OLACMS: Comparisons and Applications in Chinese and Formosan Languages

Ru-Yng Chang
Institute of Linguistics, Academia Sinica
130 Sec.2 Academy Rd.
Nankang, Taipei, Taiwan, 115
ruyng@gate.sinica.edu.tw

Chu-Ren Huang
Institute of Linguistics, Academia Sinica
130 Sec.2 Academy Rd
Nankang, Taipei, Taiwan, 115
churen@gate.sinica.edu.tw

Abstract

OLACMS (stands for Open Language Archives Community Metadata Set) is a standard for describe language resources. This paper provides suggestion to OLACMS 0.4 version by comparing it with other standards and applying it to Chinese and Formosan languages.

1 Introduction[1]

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: (1) developing consensus on best current practices for the digital archiving of language resources; (2) developing a network of interoperating repositories and services for housing and accessing such resources.

Three primary standards are the foundational basis of the OLAC infrastructure that serve to bridge the multiple gaps which now lie in between language resources and users: (1)OLACMS: the OLAC Metadata Set (Qualified DC, Dublin Core), (2) OLAC MHP: refinements to the OAI (Open Archives Initiative, http://www.openarchives.org) protocol, and (3) OLAC Process: a procedure for identifying Best Common Practice Recommendations.

It is crucial to note that the OLAC standards are not standards for the language resources community alone. They are based on two broadly accepted standards in the digital archives community. First, the Dublin Core Metadata Initiative (DCMI) is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. There are fifteen Doblin Core Metadata Elements (DCMS) and their qualifiers. OLACMS extends the DC minimally to anwer the needs of the language archives community (Bird, Simons, and Huang 2001).

Second, the Open Archives Initiative (OAI) was launched in October 1999 to provide a common framework across electronic preprint archives, and it has since been broadened to include digital repositories of scholarly materials regardless of their type. The OAI infrastructure requires compliance with two standards: the OAI Shared Metadata Set (i.e. DCMS), and the OAI Metadata Harvesting Protocol (MHP). The OAI MHP allows software services to query a repository using HTTP requests, also an important feature of the recently hyped Semantic Web (http://www.w3.org/2001/sw/). Using the OAI infrastructure, the community's archives can be federated and become a virtual meta-archive collecting all available information. The federeated structure allows end-users to query multiple archives simultaneously. Currently, the Linguistic Data Consortium has harvested the catalogs of over 20 participating archives on behalf of OLAC, and created a search interface which permits queries over all 30,000+ records. A single search typically returns records from multiple archives. The prototype can be accessed via the OLAC website.

[1] We are indebted to Steven Bird and reviewers of the 3rd Workshop on Asian Language Resources and International Standardization for their valuable comments and corrections. Colleagues of the Language Archives project at Academia Sinica provided data and suggestions. Any remaining errors are ours.

In this paper, we trace the version changes of OLACMS, especially in comparison with other (often related) international standards. We will then concentrate on the application of OLACMS to Chinese language resources. In conclusion, we will make some suggestions for OLACMS to account for the characteristics of Chinese language archives.

2 Mapping with other international standards

2.1.Mapping with IMDI

ISLE Meta Data Initiative (IMDI) is a cousin of OLACMS. IMDI proposes a metadata set for natural language processing under the broader International Standards for Language Engineering (ISLE) project. ISLE is co-sponsored by the European Commission of the EU and National Science Foundation of the USA. It aims to develop a set of internationally accepted standards for natural language processing base on the result of the earlier European standard building project (EAGLES, http://www.ilc.pi.cnr.it/EAGLES96/home.html). On one hand, IMDI is an elaboration of OLACMS since it deals specifically with recording sessions. They can also be considered a complimenting each other since they are both devised under the aegis of ISLE.

IMDI Metadata Elements for Session Descriptions, Version 2.5 was completed in June 2001. The elements evolved from the previous EAGLES metadata set described in Wittenburg et al. (2000). Both metadata sets share the aim to improve the accessibility/availability of Language Resources (LR) on the Internet. To achieve this goal, they created a browsable and searchable universe of meta-descriptions similar to those devised by other communities on the Internet.

The focus on Session Description was motivated in Broeder et al. (2000). They observed that individual linguistic resource usually exists in clusters of related resources. For instance, a field video recording of an informant who describes a picture sequence involves several resources. By his definition, an (linguistic) event that called a session is the top element and there results a number of related linguistic resources: Video tape, Photographs, Digitised video file, Digitised photographs, Digitisations of the images used as stimuli, One electronic transcription file, One or more electronic analysis files, Field notes and experiment descriptions (in electronic form). However, since not all linguistic resources come to existence directly through sessions, hence not all linguistic resources can be described by IMDI.[2]

In principle, IMDI metadata can be mapped to OLAC metadata, just as OLAC metadata can be mapped to DC. IMDI Team (August 2001) mapped IMDI Session Descriptions with OLAC 0.3 Version. IMDI Team also use existing description formalisms used by institutions that deal with "published corpora" such as [ELRA] and [LDC]. The set of metadata elements that describe "published corpora" are called "catalogue" metadata elements. The IMDI Team (Gibbon, et al. 2001) launched IMDI Metadata Elements for Catalogue Descriptions, Version 2.1. It also includes Metadata Elements for Lexicon Descriptions.

OLACMS has been updated since December 2001. Hence we did an updated comparison and present the result in this section. Note that since IMDI is an elaboration of OLACMS, we concentrate on the IMDI elements that are not specified in OLACMS and are likely to find wider application. Please note that the section contains our own recommendations inspired by the IMDI/OLAC comparison. We try to add our motivation even for the items that are directly adopted from IMDI. In terms of OLAC scheme, these suggested revision/addition can be assigned the status of attributes (for use by sub-communities), and can be incorporated into the OLACMS later if the community find such addition necessary.

2.1.1. Controlled Vocabulary

Controlled vocabulary defines the basic concepts of the metadata set and any addition to the controlled vocabulary should be motivated by

[2] It is possible to conceive language resources such as lexica and grammars as created through a very large set of (non-planned and non-documented) sessions. But this consideration is beyond the scope of this paper and will not be pursued further here.

the essentiality of the concept.

- **Controlled Vocabulary for *Logical Structure* of linguistic resources**: Language resources come in different forms and various units. A critical piece of information in cataloguing language resource is a description of the composition of the resources. For instance, any English lexicon can be conventionally and naturally viewed as composed of 26 sections defined by shared initial alphabet. Having an element of Logical Structure: alphabetically ordered would give us vital information of how to manipulate the resource. Other vocabularies such as 'sequential chapter', 'dialogue turns', or 'sequential phonemes' would also offer crucial information. In addiiton, if sequential database is indeed the future of language resources, the description of the sequencing logic will play an essential role.
- Add Annotator to [OLAC–Role]. By annotator, we do not refer to the natural person or an automatic program who puts the tags on. By annotator we refer to the institution that implemented the annotation. This information is crucial since this annotator 1) has at least partial IP right on the resource; 2) often set/defines the tagset standard adopted (e.g. Brown, LOB, Penn TreeBank). In other words, annotator can differentiate a new version of resource or even identify totally new resource.
- Add values of archiving Quality to the refine controlled vocabulary of Format.

2.1.2. Elements

One existing elements may need further refining with existing mechanisms.

- Refining the element Project: Many language resources are developed under or partially supported by a project grant. For now, a project can be the value of Creator or Contributor. But just like all other individual creators and contributors, a project needs to be described in fuller details. We need to use attributes to describe the Founder, PI's, Host Institutes, etc. of a project. An umbrella project, such as EAGLES, ISLE, or at a even more complex level, ESPRIT, requires elaboration of contributors and funding timelines themselves.

2.1.3. Updating and Revising the Attributes

- Add sub type to the Space attribute : Coverage of the language resources often calls for geographical information. Hence we need to define the subtypes that include Continent, Country, Administrative division, longitude, latitude, address, etc.
- Add subtype for non-standard Identifier : There are many sets of identifers are defined locally and do not follow URL. In this case, we can add the name of the identifier system (or cataloguer) under schme. For instance, each libary often has its own set of call numbers. Other well-known identifiers arre LCC Catalog No (<Identifier sceeme="LCC"> LCC Catalog No</Identifier>). This could also apply to well-established identifiers such as ISSN and ISBN.
- Although OLAC:Format does not stipulate any refine attributes, however, it is already stipulated in DC:Format. The DC format refine has two control vocabulary entries: Medium specifies the material that the cataloguer uses; while extent records size and duration of the archive. We suggest that OLAC can simply adopt these two refine attributes.

2.2. Mapping with Linguistic Documentation Archives

In addition to IMDI Metadata, Gary Holton (2000) also proposes a system of metadata for the description of language documentation resources following OLACMS. While the system described here should be sufficient for any linguistic resource, it is motivated by the specific ongoing need to describe linguistic documentation materials contained in the Alaska Native Language Center (ANLC) Archive. Particular attention is paid to description of first-hand documentation materials such as field notes, grammatical notes, and phonological descriptions, many of which currently exist only in written form. Existing resources are in the process of being digitized, and new digital resources continue to be acquired. The ANLC collection presently contains more than ten thousand items. While much of the material consists of original manuscripts of archival quality, the collection also includes published materials and materials existing in other archival collections, duplicated in whole or in part. The ANLC Archive thus combines both archival and library functions.

The unique need described in Holton (2000) is that he wants the Metadata set to be applied simultaneously to non-digital archives, such as manuscript, reel-to-reel cassettes, CD recordings etc. This can be done by adopting the DC:Format refine attribute of Medium. In order to descibe the archives more felicitously, we also need to add speaker, interviewer Holder, and Guardian to the value of controlled vocabulary of refine of Creator and Contributor. However, there does not seem to be any straightforward way to transfer Target Dialect.

2.3.Summary

Based on the two comparison of different metadata sets, we found that the DC qualifier can be applied effectively to solve the bridging and conversion problems between different DC-based extension metadata sets. This should be exactly what OLACMS design has in mind. The attributes that were not stipulated in OLACMS 0.4, if found in DC and motivated by actual need to describe language resources, can be easily adopted. One way to ensure the versatility is to keep all DC attribute in OLACMS, even though some of the attributes may be dormant and not actively applied now. Another issue worth noting is that any cataloguer may add sub-elements to achive more comprehensive description. However, such addition should, follow the extension and adaptability of the DC.

3 Use Controlled Vocabulary for Temporal and Geographic Location

Constable, and Simons (2000) listed all the causes for language changes, which basically involve the change in the temporal-spatial location of the poeple. Since China used a different calendar system until late in early 20th century, all inherent temporal description of inherited Chinese archives do not conform to the current DC standard. In order to identify Western and Chinese chronology, we may stipulate that the primary types of the scheme element to be Western (W_Calendar) or Chinese (C_Calendar). We may also add other chronological methods, such as lunar or solar calendar. The sub_type of Chinese calendar will then include time, dynasty name, state name, emperor's reign, and the reign name of the emperor. Take the Academia Sinica Ancient Chinese Corpus for example. Its coverage is Early Mandarin Chinese, and will marked as such in the metadata: <Coverage scheme="C_calendar/phase">EarlyMandarin </Coverage>. The users will be able to refer to a historical linguistic calendar and find that the time equals to the dynasties of Yuan, Ming, and Ching. And will be able to convert the time to western calendar using the conversion table of [Sinica Calendar].

When Coverage has a spatial refinement, a location can have different names because of the unit used in cataloguing, as well as because of temporal or regional and linguistic variaions. Hence, the spatial value of Coverage must be defined by a scheme. A scheme must stipulate temporal reference as unit of catalogue. For instance, the Sinica Corpus covers the language of the Republica of China in Taiwan. Its metadata will have the following value <Coverage refine= "spatial" scheme= "ROC/Taiwan">. As mentioned above [Sinica Calendar] offers conversion table for the past 2000 years between Chinese and Western calendars. As for the units for cataloguing of spatial location, OLAC 0.4 Version adopts [TGN](Getty Thesaurus of Geographical Terms). And many other digital archives follow Alexandria Digital Library Feature Type Thesaurus [ADL]. The ADL type thesaurus have been adopted by the digital archives project in Taiwan and translated into Chinese by Academia Sinica Metadata Architecture and Application Team [Sinica MAAT].

4 Applying OLACMS to Language Archives in Taiwan

Each text in Academia Sinica Balanced Corpus of Modern Chinese (Sinica Corpus) is marked up with five textual parameters: Mode, Genre, Style, Topic and Medium. These are important textual information that needs to be catalogued in metadata. The following shows how we transfer and represent these (legacy) textual information to OLACMS:

4.1.Mode and Genre

Table 1 The relation between Mode and Genre of Sinica Corpus(Ckip Technology Report 93-05)

Mode	Genre
Written	Reportages Commentary Advertisement Letters Announcement Fiction Prose Biography & Diary Poem Manual
written-to-be-spoken	Script Speech
Spoken	Conversation
spoken-to-be-written	analects Speech Meeting Minute

We add a refine attribute under Type. Mode is added in the controlled vocabulary as Primary type, and Genre is added as sub type. For instance, a recorded and transcribed speech is catalogued as <Type code="Sound" refine="spoken-to-be-written/Speech"/>.

4.2.Style

There are four styles that are differentiated in Sinica Corpus: Narrative, Argumentative, Expository, and Descriptive. We add a new refine attirbute under Descriptio, with Style as a controlled vocabulary. For instance, a diary will be catalogued as: <Description refine="Style"> Narration </Description>.

4.3.Medium

Sinica Corpus specifies the media of the language reources as: Newspaper, General Magazine, Academic Journal, Textbook, Reference Book, Thesis, General Book, Audio/Visual Medium, Conversation/Interview. We may also add other audio-video media such as CD,V8…etc. As mentioned above, this can be easily described with DC: Format refine attribute of Medium.

4.4.Topic

The Topic parameter of Sinica Corpus has the same content as the element Subject. This can simply be transferred through a table.

Table 2 Topic of Sinica Corpus(Ckip Technology Report 93-05)

Primary	Sub
Philosophy	Thoughts | Psychology | Religion |
Natural Science	Mathematics | Astronomy | Physics | Chemical | Mineral | Creature | Agriculture | Archeology | Geography | Environmental Protection | Earch Science | Engineering |
Social Sciences	Economy | Finance | Business & Management | Marketing | Politics | Political Party | Political Activities | National Policy | International Relations | Domestic Affairs | Military |Judicature | Education | Transportation | Culture | History | Race | Language | MassMedia | Public Welfare | Welfare | Personnel Matters | Statistical Survey | Crime | Calamity | Sociological Facts |
Arts	Music | Dance | Sculp | Painting | Photography | Drama | Artistry | Historical Relics | Architecture | General Arts |
General /Leisure	Travels | Sport | Foods | Medical Treatment | Hygine | Clothes | Movie and popular arts | People | Information | Cunsume | Family |
Literature	Literary Theory | Criticism | Other literary work | Indigenous Literature | Childern's Literature | Martial Arts Literature | Romance |

An example for the adoptation follows: for a Sinica Corpus text with a Topic of Arts and a sub-topic of Music, it will be catalgued as follows: <Subject>Arts/Music</Subject>.

4.5.Additional Controlled Vocabulary

- Proofreader: Since both manually and automatically digitized materials must be proofread to ensure quality, we suggest that [OLAC-Role] be enriched by a new value: Proofreader. For inherited texts with no IP restrictions, this may be the critical information piece of information to identify who is the rightful owner/creator of the electronic version.
- There are many Medium values old (procelain, rubbing, bamboo engraving, silk scroll, etc.) and new (DVD, MO, ZIP...etc). Hence the controlled vocabulary of attributes such as Medium and SourceCode often has quick and drastic changes. In order to maintain versatility and comprehensive coverage, this set of controlled vocabulary must be open and allows each participant to register, subject to the approval by OLAC.

5 Language Identification

Constable and Simons (2000) noted that a computer, unlike human beings, cannot automatically identify the language of a text that it is reading yet. Hence metadata must play a central role in identifying the language that each resource uses. For instance, Malay and English uses the same 26 letters. And Archaic Chinese 2000 years ago and Modern Mandarin can be expressed by pretty much the same set of Chinese characters. These are all different languages and need to be identified before a language resource can be used. SIL (Summer Institute of Linguistics, in its white-paper identified five major issues for language identification: Change, Categorization, Inadequate definition, Scale, and Documentation. SIL has produced an online searchable database: Ethnologue that provides a comprehensive system of language identification covering more than 6,800 languages. This is adopted by OLAC as an obvious improvement over the very small set covered in DC.

Bird et al. (2001), however, pointed out some problems of coverage if the Enthlogue system is adapted without further means of enrichment. The three broad categories of problem are: over-splitting, over-chunking and omission. Over-splitting occurs when a language variety is treated as a distinct language. For example, Nataoran is given its own language code (AIS) even though the scholars at Academia Sinica consider it to be a dialect of Amis (ALV). Over-chunking occurs when two distinct languages are treated as dialects of a single language (there does not appear to be an example of this in the Ethnologue's treatment of Formosan languages). Omission occurs when a language is not listed. For example, two extinct languages, Luilang and Quaquat, are not listed in the Ethnologue. Another kind of omission problem occurs when the language is actually listed, but the name by which the archivist knows it is not listed, whether as a primary name or an alternate name. In such a case the archivist cannot make the match to assign the proper code. For instance, the language listed as Taroko (TRV) in the Ethnologue is known as Seediq by Academia Sinica; several of the alternate names listed by the Ethnologue are similar, but none matches exactly.

The above problems may prove to be a stumbling block for archives that attempt to integrate linguistic resources with GIS (Geographic Information System), such as the [Formosan Language Archive] at Academia Sinica. A GIS-based language atlas will most likely be very concerned with fine-grained changes and variations among languages and dialects within a geographic area. In other words, these kind of archives may either discover yet unrecorded language or sub-language differentiations or need even finer classification in Ethnologue or any language identification system. Hence the solution proposed in Bird et al. (2001) of allowing local language classification systems to register must be implemented under OLAC.

6 Conclusion

We looked at a couple of OLAC derived metadatasets, as well as applied OLAC version 0.4. to three different language archives in Taiwan. We proposed some suggestions for enriching of OLACMS based on the study. There are two general directions to bear in mind. First, as the number and complexity of language resources becomes higher and higher, the need to have a uniform standard or to easy access to the owner of each resource becomes even greater. Therefore, we envision that the element of Creator, Contributor etc. needs further elaboration, which may include practical information such as email addresses etc. Second, as the language archives get richer, the need to note language variation grows even bigger. Simple language identification of allotting a resource a unique language code is not enough. There will be great need to infer linguistic relations from these codes. Since it is impossible to build a complete reportiore of resources for all languages, it is very often that a resources from the closest related language must be borrowed. The representation of linguistic relations will be the next challenge of language identification.

References

I. Bibliography

Bird, S. 2000. ISLE: International Standards in Language Engineering Spoken Language Group, http://www.ldc.upenn.edu/sb/isle.html

Bird, S., G. Simons, and C.-R. Huang 2001. The Open Language Archives Community and Asian Language Resources, 6th Natural Language Processing Pacific Rim Symposium Post-Conference Workshop, Tokyo, Japan.

Broeder, D., P. Suihkonen, and P. Wittenburg. 2000. Developing a Standard for Meta-Descriptions of Multimedia Language Resources, Web-Based Language Documentation and Description workshop, Philadelphia, USA.

CKIP. 1993. An Introduction to Sinica Corpus. CKIP Technology Report 93-05. IIS, Academia Sinica.

Constable, P. and G. Simons. 2000. Language identification and IT: Addressing problems of linguistic diversity on a global scale, SIL Electronic Working Papers 2000-001.http://www.sil.org/silewp/2000/001/

EAGLES/ISLE. ISLE Meta Data Initiative, http://www.mpi.nl/world/ISLE/

Gibbon, D., Peters, W., and Wittenburg, P., 2001. Metadata Elements for Lexicon Descriptions, Version 1.0, MPI Nijmegen, http://www.mpi.nl/ISLE/documents/draft/ISLE_Lexicon_1.0.pdf

Holton, G. 2000. Metadata for Linguistic Documentation Archives, Web-Based Language Documentation and Description workshop, Philadelphia, USA.

IMDI Team. 2001. IMDI Metadata Elements for Session Descriptions, Version 2.5, MPI Nijmegen, http://www.mpi.nl/ISLE/documents/draft/ISLE_MetaData_2.5.pdf.

IMDI Team. 2001. Mapping IMDI Session Descriptions with OLAC, Version 1.04, MPI Nijmegen. http://www.mpi.nl/ISLE/documents/draft/IMDI%20to%20OLAC%20Mapping%201.04.pdf

IMDI Team. 2001. IMDI Metadata Elements for Catalogue Descriptions, Version 2.1, MPI Nijmegen, http://www.mpi.nl/ISLE/documents/draft/IMDI_Catalogue_2.1.pdf

Palmer, M. 2000. ISLE: International Standards for Language Engineering: A European/US joint project, http://www.cis.upenn.edu/~mpalmer/isle.kickoff.ppt

Wittenburg, P., D. Broeder, and B. Sloman. 2000. EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources, White Paper. LREC 2000 Workshop, Athens.

II. Websites

[OLAC] Open Language Archives Community, http://www.language-archives.org

[OLACMS] OLAC Metadata Set, http://www.language-archives.org/OLAC/olacms-20011022.html

[DCMI] Dublin Core Metadata Initiative, http://dublincore.org/

[DCMS] Dublin Core Element Set, Version 1.1 - Reference Description, http://dublincore.org/documents/dces/.

[DC-Q] Dublin Core Qualifiers. http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/

[ISLE] International Standards for Language Engineering, http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

[ELRA] European Language Resources Association, http://www.icp.grenet.fr/ELRA/

[LDC] Linguistic Data Consortium, http://morph.ldc.upenn.edu/

[Sinica Calendar] Western Calendar and Chinese Calendar Conversion Table of Academia Sinica Computing Centre.
http://www.sinica.edu.tw/~tdbproj/sinocal/luso.html.

[Academia Sinica Ancient Chinese Corpus] Academia Sinica Tagged Corpus of Early Mandarin Chinese, http://www.sinica.edu.tw/Early_Mandarin/

[TGN] Getty Thesaurus of Geographical Terms, http://www.getty.edu/research/tools/vocabulary/tgn/index.html

[ADL] Alexandria Digital Library Feature Type, http://alexandria.sdc.ucsb.edu/gazetteer/gaz_content_standard.html

[Sinica MAAT] Metadata Architecture and Application Team, http://www.sinica.edu.tw/~metadata/standard/place/ADL-element.htm

[Sinica Corpus] Academia Sinica Balanced Corpus of ModernChinese, http://www.sinica.edu.tw/SinicaCorpus/

[Ethnologue] http://www.ethnologue.com

[Formosan Language Archive] Academia Sinica Formosan Language Archive, http://www.ling.sinica.edu.tw/Formosan/