

主題計畫名稱 (中/英文)： 台灣南島語數位典藏 The Formosan Languages Digital Archive	計畫主持人：(姓名、電話、電子郵件) 齊莉莎 26523191 hsez@ccvax.sinica.edu.tw
主題計畫所屬單位： 中央研究院語言學研究所	
參與 Metadata 作業的主題計畫同仁：(姓名、電話、電子郵件) 邢天馨(27863300*324 / thsin@gate.sinica.edu.tw) 博士後研究 余清華(27863300*362 / harryu@gate.sinica.edu.tw) 工程師 華加婧(27863300*362 / taligu@gate.sinica.edu.tw) 助理	
Q1、是否希望將 Metadata 需求納入計畫內？ <input checked="" type="checkbox"/> 是 否 (原因：_____)	
Q2、是否希望中研院「Metadata 工作組」協助規劃與分析典藏品 Metadata？ <input checked="" type="checkbox"/> 是 (請至 Q3) 否 (原因：_____)	
Q3、預計時程與期限？ 期望在 8 月完成 Metadata 分析，在 12 月有系統雛形可進行著錄測試	
Q4、預期的需求重點與成果？(GIS system? Exhibition Management.....) 1. 南島語 metadata 規格及需求 2. 完成語料的基本資料庫雛形 3. 線上更新及轉檔 4. 南島語同源詞與非同源詞地理分佈之 GIS 查詢	
Q5、藏品資料類型與數量(若不同資料類型需不同的 Metadata 著錄格式，請註明，並標明計畫進行的優先順序) 資料類型：魯凱語/雅美語/鄒語/賽夏語/泰雅語/排灣語/布農語/阿美語/卑南語 數量(指方言)：魯凱語: 6 雅美語: 1 鄒語: 3 賽夏語: 2 泰雅語: 2 排灣語: 2 布農語: 5 阿美語: 5 卑南語: 2	
Q6、目前典藏品目錄資訊的狀態？ 6-1 使用的系統 系統名稱、鍵檔數量、資料庫系統相關文件(規格書, Schema...等) 6-2 是否使用任何 metadata 標準？ 6-3 是否有任何既有之著錄表單？(如：文物基本資料表、登記表...等) 6-4 是否有公開網站，可供查詢使用？ 6-5 是否有其他相關參考資訊(如：其他單位網站、著錄規範...等) 6-1 使用系統：使用 VB 語言撰寫的系統轉換工具，將語料轉換成資料庫檔案 鍵檔數量：魯凱語 萬山方言已收集 764 句 相關文件：RDBMS, XML, HTML 6-2 OLAC 6-3 無 6-4 http://www.ling.sinica.edu.tw/formosan 6-5 參考國外 OLAC http://www.language-archives.org SIL http://www.sil.org LACITO http://lacito.archivage.vjf.cnrs.fr/	
Q7、藏品數位化規格需求？(解析度、檔案類型、開放程度...等) 典藏文件多含 IPA 符號，故使用者系統須安裝支援 Unicode 的字型檔： 字型檔：Lucida Sans Unicode, Arial Unicode MS, Code2000 語音檔為 MP3 格式	
其他說明與建議：	

01 Metadata 需求確認表單

填表者：____余清華_____

02 Metadata 藏品元素需求表單

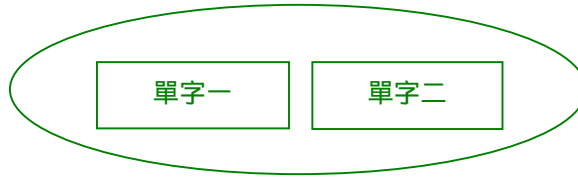
元素中文名稱		元素英文名稱		定義	著錄規範
Element	Subelement	Element	Subelement		
發音人	族語別		Language	本段語料發音人所說的語言所屬的語族	根據 XX 文獻的分類
	方言別		Dialect	本段語料發音人所說的語言所屬的方言別	
	姓名		Name	發音人的姓名	
	性別		Sex	發音人的性別	
	出生年月日		Birthdate	發音人的出生年月日	
	出生地		Birthplace	發音人的出生地	
	成長地		GrowingPlace	發音人的成長地	
	現址		CurrentAddress	發音人現在居住的地點	
	在現址居住年數		CurrentYears	發音人在現址所居住的年數	
	教育程度		Education	發音人的學歷背景	
文章	族語別		Language		
	方言別		Dialect		
	標題		Title		
	發音人		Informant		
	紀錄人		Transcriber		
	資料收集時間		Recording Time	西元年/月	
	校對編輯時間		Editing Time	西元年/月	
譯註時間		Glossing Time	西元年/月		
句子	族語別		Language		
	方言別		Dialect		
	文章別		Text	指章節、故事、談話等	
	行號		SentenceID		
	中文翻譯		Chinese translation		
	英文翻譯		English translation		
	聲音檔路徑		Soundpath	聲音檔的完整路徑	
單字	族語別		Language		
	方言別		Dialect		
	文章別		TextID		
	行號		SentenceID		
	字序		WordOrder	單字在句中的順序	
	單字拼寫		Orthog	母語的拼音	

	中文註解		Chinese gloss		
	英文註解		English gloss		
詞類	族語別		Language		
	方言別		Dialect		
	詞素		Morpheme		
	是否為附著 詞根		Bound morpheme		
	是否為自由 詞根		Free morpheme		
	詞根		Root		
	詞幹		Stem		
	註解		Gloss		
	標記		Tagging		

填表者：_____余清華_____

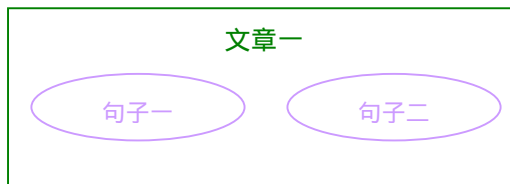
03 藏品單元(unit)層級關係圖

群組關係一：句子



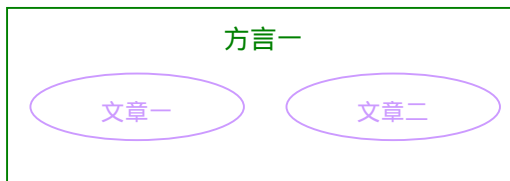
說明：最小之藏品單元(unit)為單字，兩單字或更多會組合為有意義之一組單字組(即為句子)，最小之藏品單元與句子皆各需著錄一筆紀錄。

群組關係二：文章



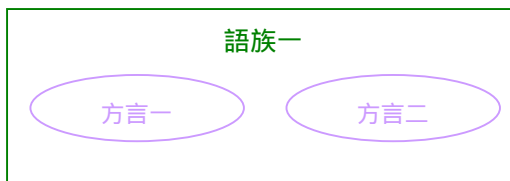
說明：最小之藏品單元(unit)—單字組成一個句子，若干句子進而組成一篇文章。

群組關係三：方言



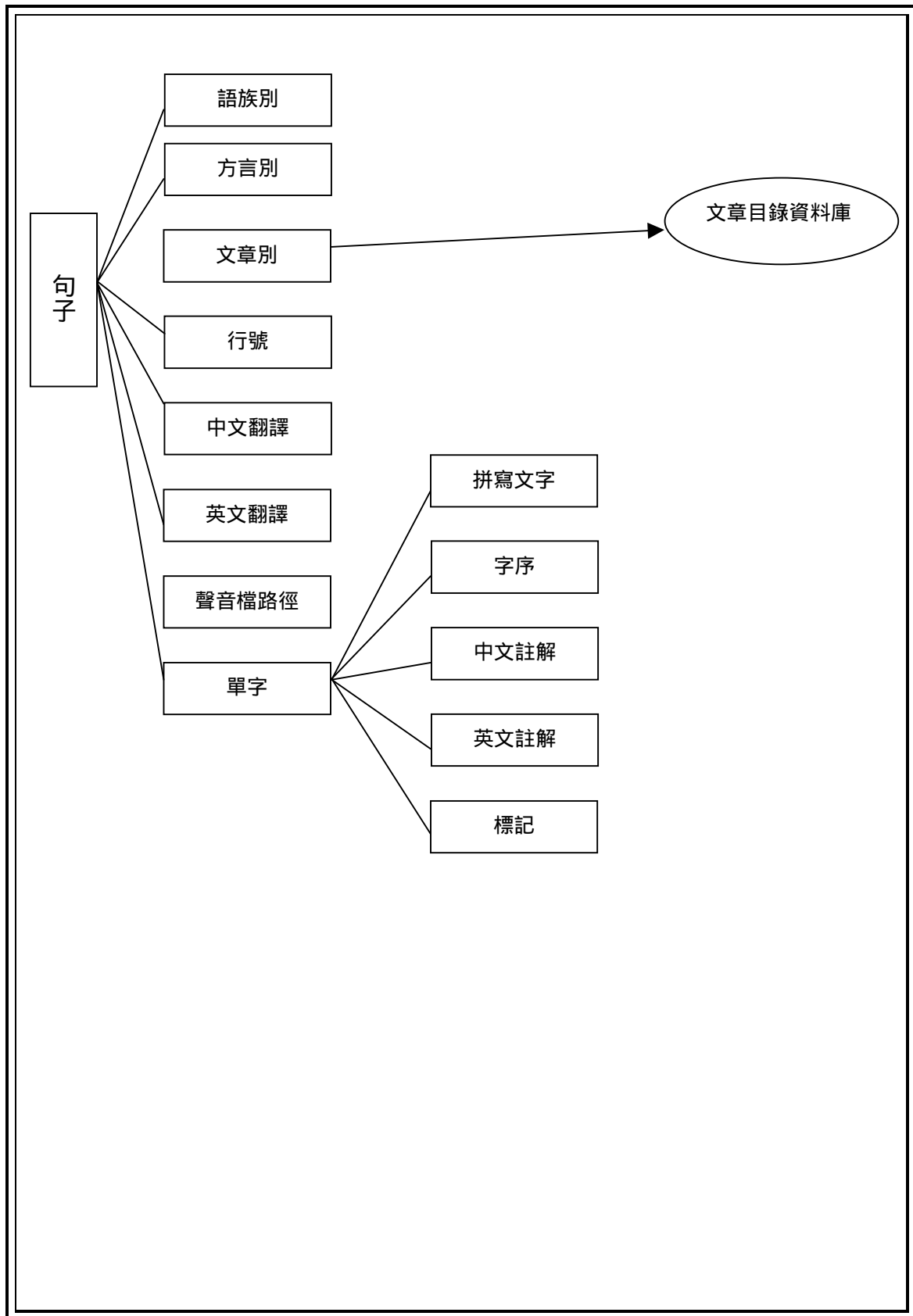
說明：一個方言將包含若干文章。

群組關係四：語言



說明：一個語族將包含若干方言。

04 元素關係結構圖



填表者：_____ 余清華 _____

06 Metadata 著錄範列表單

元素中文名稱		實例
Element	Subelement	
發音人	族語別	魯凱族
	方言別	萬山
	姓名	呂玉枝
	性別	女
	出生年月日	1924/7
	出生地	高雄現舊萬山
	成長地	高雄現舊萬山
	現址	高雄新萬山
	在現址居住年數	41
	教育程度	國小
文章	族語別	魯凱語
	方言別	萬山
	標題	我們的話
	發音人	呂玉枝
	紀錄人	齊莉莎、林惠娟
	資料收集時間	1992, 1997-1999
	校對編輯時間	1999-2001
	譯註時間	1992-2001
句子	族語別	魯凱語
	方言別	萬山
	文章別	我們的話
	行號	001a
	中文翻譯	我們的祖先萬山（自己是）萬山人。
	英文翻譯	Our ancestors used to call (themselves) Mantauran.
	聲音檔路徑	Mantauran/001a.mp3
	單字	魯凱族
單字	方言別	萬山
	文章別	我們的話
	行號	001a
	字序	0 (表示某單字在句中第一個位置)
	單字拼寫	onaʔi
	中文註解	那
	英文註解	that

填表者： 余清華

填表者：____余清華_____

08 Metadata 系統查詢功能需求表單

元素中文名稱		簡要查詢 (Simple search)	進階查詢 (Advanced search)	簡要顯示款 目	詳細顯示款目
Element	Subelement				
發音人	族語別	Y	Y	Y	Y
	方言別	Y	Y	Y	Y
	姓名	Y	Y	Y	Y
	性別				Y
	出生年月日				Y
	出生地				Y
	成長地				Y
	現址				Y
	在現址居住 年數				Y
	教育程度				Y
句子	族語別	Y	Y	Y	Y
	方言別	Y	Y	Y	Y
	文章別	Y	Y	Y	Y
	行號				Y
	中文翻譯		Y	Y	Y
	英文翻譯		Y	Y	Y
	聲音檔路徑			Y	Y
單字	族語別		Y		Y
	方言別		Y		Y
	文章別		Y		Y
	行號				Y
	字序				Y
	單字拼寫	Y	Y	Y	Y
	中文註解		Y	Y	Y
	英文註解		Y	Y	Y

填表者：____余清華_____

09 Metadata 系統紀錄建檔流程說明

著錄人員要進行紀錄建檔時，進入著錄介面之前需經過身份的確證，著錄人員必須輸入自己的姓名與密碼，方可登入 Metadata 系統進行建檔，系統需記錄建檔人姓名與建檔時間。同樣的，當著錄人員需修改紀錄內容時，也需經過身份的確證，藉由確證的動作可知該筆記錄是經由誰人所修改，系統亦需記錄修改人姓名與修改時間。系統不需記錄每次的修改資訊，只需記錄最近一次的修改資訊即可。

10 Metadata 系統主鍵元素架構格式

系統可將「句子行號」作為主鍵元素的一部份，其架構格式，如下：
兩個阿拉伯數字(文章代碼) + 三個阿拉伯數字(段落代碼) + 一個英文字元(段落中的句子代碼)
如：01001A

填表者：____余清華____