021004
後設資料工作組　翁翠霞

台灣南島語數位典藏計畫第二次分析報告

時間：2002 年 10 月 4 日 2 pm
地點：中研院語言所

大綱

一、欄位分析

二、語言典藏後設資料的標準

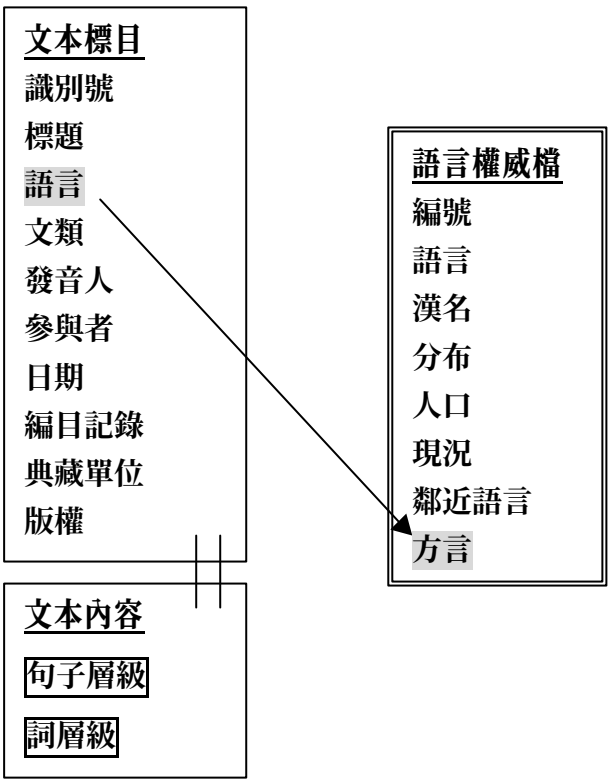
三、比對 OLAC, CGN

四、釋疑

一、欄位分析

- **後設資料工作組根據** 5 月 31 日**南島語計畫所提供的後設資料欄位需求表單和** 6 月 19 日**齊莉莎老師所提供的欄位內容，草擬出南島語計畫後設資料欄位。請參見附件一「『台灣南島語數位典藏』計畫元素需求表」。**

- **後設資料欄位結構特點如下：**
  1. **『標目』的最小描述單位：語料庫裡的一篇文本**
  2. **『文本』內欄位分成兩個層次：句子層級和詞層級**
  - **欄位結構圖如下：**



二、語言典藏後設資料的國際標準

1. 語言開放典藏設群後設資料群(Open Language Archives Community Metadata Set, OLACMS)

(1) 發展機構：語言開放典藏設群

(2) 網址：http://www.language-archives.org/

(3) 特點

- 公眾檢索導向的後設資料
- 以 OAI 為架構設計的概念；後設資料集則根據 DC 的元素參酌語言資源的特性增修而成
- 為語意最精簡，但描述資源最全面(包括：資料、工具和建議)的後設資料

(4) 最新版本：20011022 版

(5) 後設資料欄位：共 23 個

2. 文本編碼先導計畫(Text Encoding Initiative, TEI)

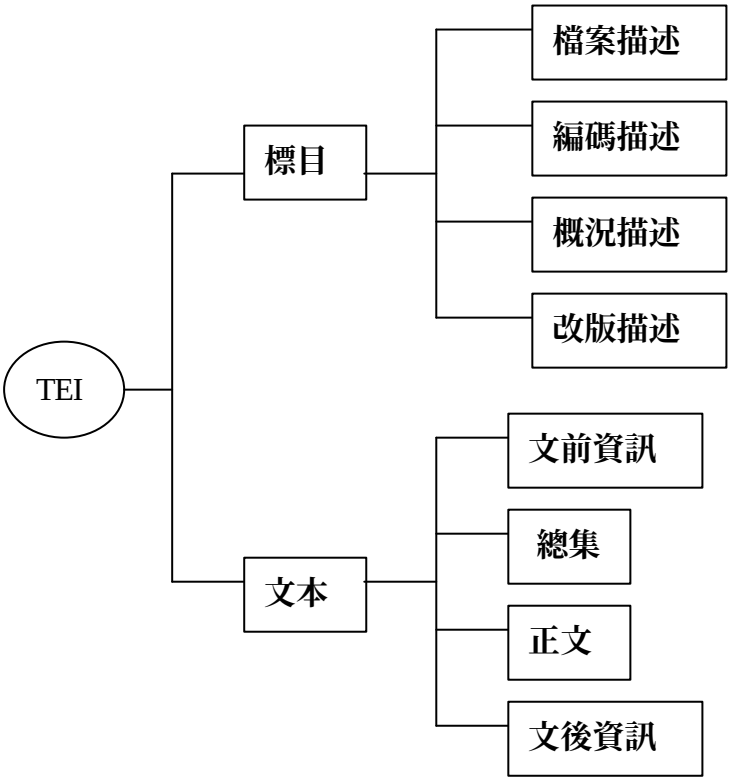
(1) 發展機構：TEI Consortium

(2) 網址：http://www.tei-c.org/

(3) 特點：

- 為電子文本的交換的編碼標準
- 初始以 SGML 為編碼語法，2002 年已發展出 XML DTD
- 欄位結構分成兩部分：標目(header)和文本(text)；其中標目部分可獨立存在，成為「獨立標目(the independent header)」作為交換的文件。整體欄位架構如下。
- 另出版 TEI 教學文件電子版—TEI Lite (TEI U5, 1995) (URL: http://www.tei-c.org/Lite/)，為 TEI 元素應用的入門子集

(4) 最新版本：TEI P4 XML 版本(2002)



TEI 欄位結構圖

3. 語料庫編碼標準(Corpus Encoding Standard, CES)

(1) 發展機構：Expert Advisory Group on Language Engineering Standards (EAGLES)

(2) 網址：http://www.cs.vassar.edu/CES/

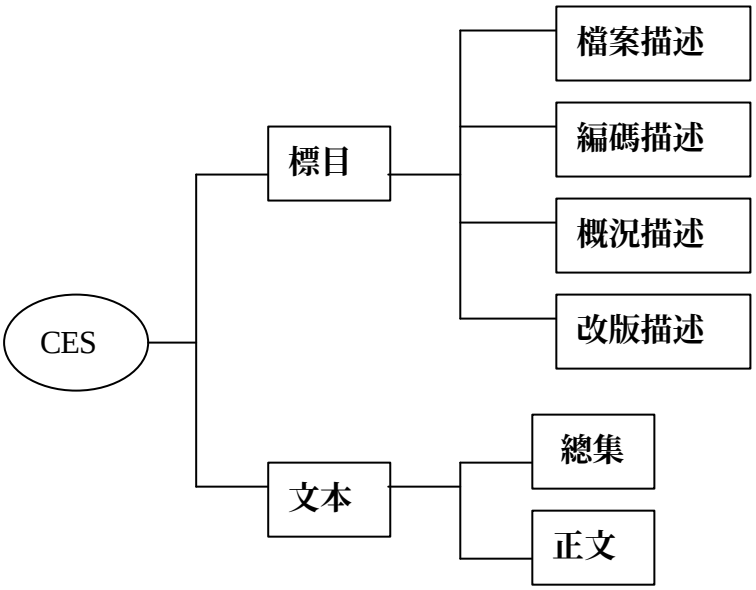
(3) 特點

- 設計目的是為自然語言處理所用到的語料庫而訂定的一套編碼標準
- 內容架構上符合 TEI 的指引手冊，但為了語料庫研究的需要對 TEI 的欄位有所調整，所以在文本層次的標碼目標著重在「語言學上的物件

(linguistic objects)」，例如言談(discourse)的最大單位，如段落，章節，以及語言分析上的基本物件，如句子、子句、片語、詞、詞素、音素等。

- 描述目標－電子文本
- 以 SGML 為編碼語法，但在 2002 年已制定出 XML DTD

(4) 欄位結構依據 TEI，分成標目(header)和文本(text)兩部份。



CES 欄位結構圖(欄位總表參見附件二)

4. 荷蘭口語語料庫(Spoken Dutch Corpus, CGN)

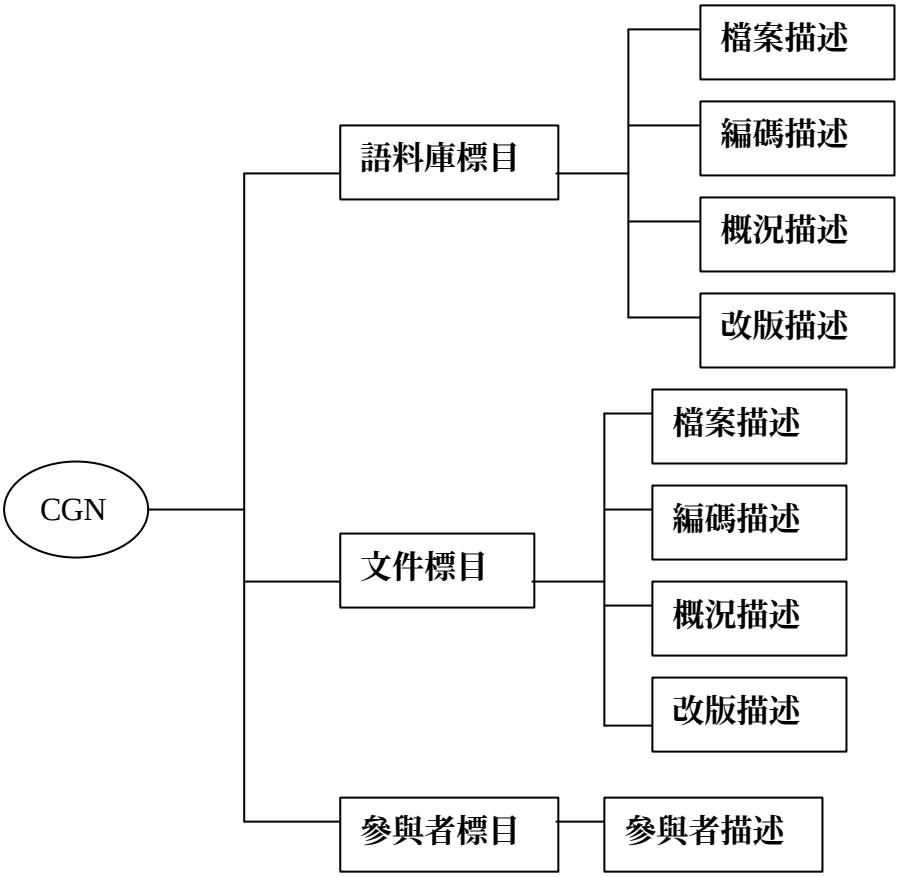
(1) 發展單位：荷蘭口語語料庫計畫(The Spoken Dutch Corpus Project)

(2) 網址：http://lands.let.kun.nl/cgn/ehome.htm

(3) 特點

- 針對荷蘭口語語料庫而建立的一套後設資料
- 標籤集(tagset)依照 EGALES 也就是 CES 的指引手冊，但仍作一些修改更動
- 描述目標－口語語料轉寫而成的電子文本

(4) 標目欄位結構如下：



CGN 後設資料欄位結構(欄位總表參見附件三)

5. 場次描述後設資料元素(Metadata Elements for Session Description)

(1) **發展單位：語言工程國際標準後設資料先導計畫**(International Standard for Language Engineering Metadata Initiative, IMDI)

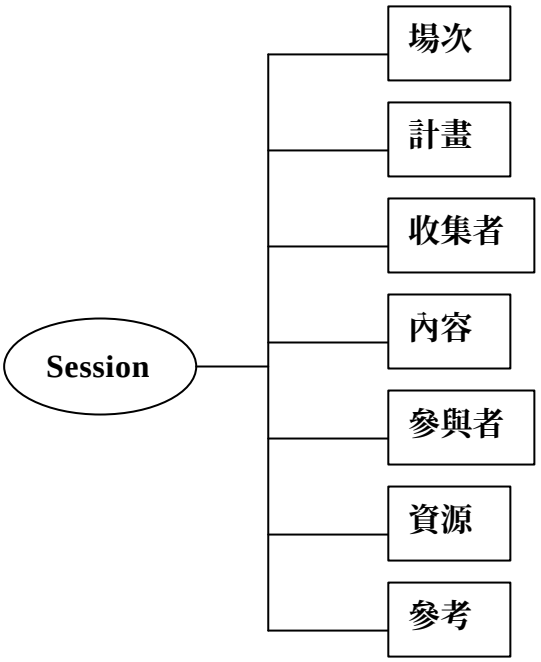
(2) **網址：**http://www.mpi.nl/world/ISLE/

(3) **特點**

- **針對多媒體／多模組**(multimedia/multimodal)**的語言資源而建立的後設資料標準**
- **以錄音或錄影時一個場次**(session)**為後設資料的最小描述單位**

(4) **最新版本：**2.5 **版**(2001 **年** 6 **月**)

(5) **主要欄位結構如下(欄位總表參見附件四)：**



三、欄位比對

- 台灣南島語數位典藏計畫所典藏的資料形態是從口語轉寫的電子文本，由於 IMDI 欄位所描述的最小單位是錄音／影的一個 session，所以並不符合。CES 是描述書面的電子文件，欠缺台灣南島語所要描述的發音人資料。因此我們選擇 OLAC 和 CGN 作為台灣南島語的文本標目比對標準。

1. 「文本標目」比對 OLAC

元素中文名稱			OLAC
Element	Subelement		
識別號			<identifier>識別號</identifier>
標題	原名		<title>標題</title>
	中文		<title refine="Alternative">中文</title>
	英文		<title refine="Alternative">英文</title>
語言	名稱		<subject.language code="名稱"/>
	漢名		<subject.language code="名稱">漢名</subject.language>
文類			<subject>文類</subject>
發音人	姓名	族名	<contributor refine="Informant">族名</contributor>
		漢名	<contributor refine="Informant">漢名</contributor>
	性別		<contributor refine="Informant">性別</contributor>

元素中文名稱			OLAC
Element	Subelement		
	出生年月日	年	<contributor refine="Informant">年</contributor>
		月	<contributor refine="Informant">月</contributor>
		日	<contributor refine="Informant">日</contributor>
	存歿		<contributor refine="Informant">存／歿</contributor>
	籍貫		<contributor refine="Informant">籍貫</contributor>
	語族		<contributor refine="Informant">語族</contributor>
	母語	類別	<contributor refine="Informant">類別</contributor>
		語言	<contributor refine="Informant">語言</contributor>
	語言能力	語言	<contributor refine="Informant">名稱</contributor>
		流利度	<contributor refine="Informant">流利度</contributor>
	出生地		<contributor refine="Informant">出生地</contributor>
	成長地		<contributor refine="Informant">成長地</contributor>
	居住地	現址	<contributor refine="Informant">現址</contributor>v
		居住年數	<contributor refine="Informant">居住年數</contributor>
	職業	現職	<contributor refine="Informant">現職</contributor>
		經歷	<contributor refine="Informant">經歷</contributor>
	教育程度		<contributor refine="Informant">教育程度</contributor>
參與者	類別		<creator refine="類別"/>
	姓名	原名	<creator refine="類別"/>原名</creator>
		他名	<creator refine="類別"/>他名</creator>
	職業	類別	<creator refine="類別"/>類別</creator>
		內容	<creator refine="類別"/>內容</creator>
日期	類別		<date refine="類別"/>
	起	年	<date refine="類別" code="年">起</date>
		月	<date refine="類別" code="月">起</date>
		日	<date refine="類別" code="日">起</date>
	迄	年	<date refine="類別" code="年">迄</date>
		月	<date refine="類別" code="月">迄</date>
		日	<date refine="類別" code="日">迄</date>
編目記錄	填表日期		<date refine="填表日期" code="填表日期"/>
	更新日期		<date refine="更新日期" code="更新日期"/>
	填表者		<creator refine="填表者">填表者</contributor>
典藏單位			<Contributor refine="Copyright claimant">典藏單位</right>
版權			<right code="版權">版權</right>

★　符合率：OLAC 基本上皆可以符合台灣南島語的欄位

★ **缺點：台灣南島語多個欄位常對應到 OLAC 一個欄位，OLAC 無法作更細緻的欄位劃分**

2. 「文本標目」比對 CGN

元素中文名稱			CGN Header
Element	Subelement		
識別號			
標題	原名		<Text Header><fileDesc><titleStmt><title lang=‘ 語言’ >原名 </title></titleStmt></fileDesc></Text Header>
	中文		<Text Header><fileDesc><titleStmt><title lang=‘ 中文’ >中文 </title></titleStmt></fileDesc></Text Header>
	英文		<Text Header><fileDesc><titleStmt><title lang=‘ 英文’ >英文 </title></titleStmt></fileDesc></Text Header>
語言	名稱		<Corpus Header><profileDesc><langUsage>名稱 </langUsage></profileDesc></Corpus Header>
	漢名		<Corpus Header><profileDesc><langUsage>漢名 </langUsage></profileDesc></Corpus Header>
文類			<Text Header><profileDesc><textClass>文類 </textClass></profileDesc></Text Header>
發音人	姓名	族名	<Participant Header><particDesc><person>族名 </person></particDesc></Participant Header>
		漢名	<Participant Header><particDesc><person>漢名 </person></particDesc></Participant Header>
	性別		<Participant Header><particDesc><person sex= ‘ 性別’ />
	出生年月日	年	<Participant Header><particDesc><birth year= ‘ 年’ />
		月	<Participant Header><particDesc><birth>月 </bitrh></particDesc></Participant Header>
		日	<Participant Header><particDesc><birth>日 </bitrh></particDesc></Participant Header>
	存歿		<Participant Header><particDesc><note>存歿 </note></particDesc></Participant Header>
	籍貫		<Participant Header><particDesc><note>籍貫 </note></particDesc></Participant Header>
	語族		<Participant Header><particDesc><language>語族 </language></particDesc></Participant Header>
	母語	類別	<Participant Header><particDesc><language/>

元素中文名稱			CGN Header
Element	Subelement		
		語言	<Participant Header><particDesc><language><firstLang>語言</firstLang></language></particDesc></Participant Header>
	語言能力	語言	
		流利度	
	出生地		<Participant Header><particDesc><birth place>出生地</bitrh></particDesc></Participant Header>
	成長地		
	居住地	現址	<Participant Header><particDesc><residence place>現址</residence></particDesc></Participant Header>
		居住年數	
	職業	現職	<Participant Header><particDesc><occupation job>現職</occupation></particDesc></Participant Header>
		經歷	
	教育程度		<Participant Header><particDesc><education opleiding>教育程度</education></particDesc></Participant Header>
參與者	類別		<Text Header><fileDesc><titleStmt><respStmt><respType>類別</respType></respStmt></titleStmt></fileDesc></Text Header>
	姓名	原名	<Text Header><fileDesc><titleStmt><respStmt><respName lang=‘語言’>原名</respName></respStmt></titleStmt></fileDesc></Text Header>
		他名	<Text Header><fileDesc><titleStmt><respStmt><respName lang=‘語言’>他名</respName></respStmt></titleStmt></fileDesc></Text Header>
	職業	類別	<Text Header><fileDesc><titleStmt><respStmt>類別</respStmt></titleStmt></fileDesc></Text Header>
		內容	<Text Header><fileDesc><titleStmt><respStmt>內容</respStmt></titleStmt></fileDesc></Text Header>
日期	類別		
	起	年	<Text Header><fileDesc><sourceDesc><rec date>年</rec date></sourceDesc></fileDesc></Text Header>
		月	<Text Header><fileDesc><sourceDesc><rec date>月</rec date></sourceDesc></fileDesc></Text Header>
		日	<Text Header><fileDesc><sourceDesc><rec date>日</rec date></sourceDesc></fileDesc></Text Header>
	迄	年	<Text Header><fileDesc><sourceDesc><rec date>年</rec date></sourceDesc></fileDesc></Text Header>

元素中文名稱			CGN Header
Element	Subelement		
		月	<Text Header><fileDesc><sourceDesc><rec date>月</rec date></sourceDesc></fileDesc></Text Header>
		日	<Text Header><fileDesc><sourceDesc><rec date>日</rec date></sourceDesc></fileDesc></Text Header>
編目記錄	填表日期		
	更新日期		<Text Header updated= ' 日期' >更新日期</Text Header>
	填表者		<Text Header creator>填表者</Text Header>
典藏單位			<Corpus Header><fileDesc><titleStmt><respStmt><respName>典藏單位</respName></respStmt></titleStmt></fileDesc></Corpus Header>
版權			<Corpus Header><fileDesc><publicationStmt><copyright>版權</copyright></publicationStmt></fileDesc></Corpus Header>

★ **符合率：**82%

★ **優點：**CGN **為說話者獨立建立一組標目欄位，可以滿足台灣南島語著重記錄發音人資料的需求**

★ **缺點：**(1) CGN **為荷蘭口語語料庫的後設資料，不是國際通用的後設資料標準**

(2) CGN **的欄位結構分成語料庫標目、文本標目和參與者標目三大部分，但台灣南島語的欄位結構僅為文本標目，發音人的資料是含在文本標目內，因此在比對上會出現台灣南島語的欄位分散在** CGN **的三個標目層次內。**

結論

到目前為止，語言典藏國際上的後設資料標準並不能和台灣南島語的標目欄位作很好的比對，除了 OLAC **外，建議可以採用** CGN/CES **作為較為特定語料庫的國際互通標準。**

四、釋疑

1. **是否文章內的每一個句子或每一個詞都要是後設資料的一筆記錄？(就「詞層級」來說，不同的文章可能會出現相同的詞彙，如此可能會有重複的記錄，建議可以把「詞層級」單獨建立「詞彙資料庫」。)**
2. **如果文章內的每一個句子或詞都須成為後設資料的一筆記錄，原有資料庫(台**

灣南島語數位典藏資料庫)是否已有現成的欄位可供連結？還是需要由著錄者自行輸入？

3. 在台灣南島語數位典藏資料庫裡，分有段落顯示的部份，並有聲音檔和中英文翻譯，因此後設資料欄位是否應增加「段落層級」的後設資料？