The META-SHARE Schema for the Description of Language Resources

Μαρία Γαβριηλίδου, Πένυ Λαμπροπούλου, Στέλιος Πιπερίδης

ABSTRACT

This paper presents the metadata schema for the description of language resources and technologies (LRTs) currently under development for the needs of META-SHARE, an open distributed facility for the exchange and sharing of LRTs The description of LRTs is granular and abstractive, combining the taxonomy of LRTs with an inventory of a structured set of descriptive elements, of which only a minimal subset is obligatory. The schema also proposes recommended and optional elements, while it includes a set of relations for the linking of resources with other resources and with related material. The current paper presents the main principles of the metadata schema, focusing on the description of text corpora.

Το σχήμα τεκμηρίωσης γλωσσικών πόρων META-SHARE

Μαρία Γαβριηλίδου, Πένυ Λαμπροπούλου, Στέλιος Πιπερίδης

ΠΕΡΙΛΗΨΗ

Στην ανακοίνωση αυτή παρουσιάζεται ένα σχήμα μεταδεδομένων για την περιγραφή Γλωσσικών Πόρων και Τεχνολογιών (ΓΠΤ), το οποίο υποστηρίζει το META-SHARE, μία ανοιχτή κατανεμημένη υποδομή για την ανταλλαγή και διανομή ΓΠΤ. Η ανάπτυξη του σχήματος μεταδεδομένων, της υποδομής META-SHARE καθώς και του META-NET (http://www.meta-net.eu/meta/about-el) ενός δικτύου αριστείας 47 οργανισμών από 31 κράτη αποτελούν μέρη του ερευνητικού έργου T4ME (http://t4me.dfki.de/).

Το σχήμα μεταδεδομένων χρησιμοποιείται για την περιγραφή γλωσσικών πόρων και τεχνολογιών. Ένα υποσύνολο των στοιχείων περιγραφής αποτελούν κριτήρια αναζήτησης και εντοπισμού των πόρων, γεγονός που καθιστά το σχήμα βασικό συστατικό του μηχανισμού αναζήτησης που περιλαμβάνει η υποδομή. Το σχήμα απευθύνεται σε ειδικούς της Γλωσσικής Τεχνολογίας, δηλαδή σε ειδικό κοινό - ωστόσο, στοχεύει να είναι σαφές και εύχρηστο στην ευρύτερη επιστημονική κοινότητα.

Η διαδικασία ανάπτυξης του σχήματος βασίστηκε σε μελέτη ανάλογων προηγούμενων δράσεων καθώς και σε μελέτη των αναγκών των χρηστών, όπως αυτές προέκυψαν από σχετικές συνεντεύξεις που διοργανώθηκαν στο πλαίσιο του έργου. Το σχήμα στοχεύει να είναι χρήσιμο τόσο σε δημιουργούς και παρόχους γλωσσικών πόρων και τεχνολογιών που επιθυμούν να περιγράψουν τους πόρους τους ώστε να είναι εντοπίσιμοι, όσο και στους χρήστες, που επιθυμούν να εντοπίσουν πόρους κατάλληλους για την έρευνά τους. Βάσει αυτού, οι αρχές σχεδίασης του σχήματος είναι

- σαφής και σημιοσιολογικά διαφανής ορολογία, υποστηριζόμενη από ορισμούς και παραδείγματα,

- εκφραστική ισχύς με χρήση στοιχείων κατάλληλων για την περιγραφή κάθε είδους γλωσσικών πόρων και τεχνολογιών αλλά και κάθε σταδίου ανάπτυξής τους,

- υιοθέτηση των σχετικών προτύπων, όπου υπάρχουν,

- ευελιξία, με την παροχή δύο επιπέδων περιγραφής (ελάχιστης και πλήρους),

- διαλειτουργικότητα με άλλα αποθετήρια πόρων και σχετικά εργαλεία,

- δυνατότητα συγκομιδής των μεταδεδομένων από άλλα αποθετήρια,

- επεκτασιμότητα, ώστε να είναι σε θέση να καλύπτει μελλοντικά νέους τύπους πόρων.

Το σχήμα αποτελείται από στοιχεία που οργανώνονται σε συστατικά και συνδέονται μεταξύ τους με σχέσεις. Τα στοιχεία κωδικοποιούν χαρακτηριστικά περιγραφής των γλωσσικών πόρων ενώ οι σχέσεις συνδέουν πόρους του αποθετηρίου μεταξύ τους (π.χ. πρωτογενή με επισημειωμένα δεδομένα, πόρους με τα εργαλεία επεξεργασίας τους) αλλά και με περιφερειακούς πόρους (όπως κείμενα τεκμηρίωσης, εγχειρίδια χρήσης του εργαλείου κ.λπ.) Το σύνολο των στοιχείων και των συστατικών που περιγράφουν έναν τύπο γλωσσικού πόρου αποτελεί το προφίλ αυτού του τύπου. Ορισμένα συστατικά περιλαμβάνουν πληροφορία που αφορά κάθε τύπο πόρου – για παράδειγμα, τα συστατικά που παρέχουν πληροφορία για την ταυτότητα, τη διαθεσιμότητα ενός πόρου κ.λπ. Συστατικά που περιλαμβάνουν πληροφορία π.χ. για το περιεχόμενο, την επισημείωση, τις χρήσεις ενός πόρου εξαρτώνται από τον τύπο του πόρου.

Το σχήμα προβλέπει δύο επίπεδα

- το αρχικό επίπεδο, που περιλαμβάνει τα απαραίτητα βασικά χαρακτηριστικά για την περιγραφή ενός πόρου (από την πλευρά του παρόχου) και τον εντοπισμό του (από την πλευρά του χρήστη (**βασικό σχήμα**)
- το δεύτερο επίπεδο, που παρέχει μεγαλύτερο βαθμό λεπτομέρειας και καλύπτει όλα τα στάδια παραγωγής και χρήσης ενός πόρου (**πλήρες σχήμα**).

Τα δύο αυτά επίπεδα περιλαμβάνουν τέσσερεις τάξεις στοιχείων: το πρώτο περιλαμβάνει Υποχρεωτικά και Υποχρεωτικά-υπό-συνθήκη στοιχεία, ενώ το δεύτερο περιλαμβάνει Προτεινόμενα και Προαιρετικά στοιχεία.

Στην ανακοίνωση αυτή παρουσιάζονται οι γενικές αρχές του σχήματος, τα στοιχεία που αφορούν όλους τους τύπους και αναλυτικά το υποσύνολο που αφορά τους πόρους γραπτής γλώσσας.

0 Introduction[1]

The diverse and heterogeneous landscape of huge amounts of digital resources' collections (publications, datasets, multimedia files, processing tools, services and applications) has drastically transformed the requirements for their publication, archiving, discovery and long-term maintenance. Digital repositories provide the infrastructure for describing and documenting, storing, preserving, and making this information publicly available in an open, user-friendly and trusted way. Repositories represent an evolution of the digital libraries paradigm towards open access, advanced search capabilities and large-scale distributed architectures. META-SHARE (www.meta-share.eu) is a network of repositories of ***language data, tools and related web services*** documented with high-quality ***metadata***, aggregated in central inventories allowing for uniform search and access to resources.

In the context of META-SHARE, the term ***metadata*** refers to descriptions of Language Resources (LRs), encompassing both data sets and tools / technologies / services used for their processing, also referred to as Language Resources and Technologies (LRTs).

1 Design principles for the metadata model

The metadata descriptions constitute the means by which LR users identify the resources they seek. Thus, the META-SHARE metadata model [Gavrilidou et al., 2010] forms an integral part of the search and retrieval mechanism, with a subset of its elements serving as the access points to the LRs catalogue. Although META-SHARE aims at an informed community (HLT specialists), this is by no means interpreted as a permission to create a complex schema; user-friendliness of the search interface should be supported by a well grounded schema. In this effort, we have built upon three main building blocks:

(a) study of widespread metadata models in HLT and LR catalogue descriptions[2].
(b) user requirements, collected through a survey conducted in the framework of the project [Federmann et al., 2011].

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

[2] The schemas taken into account include: Corpus Encoding Initiative (CES XCES), Text Encoding Initiative (TEI), Open Language Archives Community (OLAC), ISLE MetaData Initiative (IMDI), European National Activities for Basic Language Resources (ENABLER), Basic Metadata Description (BAMDES), Dublin Core Metadata Initiative (DCMI), ELRA Catalogue, ELRA Universal Catalogue, LRE map, LDC catalogue, CLARIN metadata activities and the ISO 12620 – DCR.

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

The basic design principles of the METASHARE model are:

- semantic clarity: clear articulation of a term's meaning and its relations to other terms
- expressiveness: successful description of any type of resource
- flexibility: possibility for exhaustive but also for minimal descriptions
- interoperability: mappings to widely used schemas
- extensibility: allow for future extensions, as regards the coverage of more resource types as they become available.
- harvestability: allow harvesting of the metadata (OAI-compatible).

2 The metadata model essentials

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

The set of all the components and elements describing specific LR types and subtypes represent the ***profile*** of this type. Obviously, certain components include information common to all types of resources (e.g. identification, contact, licensing information etc.) and are, thus, used for all LRs, while others (e.g. components including information on the contents, annotation etc. of a resource) differ across types. Profiles for each type will be proposed as templates or guidelines for the description of the resource. 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 detailed information on a resource and covering all stages of LR production and use.

The minimal schema contains those elements considered indispensable for LR description (from the provider's perspective) and identification (from the consumer's perspective). It takes into account the views expressed in the user survey concerning which features are

considered sufficient to give a sound "identity" to a resource.

These two levels contain four classes of elements: the first level contains Mandatory (M) and Condition-dependent Mandatory (MC) elements (i.e. to be filled in when specific conditions are met), while the second level includes Recommended (R) and Optional (O) elements.

3 The META-SHARE ontology

META-SHARE takes a 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 guidelines are informative for LR users and advisory for prospective LR producers.

In the proposed META-SHARE ontology, a distinction is made between LR per se and all other related resources/entities, such as reference documents related to the resource (papers, reports, manuals etc.), persons / organizations involved in their creation and use (creators, distributors etc.), related projects and activities (funding projects, activities of usage etc.) and licenses (for the distribution of the LRs).

4 Proposed LR taxonomy

Central to the model is the LR taxonomy, which allows us to organize the resources in a structured way, taking into consideration the specificities of each type. The *resourceType* is the basic element according to which the LR types and subsequently the specific profiles are defined and take one of the following values:

- **corpus** (including written/text, oral/spoken, multimodal/multimedia corpora)
- **lexical / conceptual resource** (including terminological resources, word lists, semantic lexica, ontologies etc.)
- **language description** (including grammars, language models, courseware etc.)
- **technology / tool / service** (including basic processing tools, applications, web services etc. required for processing data resources)
- **evaluation package** (for packages of datasets, tools and metrics used for evaluation purposes).

5 Contents of the model

The core of the model is the *ResourceInfo* component (Figure 1), which contains 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.

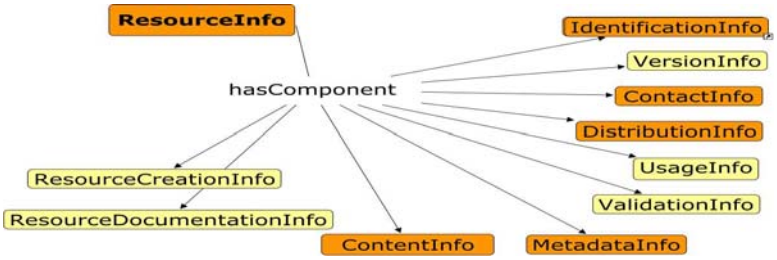

Figure 1

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

- the *IdentificationInfo* component includes all elements required to identify the resource, such as the resource full and short name, the persistent identifier (PID, to be assigned automatically by the system) 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
- the *ValidationInfo* component provides an indication on whether the resource has been validated and the relevant details
- the *ResourceCreationInfo* and its dependent components group together information regarding the creation of a resource (creation dates, relevant project name etc.)
- the *UsageInfo* component provides information on the intended use of a resource (i.e. the application(s) for which it was originally designed) and its actual use (i.e. applications for which it has already been used).
- the *MetadataInfo* is responsible for all information relative to the metadata record creation, such as the catalogue from which the harvesting was made (in the case of harvested records) or the creation date and metadata creator (in case of records created from scratch using the metadata editor) etc.

- the *ResourceDocumentationInfo* provides information on publications and documents describing the resource; basic documents (manuals, guidelines etc.) can be included in the META-SHARE repository and linked to the resource;
- finally, the *ContentInfo* component describes the essence of the resource, specifying the *resourceType* and the *mediaType* elements, which give rise to specific components, distinct for each LR type, presented below.

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

The *ContentInfo* component (Figure 2) groups together information on the contents of the resource. The elements included are:

- *description*: free text describing the resource
- *resourceType* having the values *corpus, lexical/conceptual resource, language description, technology/tool/service, evaluation package*
- *mediaType* (used for data resources)
- *mediaTypeCompatibility* (used for tools)

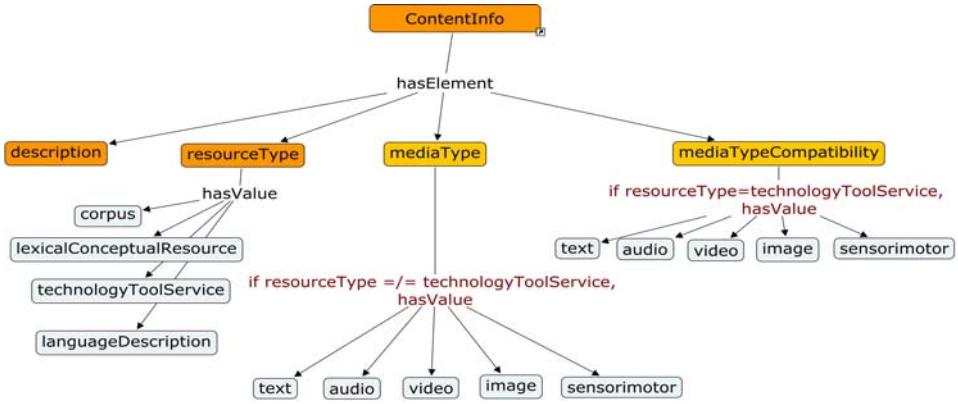

Figure 2

The following *mediaType* values are foreseen:

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

- audio (+ text): the audio feature set will be used for a whole resource or part of a resource that is recorded as an audio file; its transcripts will be described by the relevant Text feature set
- image (+ text): the Image feature set is used for photographs, drawings etc., while the Text set will be reserved for its captions
- video: moving image (+ text) (+ audio(+ text)): used for multimedia corpora, with Video for the moving image part, Audio for the dialogues, and Text referring to the transcripts of the dialogues and/or subtitles
- sensorimotor: used for resources which contain data collected through the use of relevant equipment (gloves, helmets, body suits, laryngographs, etc.) and used to measure the activity of non-verbal modalities (gestures, facial expressions, body movements, articulatory activity, etc.) and their interaction with objects.

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

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

- *CorpusInfo*, *LexicalConceptualResourceInfo*, *LanguageDescriptionInfo*, *TechnologyToolServiceInfo* and *EvaluationPackageInfo* include information specific to each LR type
- *TextInfo*, *AudioInfo*, *VideoInfo*, *ImageInfo*and *SensorimotorInfo* provide information depending on the media type of a resource.

The mandatory generic components and elements in the **minimal schema** are:

- *IdentificationInfo*, including the name of the resource and persistent identifier
- *ContentInfo*: all elements (*description*, *resourceType, mediaType*) are mandatory
- *DistributionInfo*: *availability* must be filled in and depending on the type of availability, further elements are mandatory (license, types of restrictions if any etc.)
- *MetadataInfo*: depending on the way the metadata record has been created (harvesting vs. manual creation), a different set of elements must be filled in

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

In the next section, we provide a detailed view of the subset of the model for text corpora.

6 Text corpora

Text corpora are marked as such by the element *resourceType=corpus, mediaType=text* and their description must include a *CorpusInfo* and a *TextInfo* component (Figure 3). As already mentioned, alongside the text corpora, the textual parts of audio corpora (transcriptions) and video corpora (subtitles) are also included.

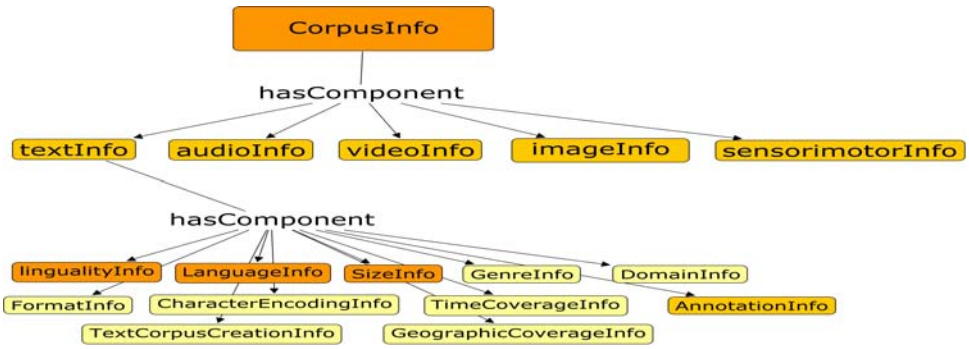

Figure 3

The type dependent information for text corpora is represented in the following components:

- *LingualityInfo* provides information on the linguality type (mono-/bi-/multilingual corpora) and multilinguality type of text resources (parallel vs. comparable corpora)
- *LanguageInfo* comprises information on the language(s) of a resource; the *LanguageVarietyInfo* component supplies information for data in regional language varieties, dialects, slang etc., if contained in the resource
- *SizeInfo* provides information on the size of the whole resource but can also be attached to any other component that needs a specification of size
- *AnnotationInfo* groups information on the annotation of text corpora, such as annotation types (lemmatization, PoS tagging, semantic annotation etc.), methods etc.

The above four components are obligatory for all text corpora. A further set of components (*FormatInfo*, *CharacterEncodingInfo*, *TextCorpusCreationInfo)* are recommended, while optional components used to give information on the corpus are *TimeCoverageInfo, GeographicCoverageInfo*, *DomainInfo* and *TextGenreInfo*.

7 Conclusions and future work

The current version of the model contains the general model, the application of the model to text corpora (as presented in this paper), but also the LR types lexical conceptual resources, audio, video, image, and sensorimotor. Work is ongoing for the rest types, namely LanguageDescription, technologyToolService and evaluationPackage. In this process, it is expected that new components and elements will arise. A set of resources selected to represent all LR and media types is being described according to the model, in order to test its functionality; these resources with their descriptions will be uploaded in the prototype infrastructure for testing and exemplification purposes.

8 References

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

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

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

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

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

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

Αγγλο-ελληνικό γλωσσάριο όρων

Αγγλικός όρος	Ελληνικός όρος
• component	• συστατικό
• element	• στοιχείο
• interoperability	• διαλειτουργικότητα
• language resource	• γλωσσικός πόρος
• lemmatization	• λημματοποίηση
• metadata harvesting	• συγκομιδή μεταδεδομένων
• PoS tagging, Part of Speech tagging	• επισημείωση μέρους του λόγου
• repository	• αποθετήριο
• sensorimotor	• αισθησιοκινητικός
• web service	• διαδικτυακή υπηρεσία

Μαρία Γαβριηλίδου

Γλωσσολόγος,
Αρτέμιδος 6 & Επιδαύρου, 151 25 Μαρούσι
Ηλ.ταχ.: maria@ilsp.gr

Πένυ Λαμπροπούλου

Αρτέμιδος 6 & Επιδαύρου, 151 25 Μαρούσι
Ηλ.ταχ.: penny@ilsp.athena-innovation.gr

Στέλιος Πιπερίδης

Ηλεκ/γος Μηχανικός ΕΜΠ
Αρτέμιδος 6 & Επιδαύρου, 151 25 Μαρούσι
Ηλ.ταχ.: spip@ilsp.gr