SHACHI
LANGUAGE RESOURCES METADATA DB

SHACHI Metadata Set

Updated on 27 Jan. 2008

Table of Contents

SHACHI Metadata Set

This document defines the metadata set used by SHACHI to describe language resources. SHACHI metadata set is based on DCMI Metadata Terms and conforms to OLAC metadata forms. The definition of all DC elements, DC refinements, OLAC metadata set, and OLAC extensions are found at the following URLs.

DCMI (The Dublin Core Metadata Initiative)
http://dublincore.org/index.shtml
DCMES (The Dublin Core Metadata Element Set)
http://dublincore.org/documents/dcmi-terms/

OLAC (Open Language Archives Community)
http://www.language-archives.org/
OLAC Metadata set
http://www.language-archives.org/OLAC/metadata.html

SHACHI conforms to 15 DCMES. SHACHI also uses DC Element refinements and The DC Type Vocabularies. In addition to these DC qualifiers, SHACHI utilizes OLAC extensions and SHACHI extensions so as to achieve more specific description of the language resources.

1. contributor

1.1. DCMES: Contributor

1.1.1. OLAC: Participant Roles [olac:role]

http://www.language-archives.org/REC/role.html

* annotator
* author
* compiler
* consultant
* data_inputter
* depositor
* developer
* editor
* illustrator
* interpreter
* interviewer
* participant
* performer
* photographer
* recorder
* researcher
* research_participant
* responder
* signer
* singer
* speaker
* sponsor
* transcriber
* translator

1.1.2. SHACHI: Attribute of Speaker / Author

Definition: Attribute of a contributor

Comments: There are 5 controlled vocabularies as a value of the speaker and author (OLAC role) as a contributor.

1.1.2.1. SHACHI: mother-tongue

Definition: The speaker is whether a native or non-native speaker of the language.

Comment: Select "native" unless otherwise specified. Typical case to select "non-native" is such as "English spoken by Japanese" for a research purpose.

1.1.2.2. SHACHI: intonation

Definition: Dialectal status whether the performer uses a standard language or a dialect.

Comment: It describes whether the speaker of the language is a standard language speaker. Unless the resource collects a dialect, select "standard dialect". Only when the resource collects a particular dialect, specify the name of the dialect.

1.1.2.3. SHACHI: level

Definition: Whether the speaker has received a professional level of linguistic(speaking or writing) training or no such training.

Comment: If the resources contain languages of trained speakers, describe their professions. If the speakers of the languages are ordinary people who are not particularly trained in speaking those languages, select "amateur". Specify other attributes such as "English learner". For example, announcers or journalists are "professional". When there is no such information on the speakers, select "amateur".

1.1.2.4. SHACHI: age

Description: An age group the performer is in. When there are many performers in a resource, the ratio between all the age groups to which the performers belong.

Comment: "It describes which age group the speakers of languages are in. e.g. people in their twenties, children from 13 to 15 years old. Select "adult" unless otherwise stated. Select "child" if the resource contains children's language such as infant speech development.

1.1.2.5. SHACHI: gender

Description: Sex of the speaker. When there are many performers in the resource, the ratio of males and females.

Comment: Select "male", "female", or "male & female". If there is the data of the ratio between males and females, describe it. If there is no information regarding the gender of the speakers, and when it is considered that both males and females are apparently participated, select "male & female"

2. coverage

2.1. DCMES: Coverage

2.1.1. DC refinement: Temporal

2.1.2. DC refinement: Spatial

3. creator

3.1. DCMES: Creator

4. date

4.1. DCMES: Date

4.1.1. DC refinement: Created

4.1.2. DC refinement: Issued

4.1.3. DC refinement: Modified

5. description

5.1. DCMES: Description

5.1.1. SHACHI: Price

Definition: The price information of the resource in a particular medium.

Comment: If the resource is distributed at fixed price, mention the price. If they are free of charge, indicate "free".

6. format

6.1. DCMES: Format

6.1.1. DC refinement: Extent

6.1.2. DC refinement: Medium

6.1.3. SHACHI: functionality

Definition: Software Functionality

Comments: The software tools to manipulate the resource included in or attached to the language resource. If the name of the tool is not mentioned and only the description of the tool is found, summarize the description. If the description has only a vague idea, write the URL.

Examples: A parser, word sorting tool, audio visualizer, etc.

6.1.4. SHACHI: markup

Definition: A markup scheme used by a digital resource.

Comments: HTML SGML XML, etc.

6.1.5. SHACHI: encoding

Definition: An encoded character set used by a digital resource.

Comments: ASCⅡ, EUC, Shift JIS, etc.

7. identifier

7.1. DCMES: Identifier

8. language

8.1. DCMES: Language

8.1.1. OLAC-Language extension

The ISO639-3 extension

9. publisher

9.1. DCMES: Publisher

10. relation

10.1. DCMES: Relation

10.1.1. DC refinement: isVersionOf

10.1.2. DC refinement: hasVersion

10.1.3. DC refinement: isReplacedBy

10.1.4. DC refinement: Replaces

10.1.5. DC refinement: isRequiredBy

10.1.6. DC refinement: Requires

10.1.7. DC refinement: isPartOf

10.1.8. DC refinement: hasPart

10.1.9. DC refinement: isReferencedBy

10.1.10. DC refinement: References

10.1.11. DC refinement: isFormatOf

10.1.12. DC refinement: hasFormat

10.1.13. DC refinement: conformsTo

10.1.14. SHACHI: Utilization

Definition: Applicability of the resource.

Comment: The described resource is utilized for the described purpose at the phase of research and development.

11. rights

11.1. DCMES: Rights

12. source

12.1. DCMES: Source

13. subject

13.1. DCMES: Subject

13.1.1. OLAC: Linguistic Subject Vocabulary

* anthropological_linguistics
* applied_linguistics
* cognitive_science
* computational_linguistics
* discourse_analysis
* forensic_linguistics

* general_linguistics
* historical_linguistics
* history_of_linguistics
* language_acquisition
* language_documentation
* lexicography
* linguistics_and_literature
* linguistic_theories
* mathematical_linguistics
* morphology
* neurolinguistics
* philosophy_of_language
* phonetics
* phonology
* pragmatics
* psycholinguistics
* semantics
* sociolinguistics
* syntax
* text_and_corpus_linguistics
* translating_and_interpreting
* typology
* writing_systems

13.1.2. OLAC-Language extension
The ISO639-3 extension

13.1.3. SHACHI: mono_multi_lingual
Definition: The value may either Monolingual or Multilingual.

13.1.3.1. monolingual <SHACHI value>
Definition: A resource using only one language.
Comment: The same language is used for the subject language and to describe the subject language.
Example: Mainichi Newspaper 1991-2006 data files

13.1.3.2. multilingual <SHACHI value>
Definition: A resource in several different languages.
Comment: Two or more languages are collected in the resource. One language is translated into another language.
Example: ECI/MCI (European Corpus Initiative/Multilingual Corpus)

13.1.4. SHACHI: Resource Subject
Select a resource type using a controlled vocabulary; corpus, dictionary, thesaurus, or glossary.

13.1.4.1. corpus <SHACHI value>
Definition: A collection of written or spoken texts.
Example: A corpus of 1 million words of spoken English/ the whole corpus of Renaissance poetry.

13.1.4.2. dictionary <SHACHI value>
Definition: A list of the words of a language in which the definitions or meanings of the words are explained either in the same language or in a different language.

Example: Oxford English Dictionary, Eijiro on the Web, English Japanese Learner's Dictionary

13.1.4.3. thesaurus <SHACHI value>

Definition: A list of the words of a language in which the words are arranged in groups that have similar meanings.

Example: COBUILD Thesaurus, Oxford Concise School Thesaurus

13.1.4.4. glossary <SHACHI value>

Definition: A list of words on a particular subject.

Example: National Information Assurance Glossary, Free Online Dictionary of Philosophy

14. title

14.1. DCMES: Title

14.1.1. DC refinement: alternative

15. type

15.1. DSMES: Type

15.1.1. The DCMI Type Vocabulary

SHACHI uses three of these vocabularies: Image, Sound, and Text (underlined)

* Collection
* Dataset
* Event
* Image
* Interactive Resource
* Service
* Software
* Sound
* Text

15.2. OLAC: Linguistic Data Types

[olac:linguistic-type]

http://www.language-archives.org/REC/type.html

* lexicon
* primary_text
* language_description

15.3. Discourse Types

[olac:discourse-type]

http://www.language-archives.org/REC/discourse.html

* drama
* formulaic_discourse
* interactive_discourse
* language_play
* oratory
* narrative
* procedural_discourse
* report

* singing

* unintelligible_spee

15.4. SHACHI: Purpose

The purpose of the creation of the language resource.

* lexicography
* analysis
* developing_technologies
* education

15.4.1. lexicography <SHACHI value>

Definition: The creation of the resource intended for lexicography.

Example: Longman Corpus Network

15.4.2. analysis <SHACHI value>

Definition: The creation of the resource intended for analysis.

Example: The resource for the study of first and second language acquisition.

15.4.3. developing_technologies <SHACHI value>

Definition: The creation of the resource intended for developing technologies.

Example: The purpose of simultaneous interpretation corpus of Nagoya University is "to develop machine interpretation technology".

15.4.4. education <SHACHI value>

Definition: The creation of the resource intended for the use in education.

Examples: Longman Learners' Corpus

15.5. SHACHI: Style

Style describes whether the resource collects sentences in a spoken style (colloquial style) or in a written style.

15.5.1. speech <SHACHI value>

Definition: The resource collects sentences in spoken (colloquial) style.

Example: Longman Spoken American Corpus

15.5.2. written <SHACHI value>

Definition: The resource collects sentences in written style.

Example: Longman Written American Corpus

15.6. SHACHI: Form

A type of form describes whether sentences recorded in the language resource are composed in a fixed form or in an unfixed form.

15.6.1. fixed <SHACHI value>

Definition: A collection of fixed forms of expressions.

Examples: Greetings: Good morning, Goodbye./ Fast food shop: Is this for here, or to go?, One moment, please./ In a letter or email: "Should you have further questions, please do not hesitate to contact us."/ Automatic Teller Machine: Withdrawal? / Written regulations, acts, or rules as their styles are fixed to some extent.

15.6.2. unfixed <SHACHI value>

Definition: The resource collects various forms of expressions.

Examples: The unlimited sentences composed in an unfixed form. For example, prose, newspaper database, simultaneous interpretation database.

15.7. SHACHI: Sentence

Sentence describes the length of each sentence.

15.7.1. short <SHACHI value>

Definition: A collection of short sentences

long <SHACHI value>

Definition: A collection of long sentences

15.7.2. mixed <SHACHI value>

Definition: A collection of varying length of sentences

15.8. SHACHI: Annotation

Added notes to the text of the recourse, giving extra information to the text, especially for machine reading and parsing.

15.8.1. annotated <SHACHI value>

Definition: Tagged corpus

Example: Lancaster Parsed Corpus / A syntactic analysis in the form of a phrase marker. / The texts are tagged with part-of-speech and morphological annotation.

15.8.2. plain <SHACHI value>

Definition: A corpus without annotations.

15.9. SHACHI: Annotaion_sample

Definition: A sample of tagged data

15.10. SHACHI: Sample

Definition: A sample of the language resource