語言典藏語言開放典藏社群之雙語元素集及控制詞彙中文版

張如瑩
中央研究院語言所籌備處
115 台北市南港區研究院路二段 128 號
+886-2-27883799 ext.1562
ruyng@gate.sinica.edu.tw

黃居仁
中央研究院語言所籌備處
115 台北市南港區研究院路二段 128 號
+886-2-26523108
churen@gate.sinica.edu.tw

1 簡介

OLAC（Open Language Archives Community）於 2000 年 12 月的一個語言資源工作營中，由來自北美、南美、歐洲、非洲、中東、亞洲、澳洲的語言學家與軟體發展者所創，希望藉由下列步驟進行創造世界性語言資源的虛擬圖書館：(一)、針對語言資源數位典藏發展一致性的實踐指引。(二)、發展一網路上具有互通性且提供存取相關語言資源的儲存器和服務中心。OLAC 以 OAI(Open Archives Initiative)爲典藏架構的基礎，並針對語言資源的特質，以 Dublin Core15 個元素(elements)爲基礎，制定出了元素集－OLACMS（OLAC Metadata Set），2001 年 10 月推出 0.4 版本，包含：貢獻者/單位（Contributor）、涵蓋範圍（Coverage）、創造者（Creator）、日期（Date）、資源描述（Description）、資源格式（Format）、資源 cpu 格式（Format.cpu）、資源編碼格式（Format.encoding）、標誌語言（Format.markup）、作業系統需求（Format.os）、程式語言（Format.sourcecode）、資源識別碼（Identifier）、語言（Language）、出版者（Publisher）、關聯性（Relation）、權利管理（Rights）、來源（Source）、主題（Subject）、主題使用語言（Subject.language）、資源標題（Title）、資源型態（Type）、軟體資源的功能（Type.functionality）、語言學上的資源型態（Type.linguistic）。OLACMS 可經由 XML 的 DTD 或 Schema 編碼加以驗證，它是經由 OAI (Open Archives Initiative)與 Dublin Core 元素集的搭配使用得以實踐之。

OLAC 後設資料用四個屬性：refine, code, lang, scheme 對 Dublin Core 提出特定特性進一步定義。還有另一個屬性爲 langs，再以控制詞彙（Controlled Vocabulary）規範屬性的值。(一)、refine：用來識別元素較精細的意義或更多的特定特性。(二)、code：用來標記元素後設資料中 OLAC 特有的某些標誌系統，其中所使用的控制詞彙及其定義已在其他 OLAC 正式文件中進行定義，而所有的服務中心皆可編譯一致且具有意義的 code 的內容值。每個 code 的控制詞彙在 OLAC 元素中經 XML Schema 合併，且皆是由 OLAC 註冊中心認證，符合 OLAC 標準測試過，可到 OLAC 網站（OLAC, http://www.language-archives.org）獲取相關資料。(三)、scheme：元素內容文字是已經標準化的名稱，也就是說元素內容所採用的是屬於標準的編碼，也許是用控制詞彙，也許是用正規註解(Formal Notation)，也許是 Dublin Core Metadata Initiative 已註冊

的 scheme 或者是由 OLAC 小組自行註冊相關的 scheme。（四）、lang：每個 OLAC 後設資料中必有此屬性，規範元素內容(element content)所使用的語言，預設值爲 en 英文 English。當資料具有多語言的特性，則新增以多個元素搭配屬性值加以表達，如此一來若是多語的後設資料，使用者可選擇用其中一種語言來解讀。（五）、langs：屬於 OLACMS 根元素<olac>元素的屬性，規範閱讀該後設資料時的語言。[1]

本研究配合 OLAC 的版本更新與國際接軌上的變動、因應中國語料資料的特性以及以不同媒材著錄資料時所產生的特性，針對 OLAC 提出修改建議以及中文版雛型。

2 與國際標準對應

2.1 與 IMDI 對應

ISLE IMDI（ISLE Meta Data Initiative）和 OLAC 一樣同屬於 ISLE（The International Standards for Language Engineering）專案所贊助的計畫[2][3]，其目的是要在歐洲研擬的自然語言處理標準（EAGLES；http://www.ilc.pi.cnr.it/EAGLES96/home.html）的基礎上，發展出計算語言學資源與程序的真正國際標準。

IMDI 是由先前的 EAGLES 後設資料集發展成形的[4]。EAGLES/ISLE 專案的目標爲促進語言資源在網際網路上的擷取和獲取。爲達此目標，如同全球網際網路上其他類似的語言資源協會一樣，他們提出藉由創造一個全球性可瀏覽、可搜尋的後設描述（Meta-description），對該資源進行典藏。當全球各地越來越多搭配著適當後設資料的語言資源被創造，使得公司或研究者較容易挖掘到她們所需要的。IMDI 於 2001 年 6 月提出 IMDI 集會描述（Session Descriptions）後設資料元素 2.5 版本，。[5]IMDI 工作小組並依其爲基礎，處理像 ELRA [6]和 LDC[7] 此類“已出版的語料庫”（published corpora），於 2001 年 6 月發布 IMDI 編目描述後設資料 2.1 版[8]，並 2001 年 12 月由 Gibbon 等人公佈詞彙的後設資料元素[9]。

IMDI 工作小組在 2001 年 8 月(August 2001)已經針對 IMDI 集會描述及 OLACMS 0.3 版進行對應[10]，但由於 OLAC 已經修正過，加上編目（Catalogue）和詞彙（Lexicon）後設資料的出現，本研究將進行重新對應。

2.1.1 控制詞彙

根據 OLACMS 與 IMDI 比所提出控制詞彙的建議如下：

- 語言資源的邏輯結構：語言資源可能以不同單位組成。而每個單位內的各分子組成訊息又是相當重要的資訊。例如：英文詞彙庫是由 26 個英文字母所組成，字母呈現的順序有其一定的隱藏涵義、章節的順序以及對話的聲調都是相當重要的訊息，所以未來必須在資料庫中表達該訊息。
- 註解者（Annotator）新增至[OLAC-Role]。必須強調並非由自然人或是由程式產生註解，而是由執行單位所進行的產生的，註解者的訊息可提供：註解所採用的標記集訊息，如：Penn Treebank, LOB。另外，註解者所隱藏的含意是指該

語言資源有了新版本，或者是有了新的語言資源。

- 著錄時的品質（Quality）在 Format 的 refine 之控制詞彙中增加 Quality 的值。

2.1.2 元素

對 OLCAMS 的元素變更建議如下：

我們發現許多語言資料往往是由一個專案計畫所執行或支援，當只填入 Project 名稱，則直接將填如 Creator 或 Contributor。但是有關 Project 的詳細描述，例如：創建者、協助單位…等，如同傘狀結構般的組織，諸如：EAGLES,ISLE 還有更複雜層次的 ESPRIT 等，必須由各子協會依照需求自行發展子元素加以描述著錄。

2.1.3 屬性更新

語言資源的涵蓋範圍（Coverage）通常需要紀錄經緯度，所以綜合上數與國外情況比對結果，在空間上加上子型態（sub type），內容涵蓋洲（Continent）、國家（Country）、行政區域（Administrative Division）、經緯度（Longitude and Latitude）、地址（Address）…等。

若識別碼（Identifier）既非國際標準，也不是 URL，而是由典藏單位自行定義，可在 scheme 內加入著錄單位簡稱輔助，例如：一般圖書館都會擁有各自館藏內的索書號，又例如：美國國會圖書館分類法（Library of Congress Classification）的 LCC 分類號，則可以<Identifier scheme=LCC> LCC 分類號</Identifier>標著。其他未出現在 OLACMS 中卻十分著名的辨識碼，如：國際標準期刊號 ISSN(International Standard- Serial Number)。

OLAC:Format 雖沒有 refine 的屬性，但是 DC:Format 有，且具有兩個控制詞彙：媒材（Medium）和長度（Extent）。Medium 記載著錄時所使用的媒材，Extent 紀錄尺寸和持續期間（Size and Duration），皆適合於描述語言資源，所以建議 OLAC 仍沿用該項屬性。

2.2 與典藏語言文件的對應

除了 MIDI Metadata 之外，Gary Holton (2000) [11]提出典藏語言文件資源描述的後設資料系統。該系統目前描述的實驗資料是美國阿拉斯加本土語言中心（Alaska Native Language Center, ANLC）典藏的語言文件。

我們將其所列出的元素與 OLACMS 比對，發現 Gary Holton 希望針對非數位化的資料格式，包括：手稿（Manuscript）、開盤式錄音機（reel-to-reel）、卡帶式（Cassette）、CD 唱片（CD recording）加以著錄，實則可用 DC 中 Format 原有的 refine 值－Medium 去概括。另外，增加演說者（Speaker）、面談者（Interviewer）、所有者（Holder）、管理者（Guardian）於 Creator 和 Contributor 的 refine 控制詞彙中。但文中所提的是目標同源語（Target Dialect）則尚未發覺有適當方式轉換。

2.3 小結

從上述比對情況可知，沿用原來 DC 中的修飾詞（Qualifiers）[12]，除

了可解決以 DC 爲基礎延伸出的後設資料之間銜接的問題，並具有國際性，也可解決 OLAC 0.4 版中缺乏某些屬性而造成無法對照的問題，所以建議最好保留原來 DC 裡頭已經擁有的屬性。另外，若是著錄單位須針對某元素更詳盡的描述往往必須由各子協會依照需求，在仍符合 DC 的精神下發展出子元素，但且記注意彈性與延伸性的重要性。

3 時空範圍的控制詞彙

語言會因爲時空轉變而產生變化，由於中國地區的涵蓋範圍在現代之前的紀年方式的不同。爲了辨別西元年或是中國年或是其他方式的紀年，scheme 可限定主型態（primary type）爲西元(E_Calendar)或中國曆法(C_Calendar)或其他曆法，諸如：陰曆（Lunar）和陽曆（Solar Calendar）。其中，中國曆法的子型態則包含：時期（Era）、朝代（Dynasty name）、國號（State name）、帝號（Emperor's reign）、年號（Reign's name）。例如：中央研究院近代漢語標記語料庫（簡稱近代漢語語料庫，Academia Sinica1 Ancient Chinese Corpus）[13]語料所涵蓋的範圍爲近代，則標示方法如下：<Coverage scheme="C_calendar/phase">EarlyMandarin </Coverage>或<涵蓋範圍 scheme="中國曆法/時期">近代</涵蓋範圍>，則使用者可透過時間轉換表[14]搜尋到朝代範圍涵蓋了元（Yuan）、明（Ming）、清（Ching）三個屬於近代的資料。

當 Coverage 的 refine 是屬於空間上，則由於不同地區的名稱，除了著錄單位不同，也會因爲時間的影響在同時期或朝代而擁有不同的名稱，所以整體而言 Coverage 屬於空間的控制值時，他必須搭配 scheme 使用。scheme 必須規範其著錄的方式，必須是：時期或者朝代/空間著錄單位/，譬如：著錄中央研究院現代漢語平衡語料庫（現代漢語平衡語料庫，Sinica Corpus）[15]涵蓋範圍是屬於中華民國時代的中國地區，則著錄方式爲 <Coverage refine= "spatial" scheme= "ROC/Taiwan">或<涵蓋範圍 refine="空間" scheme="民國/中國">，其中國的朝代可參考中央研究院計算中心兩千年中西曆轉換系統[14]。空間著錄單位可參考 OLAC 0.4 版所提供的 TGN(Getty Thesaurus of Geographical Terms)[16]或者參考 ADL（Alexandria Digital Library Feature Type Thesaurus）[17]，而中研院的數位典藏國家型計劃後設資料工作組 MAAT 已經於日前將 ADL 翻譯成中文版本[18]。

4 將台灣語言套用於 OLACMS

在現代漢語平衡語料庫中有語式（Mode）、文類（Genre）、文體（Style）、主題（Topic）和媒體（Medium）需要著錄[19]。各項的資料以及當套用在 OLACMS 時的使用方法如下：

4.1 語式及文類

表 1 現代漢語平衡語料庫中語式及文類間的關係

語式	Mode	文類	Genre
書面語	written	報導	Reportages
		評論	Commentary

		廣告或圖文	Advertisement
		信函	Letters
		公告啓事	Announcement
		小說故事寓言	Fiction
		散文	Prose
		傳記日記	Biography & Diary
		詩歌	Poem
		說明手冊	Manual
演講稿/劇本/腳本	written-to-be-spoken	劇本	Script
		演講	Speech
口語	Spoken	會話	Conversation
正式演說紀錄	spoken-to-be-written	語錄	Analects
		演講	Speech
		會議記錄	Meeting Minute

Type 新增 refine 屬性，語式（Mode）爲控制詞彙的主要型態（Primary type），文類（Genre）爲子型態（sub type）。例如：文本資料，原本是一場演講（Speech）所紀錄下來爲文字資料的，則以<資源型態 code="聲音" refine="口語紀錄/演講" lang="x-sil-CHN"/>著錄。

4.2 文體

現代漢語平衡語料庫所規劃的文體（Style）包含：記敘（Narration）、論說（Argumentation）、說明（Exposition）、描寫（Describe）。資源描述（Description）新增屬性 refine。其中有一控制詞彙爲文體（Style），例如：著錄的文章內容爲日記，屬於記敘文，則著錄方式爲<資源描述 lang="x-sil-CHN">記敘</資源描述>。

4.3 媒體

現代漢語語料庫中包含的媒體（Medium）有：報紙（Newspaper）、一般雜誌（General Magazine）、學術期刊（Academic Journal）、教科書（Textbook）、工具書（Reference Book）、學術論著（Thesis）、一般圖書（General Book）、視聽媒體（Audio/Visual Medium）、會話訪談（Conversation/Interview）、其他（Elsewhere）。如同以 CD,V8…等紀錄是一樣的，可用 DC 中 Format 原有的 refine 值－Medium 去概括。

4.4 主題

在現代現語語料庫中的主題（Topic）就是 Subject 的元素內容。現代漢語語料庫的主題內容與層級關係如下表 2。若針對文章主題爲藝術（Arts）之下的音樂（Music）則著錄方式爲<主題 lang="x-sil-CHN">藝術/音樂</主題>。

表 2 現代漢語平衡語料庫的主題及子主題

主題	子主題
哲學（Philosophy）	思想（Thoughts）、心理（Psychology）、宗教（Religion）
科學（Natural Science）	數學（Mathematics）、天文（Astronomy）、物理（Physics）、化學（Chemical）、礦冶（Mineral）、生物（Creature）、農漁牧業（Agriculture）、考古（Archeology）、地理

	（Geography）、環保（Environmental Protection）、大學科學（Earch Science）、工程（Engineering）
社會（Social Sciences）	經濟（Economy）、財政（Finance）、商管（Business & Management）、行銷（Marketing）、政治學（Politics）、政黨（Political Party）、政治現象（Political Activities）、國家政策（National Policy）、國際關係（International Relations）、內政（Domestic Affairs）、軍事（Military）、司法（Judicature）、教育（Education）、交通運輸（Transportation）、文化（Culture）、歷史（History）、民族（Race）、語文（Language）、傳播（MassMedia）、公益（Public Welfare）、福利（Welfare）、人事（Personnel Matters）、統計調查（Statistical Survey）、犯罪（Crime）、災禍（Calamity）、社會現象（Sociological Facts）|
藝術（Arts）	音樂（Music）、舞蹈（Dance）、雕塑（Sculp）、美術（Painting）、攝影（Photography）、戲（Drama）、技藝（Artistry）、文物（Historical Relics）、建築（Architecture）、藝術總論（General Arts）
生活（General/Leisure）	旅遊（Travels）、體育（Sport）、食物（Foods）、醫療（Medical Treatment）、衛生保健（Hygine）、衣飾（Clothes）、影藝（Movie and popular arts）、人物（People）、訊息（Information）、消費（Consume）、家庭（Family）
文學（Literature）	文學理論（Literary Theory）、批評與鑑賞（Criticism）、其他文學創作（Other literary work）、鄉土文學（Indigenous Literature）、兒童文學（Childern's Literature）、俠義文學（Martial Arts Literature）、言情文學（Romance）

4.5 其他控制詞彙

由於語言資源的資料可能有謬誤，必須進行校對，所以[OLAC-Role]控制詞彙必須再加上校對者（Proofreader），尤其對於資源使用情況沒有身分限制時將可成爲辨識 Creator 和 Owner 的有益資訊。

此外，Medium 還包括中國古代特殊媒材，如：瓷器（Porcelain）、拓片（Rubbing）、簡牘（Bamboo engraving）、娟繡（Silk）、畫軸（Scroll）...等，新興媒材則包括：DVD, MO, ZIP...等，所以像 Medium 和 SourceCode 這種控制詞彙變化巨大且迅速的，爲了具有一定的彈性，最好爲開放式控制詞彙（open controlled vocabulary）以供各協會自行加入適當詞彙或者由一統一註冊機構提供服務。

5 語言識別

Constable 等人 [20]提到由於電腦不似人腦可自動判斷使用的詞彙所應用的語言，所以在後設資料中必須規範所描述的語言資源以及元素內容所採用的語言，因此語言識別便佔有舉足輕重的地位。舉例說明，先秦漢語

資料與現代白話、或北京土話與台灣中南部的台灣國語，雖說是同一「語言」，但絕不能用同樣的詞庫、文法、或概念索引典處理。然而諸如 SILE 有關語言識別與資訊科技的白皮書提到，在定義全球性規模的語言識別時產生的五大議題：變化（Change）、目錄化（Categorization）、不適當的定義（Inadequate definition）、規模不足（Scale）以及缺乏完整的文件說明（Documentation），即使 SIL（Summer Institute of Linguistics）[21]經過專家學者共同研討定義，成立 Ethnologue（http://www.ethnologue.com/）[22]目前已針對全球 6800 種語言提供線上語言識別查詢的功能。

但 Bird 等(2001)[23]提到三種語言進行目錄化最廣泛的問題（也是 Ethnologue 引起相關學者爭議的原因）：過於分散零碎（over-splitting）、過於厚重（over-chunking）以及遺漏（omission）。過於分散零碎是因語言變化卻被視爲另一種語言，例如：Nataoran 的語言代碼是 AIS，但中研院的學者齊麗莎小姐視爲 ALV(Amis 阿美語)。過於厚重，則是將兩種有區別的語言視爲某一語言的同源語。遺漏，並未列出某一種語言，情況一：絕種的語言雷朗（Luilang）並未被列在 Ethnologue。情況二：已經被列出，但 Ethnologue 無適當的代碼，例如：Taroko（TRV），但中研院則視爲賽德克（Seediq）。對於試圖以 GIS(Geographic Information System)方式去呈現語言資料，並重視同源語分類的語言典藏－台灣南島語數位典藏[24]將是一個很嚴重的絆腳石。除了南島語分類過於粗略，也未包含一般臺灣地區客家話、閔南語語音上的變化，例如：一般人所知的苗栗客家人使用的是四縣腔調。新竹用海陸腔、桃園市海陸與四縣皆通，較特殊的則是新竹竹北到芎林的部分客家人，他們說的是少見的饒平客家話[25]。日前 Simons[26]已提出由各使用者自行應用 Ethnologue 的語言識別碼資料、提出修改建議和修改的標準程序，以及語言或同源語之間如何區分定義的根據，爲了解決台灣地區各種語言套用在後設資料時所產生的問題，接下來必須依據該標準確認包含同源語的分類，並回饋於 Ethnologue，以便於之後後設資料的應用。

6 結論

本研究針對 OLAC 藉由國外其他標準比對以及語言典藏計畫中三種不同的典藏資料對 OLACMS 提出初步修改建議以及中文版本雛型。未來當後設資料隨著所考量的資料型態越來越複雜以及內容標記（Content markup）的加入，配合語意、語法上的考量，爲了維持一致性，勢必有更多需要衡量的，而若需要著錄更詳細的資料，目前 OLAC 未有適當欄位可供填入，各子協會必須再發展適當的子元素。另外，未來必須朝著依據語言分類標準針對台灣地區的語言加以定義的方向著手。

參考文獻

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

[2] Martha Palmer, ISLE: International Standards for Language Engineering: A European/US joint project, http://www.cis.upenn.edu/~mpalmer/isle.kickoff.ppt <2000>

[3] EAGLES/ISLE ISLE Meta Data

Initiative, http://www.mpi.nl/world/ISLE/
[4] Wittenburg, P., Broeder, D., and Sloman, B., EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources, White Paper. LREC 2000 Workshop, Athens, 2000.
[5] IMDI Team, IMDI Metadata Elements for Session Descriptions, Version 2.5, MPI Nijmegen, http://www.mpi.nl/ISLE/documents/draft/ISLE_MetaData_2.5.pdf. <June 2001>
[6] European Language Resources Association, http://www.icp.grenet.fr/ELRA/
[7] Linguistic Data Consortium, http://morph.ldc.upenn.edu/
[8] IMDI Team, IMDI Metadata Elements for Catalogue Descriptions, Version 2.1, MPI Nijmegen, http://www.mpi.nl/ISLE/documents/draft/IMDI_Catalogue_2.1.pdf, <June 2001>.
[9] Gibbon, D., Peters, W., Wittenburg, P., Metadata Elements for Lexicon Descriptions, Version 1.0, MPI Nijmegen, http://www.mpi.nl/ISLE/documents/draft/ISLE_Lexicon_1.0.pdf,<December 2001>.
[10] IMDI Team, 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 , <August 2001>.
[11] Gary Holton, Metadata for Linguistic Documentation Archives, Web-Based Language Documentation and Description workshop, Philadelphia, USA, 2000.
[12] Dublin Core Qualifiers, http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/
[13] 中央研究院近代漢語標記語料庫, http://www.sinica.edu.tw/Early_Mandarin/
[14] 中央研究院計算中心兩千年中西曆轉換系統, http://www.sinica.edu.tw/~tdbproj/sinocal/luso.html.
[15] 中央研究院現代漢語平衡語料庫, http://www.sinica.edu.tw/SinicaCorpus/
[16] Getty Thesaurus of Geographical Terms, http://www.getty.edu/research/tools/vocabulary/tgn/index.html
[17] Alexandria Digital Library Feature Type, http://alexandria.sdc.ucsb.edu/gazetteer/gaz_content_standard.html
[18] 數位典藏國家型計劃後設資料工作組, http://www.sinica.edu.tw/~metadata/standard/place/ADL-element.htm
[19] 中央研究院資訊所詞庫小組技術報告 92-05,現代漢語平衡語料庫簡介.
[20] Constable, Peter and Gary F. Simons, 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/,<2001>.
[21] SIL Homepage, http://www.sil.org/
[22] Ethnologue, Languages of the World, http://www.ethnologue.com
[23] Steven Bird, Gary Simons, Chu-Ren Huang, The Open Language Archives Community and Asian Language Resources, 6th Natural Language Processing Pacific Rim Symposium Post-Conference Workshop, Tokyo, Japan, 2001.
[24] 台灣南島語數位典藏, http://www.ling.sinica.edu.tw/Formosan/
[25] 陳子祺，“新竹海陸腔客家話音韻研究”，國立新竹師範學院，臺灣語言與語文教育研究所碩士論文，2001年
[26] Gary F. Simons, SIL Three-letter Codes for Identifying Languages: Migrating from in-house standard to community standard .International Workshop on Resources and Tools in Field Linguistics (LREC 2002), Las Palmas, Canary ,Islands26-27 May 2002.