Integration of language attributes in Metadata and Customization for Domain Specific Repositories

Devika P. Madalli[1]
A.R.D.Prasad[1]
Imma Subirats[2]
[1]Documentation Research and Training Centre,
Indian Statistical Institute, Bangalore, India
{devika, ard}@drtc.isibang.ac.in

[2]UN Food and Agricultural Organization, Rome, Italy

Abstract. The paper presents a system for customized metadata facility adapted to expose domain specific metadata along with language attributes. Quality metadata is crucial for efficient information retrieval. Use of standard and common vocabulary leads to normalization of differences that arise due geographic, cultural, domain specific environments even amongst digital libraries with same or similar collection scopes. The present work demonstrates implementation of language attributes (which is missing in the official version of DSpace) for metadata elements. As a case study we demonstrate the AGRIS (Agricultural Information System) application profile customized for metadata in DSpace repositories. It serves as a full text content management platform for the agricultural community and especially the language-wise disparate and geographically distributed AGRIS network through enriched domain metadata in standard format for the description, management and better metadata sharing.

Keywords. Multilingual Digital Libraries, DSpace, AGRISAP, Dublin Core

1 Introduction

Digital Libraries (DLs) with especially in online network environment are expected to serve multi-cultural and multi lingual communities [1]. Digital Libraries collect and organize resources and provide metadata to facilitate retrieval. There are different metadata schema adopted according to the nature of collections and services of DLs. However, language attribute in metadata is often only set to a default value and in general metadata is provided in a single language. However, digital libraries today cater to diverse communities with different language backgrounds and requirements. It becomes essential therefore to include and indicate in what language data in metadata elements is provided. The present work is an attempt to illustrate implementation of Agricultural Information Service Application Profile[1] (AGRISAP)

[1]http://aims.fao.org/website/The-AGRIS-AP/sub

and language attributes at different levels in a digital library in agricultural domain. The implementation is done for AGRIS data centres of UNFAO.

The AGRISAP was created specifically to enhance the description, exchange and subsequent retrieval of agricultural Document-like Information Objects (DLIOs) [2]. It is a metadata schema which draws elements from Metadata standards such as Dublin Core[2] (DC), Australian Government Locator Service Metadata[3] (AGLS) and Agricultural Metadata Element Set[4] (AgMES) namespaces. It is a format that allows sharing of information across dispersed bibliographic systems [2]. AGROVOC [5] is a multilingual structured thesaurus of all subject fields in Agriculture, Forestry, Fisheries, Food security and related domains (e.g. Sustainable Development, Nutrition, etc). Its main role is to standardize the indexing process in order to make searching simpler and more efficient, and to provide users with the most relevant resources. AGROVOC is developed by FAO[3].

DSpace was built as a digital repository suit with the objective of storing, indexing, preserving and disseminating an organization's research material in digital formats [4]. MIT and Hewlett Packard designed the system between 2000 and 2002. At present it is in version 1.6. DSpace integrates a user community orientation into the system's structure. This design supports the participation of different communities within universities, research institutes, departments and any other unit of research institutions. As their requirements might be different, DSpace allows the workflow and other policy-related aspects of the system to be customized. The default metadata format in DSpace is Qualified Dublin Core. As already mentioned for Agricultural resources, AGRISAP is a world standard for resource description and bibliographic data exchange. Accordingly, customization includes creation of additional element set, modified data input forms with language attributes to indicate language of the resource and also language of data in a few metadata elements and alteration of the database tables in compliance with the AGRIS standard. Further indexing by changed or added elements and corresponding changes for search facility by specific metadata elements are incorporated. To facilitate data exchange the next step involves achieving OAI-PMH compliance of the DSpace resources using AGRISAP.

The system developed facilitates multilingual metadata provision with specific attributes that indicate the language used. It serves as a full text content management platform for the agricultural community and especially the language-wise disparate and geographically distributed AGRIS network through enriched domain metadata in standard format for the description, management and better metadata sharing.

DSpace is OAI compliant and thus bibliographic records can be easily harvested by service providers for inclusion into their systems. Bibliographic resources are described with qualified Dublin Core (QDC) metadata. By default only 3 fields (title, language and date) are mandatory. However, DSpace can be customized to use other application profiles for metadata like the AGRISAP which is meant for description of documents in the agricultural domain. This allows the use of more sophisticated and specialized metadata elements. For example, it is possible to describe the resource

[2]http://dublincore.org/

[3]http://www.agls.gov.au/

[4]http://aims.fao.org/agmes-metadataset

[5]http://aims.fao.org/website/AGROVOC/sub

with one or more subject keywords and/or choose from a vocabulary device such as the AGROVOC thesaurus.

2 DSpace and AGRISAP Metadata

DSpace has adapted QDC (Qualified Dublin Core) metadata for description of resources. But it is not fully compliant and deviates from the standard in a few elements. However, DSpace input form is only a default format and it can be extended through 'input-forms.xml' file. The main deviation is that DSpace uses contributor.author for creator more as a legacy. However, when exposing data through OAI-PMH, the data in contributor.author appears as creator which is the actual QDC element.

AGRISAP also does not incorporate all the elements defined in Qualified Dublin Core. AGRISAP, though based on QDC, additionally borrows a few elements from AGS and AGLS. ARN (AGRIS Record Number) is unique to AGRISAP. Further, AGRISAP uses elements 'creator' and 'refinedby', adding extra qualifiers like, creatorPersonal, creatorCorporate, creatorConference. The element publisher in AGRISAP is further refined with qualifiers into publisherName, publisherPlace as in the case of a typical library catalogue. 'Subject' (keyterms) accommodates both the subject classification and also subject keywords using controlled vocabularies of agriculture related thesauri. A major deviation of AGRISAP from QDC is with regard to type/medium of a digital object. QDC[6] uses the element type for the medium of a digital object, AGRISAP uses format.medium instead. DCMI[7] type vocabulary of DSpace includes: animation, article, book, book chapter, dataset, learning object, image, image-3D, map, musical score, plan of blueprint, preprint, presentation, recording acoustical, recording musical, recording oral, software, technical report, thesis, video, working paper, other. However, QDC lists relatively fewer, they are: collection, dataset, event, image, interactiveResource, movingImage, physicalObject, service, software, sound, stillImage, text. In contrast, AGRISAP uses format.medium to enter the information about the medium, following IMT which can include: Microfilm, Microfiche, VCD, DVD, Audiotape, Reel, Film, Tape, CD-ROM, Videocassette, Videodisc, Videotape. A detailed mapping exercise to compare compatibility of DSpace default metadata and AGRISAP with Qualified Dublin Core elements was conducted. The mapping is presented in the table below:

Table 1. Mapping of DSpace and AGRISAP metadata element sets

DSpace			**AGRISAP**		
Element	Qualifiers	Schema	Element	Qualifiers	Schema
contributor	Author				
title			Title		
	alternative			alternative	
date	Issued		Date	Issued	W3CDTF

[6]http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/
[7]http://dublincore.org/documents/dcmi-type-vocabulary/

publisher			publisher	publisherName,	
				publisherPlace	
identifier		ISBN, ISSN, URI, ISMN, Govt.Doc, other	identifier		URI,ISBN RN,JN,PN IPC,DOI
	Citation				
Relation	ispartofseries		relation	isVersionOf	URI ISBN RN JN PN IPC DOI
				hasVersion	
				isReplacedBy	
				Replaces	
				isRequiredBy	
				Requires	
				isPartOf	
				hasPart	
				isReferencedBy	
				references	
				isFormatOf	
				hasFormat	
				isTranslation Of	
				hasTranslation	
type		DCMI Type Vocabulary			
language		ISO 639-2 RFC 1766	language		ISO639-2 ISO639-1
subject				subjectClassification	ASC, CABC DDC, LCC UDC
				subjectThesaurus	AGROVOC CABT, ASFAT NALT, MeSH LCSH

description			description		
	Abstract			Abstract	
	sponsorship			descriptionNotes	
				descriptionEdition	
			creator	creator Personal	
				creatorCorporate	
				creatorConference	
			availability	availabilityLocation	
				availabilityNumber	
			rights	rightStatement	
				termOfUse	
			coverage	Spatial	POINT,ISO3166 TGN, Box
				Temporal	Period, W3CDTF
			citation	citationTitle	
				citationIdentifier	ISSN, CODEN
				citationNumber	
				citationChronology	
			format	Extent	
				Medium	IMT
			ARN		

3 Multi-lingual Digital Libraries

Language issues in Digital libraries are multifarious [5]. Over the years we have observed the growth of different types of digital libraries distinct by their user communities, resources, formats and services.This proves that an increasing amount of the world's knowledge is being organized in domain-specific compartments and stored in digital form, accessible over the Internet. DLs have included in their objectives though gradually the need for multilingual services. Accordingly they include many languages and a lot of documents are being emanating from the non-english speaking countries. This has opened avenues for investigations into the issues of multilingualism in the digital library world. The present state of the research in multilingual, or cross language access can be broadly categorized as:

1. Multiple language recognition, manipulation and display; and
2. Multilingual or cross-language search and retrieval.

Significant work has been carried out in the area of multiple language recognition and representation. Such as, internationalization of hypertext markup language [6],

provision of multilingualism in HTML, localization and presentation issues using UNICODE[8]. With these developments it can be said that the WWW has matured to provide the platform for representing and manipulating multilingual resources in DLs and to encourage the monolingual DLs go from local to global. The Cross language search and retrieval is rather more challenging task and it is the main basis for the recent ongoing works. The machine translation groups [7] working for the automatic translation of queries send by users to the native language of the system is proof for demand cross-lingual services. Cross-Language Information Retrieval (CLIR), in which methodologies and tools developed for Natural Language Processing (NLP) are being integrated with techniques and results coming from the Information Retrieval (IR) field [8]. The other approaches are from knowledge organization field using thesauri, ontologies and dictionaries in different languages. Dagobert Soergel [9] discusses how in information retrieval a thesaurus can be used in two ways, controlled vocabulary indexing and searching and knowledge-based support of free-text searching. Other issues that relate to dealing with multilingual resources in a single Digital Library have also been discussed [10]. It is however to be noted that various digital library software and tools are also now compatible with the multilingual documents. Metadata scheme like Dublin Core and MODS etc are also well supported with the language attributes for better description of the digital documents for quick access and retrieval.

4 Language facilitation issues in DSpace

One of the major issues in DSpace is that it does not allow entering language attribute of an element though, QDC (which DSpace claims to have implemented) has some elements with languages attributes. Also in DSpace language value-pairs are listed. These can be used only with the element 'language' as data. However, there is no facility to use language as an attribute value with other elements. Hence, if one enters the title.alternative in language other than the default language mentioned in dspace.cfg file (config file), DSpace still displays the default language (English) rather than the language of the alternative title. To overcome this, a patch has been developed to DSpace to facilitate provision of the language choice for any element.

5 System Implementation

For the present work, a test bed is created using DSpace repository software. As stated, the default metadata format in DSpace is Qualified Dublin Core. However, we develop a patch for incorporating AGRIS metadata. Within AGRIS, language attribute is applicable for specified elements. The qualifiers are mentioned using a '.' separator as shown below:

[8]http://www.unicode.org/

```
title.alternative
title.supplement
subject.subjectClassification
subject.subjectThesaurs
description.abstract
identifier
coverage.spatial
citation.citationTitle
```

Accordingly the Dublin Core representation of the elements is as shown below:

```
 <field>
       <dc-schema>dc</dc-schema>
       <dc-element>title</dc-element>
       <dc-qualifier>alternative</dc-qualifier>
       <repeatable>true</repeatable>
        <language>iso639_2_languages</language>
       <label>Other Titles</label>
       <input-type>onebox</input-type>
       <hint>Enter titles in other languages, please
enter</hint>
       <required></required>
 </field>
```

6 Language at input stage

The input screens in DSpace are designed using 'input-forms.xml' file. This file describes for each metadata element along with the characteristics like the caption/label, qualifier, input data type, help message etc. However, with the official version one cannot add language attribute. The language attribute patch to DSpace facilitates extending the input-forms.xml file by simply entering a line `'<language>iso639_2_languages</language>'` under each element definition. One can achieve the desired effect of having 'language' button appearing for any element irrespective of the 'input-type'. However, if it is not required to have language attribute for an element, one can either ignore entering the line or may enter with empty value like the following
`<language></language>`.

The following sample entry for title element demonstrates the usage of language attribute where language choice can be entered as shown below:

```
 <value-pairs value-pairs-name="common_iso_languages" dc-
term="language_iso">
      <pair>
        <displayed-value>English</displayed-value>
        <stored-value>en</stored-value>
      </pair>
      <pair>
        <displayed-value>Spanish</displayed-value>
        <stored-value>es</stored-value>
      </pair>
      <pair>
        <displayed-value>German</displayed-value>
        <stored-value>de</stored-value>
      </pair>
      <pair>
        <displayed-value>French</displayed-value>
        <stored-value>fr</stored-value>
      </pair>
      <pair>
        <displayed-value>Italian</displayed-value>
        <stored-value>it</stored-value>
      </pair>
      <pair>
        <displayed-value>Japanese</displayed-value>
        <stored-value>ja</stored-value>
      <pair>
        <displayed-value>Chinese</displayed-value>
        <stored-value>zh</stored-value>
      </pair>
    </value-pairs>
```

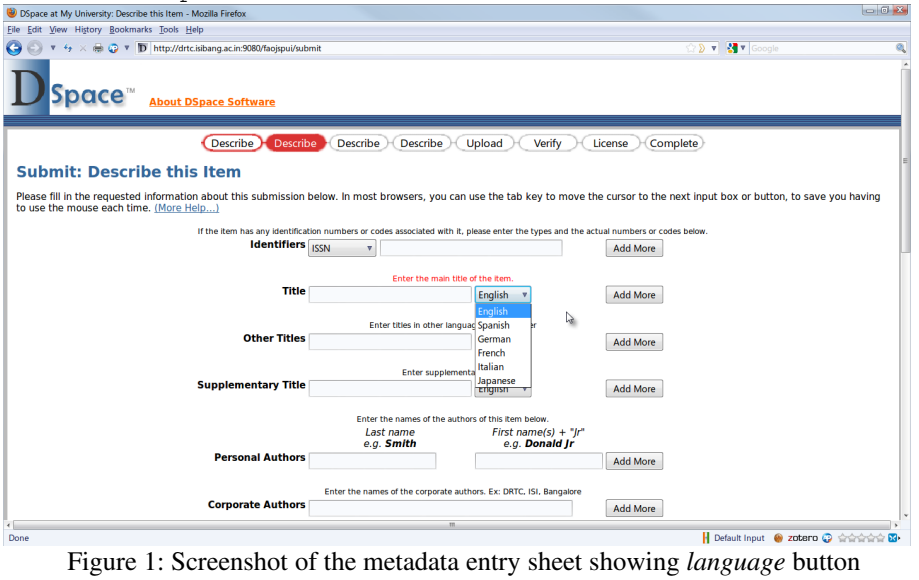

Figure 1: Screenshot of the metadata entry sheet showing *language* button

Though the examples (figure 1) uses two letter language codes like 'fr' for French, 'it' for Italian, following the ISO 639-1, one can change the two letter codes to three letter codes as prescribed by ISO 639-2 or even ISO 639-3[9].

7 Language at Database level

The postgresql database of DSpace contains several tables and metadata values are stored in the table 'metadatavalue' with the following columns,

```
metadata_value_id
item_id
metadata_field_id
text_value
text_lang
place
authority
confidence
```

Fortunately, the language (*text_lang*) column is provided for the elements in the tables, where postgresql stores the 'language' data. But again, normally this column puts language values taken from default language set in the DSpace configuration file, *dspace.cfg*. However, the dspace patch allows the data entry personnel to choose a language from the pull-down menu and that allows storing the chosen value in the *text_lang* column.

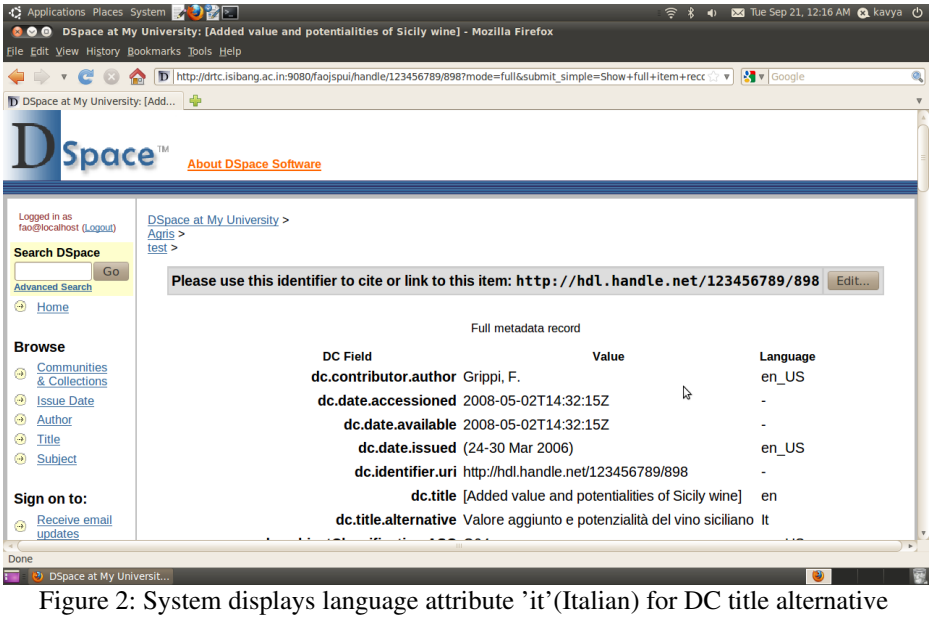

Figure 2: System displays language attribute 'it'(Italian) for DC title alternative

This is displayed to user in full view record as in figure 2.

[9]http://www.sil.org/iso639-3/codes.asp

8 Limitations and further work

At present, for list input data type, language attribute is not provided by default. Normally, the data values of an element for which list input data type is used will have enumerated values and these enumerated values are mentioned as 'value-pairs'. The values in value-pairs can be entered in whatever language the enumeration is provided. AGRISAP does not prescribe language attribute for elements that require enumerated values. Similar arguments can be applied to series input-data type. However, future releases of the patch will attempt to provided 'language' attribute to these two input-data type (list and series), so that the end user/administrator will have the liberty to use them.

Further, for metadata schema which prescribes more than one attributes neither the existing database tables nor the patch developed provides any solution. In such cases, one has to extend the 'metadatavalue' table definition to accommodate more than one attribute. But as the patch aims at least invasive code, it is not attempted to change the structure of the database.

9 Conclusion

The introduction of language attribute makes DSpace fully compliant to QDC and AGRISAP. One of the main goals of the present work, is to enrich metadata from just English language (default) capabilities to multilingual metadata that incorporates language attributes for the elements where AGRISAP has prescribed language as a mandatory attribute. However, it is also important to note that any modification to the official source code release of DSpace, should be least invasive. In fact, if one does not wish to have language attributes for all the elements, they can use the existing input-forms.xml file as such even after using the patch. The work presented here is especially relevant for agricultural content as resources are in different languages with AGRIS community spanning over 200 data centres world over. A liveCD having Ubuntu OS along with DSpace enriched with language attribute is also made available[10].

References

1. Chen, Ching-chih.: Delivery of Web-based Multilingual Digital Collections and Services to Multicultural Populations: The Case of Global Memory Net. News Letter 73, IFLA (2007)
   http://archive.ifla.org/IV/ifla73/papers/097-Chen-en.pdf
2. Food and Agricultural Organisation (2010),
   http://aims.fao.org/printpdf/website/The-AGRIS-AP/sub
3. Lauser, Boris., Sini, Margherita., Liang,Anita., Keizer, Johannes., Katz ,Stephen. : From AGROVOC to the Agricultural Ontology Service / Concept Server An OWL model for creating ontologies in the agricultural domain. In OWLED 06 Workshop on OWL: Experiences and Directions, Athens,Georgia, USA, (2006)
4. Dspace. http://www.dspace.org/.
5. Borgman, Christine L.: Multi-Media, Multi-Cultural, and Multi-Lingual Digital Libraries Or How Do We Exchange Data In 400 Languages?. D-Lib Magazine (1997)

[10]http://drtc.isibang.ac.in/livecd/dspaceagrisap.iso

http://www.dlib.org/dlib/june97/06borgman.html
6. Yergeau, F., Adams, G., Duerst , M.: Internationalization of the Hypertext Markup Language. RFC 2070, Network Working Group (1997)
7. Peters, Carol. , Picchi, Eugenio. : Across Languages, Across Cultures: Issues in Multilinguality and Digital Libraries. D-Lib Magazine (1997)
8. Oard, D.W., Dorr, B.J.: A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies (1996)
9. Soergel, D. : Multilingual thesauri in cross-language text and speech retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford, CA (1997)
10. Mizera-Pietraszko,Jolanta. : Model Design of User Interfaces for Multilingual Digital Libraries. TCDL Bulletin, Vol. 3(3) (2007) http://www.ieee-tcdl.org/Bulletin/v3n3/mizera-pietraszko/mizera-pietraszko.html