知識組織系統的多語詞彙語意對應分析

The Analysis of Mapping Multilingual Lexical Semantics for Knowledge Organization Systems

陳淑君
中央研究院資訊科技創新研究中心

城菁汝(代宣讀)

2010年9月6日至9月8日
第八屆（2010）兩岸三院資訊技術交流與資源共享研討會

前言

- 數位圖書館的重要特色之一，是可以跨越時間與空間，讓全球皆有機會近用(access)數位資源。
- 但是，如何跨越語言的障礙，讓不同語言的使用者可以檢索到典藏在世界各地的數位內容，是目前數位圖書館服務面臨的重要研究課題。

Art & Architecture Thesaurus (AAT) 中文化研究發展計畫

- 中央研究院2008年起開始嘗試，找出一套方法，可以落實數位典藏（數位圖書館）的多語化。特別是以控制詞彙(Controlled Vocabularies)/知識組織系統(Knowledge Organization Systems)的方式進行多語化索引典(Thesaurus)，並將之與典藏系統整合，提供多語查詢及瀏覽。
- 在此情境下，幾項重要研發任務開始進行，包括：
  - 合作對象：與美國Getty Research Institute進行 AAT中文化合作關係
  - 方法論：衍生/模仿 (Derivation/Modeling) ，翻譯/改作(Translation/Adaptation)，直接對照(Director Mapping)
  - 任務小組: equivalence mapping, 多語索引典建構與系統開發，跨分項合作與團隊形成，Translation team與數位典藏系統的檢索整合

AAT的7個層面與21個層級

- 相關概念層面
  - 層級：相關概念
- 物理屬性層面
  - 層級：屬性和特性、狀態和結果、設計元素、顏色
- 風格和時期層面
  - 層級：風格和時期
- 代理者層面
  - 層級：人、機構、生物
- 活動層面
  - 層級：學科、功能、事件、體育和心智活動、過程和技術
- 質材層面
  - 層級：材質
- 物件層面
  - 層級：
    - 物件群組和系統
    - 物件種類
    - 組件
    - 建築環境
    - 裝潢和裝備
    - 視覺和聽覺溝通

AAT記錄

Click the icon to view the hierarchy.

ID: 300151343 **Record Type:** concept

ceramics (objects) (<object genres by material>, <object genres (Guide Term)>, Object Genres (Hierarchy Name))

Note: Refers in general to articles made of ceramic, which is any of various hard, brittle, heat-resistant and corrosion-resistant materials made by shaping and then firing a nonmetallic mineral, such as clay, at a high temperature. Earthenware, porcelain, and brick products are examples of ceramics.

同義關係

Terms:

ceramics (objects) (**preferred**,C,U,LC,English-P,D,U,PN)
keramiek (object) (C,U,English,AD,U,SN)
ceramica (object) (C,U,English,AD,U,SN)
ceramic ware (C,U,English,UF,U,N)
keramics (C,U,English,UF,U,N)
céramique (object) (C,U,French,AD,U,SN)
Keramik (C,U,German-P,D,U,PN)
cerámica (object) (C,U,Spanish-P,D,U,PN)
cerámicas (C,U,Spanish,AD,U,PN)
keramik (object) (C,U,Swedish,AD,U,SN)

階層關係

Facet/Hierarchy Code: V,PE

Hierarchical Position:

Objects Facet
.... Object Genres (Hierarchy Name) (G)
..... <object genres (Guide Term)> (G)
..... <object genres by material> (G)
..... ceramics (objects) (G)

Additional Parents:

Materials Facet
.... Materials (Hierarchy Name) (G)
..... materials (matter) (G)
..... <materials by composition> (G)
..... inorganic material (G)
..... clay (G)
..... <clay products> (G)
..... <ceramic and ceramic products> (G)
..... ceramics (objects) (G)

相關關係

Additional Notes:
Spanish ..... Usado generalmente para artículos hechos de cerámica.

Related concepts:
condition is .... **grit-tempered**
..... (<fabrication attributes: ceramics>, <fabrication attributes>, ... Attributes and Properties) [300263316]
condition is .... **shell-tempered**
..... (<fabrication attributes: ceramics>, <fabrication attributes>, ... Attributes and Properties) [300265126]

Sources and Contributors:
cerámica (object)..... [CDBP-DIBAM Preferred, VP]
..... Cassell's Spanish Dictionary (1990)
..... Fleming y Honour, Diccionario de Artes Decorativas (1987) 169
..... IFLA Glossary for Art Librarians (1984)
ceramica (object)..... [VP]
..... IFLA Glossary for Art Librarians (1984)
cerámicas..... [CDBP-DIBAM]
..... Comité, Plural del término en singular
ceramics (objects)..... [VP Preferred]
..... Boger, Dictionary of World Pottery and Porcelain (1971)
..... Candidate term Candidate term - Program for Art on film - 5/89
..... CDMARC Subjects: LCSH (1988-)
..... Hamer, Potter's Dictionary (1986)
..... IFLA Glossary for Art Librarians (1984)
..... Random House Dictionary of the English Language (1987) 2. w/plural verb, articles of earthenware, porcelain, etc.
..... RILA, Subject Headings (1975-1990)
..... Worcester Art Museum Library, List of subject headings, unpub. (1976)
ceramic ware..... [VP]
..... Savage and Newman, Illustrated Dictionary of Ceramics (1974)
céramique (object)..... [VP]
..... IFLA Glossary for Art Librarians (1984)
keramics..... [VP]
..... Hamer, Potter's Dictionary (1986)
keramiek (object)..... [VP]
..... IFLA Glossary for Art Librarians (1984)
Keramik..... [VP]
..... Cassell's German Dictionary (1978)
..... IFLA Glossary for Art Librarians (1984)
keramik (object)..... [VP]
..... IFLA Glossary for Art Librarians (1984)

Subject: ..... [CDBP-DIBAM, VP]
..... RILA, Subject Headings (1975-1990) Porcelain; Pottery

Note:
English ..... [VP]
..... IFLA Glossary for Art Librarians (1984)
Spanish..... [CDBP-DIBAM]
..... TAA database (2000-)

等同關係對應之研究
Equivalence Mapping

- 了解以對應方法操作不同語言（英文與中文）的知識組織系統，詞彙等同關係的類型、概念結構匹配關係的類型，及在對應的操作過程之相關問題

- 等同關係對應，是本研究核心的分析方法
  - 詞彙等同關係的對應，是目前解決不同語言及不同知識組織系統的互通性研究領域，最重要的一種研究方式
  - 在索引典的情境下，係指識別出詞彙、概念及層級關係的等同性之過程

資料蒐集

- 本研究以美國蓋提研究所（The Getty Research Institute）發展的「藝術與建築索引典」（Art & Architecture Thesaurus，以下簡稱AAT）詞彙，及故宮博物院（以下簡稱故宮）參與數位典藏與數位學習國家型計畫（Taiwan e-Learning and Digital Archives Program, 以下簡稱TELDAP）的「器物」子計畫之控制詞彙爲主要研究對象。

- 以「故宮器物後設資料需求規格書」選擇其中60個控制詞彙，進行與AAT詞彙等同關係對照實作。此60個控制詞彙，含括AAT層面架構中的6個層面，分別是「物理屬性」、「風格與時期」、「代理者」、「活動」、「材質」及「物件」層面；只有「相關概念」層面尚未包括在本次先導研究範圍。

詞彙選擇範例

Selected terms of TELDAP	**AAT FACET and Hierarchy**
全器形制－植物－蔬果式－葫蘆形 全器形制－幾何－平面－三角形 局部形制－頸－長頸	PHYSICAL ATTRIBUTES FACET **Attributed and Properties Hierarchy**
紋飾－幾何－雷紋－雲雷紋 紋飾－幾何－點狀紋－乳丁紋	PHYSICAL ATTRIBUTES FACET Design Elements Hierarchy motifs
窯系－明代官窯系 窯系－景德鎮窯系 窯系－定窯系	STYLES AND PERIODS FACET Styles and Periods Hierarchy <Chinese ceramics styles>
考古學文化－仰韶文化 考古學文化－良渚文化	STYLES AND PERIODS FACET Styles and Periods Hierarchy <Chinese Neolithic periods>
紋飾－人物－宮廷人物－后妃 紋飾－人物－佛道人物－彌勒佛	AGENTS FACET People Hierarchy religious (people)

資料分析

- 辨識不同語言詞彙的等同關係對應之匹配類型
  - 確認**TELDAP**英譯詞彙
  - 判斷與確認此詞彙在**AAT**可能的等同概念之詞彙
  - 賦予等同關係的匹配類型
- 識別不同來源知識組織系統的概念結構異同性

研究結果

- 詞彙等同關係匹配類型：頻率
  - 「完全等同關係」（35個，58%）是本研究結果出現最高頻率的匹配類型，其次依序是「部份等同關係（種－屬附屬關係）」（18個，30%）、「不等同關係」（6個，10%）、「不完全等同關係」（1個，2%）。
- 詞彙等同關係的有效匹配類型
  - 根據本研究發現，TELDAP與AAT詞彙等同關係，具有6種匹配類型。分別爲：完全等同關係、完全等同關係（cross ref.）、不完全等同關係、部分等同關係（種－屬附屬關係）、不等同關係（文化獨特性）、不等同關係（超越收錄範圍）
- 概念結構異同性分析
  - 根據本研究初步發現，不同系統的概念結構相似性範圍可依程度分爲4種類型，包括：
    - 架構相似，可採用模仿方法（modeling）將來源詞彙(AAT)的部份架構移植到目標詞彙(TELDAP)的系統；
    - 架構相似，但需再擴充或修訂來源詞彙(AAT)的架構；
    - 架構不相似，目標詞彙(TELDAP)可部份等同對應至來源詞彙(AAT)；
    - 架構缺乏，目標詞彙(TELDAP)無法等同對應至來源詞彙(AAT)。

等同關係匹配類型

Match Type code	Definition	Sample	
		TELDAP's examples	AAT's examples
1-1	Exact Equivalence	仰韶文化	**Yangshao**
1-2	Exact cross-reference match	掐絲	**filigree enameling**
2-1	Inexact Equivalence	香爐	**censers**
3-2	Partial Equivalence (species-genus relationship …Subordination)	乳丁紋	**dots**
5-1	Non-equivalence (culture-dependent)	龍泉窯	X
5-3	Non-equivalence (beyond scope)	彌勒佛	X

匹配編號	匹配類型	例證說明	
		TELDAP詞彙	**AAT詞彙**
1-1	**Exact Equivalence**	仰韶文化**(<**考古學文化**)**	**Yangshao** (<Chinese Neolithic periods>, <Chinese prehistoric periods>, ... Styles and Periods)

仰韶文化 玉鑿
5000 B.C.-3000 B.C.

匹配編號	匹配類型	例證說明	
		TELDAP詞彙	AAT詞彙
2-1	Inexact Equivalence	香爐	**Censers** (<ceremonial containers>, <containers by function or context>, ... Furnishings and Equipment) Note: Refers to containers **with perforated covers** used for burning incense in a ritual context, **especially ecclesiastical**; usually of metal or ceramic. ID: 300198814

作者不詳（-）。[明 葡萄紋香爐]。《數位典藏聯合目錄》。http://catalog.ndap.org.tw/?URN=839062（2009/02/22瀏覽）。

不完全等同關係
Inexact Equivalence

- 編號: 300198814

  ID: 300198814
- 香爐 (<儀式容器>, <依功能或使用情境區分之容器>, ...裝飾與設備)

  censers (<ceremonial containers>, <containers by function or context>, ... Furnishings and Equipment)
- 註釋: 用指蓋上有孔的薰爐，尤其是基督教儀式中所使用的薰香容器，多用金屬或陶瓷做成 。

  Note: Refers to containers with perforated covers used for burning incense in a ritual context, especially ecclesiastical; usually of metal or ceramic.

不完全等同關係
Inexact Equivalence

- TELDAP的「香爐」代表的其中一種概念，是具有「宗教或儀式上的功用」置於廟前(體型較大)，爲插香所用；或是放在神桌上(體型較小)插香用。
- 此時，「香爐」與censers具有「不完全等同關係」，因爲TELDAP「香爐」所代表的概念，包括沒有蓋子，也非用於基督教儀式。

台中孔廟銅製香爐
文建會，台灣大百科

匹配編號	匹配類型	例證說明	
		TELDAP詞彙	**AAT詞彙**
3-2	部分等同關係（種－屬附屬關係） Partial Equivalence (species-genus relationship ...Subordination)	乳丁紋(＜點狀紋＜幾何＜紋飾)	**dots** (<geometric motifs>, motifs, ... Design Elements) **ID: 300010145**

商後期 鉤連乳丁紋羊首罍

部分等同關係

- 部分等同關係可以有二種類型：
  - 屬種關係(也就是對照到更狹義的概念詞彙，Narrower term)
  - 種屬關係(也就是對照到更廣義的概念詞彙，Broader term)
- 本研究的初步結果，發現所有18個具有部分等同關係的詞彙，皆是種屬關係(species-genus relationship)，因此可以得知，TELDAP詞彙的專指度較爲深度，許多概念詞彙只能暫時對照到AAT較廣義的詞彙．如：與幾何相關的基本圖案(motifs)，AAT在「幾何」之下，有「圓點」，但僅至此層次；TELDAP則在「幾何」之下，有「圓點」，之下再有「乳丁紋」、「穀紋」、「梅花點紋」等不同的區別(如下圖)．這些詞彙未來需進一步識別，是否分別代表很重要的概念? 若是，則需爲這些詞彙也建立一個概念詞彙，並新增至目前AAT 「圓點」之下。

匹配編號	匹配類型	例證說明	
		TELDAP詞彙	AAT詞彙
5-1	不等同關係（文化獨特性） Non-equivalence (culture-dependent)	龍泉窯(<窯系) Longquan Kilns	×

作者不詳（1101 A.D.-1300 A.D.）。[南宋 龍泉窯 青瓷鳳耳瓶]。《數位典藏聯合目錄》。http://catalog.ndap.org.tw/?URN=835546（2009/02/22瀏覽）。

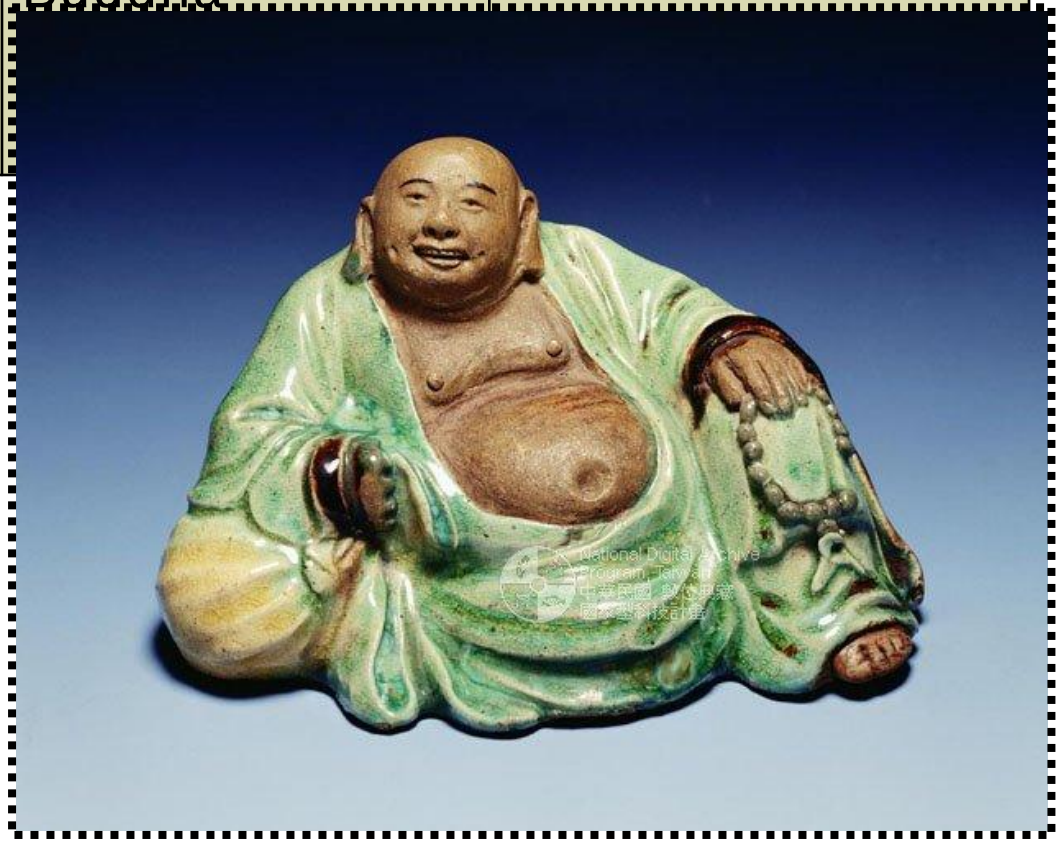
不等同關係（文化獨特性）
Non-equivalence (culture-dependent)

- 所謂不等同關係（文化獨特性），是指TELDAP詞彙，未包含任何能夠與AAT意義相當的詞彙（不論是部份等同或不完全等同皆未在內）。主要情況，是TELDAP（中文）詞彙表達的概念具有文化依賴性，此對AAT而言是陌生的概念。

- 窯系(Kiln system)是中國陶瓷根據各地窯場產品、工藝、釉色、造型與裝飾的不同，所劃分的瓷窯體系。其中，該詞彙項下包括「龍泉窯系」(Longquan Kilns)，其產品歷史悠久，暢銷亞洲、非洲、歐洲三大洲許多國家，深受人們喜愛，被讚爲「世界最佳者」，是中國歷史上的一個名窯。但目前尚無法在AAT找到等同關係的詞彙（如下圖）。

Styles and Periods Facet
.... Styles and Periods
..... <styles and periods by region>
..... Asian
..... East Asian
..... Chinese
..... <Chinese styles and periods>
..... <Chinese styles (guide term)>
..... <Chinese ceramics styles>
..... Ru

匹配編號	匹配類型	例證說明	
		TELDAP詞彙	AAT詞彙
5-3	不等同關係（超越收錄範圍）Non-equivalence (beyond scope)	彌勒佛(<佛道人物<人物<紋飾) Maitreya Buddha	X



作者不詳（清）。[中文品名：彌勒佛（72-00407）]。《數位典藏聯合目錄》。http://catalog.ndap.org.tw/?URN=1197865（2009/02/07瀏覽）。

不等同關係（超越收錄範圍）

Non-equivalence (beyond the scope)

- 不等同關係的另外一種類型，是TELDAP該詞彙並不在AAT收錄範圍．例如：本研究樣本「彌勒佛」是屬於個人的專有名詞(proper name)，因此不列入未來新增至AAT的考量．但是，可以考慮在另外的人名權威檔（Name Authority File）建立

研究發現
概念結構異同性

- 根據本研究初步發現，不同系統的概念結構相似性範圍可依程度暫時分爲4種類型，依序是：
- Type 1　架構相似：可採用模仿方法（modeling）將來源詞彙（AAT）的部份架構移植到目標詞彙（TELDAP）的系統
- Type 2　架構相似：但需再擴充或修訂來源詞彙（AAT）的架構
- Type 3　架構不相似：目標詞彙可部份等同對照至來源詞彙
- Type 4　架構缺乏：目標詞彙（TELDAP）無法等同對照至來源詞彙（AAT）

Type 1 架構相似
Similar Structure

研究發現

代理者層面 > 人物(層級) > 人物 > (人物，依據職務區分) > (公部門的人物) > 統治者 > 宮廷人物 > 后妃
Agents Facet > people > people (agents) > <people by occupation> > <people in government and administration> > rulers (people) > monarchs > empresses

Agents Facet代理者層面
.... People人物層級
..... people (agents) 人物
..... <people by occupation> <人物，依據職務區分>
..... <people in government and administration> <政府與管理的人物>
..... rulers (people) 統治者
..... monarchs 宮廷人物
..... empresses 后妃

研究發現

Type 2 架構相似
Similar Structure

架構相似－但需再擴充或修訂目標詞彙的架構

TELDAP詞彙代表的概念常出現中國文化的獨特性，因此需在**AAT**現有的概念結構中，再更深入往下層延伸細部結構、或往寬廣的層級新增整套概念結構。

具中國文化獨特性或表達性概念的詞彙，包括：「窯系」、「考古文化」、「全器形制」、「書體」及「紋飾」等概念結構。

Type 2 架構相似
Similar Structure

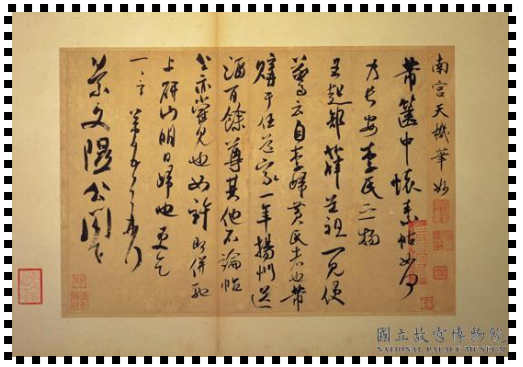
- 以書體(script) 爲例，在本次研究個案，包括 clerical script 隸書(clerical script)、楷書(regular script)、篆書(seal script)等皆是漢字重要的形體，也經常表現於書畫作品之中。
- 而AAT <Arabic scripts>阿拉伯字體的下面，有分出十二種字體，其分類概念，與跟漢字形體演變的歷史過程(如：甲骨文、金文、篆書、隸書、楷書、行書、草書等)是相似的，因此建議可以在AATscripts (writing)- <scripts by form> 項下，以相似架構新增一組<Chinese scripts>（中文書體）　　如下圖

Type 2 架構相似
Similar Structure

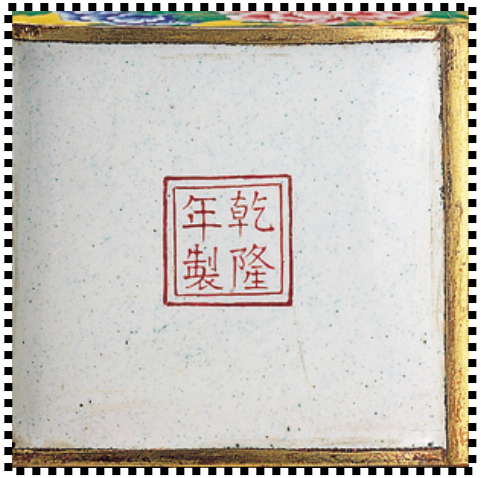
研究發現

clerical script (隸書, lishu)

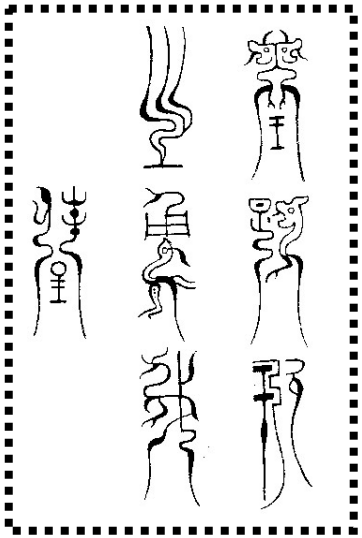
seal script （篆書）



semi-cursive script (行草書)



regular script
(楷書,Kaishu)



bird-and-insect script
（鳥蟲書）

Type 2 架構相似
Similar Structure

研究發現

.... Objects Facet
..... Components (Hierarchy Name)
..... components (objects)
..... <components by specific context>
..... <information form components>
..... <script and type forms>
..... **scripts (writing)**
..... <scripts by form>
..... **<Arabic scripts>** 阿拉伯書體
..... **<Chinese scripts>** （中文書體）
..... **bird-and-insect script**（鳥蟲書）
..... **clerical script**（隸書，**lishu**）
..... **regular script**（楷書，**kaishu**）
..... **seal script** （篆書）
..... **semi-cursive script** （行草書）

Suggest:
Create a set of new guide term and candidate terms

研究發現

Type 3 架構不相似
Dissimilar Structure

所謂「架構不相似」，在本研究是則表示**TELDAP**詞彙呈現的概念結構與**AAT**表現的概念結構之分類邏輯不相同，但**TELDAP**詞彙可以完全等同或部分等同地對照到**AAT**詞彙。如：釉色

Type 4 架構缺乏
Lack of Structure

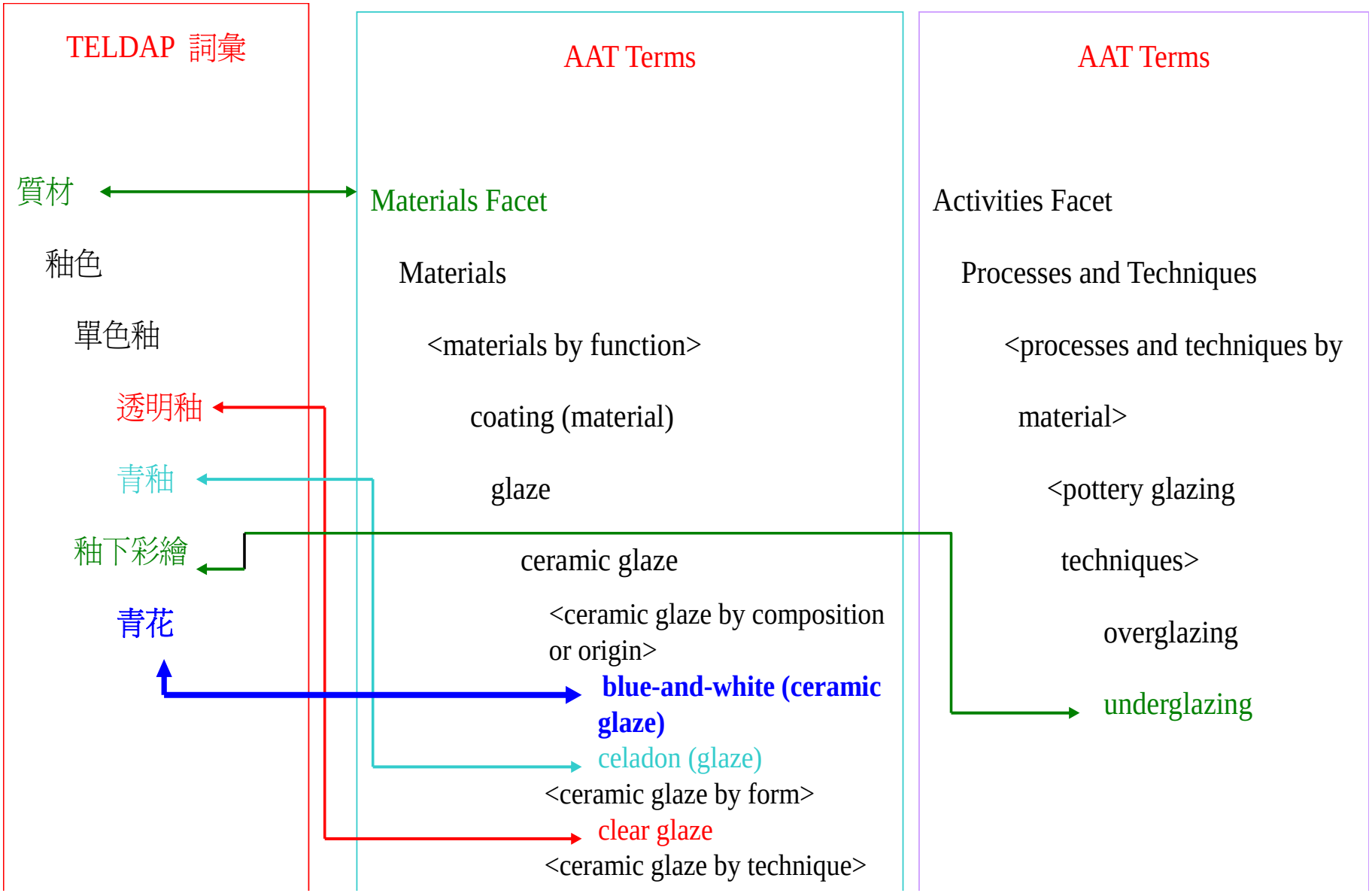
「架構缺乏」，則是指**TELDAP**詞彙呈現的概念結構無法在**AAT**發現相似結構，相關詞彙也無法於**AAT**對照到任何等同關係的詞彙。如：局部形制

ceramics glaze

- TELDAP與AAT對於陶瓷釉的分類角度不盡相同，如：
  - TELDAP以釉色作爲陶瓷釉形式描述的一種分類角度；
  - AAT則依組成(composition) 、形式(form) 、技法(technique)作爲概念分類的依據．
- 因此會形成相同概念的詞彙(如：TELDAP的青花與AAT的blue-and-white (ceramic glaze))所呈現的不同概念結構。TELDAP的「青花」隸屬於「釉下彩繪」；而AAT的blue-and-white (ceramic glaze)則隸屬於<ceramic glaze by composition or origin> 之下的詞彙。

詳以下圖示

ceramics glaze



結論1/3

- 相同領域的不同語言、不同知識組織系統之間的詞彙，具有高度相容性（包括詞彙等同關係程度，及概念結構相似性高），雖然彼此間會有文化獨特性產生的概念與詞彙，或互相尚未收入的
- 未來可進一步研究，如何發展與設計，可以容納更多視野與觀點的概念結構

結論2/3

- 本研究對於TELDAP與Getty's AAT合作案，有非常直接的助益．建議可以針對「完全等同關係」的詞彙，進行雙語的建置與互相連結。而對於「部分等同關係」的詞彙，若是被進一步確認是重要的概念詞彙，則可以產生新詞彙，並貢獻至AAT，如此經由亞洲區域觀點的加入與連結，可豐富AAT本身的完整性與多元文化性之價值。
- 中文「藝術與建築索引典」的設計發展，經由TELDAP的個案研究，初步結果發現AAT的概念架構與詞彙似乎可以含括大部分TELDAP需求，且提供更大的脈絡架構，及連結TELDAP的深度性詞彙。這對於知識組織系統的發展與互通性，具有很大的意義，並且可以節省從頭開始建置的成本。

結論3/3

- 基於本研究初步發現「完全等同」(58%)與「部分等同」(30%)關係，佔TELDAP與AAT詞彙間的大多數比例(88%)；以及AAT本身已建立及不斷建置中的多語言(如：西班牙文、法文、義大利文、德文等)，加上本研究對於中文等同關係的研究，將預期可以精進與貢獻於未來TELDAP多語言資訊檢索的環境。

謝謝!!

陳淑君
中央研究院資訊科技創新研究中心

城菁汝(代宣讀)