TIIARA: A Language Tool for Bridging the Language Gap

Nouf Khashman[1], Elaine Ménard[2(✉)], and Jonathan Dorey[3]

[1] Qatar National Library, Qatar Foundation, Doha, Qatar
nkhashman@qf.org.qa
[2] School of Information Studies, McGill University, Montreal, QC, Canada
elaine.menard@mcgill.ca
[3] School Information Studies, McGill University, Montreal, QC, Canada
jonathan.dorey@mail.mcgill.ca

Abstract. This paper presents and discusses the results of the second phase of the development of TIIARA (Taxonomy for Image Indexing and RetrievAl), a bilingual taxonomy dedicated to image indexing and retrieval. TIIARA offers indexers and image searchers innovative and coherent access points for ordinary images. Initially developed in French and English, the taxonomy has been subsequently translated in 8 languages. The preliminary steps of the elaboration of the bilingual structure are briefly described. The process used in the translation of TIIARA in Arabic language is presented, as well as the main difficulties encountered by the translator. Adding more languages in TIIARA constitutes an added value for a controlled vocabulary meant to be used by image searchers who are often limited by their lack of knowledge of multiple languages.

Keywords: Taxonomy · Controlled vocabulary · Multilingual information · Translation · Arab language

1 Introduction

The organization of visual resources such as personal photos as always been challenging. For years, traditional controlled vocabularies have been chosen for indexing and classification of images, with mixed successes. Although well developed and trying to be as enumerative as possible, these terminologies are not always precise for the pictures we take over the years with our digital cameras. Often considered too general to provide the right degree of granularity leading to precise retrieval results, most controlled vocabularies can only provide a primary subject or a broader category for images. Additionally, very good controlled vocabularies only exist in one language, limiting their use to the individuals who are familiar with this specific language.

In order to overcome this lack, TIIARA (Taxonomy for Image Indexing and RetrievAl), a bilingual taxonomy dedicated to image indexing and retrieval, was developed in order to offer a vocabulary appropriate for image indexing and retrieval. Among the many advantages of controlled vocabularies such as taxonomies, it is worth mentioning consistency and enhanced possibilities to match indexing words to search query terms. Furthermore, if the image searcher has the possibility to browse a taxonomic structure to initiate or refine queries, the retrieval will be facilitated.

A. Marcus (Ed.): DUXU 2016, Part II, LNCS 9747, pp. 386–395, 2016.
DOI: 10.1007/978-3-319-40355-7_37

For its initial development, TIIARA included two languages, French and English. As a logical follow-up to the initial development of a bilingual controlled vocabulary, it was decided to translate TIIARA in eight other languages to increase its international scope: Arabic, Spanish, Brazilian, Portuguese, Chinese, Italian, German, Hindi and Russian. This paper describes the initial steps of the elaboration of the bilingual structure, the process used in the translation of TIIARA in Arabic, as well as the main difficulties encountered by the translator. The last section concludes the paper and proposes future directions for improving the multilingual taxonomy.

2 Related Works

Text-based image indexing and retrieval have been studied extensively over the years [1–16]. These studies present the numerous challenges of image organization. The advantages and disadvantages of controlled and uncontrolled vocabularies used for documents or multimedia indexing, are extensively described in the literature [6, 10, 17–22]. Traditionally, some general terminologies such as the Library of Congress Subject Headings (LCSH) or specific vocabularies, including Getty's Art & Architecture Index (AAT) and the Thesaurus for Graphic Materials (TGM) have been chosen for describing visual resources such as images. As an alternative to these conventional vocabularies, taxonomies can be considered as innovative usable means for image indexing and retrieval. They can simplify the searching process and facilitate finding the "right" information near effortlessly. Unfortunately, very few studies described the basic processes of their development [23–26].

Most controlled vocabularies often present shortcomings. For example, they are often not exhaustive enough [13] to provide adequate descriptive information or access points suitable for all uses. In addition the neologisms and terminological changing usage will not be integrated quickly. This slow updating will be definitively frustrating for indexers or image searchers that rely on an up-to-date terminology.

Nor do most controlled vocabulary allow for the use of specific queries, so search results are less accurate than they need to be, a problem compounded when it comes to multilingual information. Some interesting projects exist, especially in Europe, where multilingualism is a requirement. For example, the UNESCO Thesaurus is a quadrilingual controlled and structured list of terms used in subject analysis and retrieval of documents and publications in the fields of education, culture, natural sciences and social and human sciences. With more than 7,000 terms in English and in Russian, 8,600 terms in French and in Spanish, this thesaurus offers terms from the fields of education, culture, natural sciences, social and human sciences, communication and information [27]. However, these multilingual vocabularies are rare and often very limited in the choice of languages offered. This can be explained by the fact that high-quality multilingual controlled vocabularies (thesauri, taxonomies, etc.) take a long time to be developed. Their construction can be long and expensive processes, and their maintenance time-consuming.

3 Objectives

Taxonomies are increasingly being used to organize content within organizations and to support navigation of digital content [28–30]. The review of the literature completed in the initial phase of this research project [31] revealed that there is a gap in our understanding of image searchers' expectations and what is available in terms of searching functionalities. Initially, TIIARA was developed in French and English, the two official languages of Canada, where this research is taking place. Once TIIARA was updated and retested, it was decided that it could be interesting to add other languages to TIIARA. For this first expansion phase, eight languages were selected: Arabic, Spanish, Brazilian Portuguese, Chinese, Italian, German, Hindi and Russian. This section summarizes the development of TIIARA, the translation process in Arabic, and the main difficulties translators faced.

Multilingual information processing has gained more and more attention in recent years. However very few research really explained the hit and miss of elaborating a bilingual controlled vocabulary and how other languages can be thereafter integrated in the making of a thoroughly multilingual vocabulary. This study proposes to fill this gap and answer the following research questions:

1. What are the general steps of the development of a bilingual taxonomy?
2. What are the main steps of the translation of TIIARA in Arabic?
3. What difficulties were encountered during the Arabic translation process?

4 TIIARA Development

4.1 The Bilingual Structure

The initial structuring of the taxonomy involved choosing top-level categories and their subcategories. Two approaches were chosen: starting from the narrowest terms possible and moving to the more generic ones (bottom-up approach) or the selection of general concepts within the taxonomy which are then subdivided (top-down approach). As mentioned previously, TIIARA was simultaneously structured in English and French to keep the taxonomy as parallel as possible. Both languages come from related Indo-European language families and have common origins [32]. The Indo-European family comprises languages largely used throughout Europe, Western and South Asia, and other parts of the world as a result of colonization. This group of languages refers to the easternmost extension of the family from the Indian subcontinent to its westernmost reach in Europe [33].

It was decided that the number of top-level categories and the depth of the taxonomic structure would be kept to a minimum. At first, TIIARA included nine top-categories. This initial version was tested in order to validate and refine the vocabulary and its organization. For the first validation phase, the card-sorting technique was used. Data gathered revealed difficulties encountered using the taxonomy structure and dynamically suggested ways to improve it [26]. Following this first evaluation, the preliminary nine main groupings were reduced to seven top categories (Table 1):

Table 1. TIIARA main categories (English and French)

Main category	Definition
Abstract Ideas Idées abstraites	Related to an idea of something formed by mentally combining all its characteristics or particulars; a concept.
Arts and Entertainment Art et divertissement	Related to people, tools, equipment and products specifically associated with dance, design, visual arts, writing, music, television and film, and stage.
Daily Life Vie quotidienne	Related to the activities and experiences that constitute a person's normal existence.
Nature Nature	Related to the phenomena of the physical world, including plants, animals, the landscape and other features and products of the Earth, as opposed to humans or human creations.
Places Lieu	Related to a building or a physical environment used for a special purpose.
Objects and Equipment Objet et équipement	Related to unique objects or pieces of equipment (not in active use by a person).
Work Travail	Related to people doing a job, other than those listed in "Arts and Entertainment."

These main categories were then developed to include second-, third- and, in some cases, fourth-level subcategories, in French and English. Two indexers, one English-native and one French-native speaker, used TIIARA to describe a small image database (IDOL [Images DOnated Liberally]. This custom-built database includes 6,015 images offered voluntarily by photographers. The indexing terms assigned by the two indexers were evaluated and compared to identify potential gaps in the taxonomy.

A second round of testing took place with a representative sample of image searchers. We asked the participants to complete retrieval tasks of images indexed using the revised taxonomy TIIARA to measure its degree of effectiveness and efficiency. During this experiment, a sample of 60 respondents were asked to indicate where in the taxonomic structure they thought they would find each one of the 30 images shown. Participants were also asked to fill a questionnaire intended to obtain their general opinion on TIIARA and to report any difficulties encountered during the retrieval process. The quantitative data was analyzed according to statistical methods, while the content of open-ended questions was analyzed and coded to identify emergent themes. The results of this phase of the research project indicated that, despite the fact that some categories still need further refining, TIIARA already constitutes a successful tool that provides access to ordinary images. The bilingual taxonomy constitutes a definite benefit for image searchers who are not very familiar with images indexed in English, which still dominates the Web.

Once fully tested and updated according to feedback received from indexers and image searchers, TIIARA was translated in eight different languages. Arabic was a logical choice since this language is "the largest member of the Semitic branch of the

Afro-Asiatic language family" that comprises all descendants of Classical Arabic spoken primarily across the Middle East and North Africa. It is one of the six languages of the United Nations, and serves as the first language for 22 Arab countries, and as a second language in over a dozen more. With over 375 million native speakers [34], it is one of the most spoken languages [33], mainly in the Middle East, North Africa, and some Muslim countries such as Iran, Malaysia, and Indonesia.

As a Semantic language, Arabic shares similarities with other Semitic languages, such as Aramaic and Hebrew. In terms of writing, several languages use the Arabic alphabet, such as Farsi, Urdu, Pashto and Kurdish [35].

4.2 The Arabic Translation Process

The Arabic alphabet contains 28 letters, most of which change form depending on whether they appear at the beginning, middle or end of a word, or on their own. Arabic uses eight main diacritic marks that can change the meaning of a word drastically based on their positions on those letters. When those diacritics are excluded or omitted, homonymy problem may arise [36].

Moreover, the language has around 5 million words that are derived from around 11,300 roots compared to 400,000 keywords in English, which has total of 1.3 million words. This makes the language rich of terminology and has a complex morphology compared to English and European languages [37]. Arabic is also a major source of vocabulary for other languages such as Kurdish, Spanish, Persian, Urdu, and Swahili.

Arabic language can be classified into three main variants: the formal Arabic language, known as Classical Arabic or Fus-ha, is the language in which the Qur'an is written. This is relatively a difficult form of Arabic, which is considered today more of a written language than a spoken one.

The second form is Modern Standard Arabic (MSA), which is similar but easier than Classical Arabic. It's understood across the Arab world and used on television and media, as well as to teach Arabic as a foreign language.

The last form is the Colloquial Arabic (local dialects), which differs from one Arab country to another, and even within each country. Those dialects differ from MSA and each other in terms of phonology, morphology, lexical choice and syntax. The translation of the TIIARA was primarily based on MSA; for example, translating clothes into ملابس (malabis) rather than هدوم (hudoum), which is used locally in Egypt.

4.3 The Difficulties Encountered

While some difficulties encountered in translating the TIIARA from English to Arabic will be discussed based on the particularities of the Arabic language previously discussed [36], others will be based on the translator's experience working with the TIIARA.

The peculiar morphology of Arabic might render methods used for English retrieval inappropriate. An example from TIIARA would be "beauty and hygiene", which translated to Arabic as جمال ونظافة"jamal w nathafa". In this case, the letter "waw" (meaning and) has been slightly transformed and linked with the following word; this would create problems were it decided to treat "and" as a stop word.

There was also the question whether the definite article the ال (Al) should be included or not in the translation, since the 'al' and a number of conjunctions and prepositions, are not separated from their following word by a space. Examples include الناس (people), إسلام (Islam), المنسوجات (textiles). These terms were treated based on the context and not one single rule was followed.

It is common to find many Arabic words that have different pronunciations and meanings but share the same written form (homonyms), making finding the appropriate semantic occurrence of a given word a problem. An example would be كتاب and كتّاب, where the first word means a book, while the second means authors. Another example would be حب (love), as it can be written as intended حُب, or as حَب (grains). The reason for this confusion is the omission or misplacement of the diacritical marks, which are also not usually indexed in online information retrieval systems.

Arabic plurals are formed more irregularly than in English depending on the root and the singular form of the word. The plural form might be produced by the addition of suffixes, prefixes or infixes, or by a complete reformulation of the word. An example from the TIIARA would be translating elephants into أفيال or فيلة, sheep into غنم or أغنام, canyons into وديان or أودية. The latter translation in these cases was chosen to simplify the process.

Every Arabic letter is pronounced as a word and cannot be used to represent one character like in English. Therefore, in Arabic, acronyms and abbreviations are not found. Therefore, the term Sport utility vehicles (SUVs) was translated into سيارات دفع رباعي, and Recreational vehicles (RVs) into سيارات ترفيهية without adding acronyms.

Moreover, it appears that MSA orthography has largely been standardized for a long time now [37]. However, few variations persist across and within different Arab countries. In TIIARA, for example, golf was transliterated into غولف, but it could also be written as جولف because the /g/sound does not occur in MSA and is replaced with the closest letters to that sound, غ and ج in this case.

There were a few instances where the politically correct terms needed to be included rather than using the literal translation of the word. For example, disabled people was translated into ذوي الاحتياجات الخاصة (people with special needs) rather than معاقون, which literally means "handicapped".

Finally, it was noted [38] that some terms related to innovations and borrowed words might not be regulated. In TIIARA, there were a number of terms that are not found in the Arabic language. To overcome this challenge, the terms were transliterated into Arabic. For example, curling was transliterated into الكيرلنج, accompanied by the word sport رياضة to indicate it is a sport, and doing the same for hockey هوكي الجليد and lacrosse رياضة الكروس. Other foreign terms were directly translated into Arabic, such as مشغلات أقراص فيديو رقمية (DVD players), ماسحات ضوئية (scanners), مسجل فيديو شخصي (personal video recorders) and مراكز تجارية (malls). However, people tend to use the English term whenever they refer to these inventions.

5 Discussion and Conclusion

According to the Working Group on Guidelines for Multilingual Thesauri [39], several possibilities may be considered in the development of multilingual controlled

vocabularies such a taxonomy: building a new vocabulary from the bottom up, starting with one language and adding one or more languages, starting with more than one language simultaneously, combining or merging existing monolingual controlled vocabularies, connecting existing controlled vocabularies to each other or translating a controlled vocabulary into one or more other languages. For TIIARA, the development was made in parallel in French and English. To add more languages, we selected the last of the aforementioned options (translation). Translated version of TIIARA now exists in multiple languages (French, English, Arabic, Spanish, Brazilian Portuguese, Italian, Chinese, German, Hindi and Russian). Other translations are also being considered (Polish, Japanese, Greek, etc.).

Several linguistic questions arose in the development of a multilingual controlled vocabulary. For its initial development, TIIARA only included French (from the Romance branch) and English (from the Germanic branch); two languages that are not different theoretically, being from the same Indo-European language family and having common origins [32]. However, problems with the structural hierarchy could arise in multilingual controlled vocabularies, particularly when the different languages show crucial discrepancies in the hierarchical levels where concepts are organized.

It is worth mentioning that the individuals that participated in the TIIARA translation, including the Arabic translation, could not be considered as "professional translators," nor "professional taxonomists". The different translations were rather produced by volunteers with sufficient knowledge of the source languages (English or French) and the chosen target languages. Consequently, the translators faced many difficulties, at many levels, but mainly at the semantic and syntactic level [40]. Semantic and syntactic ambiguities remain unquestionably one of the main problems that could have potentially serious consequences on the intrinsic structure of the taxonomy.

Several situations were reported by the Arabic translator: (1) some terms could be exactly translated; (2) some translated terms were inexact or nearly equivalent; (3) some translated terms only corresponded to a partial equivalence; (4) some translated terms could be matched to one-to-many equivalents, where to express the meaning of the preferred term in one of the languages, two or more preferred terms were needed in the other language; and finally, (5) some terms from the source language did not match equivalent term in the target language.

It is without surprise that the other translators also reported similar experiences. As all linguistic entity conceptualises the world from their own perspectives, meanings are rarely symmetrical across languages. Therefore, the aim for TIIARA has not been to pursue exact equivalence between languages but, instead, to lead the information retriever towards relevant search results regardless of which language is used. The multilingual nature of any controlled vocabulary, including TIIARA, poses a number of language- and culture-related challenges, and building harmonious and understandable hierarchy in more than one language is definitively a complex process that may require compromises. Nevertheless, constructing multilingual controlled vocabularies is a crucial factor in the context of information globalization. This cannot be achieved without acknowledging and respecting the differences that exist between the specific characteristics of different languages.

The next step of this research will be the testing of the translated versions with real image searchers. We also planed to include the multilingual TIIARA in SINCERITY (Search INterfaCE for the Retrieval of Images with a TaxonomY), a bilingual search engine that has been developed in parallel with the present project. The suggestion of integrating a taxonomy to assist image retrieval was expressed by many image searchers who participated in the exploration of the roles and usefulness of functionalities for image searching in a bilingual context [41]. For the moment, even if most image searchers prefer searching with keywords related to the content of the image they are looking for, the events taking place or people that appear in the picture, most search engines still do not offer their users the opportunity to browse a taxonomic structure to initiate their queries.

Moreover, most image searchers prefer searching in their own language. Giving them a hierarchical structure they can navigate seems the perfect solution to facilitate the retrieval process. Moreover, if the taxonomic structure includes several languages it will consequently give different linguistic communities equivalent opportunities and subsequently bridge the information divide that still exists. Especially for image searchers who have difficulties formulating a query using words from another language.

References

1. Panofski, E.: Meaning in the Visual Arts: Papers in and on Art History. Doubleday, Garden City (1955)
2. Krause, M.G.: Intellectual problems of indexing picture collections. Audiov. Librarian **14**, 73–81 (1988)
3. Markey, K.: Access to iconographical research collections. Libr. Trends **37**, 154–174 (1988)
4. Armitage, L.H., Enser, P.G.B.: Analysis of user need in image archives. J. Inf. Sci. **23**, 287–299 (1997)
5. Jörgensen, C.: Attributes of images in describing tasks. Inf. Process. Manag. **34**, 161–174 (1998)
6. Jörgensen, C.: Image Retrieval – Theory and Research. Scarecrow, Lanham (2003)
7. Markkula, M., Sormunen, E.: End-user searching challenges indexing practices in the digital newspaper photo archive. Inf. Retrieval **1**, 259–285 (2000)
8. Goodrum, A.A., Spink, A.: Image searching on the excite web search engine. Inf. Process. Manag. **37**, 295–311 (2001)
9. Choi, Y., Rasmussen, E.M.: Searching for images: the analysis of users' queries image retrieval in American history. J. Am. Soc. Inf. Sci. Technol. **54**, 498–511 (2003)
10. Matusiak, K.K.: Towards user-centered indexing in digital image collections. OCLC Syst. Serv. **22**, 283–298 (2006)
11. Enser, P.G.B., Sandom, C.J., Hare, J.S., Lewis, P.H.: Facing the reality of semantic image retrieval. J. Documentation **63**, 465–481 (2007)
12. Enser, P.G.B.: The evolution of visual information retrieval. J. Inf. Sci. **34**, 531–546 (2008)
13. Greisdorf, H.F., O'Connor, B.C.: Structures of Images Collections: from Chauvet-Pont d'Arc to Flickr. Unlimited Libraries, Westport (2008)
14. Ménard, E.: Image retrieval: a comparative study on the influence of indexing vocabularies. Knowl. Organ. **36**, 200–213 (2009)

15. Chung, E.K., Yoon, J.W.: Categorical and specificity differences between user-supplied tags and search query terms for images. An analysis of Flickr tags and Web image search queries. Inf. Res. **14** (2009)
16. Stvilia, B., Jörgensen, C.: User-generated collection-level metadata in an online photo-sharing system. Libr. Inf. Sci. Res. **31**, 54–65 (2009)
17. Markey, K., Atherton, P., Newton, C.: An analysis of controlled vocabulary and free text search statements in online searches. Online Rev. **4**, 225–236 (1980)
18. Muddamalle, M.R.: Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics. J. Am. Soc. Inf. Sci. **49**, 881–887 (1998)
19. Savoy, J.: Bibliographic database access using free-text and controlled vocabulary: an evaluation. Inf. Process. Manag. **41**, 873–890 (2005)
20. Arsenault, C.: L'utilisation des langages documentaires pour la recherche d'information. Documentation et Bibliothèques. **52**, 139–148 (2006)
21. Macgregor, G., McCulloch, E.: Collaborative tagging as a knowledge organisation and resource discovery tool. Libr. Rev. **55**, 291–300 (2006)
22. Rafferty, P., Hidderley, R.: Flickr and democratic indexing: dialogic approaches to indexing. Aslib Proc. **59**, 397–410 (2007)
23. Lambe, P.: Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness. Chandos Publishing, Oxford (2007)
24. Whittaker, M., Breininger, K.: Taxonomy development for knowledge management. In: World Library and Information Congress: 74th IFLA General Conference and Council, Québec, 10–14 August 2008
25. Hedden, H.: The Accidental Taxonomist. Information Today, Medford (2010)
26. Ménard, E., Smithglass, M.: Digital image description: a review of best practices in cultural institutions. Libr. Hi Tech. **30**, 291–309 (2012)
27. UNESCO: UNESCO Thesaurus (2016). http://databases.unesco.org/thesaurus/
28. Gilchrist, A., Kibby, P.: Taxonomies for Business: Access and Connectivity in a Wired World. TFPL Consultancy, London (2000)
29. Kremer, S., Kolbe, L.M., Brenner, W.: Towards a procedure model in terminology management. J. Documentation **61**, 281–295 (2005)
30. Uddin, M.N., Janecek, P.: Performance and usability testing of multidimensional taxonomy in web site search and navigation. Perform. Meas. Metrics **8**, 18–33 (2007)
31. Ménard, E., Smithglass, M.: Image retrieval in a bilingual context: a review of best practices. Libr. Hi Tech. **32**, 98–119 (2014)
32. Walter, H.: Honni soit qui mal y pense. Robert Laffont, Paris (2001)
33. Ethonlogue Languages of the World: Explore the Languages of the World (2015). https://www.ethnologue.com/
34. Internet World Stats (2016). http://www.internetworldstats.com/stats19.htm
35. British Broadcasting Corporation (BBC): A guide to Arabic – 10 facts about the Arabic language (2014). http://www.bbc.co.uk/languages/other/arabic/guide/facts.shtml
36. Moukdad, H., Large, A.: Information retrieval from full-text Arabic databases: can search engines designed for English do the job? LIBRI **51**, 63–74 (2001)
37. Habash, N.Y.: Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers, San Rafael (2010)
38. Azmi, A.M., Aljafari, E.A.: Modern information retrieval in Arabic - catering to standard and colloquial Arabic users. J. Inf. Sci. **41**, 506–517 (2015)
39. International Federation of Library Associations and Institutions - IFLA. Guidelines for Multilingual Thesauri (2009). http://www.ifla.org/files/assets/hq/publications/professional-report/115.pdf

40. Braschler, M.: Combination approaches for multilingual text retrieval. Inf. Retrieval **7**, 183–204 (2004)
41. Ménard, E., Khashman, N., Dorey, J.: Two solitudes revisited: a cross-cultural exploration of online image searcher's behaviors. In: Marcus, A. (ed.) DUXU 2013, Part II. LNCS, vol. 8013, pp. 79–88. Springer, Heidelberg (2013)