorpusInfo*, *LexicalConceptualResourceInfo*, *LanguageDescriptionInfo* and *TechnologyToolServiceInfo* which include information specific to each LRT

type (e.g. subtypes of corpora and lexical/conceptual resources, tasks performed for tools etc.)

- *TextInfo*, *AudioInfo*, *VideoInfo*, *ImageInfo* or *TactileInfo* which provide level two information specific to the media type of a LRT. As a common reference, these components will be called *MediaType* components altogether.

  The components attached to the *MediaType* components are less or more dependent on the media type, but should inform about the following features:

  ⇒*content*: including primarily the languages covered in the resource and additional classificatory information such as domains, geographic coverage, time coverage, setting, type of content, etc.

  ⇒*format*: file format, size, duration, character/audio/video encoding, etc.;

  ⇒*creation*: this is to be distinguished from the *ResourceCreationInfo* component which is attached to the resource (first) 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 (scanning and web crawling for texts, vs. recording methods for audio files and so on).

  ⇒*annotation*: information relative to the various annotation levels (tiers) of a LR. This component applies only to corpora, and is media type-driven in the sense that one 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.

2.3. The minimal schema

The META-SHARE metadata description proposal defines a minimal model, which includes the set of the components which are either *mandatory (M)* or so-called *condition-dependent mandatory (MC)* description of a LRT. The condition-dependent mandatory elements are obligatory only if certain conditions are met.

Other components are *recommended (R)* ones. Mandatory and condition-dependent mandatory components are selected so that either they provide basic information or are necessary for the broad categorization of the LRT, providing the minimal schema.

The CESAR community follows the recommendations of META-NET concerning the the selection of the mandatory components. As at the time of the preparation of this document there is no available META-NET proposal for non-textual corpora or main types in the first level of the taxonomy other than corpus, project partners shoul specify their needs following the guidelines provided by META-NET, and paying special attention to the compatibility of the extensions proposed. All suggested extensions must be agreed upon by partner projects and META-NET.

The minimal schema contains only first level components and elements, constituting an obligatory set of components and elements to be specified in order for a LRT to be included in the infrastructure. „The minimal schema is considered as the "guarantee level" for interoperability as regards LRT 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 mandatory elements of the metadata description model are (extract from META-NET D7.2, extended by proposed adaptation for CESAR for audio and partly for video media):

- *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 mandatory 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/).
- Medium dependent encoding info: media dependent information on encoding issues:
  - *CharacterEncodingInfo (MC):* **For text medium**. 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*
  - *AudioEncodingInfo (MC):* **For audio medium**. Groups together information on character encoding of the resource; repeated if parts of the resource have different character encodings. Includes:
    - *audioEncoding:* name of the audio encoding used in the resource or accepted by the tool/service. Recommended values: *RIFF-WAVE;RAW;FLAC;* etc.
      - ⇒ *SizeInfo* : for all audio files mandatory elements, each in a *SizeInfo* component are: sampling rate, sample bits, number of channels, etc.
      - ⇒ *quantisation* : quatisation characteristics, recommended values: A-LAW, MU-LAW, LIN, etc. Can also be a *SizeInfo* component with recommended values.
    - *SizeInfo*
  - *VideoEncodingInfo (MC):* **For video medium**. Groups together information on video encoding of the resource; repeated if parts of the resource have different encodings. Includes:
    - *videoEncoding:* name of the video encoding used in the resource or accepted by the tool/service. Recommended values: *MPEG-1; MPEG-2,* etc.
      - ⇒ *SizeInfo* : for all video files mandatory elements, each in a *SizeInfo* component are: bitrate, resolution, 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 for text: *segmentation; alignment; structural annotation; lemmatization; stemming; PosTagging; bPosTagging…*

Values for audio: transcription, phoneme-level segmentation, word-level segmentation, prosodic annotation, etc.

⇒ *annotationMode*: recommended values: automatic, hand-made or semi-automatic

3 Proposed extensions

The currently available META-SHARE templates do not treat lexica and tools main types and audio sub-types, which, however represent a considerable part of the LRTs which are planned to be involved in META-SHARE by CESAR partners. In this sections, some general considerations will be given which raised from CESAR side.

3.1 Tools

For tools and technology main types, the CESAR community proposes the same set of mandatory or condition dependent mandatory components/elements, underlying that *FormatInfo*, *EncodingInfo* or *AnnotationInfo* should be interpreted for the data that the tool accepts, needs or generates. For a more refined description, input and output eventually should have their own descriptions using these three components. An extension seems indispensable concerning the eventual dependencies, system and platform requirements the tool requires and the current version, the known bugs, whether the tool is being developed and by who etc. Therefore, the following components should be considered as mandatory:

- *RequieremenInfo:* Groups together information on system and platform requirements
- *VersionInfo:* Version information, dates.

A recommended component for tools could be the *DevelopmentInfo* component, in order to facilitate the coordination between eventual user groups who envisage further development of the tool :

- *DevelopmentInfo:* Can be also given in a *ForeSeenUseInfo* component
    - *description*: free text description of goals
    - *date*: when is planned to be ready
    - *CommunicationInfo*: groups information on communication details (address, email etc.) and can be attached to either *PersonInfo* or *OrganizationInfo*

3.2 Recommended and optional descriptive elements for audio LRs

Concerning the spoken corpora, a lot of useful information might be of interest for the future users. Some relevant components were proposed to be added to the mandatory set of descriptive components, however, there are others, that based on CESAR partners' experience can be crucial in the field of speech technology. The mandatory components inform about audio encoding format, including the audio format itself, sampling rate, sample bits, quantization characteristics, number of audio channels. Also, the size of a spoken resource can be given in a number of units: (number of speakers, number of utterances, etc.) These features fit well into the META-SHARE model using the *SizeInfo* component. The

annotation information is also mandatory and can represent various tiers of data processing (segmentation on different levels, prosody annotation and so on).

Concerning a spoken resource, the following information can also be very important:

- Recording environment (studio, office/home, booth, street, etc.), recording protocol
- The spoken units (paragraph, sentence, dates/addresses, words, numbers, CVC, diad, triad, etc). Even a simpler speech database usually contains several spoken units.
- Speech style and genre (read/spontaneous, topic, monologue/dialogue/multi-party, socio-linguistic issues, etc.)
- Properties of devices used for the creation of the corpus (microphone characteristics and type, transducer and eventual transmission characteristics, AGC, etc.)

4. Metadata edition

In order to ensure a standardized, error free and efficient metadata description, it is highly recommended that project members can use a metadata editor tool which facilitates the work. Such a tool should be characterized as follows:

- Incorporate metadata schemes for all types of LRTs (main type and medium dependent), preferably in a manner that even users with lesser knowledge on metadata model architecture be able to use it (provide an efficient interface between the underlying metadata model and the information needed)
- Support the minimal scheme (obligatory fields, error check, etc.)
- Support for updating metadata description
- Provide xml based machine readable metadata ready for harvesting
- Contain a well described help menu
- Export data both in machine and human readable formats
- Eventually support import from existing metadata description standards
- Provide a transparent view of the resource
- Provide some sort of checking and validation of the metadata

Although a full metadata description concerning at least the minimal scheme is of basic importance, in case of some resources some information might be temporary missing or unavailable in the first or second upload cycle. Therefore the tool could have a more permissive mode e.g. allowing for UNKNOWN values in case of closed enumerations. For such values, per se, a warning message should be issued during the checking process.

As of June 2011, META-NET is planning to launch the first version of the metadata editor tool by the end of July, 2011.

5. UML model schemes

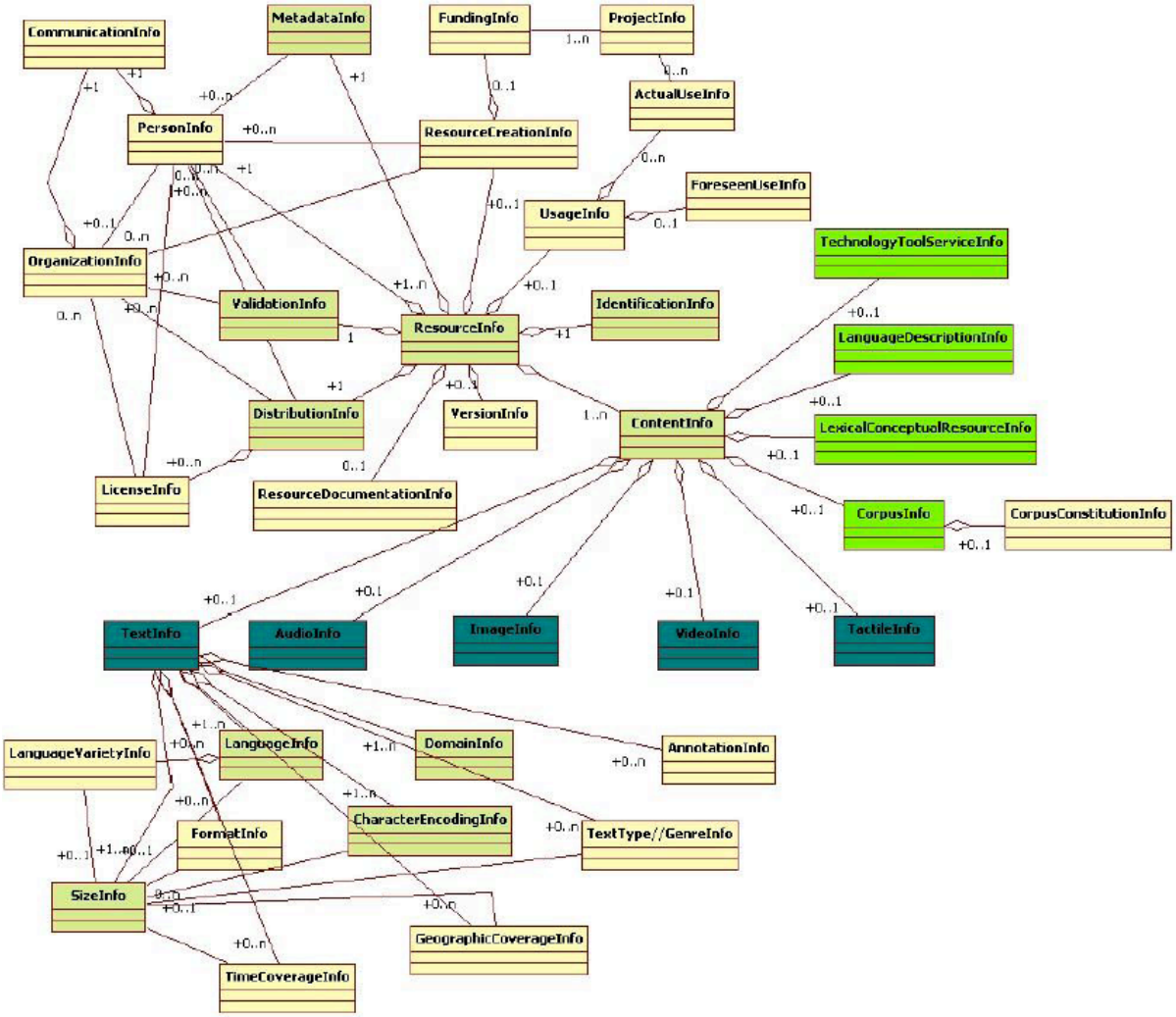

Fig. 1. Metadata description model UML scheme (copied from META-NET D7.2)