EasyChair Preprint

№ 9549

The Development of a Multilanguage Thesaurus Based on Linked Data

Magda El-Sherbini

January 6, 2023

The Development of a Multilanguage Thesaurus Based on Linked Data

A presentation and a live demonstration

Created by Magda El-Sherbini

Professor and Middle East & Islamic Studies Librarian

The Ohio State University Library for

Special Library Association – Arabian Gulf Chapter

March 7-9, 2023

Abstract:

The need for multilingual information access has been addressed in many forms for several years. The ALCST (Association for Library Collection and Technical Services) Non-English Access Committee indicated that research in the area of assigning subject headings in the language of the script will enable the user to find materials in the library more efficiently. It also indicated that LCSH (Library of Congress Subject Headings) does not sufficiently represent the culture of the Arabic and Islamic world. Several examples of the deficiency of LCSH coverage in these areas illustrate the need to develop an Arabic open-source controlled vocabulary that can be used when assigning subject headings for Arabic materials. This paper aims to address three issues:

- The LCSH and its coverage of subjects
- Enhancing the discoverability of Arabic materials by adding Arabic subject headings
- The development of an Arabic thesaurus based on the linked-data approach

I. LCSH AND ITS COVERAGE OF NON-ROMAN SUBJECTS

Many North American library catalogs support title, author, and keyword searching in scripts. However, when it comes to subject searching, these systems provide access only to controlled English-language subject headings and thesauri, such as the LCSH. In much of the cataloging of materials written in languages written in non-Roman scripts, English-language subject access provides insufficient description of the content to ensure the retrieval of the item. Some concepts that function in other languages or cultures do not have English equivalents. In such cases catalogers would select a controlled vocabulary subject heading that was "close enough or broader." On the other hand, a user who is searching for a book written in a certain language about a certain subject should be able to conduct a subject search in that language if that is preferred.

Several authors addressed the issues of providing access to the catalog by subjects in multiple languages. In his article, Agenbroad offered an extensive historical overview of Romanization in library catalogs (Agenbroad, 2006). He identified institutional policy, cataloging standards, and lack of technical feasibility, as major obstacles to implementing script access points. Aliprand also pointed out that Romanization is inadequate for providing access to materials in scripts (Aliprand, 2005). She described Romanization as "Information Distortion." (Aliprand, 2005) Aliprand pointed to the need for "locale-specific" access points, determined by the user's preferred language and written in the proper script. She also advised that authority files present multiple script access points

Researchers continued to identify alternatives to using script subject headings instead of LSCH. In Models for Multilingual Subject Access in Library Online Catalogues, Davies described the two models used by the International Labor Organization (ILO) to provide users with subject access in languages of their choice (Davis, 2003). In the first model, subject terms were translated on the fly when catalog records were accessed during search, display, or export. In the second model, descriptors were translated in a batch process after being entered into the bibliographic record and the equivalent descriptors in other languages were entered into the bibliographic record. Each of these models has advantages and disadvantages in terms of data storage, indexing, and translating. Park addressed name and subject access across languages and cultures (Paark, 2007). She examined current mechanisms for cross-lingual name

and subject access and identified major factors that hinder cross-lingual information access. For example, the author looked at converting concepts expressed in Korean into English LCSH subject headings. She pointed out that mapping Korean names and subjects to their English counterparts is very difficult due to different linguistic structures and sociocultural norms. In the case of English and Korean, these structural differences are considerable, unlike between English and other Western languages, because English and Korean are unrelated languages. Moreover, “word segmentation and transliteration schemes dealing with scripts also play a part in limiting access to cross-lingual and cross-cultural resource.” (Paark, 2007). Landry discussed this issue in his research “Multilingual Subject Access: The Linking Approach of MACS” by exploring solutions to multilingual subject access to online catalogs. The strategy for this research was to develop a Web-based link and search interface through which equivalents between three Subject Headings are established. Languages–SWD/RSWK (Schlagwortnormdatei/Regeln für den Schlagwortkatalog) for German, RAMEAU (Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié) for French, and LCSH for English–can be created and maintained, and by which users can access online databases in the language of their choice (Landry, 2004).

Providing simultaneous access to multi-script subject headings has been recognized as an issue to be solved by the Association for Library Collections and Technical Services (ALCTS) Task Force on non-English Access as far back as 2007. Although many libraries developed substantial foreign language collections, the language of the library catalogue and its subject access points has been English (Services, Task Force on Non-English Access Report, 2007). The author began to study this topic in 2010 after being appointed to serve as Chair of the ALCTS Steering Committee to Oversee the Implementation of the “Recommendations Contained in the Report of the ACLTS Task Force on non-English Access 2007-2009” (Services, Steering Committee to Oversee the Implementation of the recommendations Contained in the Report of ALCTS Task Force on Non-English Access , 2007-2209). The report included eleven recommendations. The Steering Committee was successful in implementing ten of the recommendations, but recommendation number eleven; “*Assign the ALCTS Subject Analysis Committee (SAC) working*

with appropriate library organizations to study the needs of library users for multilingual subject access in the appropriate script(s), and to propose steps to address those needs," was not implemented (ALCTS Task Force on Non-English Access Final Report, 2009). The ALCTS Steering Committee members explained the reasons for not implementing recommendation eleven in these terms: "this recommendation was extremely vague and broad; a research project that might never be completed; is outside SAC's scope; SAC doesn't do much research on end users or their needs" and finally "more research is needed".

In a survey that was conducted in 2011, the authors attempted to assess the need to provide subject access to the library online catalogue. The results of the survey indicated that:

- End users were not completely dissatisfied with the current library catalog.
- End users and librarians wanted a system that is more open to multilingual subject headings.
- Highlighted areas of opportunity for libraries to make significant improvements to the catalog (El-Sherbini). Following the above research, the author received a grant to study the Bibliotheca Alexandrina's (BA) multilingual catalogue. The BA provided access to their library collections in three languages (Arabic, English, and French). This research revealed that the BA is using the Library of Congress MARC 21 Model B, where the transcribed text in the bibliographic record is entered only in the script in which it appears. Subject access is recorded based on the authorized thesaurus in each language. For example, the French thesaurus "RAMEAU" is used for assigning subject headings for a book in the French language (El-Sherbini M. , 2015)

II. ENHANCE THE DISCOVERABILITY OF NON-ROMAN MATERIALS BY ADDING NON-ROMAN SUBJECT HEADINGS

Information gathered from the work on the ALCTS Steering Committee and the author's research in this are gave birth to the idea of internationalizing the Ohio State University (OSU) library catalogue. Three factors made this development possible:

- RDA (Resource Description and Access) was written with "internationalization" in mind
- The OCLC FAST (Faceted Application of Subject Terminology), which is derived from the LSCH but applies a simpler syntax while retaining the richness of the LCSH vocabulary.
- Users' needs to search the library catalogue by subject terms in their preferred language, e.g., Arabic, Chinese, Japanese, etc.

Detailed description of the pilot project can be found in the IFLA report "Subject Access: Unlimited Opportunity, August 11-12, 2016, Columbus, Ohio (El-Sherbini m. , 2017). The pilot project allows users to access the catalog by subject in their preferred languages. The subjects' terms are taken from International National libraries' open sources-controlled vocabulary such as:

- *Sharing* 국가서지**:**도서(서지), 온라인 자료, 기사 색인 및 주제명, 저자명, 그리고 도서관 정보 Linked Open Data 에 대한 탐색 서비스를 만나보세요! (See example fig.1)
- Ndlsh : governed by Japan National Diet Library and freely available (http://id.ndl.go.jp/auth/ndla) (see example fig. 2)
- Qā'imat ru'ūs al-mawḍū'āt al-'Arabīyah al-qiyāsīyah lil-maktabāt wa-marākiz al-'ulūmāt wa-qawā'id – print thesaurus (See example fig.3)
- List of Chinese Subject Terms from the National Central Library of Taiwan : http://catld.ncl.edu.tw/index.jsp (國家圖書館鏈結資源 = National Library Link Resources) (see example fig. 4)

In this process, the cataloger is assigning the LCSH terms as normal, converting them to FAST, and then finding the equivalents of FAST terms in the International thesaurus. Examples below illustrate this process.

LOCATION	CALL NO.	YEAR	STATUS	NOTE
Book Depository	DS915.19 .P36 1983		AVAILABLE	

Description	401 pages ; 21 cm
	text
	unmediated
	volume
Edition	Ch'ung pan
	重版
Note	Colophon inserted
	Includes bibliographical references
OCLC/Control #	27167640
Author	Pak, Sŏng-su, 1931- author.
	朴成壽, 1931- author
Series	Tonga kyoyang ch'ongsŏ ; 10.
	東亞敎養叢書 ; 10
Subject	Korea -- History -- 1864-1910 -- Historiography
	Korea -- History -- 20th century -- Historiography
Other Subject	Historiography.
	역사 기록학
	Korea.
	한국(국명)[韓國]
Genre	History.

Fig. 1: Subject in Korean

LOCATION	CALL NO.	YEAR	STATUS	NOTE
Fine Arts Library Oversize 2nd Floor	GT2915 .K52 2013		AVAILABLE	

Description	149 pages : color illustrations ; 30 cm
	still image
	text
	unmediated
	volume
Note	Catalog of exhibitions held at Itsuō Bijutsukan, October 5-November 17, 2013 and at Fukuoka-shi Bijutsukan, January 5-February 16, 2014
	Co-published by Fukuoka-shi Bijutsukan, distributed by Shibunkaku Shuppan
	Parallel title from cover
	Includes bibliographical references
Local Note	OSU copy 1 gift of Gustavus and Sidney L. Basch Merorial Fund
ISBN	9784784217267
	4784217266
OCLC/Control #	881592695
Subject	Japanese tea ceremony -- Utensils -- Exhibitions.
	Art objects, Japanese -- Exhibitions
	Matsunaga, Yasuzaemon, 1875-1971 -- Art collections -- Exhibitions
Subject	松永安左ェ門, 1875-1971 -- Art collections -- Exhibitions
Subject	Kobayashi, Ichizō, 1873-1957 -- Art collections -- Exhibitions
Subject	小林一三, 1873-1957 -- Art collections -- Exhibitions
Other Title	Gustavus and Sidney L. Basch Memorial Fund Collection
Addt'l Author	Kobayashi, Ichizō, 1873-1957.
	小林一三, 1873-1957
	Matsunaga, Yasuzaemon, 1875-1971.
	松永安左ェ門, 1875-1971
	Itsuō Bijutsukan.
	逸翁美術館
	Fukuoka-shi Bijutsukan.
	福岡市美術館
Other Title	Kobayashi Ichizō to Matsunaga Yasuzaemon
	小林一三と松永安左ェ門

Fig. 2: Subject in Japanese

LOCATION	CALL NO.	YEAR	STATUS	NOTE
Thompson Library Stacks 4th Floor	BP75.5 .M373 2017		AVAILABLE	

Description	237 pages ; 25 cm
	text
	unmediated
	volume
Edition	al-Ṭabʻah al-Ūlá
	الطبعة الأولى
Note	Originally presented as the author's thesis (master)--Kullīyat al-Ādāb wa-al-ʻUlūm al-insānīyah bi-Jāmiʻat Muḥammad al-Khāmis bi-al-Rabāṭ, 2008
	Includes bibliographical references (pages 230-236) and indexes
ISBN	9789777172981
	9777172982
OCLC/Control #	990283513
Author	Marākishī, Muḥammad Ilyās, 1983- author.
	مراكشي، محمد الياس، 1983-
Subject	Muḥammad, Prophet, -632 -- Friends and associates.
	Islamic law -- Interpretation and construction -- History
	Islamic law -- History.
	Sunna.
Other Subject	Muḥammad, Prophet, -632.
	محمد، النبي، 632-.
	Friends and associates.
	الصحابة و التابعون.
	Islamic law.
	الشريعة الإسلامية.
	Islamic law -- Interpretation and construction.
	Sunna.
	السنة.
Genre	History.
	تاريخ.
Other Subject	local/osu .الفقه الإسلامي
	qrmak .الفقه الإسلامي: أصول

Fig. 3: Subject in Arabic

LOCATION	CALL NO.	YEAR	STATUS	NOTE
Thompson Library 3M East Asian Stacks	DS735 .Z88 2016		AVAILABLE	

Description	4, 2, 326 pages ; 21 cm
	text
	unmediated
	volume
Edition	Di 1 ban
	第1版
Note	Includes bibliographical references (pages 306-324)
ISBN	9787310050352 :
	7310050355 :
OCLC/Control #	947088913
Author	Zou, Yayan, author
	邹雅艳, author
Series	Nan kai da xue Han yu yan wen hua xue yuan bo shi wen ku.
	南开大学汉语言文化学院博士文库
Subject	China -- History.
	China -- Foreign public opinion -- History
Other Subject	Public opinion.
	輿論.
	China.
	中國.
Genre	History.
	歷史.
Other Title	Evolution of the image of China in the west from 13th to 18th century
Other Subject	本书共分6章, 借助比较文学形象学的研究方法, 对13-18世纪西方中国形象演变的历史轨迹进行了一个比较系统的分析和评述. 作者以大量文献资料为依据, 在广阔的历史文化背景中考察了13-18世纪西方中国形象的演变, 并以中国学者的文化立场对其中的各种文化倾向和思想观点进行了有理有据的评判

Fig. 4: Subject in Chinese

As you can see from the list of thesauri, there is no online Arabic thesaurus. The cataloger used the terms from a print Arabic subject headings list, such as Qā’imat ru’ūs al-mawḍū‘āt al‘Arabīyah al-qiyāsīyah lil-maktabāt wa-marākiz al-‘ulūmāt wa-qawā‘id. This process is very difficult and time consuming for the cataloger, which resulted in slower processes and gradual accumulation of backlogs. In some cases, mapping of Arabic subjects to their English counterparts is very difficult, as the two languages reflect distinct and unique cultures and traditions. In the case of English and Arabic, these structural differences are considerable. For example, the term “zakat” and “daw'ah” were established in LCSH, but not “salat,” which means “ صلاة" in Arabic. Instead, the term “prayer” was used for “salat - صلاة ” and the term “prayers” was used for “Du'a - .”دعاء LCSH did not provide the term prayers for “Du'a - دعاء” but it only provided the term prayer for “salat - ”.صلاة (Ismail, 2011)13 Even with the term “Du'a - ,”دعاء there are several variations. For example, there is Du'a al-Baha (,(دعاء البهاء Du'a al-Faraj (,(دعاء الفرج etc. Organ transplant, or naql al-a'da was a contemporary issue discussed by many scholars, but it is not in LCSH.

III THE DEVELOPMENT OF THE ARABIC THESAURUS BASED ON LINKED DATA APPROACH

Reasons for creating the Arabic Thesaurus

- Enhance access to Arabic collection by subject
- Created digital thesaurus
- Make thesaurus available globally
- Active community participation in suggesting terms of their choice

Methodology

- Researching how to create a thesaurus
- The decision was made to use the English authority words. If the Arabic terms can be equivalent to the English term, then the Arabic term will be linked to English authority record. In this case, the LCSH, FAST and Wikipedia were used as authority records. Even if the Arabic term does not have an equivalent, it is still added to the thesaurus, but without linking it to any authority record.
- How to construct a thesaurus that is comprehensive
- Broad categories of knowledge have been identified and defines.
- A study of existing print subject Arabic thesauri was conducted, such as:
    - al-Maknaz al-kabīr : muʻjam shāmil lil-majālāt wa-al-mutarādifāt wa-al-mutaḍāddāt / iʻdād al-Duktūr Aḥmad Mukhtār ʻUmar ; bimusāʻadat farīq ʻamal, 2015.
    - Qāʼimat ruʼūs al-mawḍūʻāt al-ʻArabīyah al-qiyāsīyah lil-maktabāt wa-marākiz al-ʻulūmāt wa-qawāʻid al-bayānāt / tawafar ʻalayhi Shaʻbān ʻAbd al-ʻAzīz Khalīfah (2002)
    - Maknaz ruʼūs al-mawḍūʻāt lil-makhṭūṭāt al-ʻArabīyah / Muḥammad Fatḥī ʻAbd al-Hādī, Muḥammad Fatḥī ʻAbd al-Hādī, 2010
    - al-Saʻūdīyah : qāʼimat ruʼūs mawḍūʻāt lil-maktabāt wa-marākiz al-maʻlūmāt / Shaʻbān ʻAbd al-ʻAzīz Khalīfah, Muḥammad ʻAwaḍ al-ʻĀyidī, Shaʻbān ʻAbd al-ʻAzīz. Khalīfah, Muḥammad ʻAwaḍ. ʻĀyidi, 1981
    - Qāʼimat ruʼūs mawḍūʻāt al-tarbiyah / iʻdād Muḥammad Fatḥī ʻAbd al-Hādī, 1976.

- Qā'imat ru'ūs al-mawḍū'āt al-'Arabīyah / i'dād Ibrāhīm Aḥmad al-Khāzindār., Ibrāhīm Aḥmad Khāzindār, 1994.
- Qā'imat ru'ūs al-mawḍū'āt al-'Arabīyah al-muwaḥḥaddah / i'dād Maḥmūd Aḥmad Itayyim, Maḥmūd Aḥmad Itayyim, 2007.

- The initial stage of the project covers essential topics such as Islam, Quran, politics, history, social studies, education, birds, fish, languages, sociology, philosophy, family, diseases, agriculture, etc.
- For creating a database, it was very important to select the appropriate programing languages that will allow for loading, updating, and searching the database. The program should be flexible enough to achieve the objectives of the project
- Decision was made to make the thesaurus accessible from anywhere in the world.
- Finding and selecting a reliable cloud storage space was a very important step in creating the thesaurus.

- Programming
  For creating a database, it was very important to select the appropriate programing languages that will allow for loading, updating, and searching the database. The program should be flexible enough to achieve the objectives of the project.
- Identifying Arabic subjects Collecting the terms:
  - al-Maknaz al-kabīr : mu'jam shāmil lil-majālāt wa-al-mutarādifāt wa-al-mutaḍāddāt / i'dād al-Duktūr Aḥmad Mukhtār 'Umar ; bimusā'adat farīq 'amal
  - Qā'imat ru'ūs al-mawḍū'āt al-'Arabīyah al-qiyāsīyah lil-maktabāt wa-marākiz al-'ulūmāt wa-qawā'id al-bayānāt / tawafar 'alayhi Sha'bān 'Abd al-'Azīz Khalīfah
  - Ru'ūs al-mawḍū'āt al-'Arabīyah / i'dād Qism al-Fahrasah wa-al-Taṣnīf ; ishrāf Nāṣir Muḥammad al-Suwaydān
  - Oxford dictionaries. Arabic
  - Qā'imat ru'ūs al-mawḍū'āt al-'Arabīyah , Ibrahim Ahmad alKhzindar
  - Arabic Vocabs: politics / https://wisc.pb.unizin.org/lctlresources/chapter/arabic-vocabulary-politics/

- English to Arabic Government
https://www.proz.com/glossary-translations/english-to-arabic-translations/government-politics/page1
- Encyclopedia of Islam and the Muslim world
- Many other subject sources

Designing the database

- Arabic terms are linked to either or both LSCH and FAST authority record, as well as Wikipedia when available
- Use of Boolean operators
- Terms will be in English and Arabic (future Persian and Turkish)
- Search by English or Arabic and find the equivalent
- Many terms will not have equivalent in LCSH, FAST, WIKI
Sometimes the English term means two things in Arabic (record both) (VIAF model)
- The DB doesn't include corporate bodies or names
- No cross references
- Suggest/modify term (Community participation)
- Select and buy reliable cloud storage space
- Using a commercial product that would allow to host the database.
- Using Python, a general-purpose programing language that would allow de-duping the terms and merging files with the master file in the database

Conclusion:

What is described here is the reason behind the creation of an Arabic subject thesaurus and the process of creating the thesaurus. It was created in response to a need that was articulated by ALCTS and will eventually cover all fields of knowledge. The user interface that was designed for this purpose allows the user to search terms in Arabic and find links to English terms from LCSH, Fast and Wikipedia. Arabic terms are drawn either from existing print thesauri or are created where no existing terms are available. The English language

user of the database will be able to search in English to find the matching terms in Arabic. An important feature of the database is its adaptability and flexibility. Users will be able to recommend additions or corrections of terms. As the database grows it will become more comprehensive. The thesaurus will be available universally for the users of Arabic, English, and other languages.

Works Cited

Agenbroad, J. E. (2006). *Romanization is not Enough.* Cataloging & Classification Quartely 42, no. 2, p.: 21-34.

Aliprand, J. M. (2005). *The Structure and Content of MARC 21 Records in the Unicode Enviroment.* Information Technology & Libraries 24, no. 4, p.: 170-179.

Davis, R. (2003). Models for Multilingual Subject Access in Online LIbrary Catalogues: The ILO Expereince. *Annual Conference of the European Library Automation Group.* Bern, Switserland.

El-Sherbini, M. (2015). Multilingual Subject Retrieval: Biblioteca Alexandrina's Subject Authority File and Linked Subject Data. *In Data Science, Learning by Latent Structure, and Knowledge Discovery. Springer, Berlin-Herdelberg*, 535-546.

El-Sherbini, m. (2017). *Improving Discoverability for Language Collections.*

El-Sherbini, m. a. (n.d.). An Assessment of the need to provide non-Roman Subject Access to the Library Online Catalog. *Cataloging & Classification Quarterly*.

Ismail, M. I. (2011). Issues and Challenges in Cataloging Arabic Books in Malaysua Academic Libraries . *Education for Information 28,2-4*, 1510163.

Landry, P. (2004). Multilingual Subject Access: the linking approach of MACS. *Cataloging & Classification Quarterly 37.3-4*, 177-191.

Paark, J.-R. (2007). Cross-Lingual Name and Subject Access: Mechanisms and Challenges. *Library Resources & technical Services 51, no. 3*, 186.

ALCTS (Association for Library Collection and Technical Services). (. (2007). *Task Force on Non-English Access Report.* ALCTS.

ALCTS (Association for Library Collection and Technical Services (2007-2009). *Steering Committee to Oversee the Implementation of the recommendations Contained in the Report of ALCTS Task Force on Non-English Access .* http://www.ala.org/alcts/ianda/nonenglish