

Towards a Multilingual Future – MILE conference in Dublin April 2009

The aim of the EU funded project MILE, now in its third and final year, is to increase access to digital images in the EU. At the start of the one day conference in Dublin Jessica Tier, MILE's project Manager, reminded us that metadata is currently the prime means for accessing images.

Now that the funding for digitisation projects in the UK has slowed down, and the cultural heritage sector feels the bite of the recession, it is an opportune moment to take a step back and consider the importance of allocating resources to what used to be considered the 'back office' functions of cataloguing and keywording.

The deliberations of the MILE project over the past three years have signposted a way forward in three areas: metadata mapping and classification, orphan works and IP, and multilingual access to collections, the subject of the Dublin event.

It was once thought that the internet would create worldwide distribution of images from single sites. Things have turned out differently. With global markets, local representation and distribution have become the cornerstone of the stock industry. In the same way, English will not be the only language of access for images in either the commercial or the cultural sector.

Automatic translation of keywords can only be part of a solution. This was amply demonstrated at the MILE meeting in London in October 2008. Aptly called *Speaking in Tongues* the idiosyncracies of language were highlighted when automatic keywording software was demonstrated. The only way to get consistent results from keywords, it would appear, is to use controlled vocabularies, with assigned translations for the words in the tree.

Stella Dextre Clarke, a consultant specialising in the design and implementation of controlled vocabularies, kicked off the proceedings in Dublin with a look at the standards operating in the field of thesaurus creation. The standards, developed since the first thesaurus was created in 1959, were created primarily for text retrieval.

Dextre Clarke raised questions about the suitability of these standards for the retrieval of images, drawing a distinction between recall, where as many results as possible are retrieved in a wide net approach, and precision, in which the search is narrowed down for a highly focused result. She asked if perhaps the search for images sometimes requires less precision than the search for text. The nature of the search is certainly different. Not all elements of an image can be expressed in words, so a wider net may be more appropriate. Aspects causing most trouble in image search are those which relate to visual or design aspects which may express ideas or concepts which can be hard to pin down. This introduces a 'relevance' issue of a different nature to that in text search. Precision, it turns out, is just aspect of a patchwork of solutions in the search for images.

Images are currently predominantly retrieved by use of words. The larger agencies, Getty and Corbis, and others, use their own controlled vocabularies for keywording images, and these vocabularies by and large follow the logical rules set out for thesaurus control.

Qualifiers for example assign precise meanings to words which would otherwise be ambiguous. Pitch (tar) , Pitch (frequency) pitch (Sales, Pitch (sports), pitch (gradient) is an example. Preferred terms are organized in a structured hierarchy, together with their broader and narrower terms, synonyms and related terms. So 'Cars' might be the preferred term,

‘Automobile, Car,’ are synonyms or Use For terms (UF) ‘Motorised land vehicles’ is a broader term (BT), ‘Electric cars’ is a narrower term (NT), and ‘Motoring’ could be a Related Term (RT).

The logical structure of these vocabularies means there is a place for every keyword as the tree grows. In a system where logic does not prevail there is a danger of the structure toppling over when new words are added.

Other rules include the use of plurals for the preferred terms (‘Girls’ rather than ‘Girl’ in English) but here there are differences between languages. French and German rule require the singular to be used. (‘Frau’ not ‘Frauen’).

The fact that people may search in the singular is not relevant here. If the preferred term is ‘Cars’, the word car will be a synonym or UF term and the images will be retrieved if the thesaurus is built into the search system. To differentiate between a search for a single car and a number of cars, other methods will need to be used. (‘One object’ can be used to denote a single car, but problems will arise if there are multiple incidental cars as well.)

There are other rules. Adverbs and adjectives are discouraged. For example, use ‘Happiness’ not ‘Happy’ as the preferred term. (In image libraries happy may be used in an ‘Attribute’ field).

Dextre Clarke demonstrated that interoperability between vocabularies means using the keywording rules. The areas she pinpointed were; sharing of vocabularies, easy import of data into new catalogues, assimilation of images into other collections. She raised the question of whether image libraries should follow the standards set for thesaurus use, answering in the affirmative. The standards are mostly applicable she said arguing that perhaps the criteria for synonyms and equivalence could be relaxed for images. In practice, this is how stock image libraries operate.

In the current environment, with ever more competition in the market, image libraries need to look to their keywording methods if they want to thrive. The days of letting an assistant attach random keywords to an image are over. The way to achieve accurate and consistent keywording is to create a controlled keyword list and take a systematic approach. This requires resources, but with image libraries increasingly dependent on income from a variety of agents the need for good keywording is more pressing. (There has been a high demand for places on the Electric Lane Keywording course). Similarly, multilingual keywording, seen as a pipedream not so long ago, will become a run of the mill requirement before long.

Genevieve Clavet-Merrin from the Swiss National Library gave an overview of the multilingual subject mapping and search arising from her work with the MACS (Multilingual Access to Subjects) for the European Library. The need for translation is very immediate in multilingual Switzerland where the only index at the Swiss National Library is in German.

Searching in another language is more difficult than speaking. Although users may understand results in another language, translation of the search terms is not wasted.

The MAC project was originated by the Swiss National Library and run in conjunction with the British Library, the Deutsche National Bibliothek, the Library of Congress and the Bibliothèque Nationale de France. It runs on the principles of equality of languages where there would be no one pivotal language. This is challenging as it is always easier to translate from one known standard. The project has managed to remain language neutral by a process of linking between all the different languages, creating a web of translations.

The Cacao project is another pan European project designed to help users access and navigate multilingual content in public access libraries. Supported by libraries in France, Italy

Hungary and Germany, the project has identified some of the problems facing users and indexers in a multilingual environment, some of which show up the limitations of textual analysis. The speaker representing Cacao was Dr. Frédérique Segond, from the Xerox European Research Centre.

In text libraries both metadata and content are in text format. How do you differentiate between books by Dante from books about Dante, for example? If you search for the word Rome, you may bring up results for events in Rome, views of the city, people who come from Rome and so on.

Segond pointed out that in multilingual retrieval you may have to translate the query as well as the results. Proper names need to be identified. A query for 'La Fontaine Fable' translates as 'The Fountain Fable'. A single word like Avocat in French can have two meanings and translate into two words (Lawyer and Avocado). The word 'chien' in French can translate into dog or hammer or nails in various languages. Translation does not reduce to a simple dictionary search.

Cacao deals with these problems using all the techniques available; text and metadata analysis, word by word translation., multiword translation, standard machine translation, semantic analysis, and natural language processing. . The project can be accessed at www.cross-library.com:8080/cacaoUI.

The Multimatch project has identified that online resources available to the cultural heritage sector are fragmented, leaving users to discover interpret and aggregate this content themselves using general search tools. Dr Gareth Jones from Dublin City University explained that the aim of Multimatch is to provide enhanced multilingual access to this content and develop a search engine to provide access across media types and language boundaries.

Multimatch www.multimatch.org is an EU project spanning 6 countries . The consortium covers academic institutions in Dublin, Sheffield, Amsterdam, Geneva, Spain and Italy, as well as industrial and cultural heritage partners including well known image library and publisher Fratelli Alinari of Florence.

As with Cacao, the approach is to use a combination of tools – text search, search of structured metadata, and visual search. The multilingual search covers English, Dutch, Italian, Spanish, German Polish. The need to translate both query and the results was again demonstrated. A query in Dutch will bring results from Dutch documents, but has to be translated to Italian to access the Italian documents, which will need in turn to be translated for use by Dutch speakers. Searching documents in a number of languages using this model means translating all the documents into English, and all queries to English to bring a return from any other language. This assumes then that the users can read and understand the English text. Fortunately this level of complexity is not needed for images, where at least the item located, the image, is not language specific.

Machine translation, said Jones, can be sufficient for general terms but not for personal names, organisation names, location names or titles of artworks. The name *Mona Lisa*, can be translated in several different ways in the French language alone. Mutli match has taken a hybrid and practical approach to these issues, finding information where it can and not assuming perfect matches from any one source. The project mines the url's of artworks in different languages from Wikipedia- a neat way of finding translations automatically. Multimatch uses information from a number of sources to build a picture of the data in different languages. The 3 stage dictionary process involves crawling the English Wikipedia, extracting hyperlinks to query languages (Italian and Spanish) and generating translation pairs

using hyperlink base names. The project uses automatic translation from Wordlingo in combination with dictionary based phrase translation.

If you make a query in Italian for 'Vanitas Natura Morta', the automatic translation would be 'nature dead woman', but the incorrect translation is detected automatically and replaced by 'Vanitas, still life, still lifes'. Jones gave several more examples of automatic translations against domain specific translations. (statua dela liberta query in Italian would be 'statue of the freedom' by automatic translation, Statue of Liberty in a domain specific translation)

Reality Bites was the title of the talk by Fotofinder's Agnes Folaji. Fotofinder is a portal with 6.8 million images and 4000 clients. The overall aim is to bring images from sources in other to the German speaking market. How should keywording from suppliers be incorporated into a German search? Fotofinders translates between English German and French languages. The scope of the vocabulary used has increased to 6 million words, and uses a combination of dictionaries and a linguistic engine that suggests translations to keyworders. On the positive side, the linguistic engine helps identify inappropriate terms, and enhances quality. On the negative side, it returns too many synonyms, cannot handle multi-word phrases, and sometimes delivers inappropriate search results and ambiguous words.

It's an inexact science, as Folaji demonstrated. The challenges she identified were the number of keywording standards, the use of free keywording by suppliers rather than controlled vocabularies, misspelling of words, and lack of keywording altogether, even when a complete caption is provided.

This is the real world, where image libraries depend on customers finding the image they want, but where good keywording is possibly the exception rather than the rule.

In this environment multilingual keywording is a very big challenge indeed. After hearing about the difficulties with text retrieval, we can thank our lucky stars that visual search will be available to assist in the search for images. The seminar on Visual Recognition at the Copic Congress in Dresden this year will assess the development of visual search techniques in narrowing the search for images. Multilingual keywords will be needed, but there is hope that the visual search will help the user sift the results, and perhaps even assist in the keywording.

© Sarah Saunders, Electric Lane 2009

With thanks to Liisa Kaakinen for her expert advice

